Ministère de l'enseignement supérieur et de la recherche scientifique
وزارة التعليم العالي والبحث العلمي
BADJI MOKHTAR-ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR-ANNABA
جامعة باجي مختار- عنابة

Faculté des sciences de l'ingéniorat
Département d'Informatique

Année : 2014/2015

THESE

présentée en vue de l'obtention du diplôme de

Doctorat en sciences

Intitulée :

Classification non supervisée textuelle descriptive en langue arabe

Filière
Informatique

Présentée par
Kelaiaia Abdessalem

Devant le jury :

Pr. Halima Bahi	**Université Badji Mokhtar, Annaba**	**Président**
Pr. Hayet Farida Merouani	**Université Badji Mokhtar, Annaba**	**Rapporteur**
Pr. Halim Sayoud	**Université USTHB, Alger**	**Examinateur**
Pr. Hamid Seridi	**Université de Guelma**	**Examinateur**
Pr. Redha Laouer	**Université de Tébessa**	**Examinateur**
Pr. Yamina Tlili	**Université Badji Mokhtar, Annaba**	**Examinateur**

Remerciements

الله سبحانه وتعالى أشكر على القوة التي أمدني بها حتى أكملت هذا العمل

Ma grande reconnaissance va au Professeur Hayet Farida Merouani, Mon encadreur, pour ses précieux conseils, ses encouragements et sa grande compréhension ;

Je remercie chaleureusement Madame Halima Bahi, Professeur de l'enseignement supérieur à l'université Badji Mokhtar, Annaba de m'avoir fait l'honneur de présider le jury de ma soutenance ;

Un grand remerciement à Monsieur Halim Sayoud, Professeur de l'enseignement supérieur à l'université d'Alger d'avoir accepté d'être membre de mon jury ;

Un remerciement particulier à Monsieur Hamid Seridi, Professeur de l'enseignement supérieur à l'université du 8 Mai 45, Guelma d'avoir accepté d'être membre de mon jury ;

Un remerciement chaleureux à Monsieur Redha Laouer, Professeur de l'enseignement supérieur à l'université de Tébessa d'avoir accepté d'être membre de mon jury ;

Je tiens ici à adresser mes profonds remerciements à Madame Yamina Tlili, Professeur de l'enseignement supérieur à l'université Badji Mokhtar, Annaba d'avoir accepté d'être membre de mon jury ;

Un très grand merci pour mon ami et collègue Fayçal Nouar pour l'aide qu'il m'a apporté tout au long de la réalisation de ce modeste travail ;

Mes remerciements vont également à Rafik Benabda et Nadjib Djahel pour les efforts qu'ils ont fournis pour l'aboutissement de ce travail ;

A toute personne ayant contribué de près ou de loin à l'élaboration de ce travail,
Je dis merci.

A toute ma famille…

A la mémoire de mon père…

Résumé

Les problèmes causés par l'augmentation constante du volume de l'information textuelle dans la recherche et l'organisation de cette information n'est pas à démontrer. L'un des axes qui tentent de résoudre ces problèmes est la classification non supervisée textuelle (classification thématique pour certains). Cette classification, qui consiste à générer une partition de clusters homogènes, n'est pas suffisante pour subvenir au besoin de l'utilisateur tant pour l'organisation que dans l'exploration et la recherche de l'information voulue. La description des clusters générés reste un chemin incontournable pour la compréhension de ces derniers. La majorité des techniques de description présentes dans la littérature s'articulent sur le nombre d'apparitions des mots pour décrire les résultats de la classification non supervisée par les mots clés ou par les phrases, en plus, ces techniques sont généralement destinées à la description de la classification des textes courts (*snippets* ou fichiers entêtes) retournés par les moteurs de recherche sur le web.

La présente étude présente une nouvelle approche intitulée LDK-Means (*Latent Descriptive K-Means*) qui combine deux techniques très célèbres en recherche de l'information, l'une pour la classification non supervisée et l'autre pour la description des résultats obtenus par cette classification.

La description (labellisation) consiste à faire intervenir, en plus du nombre d'apparitions des mots dans le texte (approche statistique), la relation qui décrit l'apparition conjointe de ces mots dans ce texte. L'idée derrière cette approche est l'exploitation de l'une des méthodes probabilistes thématiques la plus connue à savoir LDA *(Latent Dirichlet Allocation).*

L'approche présentée a été menée sur des collections en langue arabe, une langue connue pour par ses caractéristiques morphosyntaxiques très particulières qui nécessite plusieurs opérations de prétraitements.

Mots-Clés : Classification non supervisée, Texte, Langue arabe, Description, Labelisation, LDA, K-Moyennes, Prétraitement.

Abstract

The problems caused by the constant increase in the volume of textual information in information retrieval and organization are not to be demonstrated. One of the axes that are trying to solve these problems is text clustering. This clustering, which is aim to generate a partition of homogeneous clusters, is not sufficient to meet the need of the user for organization, exploration and information retrieval. The description of the generated clusters remains an essential way for understanding them. The majority of description techniques in the literature are based on the number of occurrences of words to describe the results of the clustering by keywords or phrases, in addition, these techniques are generally serve at the description clustering short texts (header files or snippets) returned by search engines on the web.

This study presents a new approach called LDK-Means (Latent Descriptive K-Means) that combines two well known techniques in information retrieval, one for clustering, and the other if for the description of the results of this clustering.

The description (label) is to involve, in addition to the number of occurrences of words in text (statistical approach), the relationship that describes the joint appearance of these words in the text. The idea behind this approach is the exploitation of one of the most popular topic models methods namely LDA (Latent Dirichlet Allocation).

The approach presented was conducted on for collections in Arabic, a language known for it's very specific morphsyntactic characteristics that require several operations of preprocessing.

Keys-words : Clustering, Text, Arabic language, Description, Labeling, LDA, K-means, preprocessing.

الملخص

إن المشاكل الناجمة عن الزيادة المستمرة في حجم المعلومات المحتواة داخل النص في البحث عن المعلومة وتنظيمها ليس بجديد. أحد المحاور التي تحاول حل هذه المشاكل هو التصنيف دون اشراف للنص. هذا التصنيف، الذي يهدف إلى ايجاد مجموعات متجانسة من النصوص، لا يكفي لتلبية احتياجات المستخدم من تنظيم وتنقيب وبحث عن المعلومة. إن وصف هذه المجموعات يبقى وسيلة أساسية لفهم ما تحتويه من نصوص. تستند غالبية تقنيات الوصف الموجودة حاليا على عدد ظهور الكلمات لوصف نتائج التصنيف دون اشراف للنصوص بواسطة الكلمات المفتاحية او الجمل ، بالإضافة إلى ذلك، فإن غالبية هذه التقنيات موجهة لتصنيف نصوص قصيرة (مقدمات النصوص) المعادة من طرف محركات البحث على شبكة الإنترنت.

تقدم هذه الدراسة طريقة جديدة تسمى التصنيف K-معدل الوصفي الكامن(LDK-Means) الذي يجمع بين اثنين من أبرز التقنيات المعروفة، واحدة للتصنيف دون اشراف والأخرى لوصف نتائج هذا التصنيف.

تقوم تقنية الوصف هذه على إشراك علاقة الظهور المشترك بين الكلمات داخل النص مع عدد ظهور هذه الكلمات (المنهج الإحصائي). إن الفكرة وراء هذا النهج هو استغلال واحدة من التقنيات الأكثر شهرة من بين طرق تحديد المواضيع (Topic Models) ألا وهي LDA .

لقد تم تطبيق هذا المنهج الجديد على قواعد للنصوص باللغة العربية، وهي لغة معروفة بخصائصها الشكلية والنحوية الجد صعبة التي تتطلب عدة عمليات معالجة مسبقة.

الكلمات المفتاحية : التصنيف دون اشراف، النص، اللغة العربية، الوصف، LDA، K-Means، المعالجة المسبقة.

Table des matières

Liste des figures

Liste des tableaux

Acronymes et Abréviations

ASCII : American Standard Code for Information Interchange

BOW : Bag Of Word

CCA : Corpus of Contemporary Arabic

CLARA : Clustering LARge Applications

CLARANS : Clustering Large Applications based on RANdomized Search

CNST : Classification Non Supervisée Textuelle

DCF : Description Comming First

DCL : Description Comming Last

DUC : Document Understanding Conferences

LDA : Latent Dirichlet Allocation

LDC : Linguistic Data Consortium

LSA : Latent Semantic Analysis

LSI : Latent Semantic Indexing

MALLET : MAchine Learning for LanguagE Toolkit

MCMC : Markov chain Monte Carlo

MRR : Mean Reciprocal Rank

MSA : Modern Standard Arabic

NMF : Non-negative Matrix Factorisation

OSAC : Open Source Arabic Corpora

OSAc : Open Source Arabic Corpus

PAM : Partitioning Around Médoïds

PERL : Practical Extraction and Report Language

PLSA : Probabilistic Latent Semantic Analysis

SHOC : Semantic Hierarchical Online Clustering

STC : Suffix Tree Clustering

SVD *:* Singular Value Decomposition

SVM : Support Vector Machine

TALN : Traitement Automatique du Langage Naturel

TF : Term Frequency

TFC : Term Frequency Cosine

TF-IDF : Term Frequency, Inverse Document Frequency

TREC : Text REtrieval Conference

UTF-8 : Unicode Text Format-8

VSM : Vector Space Model

XML : eXtensible Markup Language

Introduction

Actuellement en réponse à une simple requête dans n'importe quel moteur de recherche disponible sur le net des millions de pages web sont retournées, on se retrouve devant un océan de flux contenu d'informations électroniques sous plusieurs formes tirées de plusieurs sources (les news groups, les courriers électroniques, les forums de discussions, les réseaux sociaux, etc). Plus de 80% de ce flux d'informations est stocké sous une forme textuelle (non structurée ou semi structurée) [Xiao, 2010], il constitue une source inépuisable pour la construction des collections d'informations gigantesques. Donc pour chercher une information il suffit de puiser dans ces collections qui, dans la majorité des cas, sont construites sans prendre en considération l'exploitation postérieure.

Dans ces collections la masse d'information est telle que l'accès à l'information voulue, devient à la fois difficile et nécessaire. Pour remédier à cela plusieurs outils, qui rentrent dans le cadre de la fouille de texte ou *text mining*, ont été conçus selon la nature de l'information et le traitement voulu : recherche d'une information particulière (systèmes de recherche d'information, systèmes question/réponse), exploration (classification automatique textuelle), analyse de l'information (détection thématique ou *topic detection*, résumé automatique) et plus. Ces outils peuvent présenter une certaine forme de dépendance, par exemple, un système de recherche d'information peut faire appel à une classification automatique textuelle pour accélérer et améliorer la qualité des résultats de la recherche, il peut encore faire appel à un outil de traitement morphologique pour diminuer les problèmes causés par la langue naturelle.

Un des axes les plus attractifs dans la communauté de la recherche de l'information est la classification automatique textuelle ou documentaire[1]. En effet la recherche d'une information dans une collection ou son exploration dans le but d'obtenir une idée globale de son contenu, sont deux tâches très délicates surtout si cette dernière n'est pas convenablement organisée. Le but d'une classification automatique textuelle et de générer une partition de groupes (clusters) homogènes (selon un type d'organisation) dans le but de faciliter la recherche et l'exploration de la collection. Les techniques de cette classification sont répertoriées selon deux principales approches, la première est dite classification supervisée ou catégorisation et est basée sur l'apprentissage supervisé, la deuxième est dite classification non supervisée ou *clustering* ou encore apprentissage non supervisée pour la classification.

La classification non supervisée ou *clustering*, qui consiste à regrouper les données (dans le cas présent des textes ou documents) dans des groupes homogènes dits clusters, est largement utilisée dans la fouille de texte et fait l'objet de plusieurs problématiques et spécialement en analyse exploratoire des bases d'informations et cela avec différentes stratégies [Zhong et Ghosh, 2005]. Des techniques de ce type de classification qui ont été élaborées au début pour des données non textuelles (en *datamining*) ont été adaptées et testées

[1] Dans la littérature on ne trouve pas une distinction entre les deux termes.

sur des langues telles que l'anglais et le français (d'ailleurs des améliorations apportées à ces méthodes classiques continuent à voir le jour régulièrement).

Au début des années 2000, la suprématie de ces techniques dans ce domaine est contestée par des nouvelles techniques orientées textes dites méthodes thématiques probabilistes (*Topics models*). Des méthodes comme PLSA (*Probabilistic Latent Semantic Analysis*) [Hofmann, 1999a; Hofmann, 1999b] qui représente une extension de la célèbre LSA[1] (*Latent Semantic Analysis*) [Deerwester *et al.*, 1990**]**, LDA (Latent Dirichlet Allocation) [Blei *et al.*, 2003] et leurs variantes ont commencé à voir le jour à partir de la fin des années 90, leurs objectifs, au départ, était la détection thématique dans le textes (*Topic detection models*). Puis, ces modèles (spécialement LDA) ont investi le domaine de la classification automatique (intégration dans le projet Mallet (*MAchine Learning for LanguagE Toolkit*) [Mccallum, 2002], le projet Gensim [Řehůřek et Sojka, 2011], la classification non supervisée hiérarchique [Rosen-zvi *et al.*, 2004]) et a donné de bons résultats.

La performance de toutes ces techniques est déterminée par le degré de satisfaction du besoin de l'utilisateur lors de son exploration des clusters générés, surtout si ces clusters sont ''muets'' c'est-à-dire sans description. Une description ou labellisation consiste à donner à chaque cluster de la partition générée un aperçu en quelques mots (selon la technique utilisée) de son contenu. Cette description donc, si elle existe, permet à l'utilisateur d'avoir un aperçu sur le cluster et lui évite de parcourir les éléments de ce dernier si ces éléments ne lui sont pas pertinents.

Plusieurs techniques de description ou labellisation existent et donnent des résultats plus au moins bons. Le problème ici est que la majorité de ces techniques sont destinées pour la description des résultats de la classification non supervisée des fichiers entêtes ou *snippets* retournés par les moteurs de recherches sur le web. Donc utiliser ces techniques sur des documents de longueur importante et sur des collections de grandes dimensions n'est pas évident.

Motivation

Que ce soit pour la classification non supervisée ou la description des résultats obtenus les problèmes causés par la langue naturelle dans laquelle sont rédigés les textes influent directement sur la qualité des résultats obtenus. Une langue comme l'anglais a été sujet de beaucoup de travaux par les pionniers dans ce domaine qui ont concentré leurs efforts sur cette langue. Ensuite d'autres travaux se sont intéressés à étudier les autres langues européennes et les langues asiatiques, notamment le chinois et le japonais.

Malgré la montée en puissance de l'utilisation de la langue arabe dans le monde numérique, où le nombre d'utilisateurs est estimé à plus de 136 millions au 31 décembre 2013[2] (via les différents médias, les réseaux sociaux, etc.), cette langue n'a pas connu le même essor. Mis à part quelques travaux dans ce domaine, elle a encore beaucoup de chemin à faire devant elle, d'où est la motivation de travailler sur cet axe et spécialement la classification non supervisée textuelle descriptive en langue arabe.

[1] Connue encore sous LSI (*Latent Semantic Indexing*)
[2] Source : *Internet Word Stat* http://www.internetworldstats.com/stats7.htm

Objectifs visés

Au cours de la présente étude nous allons essayer de répondre aux questions suivantes:

1. Les nouvelles méthodes intitulées méthodes thématiques probabilistes donnent elles des meilleures résultats que ceux obtenus par les techniques classiques en classification non supervisée en langue arabe et spécialement les collections à grandes échelle?

2. Quel est le meilleur moyen de description des clusters en langue arabe ? Ici nous exposons la nouvelle approche LDK-Means (*Latent Description K-Means*) et son apport en description des résultats de la classification non supervisée textuelle.

3. Comment traiter les nouveaux documents (affectation directe ou relancement du processus de la classification non supervisée) ?

S'ajoute à ces trois questions l'étude de l'influence du caractère morphosyntaxique de la langue arabe sur la classification non supervisée avec les deux approches.

Organisation du mémoire

Le présent mémoire présente deux parties distinctes. Dans la première, constituée des quatre premiers chapitres et intitulée état de l'art, nous commencerons par présenter les différentes approches de la classification non supervisée ainsi que les méthodes les plus connues comme premier chapitre.

Le deuxième chapitre est consacré à la classification non supervisée textuelle ou documentaire. Il commence par les domaines d'utilisation puis il étale les techniques de préparation des collections (prétraitement) et se termine par donner un aperçu sur les méthodes thématiques probabilistes et leurs fonctionnements.

Le troisième chapitre est réservé exclusivement aux techniques de description ou labellisation des résultats de la classification non supervisée et les mesures d'évaluation des performances de ces techniques.

Le quatrième chapitre est entièrement dédié à la langue arabe et ses particularités ainsi que les majeurs problèmes rencontrés en traitement automatique de cette langue et se termine par les travaux connexes.

La deuxième partie est destinée à la contribution du présent travail, elle commence par le cinquième chapitre qui décrit la méthodologie de la recherche menée ainsi que l'approche proposée et son fonctionnement.

Dans le sixième et dernier chapitre nous exposerons en détail les différentes expérimentations menées et les résultats obtenus ainsi que les discussions pour ces résultats.

Nous achèverons ce travail par une conclusion générale dans laquelle nous allons revenir sur tout ce qui a été dit et rappeler les résultats obtenus et nous évoquerons les perspectives immédiates à cette thèse.

Chapitre I

Classification non supervisée

Chapitre I
Classification non supervisée

1. Introduction, classer pour comprendre et apprendre

"Understanding our world requires conceptualizing the similarities and differences between the entities that compose it" (*Tyron et Bailey, 1970*) *in* [Rokach, 2010].

Par sa nature l'humain prouve toujours un besoin de classer tout ce qui l'entoure dans l'objectif de comprendre, d'apprendre et de réagir. Il suffit de regarder autour de nous et de réfléchir pour constater que nous sommes entrain de classer. Ce besoin de classer c'est traduit par les recherches des scientifiques dans des différents domaines avec de différentes approches. Les premières recherches faisant intervenir la classification sont attribuées à des biologistes, petit à petit les recherches se sont élargies aux historiens, aux médecins, aux sociologues, aux archéologues et bien d'autres. Ces chercheurs ont été toujours accompagnés par des statisticiens et des mathématiciens qui ont pris la charge de traduire ces recherches en modèles et formules mathématiques.

Vers la fin des années 1980, la quantité considérable des données et le temps de traitement investi dans le classement manuel et l'accroissement de la puissance de traitement des ordinateurs, ont poussé les chercheurs à ce tourné vers la classification automatique. En effet, la majorité des recherches en classification, tous domaines confondus, ont été traduits en algorithmes, et ce en parallèle avec le perfectionnement continu des techniques de la classification automatique dans des domaines considérés jusque la comme nouveaux tels que la reconnaissance des formes (*pattern recognition*), la recherche d'information (*information retrieval*), le traitement d'images (*images processing*), etc., ce qui a donné un grand essor à cette classification.

2. Classification non supervisée, présentation

La classification non supervisée ou "*Clustering*" ou encore "*Cluster analysis*", est l'une des deux branches de la classification automatique, d'ailleurs elle ne doit pas être confondue avec la deuxième qui est la classification supervisée (*Classification*[1]) ou analyse discriminante (*Discriminant analysis*).

2.1. Classification non supervisée versus classification supervisée

Il est très important de comprendre la différence entre ces deux types. Dans le premier,

[1] Dans certains ouvrages en anglais le terme *Classification* est utilisé pour désigner la classification supervisée.

basé sur l'apprentissage supervisé, on dispose d'un classifieur (ou classificateur) déjà entraîné sur une collection d'objets étiquetés (modèles) dite d'entraînement ou d'apprentissage avec un nombre de classes connu à priori, et l'objectif est de classer tous les nouveaux objets non encore étiquetés [Duda *et al.*, 2001]. Ici l'étiquetage consiste à définir, pour le classifieur, les propriétés que doit avoir un objet pour appartenir à une classe spécifique. Cette tâche, étant manuelle, est évidement très couteuse et s'articule sur l'expertise humaine, qui peut être parfois subjective, d'ailleurs c'est le principal désavantage de ce type de classification.

Pour la classification non supervisée, le regroupement de l'ensemble d'objets non étiquetés en groupes appelés clusters, selon une mesure de similarité (ressemblance), se fera naturellement, sans connaissance préalable d'affectation des objets à ces groupes, ni sur les objets eux-mêmes [Jain et Dubes, 1988; Duda *et al.*, 2001]. Le regroupement est le plus souvent formalisé par l'objectif de définir des groupes d'objets disjoints ou non (soft ou hard plus loin) de telle manière que la distance entre les objets d'un même groupe soit minimale (propriété de la compacité), et que la distance entre ces groupes soit maximale (propriété de la séparabilité). Notons que ces deux contraintes vont dans deux sens opposés et c'est le meilleur compromis qui doit être trouvé.

Ici, il faut noter qu'un autre type de classification baptisé classification semi supervisée (*semisupervised classification*) a pris de l'importance au cours des années 2000 [Chapelle *et al.*, 2006; Jain, 2009]. Dans ce type de classification, on dispose en plus d'une petite collection étiquetée d'entrainement, d'un ensemble de contraintes qui spécifient l'appartenance (*must-link)* ou non (*cannot-link*) d'une paire d'objets au même cluster. Cette notion est particulièrement intéressante et bénéfique lorsqu'il y a absence d'une définition précise des clusters. Plusieurs travaux se sont intéressés à ce type de classification (on modifiant les deux contraintes de liens), nous citons ceux de [Basu *et al.*, 2002; Law *et al.*, 2005].

2.2. Cadre formel de la classification non supervisée

D'une manière plus formelle et plus générale, la classification non supervisée consiste à créer une partition ou une décomposition de cet ensemble en groupes telle que :

Critère 1 : les objets appartenant au même groupe se ressemblent;

Critère 2 : les objets appartenant à deux groupes différents ne se ressemblent pas.

Cette vision contraint donc à disposer d'une distance définie sur le langage de description des objets (plus loin nous allons voir les distances les plus utilisées).

2.3. Synonymes

Dans la littératures plusieurs synonymes simulés à la classification non supervisée (*clustering*) peuvent être retrouvés, nous citons: Apprentissage non supervisé *(unsupervised learning),* Taxonomie ou taxinomie numérique *(numerical taxonomy),* Quantisation vectorielle *(vector quantization)* et Apprentissage par observations *(learning by observation)* [Jain *et al.*, 1999; Xu et Wunsch, 2005]. Ces appellations sont en relation avec les domaines d'utilisation de cette classification, notamment en informatique (la recherche d'information, la

reconnaissance des formes, la segmentation d'images, etc.), en sciences de la nature et de la terre (biologie, zoologie, géographie, géologie, etc.), en médecine, en économie, etc.

3. Différentes approches de la classification non supervisée

Depuis la prise en forme des premières méthodes de la classification automatique au début des années 1950 et le développement des théories de cette dernière, des milliers d'algorithmes à la base des méthodes de classification non supervisée ont vu le jour. Nous ne prétendons pas pouvoir énumérer et décrire ces algorithmes et méthodes, mais dans la littérature beaucoup de productions scientifiques en *datamining* ont été dédiées à cette classification ou elle en a fait une grande partie. Des ouvrages tels que [Jain et Dubes, 1988; Kaufman et Rousseeuw, 1990b ; Everitt *et al.*, 2001; Duda *et al.*, 2001; Bishop, 2006 ; Han et Kamber, 2006] et des synthèses (*surveys*) telles que [Jain *et al.*, 1999; Berkhin, 2002; Xu et Wunsch, 2005; Jain, 2009; Rokach, 2010] représentent des grands classiques pour décrire les fondements, les algorithmes phares ainsi que les différentes applications de la classification non supervisée**.** Il faut noter que l'ouvrage de référence pour la majorité des publications sur cette classification est celui de [Sokal et Sneath, 1963].

Il ne faut pas oublier aussi que des travaux comme ceux de [Zamir et Etzioni, 1998; Steinbach *et al.*, 2000; Andrews et Fox, 2007; Aggarwal et Zhai, 2013] mettent la lumière sur les évolutions dans des différents domaines du *textmining*.

Tous les algorithmes décrits dans les références citées auparavant ainsi que leurs performances diffèrent et dépendent de plusieurs facteurs [Jain et Dubes, 1988; Berkhin, 2002; Jain, 2009], nous citons :

- type d'attributs de données traitées ;
- capacité de traitement d'un gros jeu de données ;
- capacité de traitement des données de hautes dimensions ;
- capacité de traiter et réactions aux observations aberrantes ;
- complexité de l'algorithme de la technique (temps de calcul) ;
- dépendance de l'ordre de l'arrivage des données ;
- dépendance des paramètres prédéfinis par les utilisateurs ;
- connaissance à priori sur les données ; etc.

3.1. Taxonomies

Dans la littérature, différentes taxonomies ont été données à ces algorithmes. Ces taxonomies ont été construites en se basant soit sur le fonctionnement de la formation de la structure finale soit sur la forme de la structure finale en elle-même.

Plusieurs publications comme [Xu et Wunsch, 2005 ; Jain, 2009] distinguent deux catégories fondamentales (d'ailleurs c'est la taxonomie la plus répandue). La première est celle des méthodes dites hiérarchiques, la deuxième est celle des méthodes dites méthodes par

partitionnements. Les algorithmes de ces deux catégories sont basés sur le calcul de la similarité entre les objets pris deux à deux.

D'autres publications comme [Zhong et Ghosh, 2003 ; Andrews et Fox, 2007] considèrent que les deux approches hiérarchiques et par partitionnements représente une seule catégorie dite discriminative ou basée sur la similarité (*similarity-based approches*). A coté de cette catégorie, une autre dite générative ou basée modèle (*model-based approches*). Les algorithmes de type génératif supposent que les objets à regrouper ont été générés à partir d'un un modèle et essaient de reconstituer ce modèle d'origine à partir de ces objets.

Une troisième taxonomie considère que les algorithmes de la classification non supervisée peuvent être classés selon, à la fois, le type des données traitées, le domaine d'application et le mode de fonctionnement [Jain *et al.*,1999 ; Berkhin, 2002; Xu et Wunsch, 2005]. Plusieurs catégories sont alors considérées notamment : les algorithmes hiérarchiques, les algorithmes par partitionnement, les algorithmes basés sur la densité, les algorithmes basés sur les grilles, les algorithmes statistiques, les algorithmes basés sur la théorie des graphes, les algorithmes basés sur la recherche stochastique, les algorithmes basés sur les réseaux de neurones, les algorithmes évolutionnaires, les algorithmes flow (*fuzzy*), etc.

Notant que [Bishop, 2006 ; Manning *et al.*, 2008] offrent un bon support pour le fondement théorique de toutes ces catégories.

3.2. Aspects discriminants

Chacune des taxonomies citées ci-dessus est construite sur un ou plusieurs aspects parmi lesquels nous citons:

- **Agglomération contre division:** Cet aspect détermine comment les clusters sont formés en ascendance ou en descendance ;
- **Hard contre Fuzzy :** Selon que les clusters soient totalement disjoints (classification en dur) ou se chevauchent (classification en flou) ;
- **Déterministe contre stochastique:** Dans les algorithmes déterministes toutes les possibilités de partitionnement sont examinées en quête de la solution optimale alors que dans les algorithmes stochastiques quelques partitions seulement sont examinées.
- **Monothéique contre polythétique:** Dans les algorithmes monothéiques les attributs des objets sont pris un par un dans le calcul de la distance alors que dans les algorithmes polythétiques le calcul de la distance se fait en prenant en compte tous les attributs en mêmes temps.

Notons que d'autres aspects peuvent être retrouvé dans [Jain *et al.*, 1999].

3.3. Principales approches

Dans les trois sections qui suivent nous présenterons brièvement les trois approches ou familles d'algorithmes les plus connues.

3.3.1. Méthodes hiérarchiques *(Hierarchical methods)*

Le but de ces algorithmes est l'obtention, à la fin du processus de la classification non supervisée, une hiérarchique de clusters appelée dendrogramme (figure 1.1). Ce dendrogramme permet aussi de voir comment sont formés les clusters. Ils sont de deux types : par agglomération (*buttom up*) et par division (*top down*).

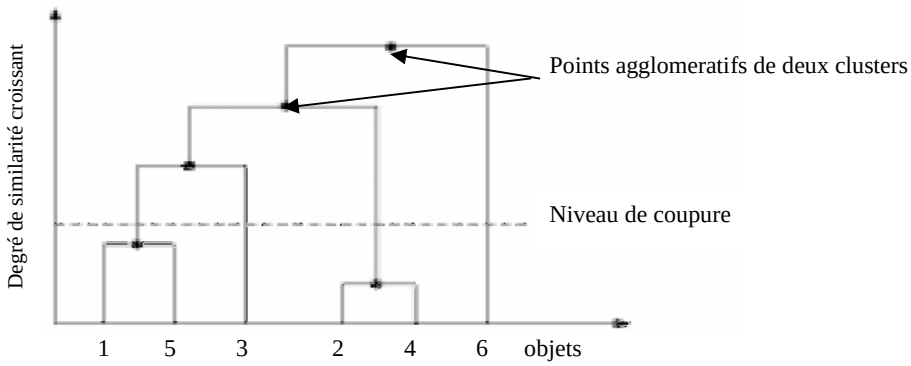


Figure 1.1: Forme type d'un dendrogramme.

a. Méthodes hiérarchiques par agglomération

Selon [Jain *et al.*, 1999], les algorithmes de la classification hiérarchique ascendante sont les plus connus des méthodes de la classification automatique. Le principe de ces algorithmes est résumé dans ce qui suit :

1. Le processus commence en considérant chaque objet comme étant un cluster et essaye de fusionner deux ou plusieurs clusters appropriés (selon une mesure de similarité) pour former un nouveau cluster.
2. Le processus est itéré jusqu'à ce que tous les points se trouvent dans un même cluster ou bien l'obtention d'un seuil pour lequel on coupe le dendrogramme.

Selon la manière avec laquelle la distance est calculée pour fusionner les clusters, quatre algorithmes sont les plus utilisés dans la plupart des méthodes hiérarchiques ascendantes: l'algorithme du lien simple (*single link*), l'algorithme du lien complet (*complete link*), l'algorithme du lien moyen (*average link*) et l'algorithme de *Ward* ou de la variance minimum (*minimum variance*).

b. Méthodes hiérarchiques par division

Dans ce type de méthodes, l'ensemble des objets sont dans un seul cluster au départ, puis, ce cluster est divisé récursivement en clusters plus raffinés selon des critères de division jusqu'à ce que chaque cluster contienne un seul point ou bien l'atteinte d'un seuil de division satisfaisant [Chavent *et al.*, 1999 ; Manning *et al.*, 2008]. Selon [Chavent *et al.*, 1999] ils existent deux principales catégories de ce type de méthodes, la première comporte les méthodes polythétiques telle que la méthode de *MacNaughton-Smith et Coll* (1964) *in*

[Chavent *et al.*, 1999], la deuxième comporte les méthodes monolithiques telle que la méthode de *Williams et Lambert* (1959) *in* [Chavent *et al.*, 1999].

Il est à noter que ce type de méthodes est conceptuellement plus complexe que celui par agglomération [Manning *et al.*, 2008], où il faut même faire appel à des techniques de partitionnement pour obtenir un niveau d'hiérarchie équilibré.

3.3.2. Méthodes par partitionnement ou à plat (*Partitional methods*)

A l'inverse des algorithmes hiérarchiques, les algorithmes de ce type produisent une seule partition d'objets, en d'autres termes, ils cherchent à diviser la population initiale en groupes disjoints [Jain, 2006], en optimisant une fonction objective qui est définie d'une façon locale (sur un sous-ensemble d'objets) ou globale (sur tous les objets) [Jain *et al.*, 1999].

a. Algorithme des K-Moyennes

Les méthodes à base de l'algorithme des K-Moyennes (K-Means) sont les plus connues en classification non supervisée avec partitionnements [Jain, 2009]. L'histoire de l'algorithme original pour ces méthodes est un peu confuse, mais généralement *J. McQueen* [MacQueen, 1967] est cité comme étant le précurseur dans ce domaine.

L'algorithme de *McQueen* est décrit par l'algorithme suivant :

1. Choisir aléatoirement *K* objets (centres) qui représentent les *K* clusters initiaux;
2. Affecter les objets aux clusters. Pour chaque objet *x*, le centre qui lui est assigné est celui qui lui est le plus proche, selon une mesure de distance[1];
3. Une fois que tous les objets sont affectés, recalculer les centres des *k* clusters;
4. Répéter les étapes 2 et 3 jusqu'à ce que plus aucune réaffectation ne soit possible (stabilisation).

Algorithme 1.1 : Algorithme de *McQueen.*

Schématiquement la figure 1.2 illustre le fonctionnement de cet algorithme. Dans cette figure nous remarquons que les centres des clusters changent de positions d'une itération à une autre, ceci est en fonction de :

- Représentation des centres de clusters ;
- Calculs des valeurs des centres de clusters.

Ces deux facteurs ont été à l'origine des modifications apportées au principe de l'algorithme de *McQueen*. En général, nous retrouvons trois principales catégories: les K-Moyennes, les K-Medoïds et les Nuées dynamiques.

La première catégorie connue sous le nom des K-Moyennes ou centres mobiles représente les centres des clusters par une moyenne *(mean)* ou une moyenne pondérée nommée "centroïde" [Steinbach *et al.*, 1999]. Plusieurs variantes de cette catégorie ont été élaborées, deux critères font la différence entre ces variantes :

[1] Habituellement la distance euclidienne est utilisée en *datamining.*

- Comment se fait la mise à jour des clusters ;
- Quand faut-il faire la mis à jour.

Pour le premier critère, les algorithmes *Standard K-Means* [MacQueen, 1967], *Lloyd* [Lloyd, 1982] et *Continuous K-Means* [MacQueen, 1967] représentent les algorithmes de références [Faber, 1994].

Pour le deuxième critère, deux variantes de l'optimisation itérative des K-Moyennes existent [Berkhin, 2002], à savoir l'algorithme de *Forgy* et l'algorithme d'optimisation itérative**.**

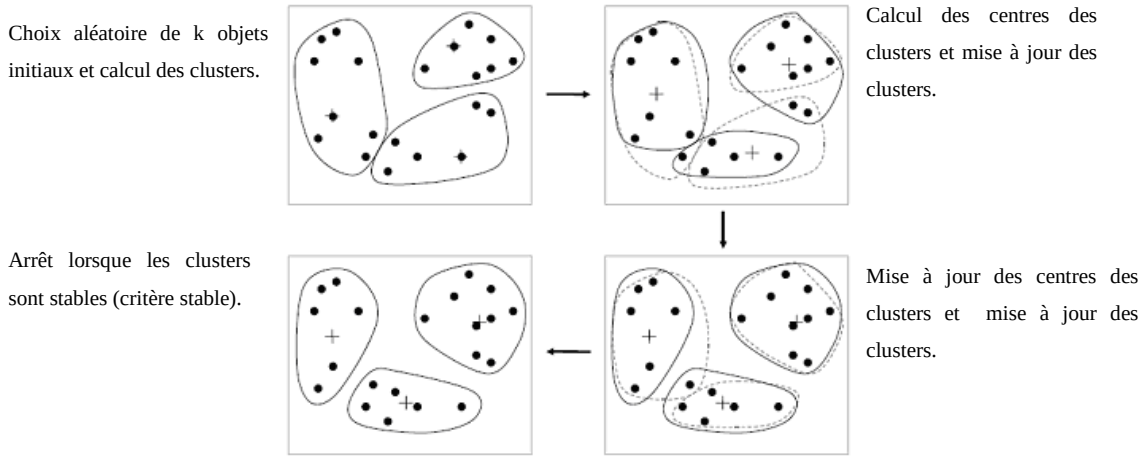


Figure 1.2 : Différentes étapes de l'algorithme *McQueen* [Pasquier, 2003].

b. Autres variantes d'algorithmes

D'autres algorithmes peuvent êtres retrouvés comme celui des K-medoïds où on représente un cluster par un de ses objets appelé médoïd qui minimise la somme des distances aux autres objets du même cluster. Des variantes de cet algorithme sont PAM *(Partitioning Around Medoids)* [Kaufman et Rousseeuw, 1990a], CLARA *(Clustering LARge Applications)* de [Kaufman et Rousseeuw, 1990b] et CLARANS *(Clustering Large Applications based on RANdomized Search)* [Ng et Han, 1994] qui représente une combinaison des deux premiers algorithmes PAM et CLARA.

L'algorithme des Nuées dynamiques [Diday, 1971] représente une autre approche des méthodes par partitionnement dans lequel le centre d'un cluster est représenté par ensemble d'objets réels appelé noyau supposé plus représentatif que le centre de gravité.

3.3.3. Méthodes génératives ou à base de modèles

A coté des deux approches exposées dans les sections 3.3.1 et 3.3.2 basées essentiellement sur le calcul de la distance entre les objets, une autre approche dite approche générative ou à base de modèle ou encore mélange de densités, est totalement différente. Les

algorithmes de cette approche supposent que les objets à regrouper ont été générés à partir d'un modèle de mélanges de probabilités et essaient de reconstituer ce modèle à partir de ces objets [McLachlan et Basford, 1988]. Le modèle de mélange gaussien est le modèle le plus populaire pour cette approche, chaque cluster est représenté par une densité de probabilité gaussienne. MCLUST [Fraley et Raftery, 1999] représente une implémentation de cette approche.

4. Différentes étapes dans le processus de classification non supervisée

Quelque soit l'approche utilisée dans un processus de classification non supervisée, ce dernier doit suivre quatre étapes essentielles :

4.1. Définition de la mesure de similarité appropriée au domaine d'application

Cette mesure est définie par une fonction de distance (similarité ou dissimilarité) entre deux objets. Cette fonction doit prendre en compte les caractéristiques des objets qui sont principalement, le type (discret, continu ou binaire) et l'échelle de mesures (Intervalle, nominale, ordinale, etc.).

Plusieurs distances ont été définies dans la littérature [Jain *et al.*, 1999; Berkhin, 2002; Rocach, 2010] et ce selon le domaine d'utilisation. En pratique, il peut s'agir de la distance euclidienne ou l'une des ses extensions telles que la distance de *Minkowski***,** la distance de *Manhattan*, ou la distance du *maximum*. On retrouve également la distance de *Pearson*, la distance du *cosinus* qui est généralement utilisée dans les représentations spatiales telles que la représentation VSM *(Vector Space Model)* [Salton *et al.*, 1975] en *textmining*.

4.2. Classification des objets

La classification peut être effectuée par l'une des approches citées dans la section 3.3. Le résultat obtenu peut être un regroupement en dur *(Hard clustering)* où chaque objet est affecté à un seul groupe (cluster), ou regroupement en flou *(Fuzzy clustering)* où chaque objet est affecté à plusieurs groupes avec un degré d'appartenance.

4.3. Abstraction des données

Consiste en l'extraction d'une simple et compacte représentation pour chaque cluster en vue d'une future analyse, cette abstraction peu être faite par la machine ou par un expert humain. Par exemple, cette représentation peut se résumer en un centroïde [Jain *et al.*, 1999].

4.4. Evaluation des résultats

Une fois les résultats sont générés par une technique de classification, la question est de prouver la pertinence ou la signification des clusters obtenus. Selon [Jain *et al.*, 1999; Steinbach *et al.*, 1999; Rocach, 2010], la validation des résultats ce fait par un examen interne *(internal quality)* qui consiste à évaluer la densité interne des clusters obtenus, ou un examen

externe *(external quality)*, qui consiste à comparer la structure obtenue avec une structure élaborée a priori, en utilisant des mesures telles que l'entropie *(Entropy)* ou la F-mesure (*F-measure)*.

5. Conclusion

Avant d'entamer les différentes parties intervenant dans notre sujet nous avons jugé utile de donner au lecteur un aperçu sur la classification non supervisée en général, pour marquer un peu la transition objets et textes.

Par ce premier chapitre nous avons voulu donner une présentation des différentes approches de la classification non supervisée en générale, nous avons survolé les trois approches principalement citées dans la littérature renforcées par des exemples de méthodes.

Nous avons ensuite détaillé le processus de la classification non supervisée par l'exposition de ses quatre phases, à savoir la définition de la mesure de similarité, la classification proprement dite, l'abstraction des données et enfin l'évaluation des résultats.

Nous espérons ainsi avoir donné un avant goût au lecteur pour entamer l'objet de l'étude en cours et le deuxième chapitre qui sera entièrement consacré à la classification non supervisée textuelle.

Chapitre II

Classification non supervisée textuelle

Chapitre II
Classification non supervisée textuelle

1. Introduction

Comme nous avons vu au cours du premier chapitre, la classification non supervisée a été sujet de plusieurs recherches dans plusieurs domaines. Avec l'augmentation du volume d'information contenu dans le support textuel, la classification non supervisée textuelle (CNST pour simplifier) ou documentaire[1] a suivi le même chemin et a été très attractive dans le monde de la recherche. L'apparition de nouvelles techniques telles que les modèles thématiques probabilistes a donné un nouveau souffle aux recherches sur cette CNST.

La majorité des recherches sur l'axe dans lequel la présente étude s'inscrit étant concentrées sur la CNST en ligne (*online*) [Navarro *et al.*, 2011], c'est-à-dire la CNST des résultats retournés par les moteurs de recherche sur le web, nous nous intéresserons dans ce qui suit à la CNST en hors ligne (*offline*) qui est totalement différente de celle en ligne que se soit en techniques de classification ou en volume de textes à classer.

2. Domaines privilégiés de la classification non supervisée textuelle

Toutes versantes dans le même axe (la recherche et l'exploration textuelle), plusieurs utilisations de la CNST sont distinguées. Ici nous nous sommes contentés de citer quelques domaines jugés les plus importants pour notre étude.

2.1. Recherche textuelle ou documentaire proprement dite

La recherche documentaire consiste à trouver les documents "**pertinents**"[2] (*relevant documents*) à une requête introduite par l'utilisateur, ordonnés dans une liste selon une estimation du degré de pertinence [Manning *et al.*, 2008]. La technique la plus utilisée dans les systèmes de recherche textuelle ou documentaire est la technique du *Ranking*, qui consiste à calculer un degré de similarité (ressemblance), mesuré par la distance entre chaque document de la collection et la requête [Salton, 1989; Harman et Voorhees, 1997; Harman et Voorhees, 1998]. Cette distance est utilisée pour donner un rang *(rank)* à chaque document qui va servir à son classement dans la liste retournée à l'utilisateur.

La liste de documents retournée par les systèmes basés sur cette technique est souvent très longue, l'utilisateur ne peut l'examiner entièrement laissant ainsi de côté certains

[1] Dans la littérature nous n'avons pas trouvé une distinction entre les deux appellations.
[2] Certains auteurs préfèrent employer le terme ressemblance entre documents et requête.

documents pertinents et mal positionnés. Deux causes sont en général à la source de cette défaillance. La première et que la requête n'exprime pas le besoin de l'utilisateur. La deuxième est que le système n'a pas su les prendre correctement en compte.

De nombreuses recherches ont essayé de trouver des solutions alternatives en se basant sur la classification non supervisée. Deux principaux axes se sont alors distingués pour l'application de cette classification pour l'amélioration du résultat d'une recherche documentaire.

2.1.1. Application de la classification non supervisée à toute la collection (approche globale)

C'est la première utilisation de la classification non supervisée en recherche documentaire. Ce genre de processus en recherche documentaire est principalement supporté par la fameuse "*Cluster Hypothesis*". Initiée par *G. Salton* (1968) *in* [Dubin, 2004] et reformulée par [van Rijsbergen, 1979], la recherche se base sur le fait que "les documents similaires sont pertinents pour les mêmes requêtes"[1]. Au départ une classification non supervisée est appliquée à la collection documentaire entière, pour regrouper les documents homogènes, avant le lancement d'une recherche sur celle-ci. Quand l'utilisateur introduit sa requête, l'algorithme de recherche renvoie les documents qui sont assortis à la requête introduite en calculant seulement la distance entre la requête et chaque représentant de cluster. La liste retournée est ensuite construite en prenant l'ordre de distance entre les documents et le représentant du cluster sélectionné à partir des clusters les plus proches de la requête jusqu'aux plus éloignés.

Le besoin de cette approche est expliqué par la limitation en puissance de calcul des ordinateurs de l'époque d'apparition de celle-ci pour le calcul de la distance entre chaque document et la requête en question[2]. Mais même avec des machines puissantes, la grande taille de représentation des documents représente un ralentisseur pour la classification non supervisée surtout si la collection est volumineuse. Néanmoins, des recherches ont essayé de contourné cet handicape soit en conduisant des expériences sur des collections avec peu de documents (travaux de [Maarek et Wecker, 1994] portant sur une bibliothèque en line), soit en réduisant la taille des vecteurs de représentation des documents (travaux de [Weiss *et al.*, 1996] qui se sont basés sur les informations fournies dans les liens hypertextes et le résumés des contenus, ou les travaux de [Anick et Vaithyanathan, 1997] qui se sont basés sur la représentation avec les concepts).

D'autres expériences se sont conduites sur des collections avec un nombre important de documents de tailles relativement petites comme celle de [Schütze et Silverstein, 1997]. Les auteurs ont comparé une recherche globale et une autre locale (application de la classification non supervisée sur la liste retournée par le système de recherche et composée d'un peu plus de 1000 documents), sur une collection de 74520 documents (de l'une des collections de la conférence TREC4[3]) répartis sur 400 clusters. Ils ont démontré que la recherche locale est meilleure que la recherche globale mais que cette dernière est plus rapide (étant donné qu'elle est effectuée une seule fois).

[1] Observé par *Keith Van Rijsbergen "closely associated documents tend to be relevant to the same requests"*
[2] *"Clearly in practice it is not possible to match each analysed document with each analysed search request because the time consumed by such operation would be excessive" G. Salton* (1968).
[3] *Text REtrieval Conference.*

Il faut noter que dans nos précédents travaux [Kelaiaia et Merouani, 2013a] nous avons montré, sur la collection CCA (*Corpus of Contemporary Arabic*) de nombre relativement petit de textes arabes [El Sulaiti, 2004], l'influence positive d'une classification non supervisée sur une recherche documentaire par rapport à une recherche classique.

2.1.2. Application de la classification non supervisée à la liste retournée par le système de recherche

Toujours basé sur le principe de la "*Cluster hypothesis*", le but de l'application de la classification non supervisée sur la liste retournée par le système de recherche est d'améliorer la position des documents pertinents, souvent dispersés sur cette liste, en les regroupant dans les mêmes clusters [Manning *et al.*, 2008].

Selon [Carpineto *et al.*, 2009], la première application est attribuée à [Cutting *et al.*, 1992; Cutting *et al.*, 1993] avec le système *Scatter/Gather*[1]. Ce dernier commence par une classification non supervisée de la collection entière sur *k* clusters, c'est la première étape, elle est appelée Dispersement *(Scattering)* et fourni un premier aperçu de la collection à l'utilisateur, qui dans la deuxième étape (*Query time*), choisi un ensemble de clusters l'intéressant. Ces clusters sont réunis par le système pour former une sous-collection de la collection originale, c'est la phase Ramassage ou Fusion *(Gathering)*. Puis le même processus *Scattering/Gathering* basé sur les choix de l'utilisateur, se réitère jusqu'à l'obtention de clusters très fins représentés chacun par un document. Ici le cycle fusion des sélections et reclassifications est considéré comment une classification des résultats retournés.

D'autres recherches qui ont suivi comme ceux de [Evans *et al.*, 1998] consistaient à parcourir les clusters et réordonner les résultats pour constituer une nouvelle liste, cela en isolant manuellement les clusters regroupant le maximum de documents pertinents.

2.1.3. Application de la classification non supervisée dans la recherche Web

A la fin des années 90, le boom Internet a orienté les recherches vers les moteurs de classification non supervisée dédiés au Web (*Web Clustering Engines*), ayant pour objectif la réorganisation des résultats retournés par les moteurs classiques depuis internet et les extranets. La principale difficulté rencontrée par les algorithmes à la base de ces moteurs et qu'en entrée, ils disposent seulement des titres et des résumés (ou entêtes) dits *snippets* des documents.

L'idée d'appliquer la classification non supervisée aux résultats de la recherche sur le web a été introduite pour la première fois dans le système *Scatter/Gather* par [Hearst et Pedersen, 1996] en se basant sur une variante de l'algorithme des K-Moyennes, puis, en 1998 [Zamir et Etzioni, 1998] présentent leur algorithme baptisé STC (*Suffix Tree Clustering*) ou classification non supervisée par arbre des suffixes, dans lequel les *snippets* (donc les documents web) partageant les mêmes phrases (séquences ordonnées de mots) sont regroupés dans les mêmes clusters. Beaucoup d'autres recherches se sont aussi intéressées à cet axe comme ceux de [Osinski *et al.*, 2004] avec l'algorithme Lingo et son principe DCF (*Description Comming First*) qui consiste à commencer par donner des descriptions aux clusters avant leurs formation, et l'algorithme SHOC (*Semantic Hierarchical Online*

[1] Certains chercheurs comme [Aggarwal et Zhai, 2013] considèrent "*Scatter/Gather*" comme une technique d'exploration.

Clustering) de [Zhang et Dong, 2004] qui construit une hiérarchie des documents en line en utilisant la décomposition en valeur singulière (SVD, *Singular Value Decomposition*).

Il faut noter que [Carpineto *et al.*, 2009; Carpineto et Romano, 2010] présentent une bonne étude comparative entre ces algorithmes et bien d'autres.

2.2. Exploration des collections textuelles

Au cours de leurs recherches sur *Scatter/Gather*, [Hearst et Pederson, 1996] ont remarqué qu'en balayant la structure exploratoire des clusters, l'utilisateur peut évaluer la pertinence des documents à l'intérieur des clusters et de trouver plus rapidement les informations intéressantes (ou du moins identifier les clusters non pertinents et les éviter).

2.3. Organisation des collections textuelles larges

La recherche textuelle et ses techniques décrites plus haut se concentrent sur la recherche des documents pertinents à une requête particulière. Ici, le défi est d'organiser la collection dans une taxonomie identique à celle crées par des humains. D'ailleurs les grandes conférences comme TREC proposent des échantillons de documents regroupés manuellement que les chercheurs participants à cette conférence essaient de reproduire le plus fidèlement possible.

Généralement la représentation hiérarchique est vue comme étant une représentation naturelle et systématique des clusters pour une bonne exploration des grandes échelles mais le coût excessif en ressources par les méthodes implémentant ce type de représentation pousse les recherches vers la représentation en partitions.

2.4. Classification supervisée textuelle

Dans le but d'améliorer les résultats de certaines recherches sur la classification supervisée, la CNST a été souvent utilisée, nous citons les travaux de [Bekkerman *et al.*, 2001] qui ont utilisé la CNST pour générer une représentation mot/cluster pour appliquer ensuite une classification basée sur les SVM (*Support Vector Machine*).

3. La classification non supervisée textuelle, du *Datamining* au *Textmining*

Les méthodes de classification non supervisée décrites dans le premier chapitre ont été développées pour les données non textuelles (en *datamining*). Beaucoup de travaux ont essayé de reconvertir plusieurs de ces méthodes (ex. les K-Moyennes) pour les utiliser avec les collections documentaires. Nous retrouvons même des plateformes (*Frameworks*) entières telles que BOW toolkit *(Bag Of Word toolkit)* [McCallum, 1996] et Lemur [Lemur project, 2014] qui implémentent un nombre important de ces méthodes. Ici nous distinguons deux principales approches:

- On considère que les textes sont des objets avec les mêmes attributs et on les traites avec les méthodes classiques avec des modifications minimes (ex. méthodes discriminantes) ;

- On considère que chaque texte est un objet à part constitué de plusieurs thèmes, chaque thème peut être traité à part (ex. modèles de mélanges thématiques).

La reconversion de ces méthodes devait s'adapter avec les propriétés de la représentation textuelle. Selon [Aggarwal et Zhai, 2013] la reconversion des méthodes orientées données non textuelles doit répondre aux propriétés particulières du texte suivantes :

- La dimension de la représentation des documents est très grande par rapport à la représentation réelle de l'information contenue dans ces documents ;
- Les corrélations qui peuvent exister entre les mots de l'espace de représentation d'un document engendrent un nombre de concepts beaucoup plus petit par rapport à cet espace ;
- Le nombre de mots représentatifs et non nuls dans les différents documents peut varier largement, d'où la nécessité d'une normalisation.

Pour rendre exploitables les textes par les méthodes de classification, ces textes doivent subir un traitement particulier décrit dans la section suivante.

3.1. Prétraitement

Pendant la construction des collections textuelles, le souci de la forme[1] dans laquelle sont rédigés les documents est loin d'être pris en considération. Avant de commencer le processus de la CNST, un prétraitement est requit pour représenter les documents de la collection sous une forme exploitable par cette CNST. Le but donc, de cette phase de prétraitement est de ne laisser, pour chaque document de la collection, que les mots représentatifs ou descripteurs. Plusieurs opérations sont généralement opérées:

3.1.1. Filtrage

Le but du filtrage est d'enlever les séquences de caractères qui pourraient bruiter et donc affecter la qualité de la CNST. Nous citons à titre d'exemple les caractères spéciaux tels que "%","$", "#", les métadatas, les éléments formatés (ex. les balises dans les documents XML), etc.

3.1.2. Tokenisation

Le processus de tokenisation est l'identification des mots et des phrases dans un texte. Une tokenisation simple pourrait utiliser l'espace blanc ou le retour chariot comme séparateur des mots. Dans le cas où l'on souhaite travailler avec des phrases les signes de ponctuation tels que '.', '?' et '!' peuvent être utilisés comme séparateurs.

3.1.3. Stemming

Le stemming (Désuffixation ou Radicalisation) est une technique morphologique largement utilisée pour la préparation des textes pour la classification. Il consiste à rechercher la racine lexicale ou stem pour des mots en langue naturelle, ceci, par l'élimination des affixes et des clitiques qui leurs sont rattachés, en d'autre terme regrouper sous un même identifiant

[1] Des pages web (ex. html, xml, etc…), des textes bruts ou encodés, des textes mis en forme (ex. documents Word), etc.

des mots dont la racine est commune. En langue arabe, par exemple, les mots : "يحْمِلُ", "حَمْلٌ", "حَمَلَة", "حَمَلَتْ" sont des flexions du stem "حَمَلَ".

Plusieurs algorithmes de stemming ont été proposés dans la littérature; pour la langue anglaise, le plus couramment utilisé est *Porter's stemmer*[1] [Porter, 1980], pour la langue arabe, on retrouve plusieurs *stemmers*, les plus connus sont : *Al-Stem*[2] **[**Darwish et Oard, 2002**]** et *StemmerLight10* [Larkey *et al.*, 2005].

Cette définition de stemming n'est pas unique dans la littérature, [Larkey, 2002] divise le stemming en langue en quatre catégories :

- Stemming à base de dictionnaires construits manuellement ;
- Stemming léger (*light stemming*) comme nous l'avons défini précédemment;
- Stemming à base d'analyse morphologique ;
- Stemming statistique qui utilise la classification pour regrouper les variantes qui découle de la même racine.

Les catégories une, trois et quatre forment en fait une seule catégorie du stemming, alors que la deuxième catégorie représente le stemming léger ou assoupli. Des recherches telles que [Chowdhury *et al.*, 2002 ; Larkey, 2002] montrent que les résultats obtenus par le stemming sur la langue arabe sont moins bons que ceux obtenus avec le stemming léger. Une autre bonne comparaison entre plusieurs techniques de stemming est dans [Eldesouki *et al.*, 2009].

Remarque importante

Il ne faut pas confondre entre le stemming et la lemmatisation *(Lemmatization)* qui consiste à retrouver l'entrée du dictionnaire (lemme) pour une forme fléchie d'un mot, en d'autre terme, rechercher des lemmes en remplaçant les verbes par leurs formes infinitives, les noms par leurs formes au singulier et regrouper des mots dont la signification est la même alors même que leurs racines sont différentes et ce en utilisant une analyse grammaticale.

En reprenant l'exemple précédant les mots : "حامِلٌ", "□محمو" en plus des mots cités sont des originaires du verbe "حَمَلَ". De cet exemple nous voyons bien que la lemmatisation est une tâche plus complexe que le stemming, puisque elle fait recours à des grandes bases de connaissances. L'algorithme *TreeTagger* [Schmid, 1994] est très connu en langues anglaise, française, allemande et italienne. Pour la langue arabe nous citons *Sebawai* de [Darwish, 2002].

Il faut noter que généralement un lemmatiseur est associer à un thésaurus qui, contrairement à un dictionnaire, ne donne pas d'informations relatives au sens et à l'emploi des mots, mais qui permet l'exploration à partir d'un concept (ou idée); les mots qui s'y rattachent et inversement.

[1] Sur le site http://www.comp.lancs.ac.uk/computing/research/stemming nous retrouvons une description détaillée du stemmer *Porter'stemmer*.

[2] *Al-Stem* a été modifié par *L.Larkey* pour qu'il puisse travailler avec l'encodage cp1256 (arabe Windows) http://www.microsoft.com/globaldev/reference/sbcs/1256.mspx.

3.1.4. Suppression des mots outils ou mots vide de sens (*Stop words*)

Cette étape consiste à éliminer les mots outils ou vides de sens (*Stop words*) qui peuvent eux aussi altérés le processus de la CNST. Généralement ces mots outils sont sauvés dans des fichiers spéciaux qui accompagnent les outils du prétraitement.

3.2. Indexation et représentation des documents

Selon la méthode de la CNST utilisée plusieurs représentations sont possibles. A partir du jeu de mots extrait de la phase du traitement appelé aussi vocabulaire, il est possible de représenter le document par un vecteur. Il est également possible d'utiliser des connaissances "a priori"[1] sur la façon dont les mots[2] sont répartis dans les documents suivant leur importance.

3.2.1. Représentation vectorielle (VSM, *Vector Space Model*)

La représentation des documents est étroitement liée à l'opération d'indexation. Cette opération consiste à attribuer un ensemble de mots appelés aussi descripteurs à chaque document de la collection. Formellement, si V est le vocabulaire (de dimension v) contenant les mots qui apparaissent au moins une fois dans la collection, un document d_j est transformé en un vecteur dont la représentation est la suivante:

$$d_j = (p_{j1}, p_{j2}, ..., p_{jv}) \quad (2.1)$$

p_{jk} est appelé poids correspond à la contribution du mot (descripteur) de rang k à la sémantique du document d_j.

En représentant chaque document de la collection sous le schéma 2.1 nous aurons une matrice dite document – descripteur. Cette représentation est la plus connue en recherche d'information, elle est connue sous le nom de représentation vectorielle ou VSM (*Vector Space Model*) [Dubin, 2004]. Ici l'emplacement des mots dans les documents n'est pas pris en considération d'où l'appellation parfois de cette représentation par "sac de mots" *(bag of words)*.

Pour mesurer la contribution d'un mot dans un document (p_{jk} dans le schéma 2.1) plusieurs techniques de pondération des mots ont été proposées :

a. Pondération binaire

Chaque document est représenté par un vecteur dans l'espace V (représentant le vocabulaire) dont les composantes informent sur la présence (valeur égale à 1) ou l'absence (valeur égale à 0) d'un mot dans un document.

D'une manière plus formelle le schéma 2.1 devient :

$$dj_{binaire} = \begin{cases} p_{ji} = 1 & \text{si le i}^{\text{ème}} \text{ mot} \in V \text{ apparait dans } d_j \\ p_{ji} = 0 & \text{sinon} \end{cases} \quad (2.2)$$

[1] Un exemple de ce type de représentation est dans les méthodes de détection thématique.
[2] Tout au long du présent mémoire nous ne faisons pas une distinction entre mot, terme et attribut.

b. Pondération fréquentielle

La pondération fréquentielle prend en compte le nombre d'apparitions d'un mot dans un document. Ainsi, un document est représenté dans l'espace V et chaque composante correspond au nombre d'apparition du mot correspondant dans le document. p_{ji} dans (2.2) devient donc le nombre d'apparitions du mot de rang i dans le document d_j.

Une normalisation de cette représentation consiste à diviser le nombre d'apparition du mot de rang i dans un document d_j par sa cardinalité.

c. Pondération TF-IDF

La pondération TF-IDF (*Term Frequency, Inverse Document Frequency*) est basée sur la loi de *Zipf* qui décrit la loi de répartition des mots d'un ensemble de documents [Salton et Buckley, 1988; Salton, 1989]. Cette pondération tente d'être plus informative que les pondérations précédentes en décrivant une relation entre les mots et leurs importances dans un document et dans toute la collection.

Formellement le schéma 2.2 devient :

$$d_{ji\ tf\text{-}idf} = p_{ji} * p_{vi} \qquad (2.3)$$

d_{ji} est la $i^{ème}$ composante du vecteur d_j;
p_{ji} poids du $i^{ème}$ mot dans le $j^{ème}$ document;
p_{vi} poids du $i^{ème}$ mot dans la collection (v représente la cardinalité du vocabulaire de la collection).

Le modèle le plus classique pour calculer ces deux poids est la fréquence du mot dans le document notée tf_i^d pour le premier et $log\ (|D|/df_i)$ pour le deuxième. $|D|$ et df_i représentent, respectivement, le nombre de documents de la collection et df_i est le nombre de documents qui contiennent le mot du rang i. En résultat nous aurons :

$$d_{ji\ tf\text{-}idf} = tf_i^{dj} * log\left(\frac{|D|}{df_i}\right) \qquad (2.4)$$

Il faut noter que plusieurs normalisations pour la représentation TF-IDF et bien d'autres ont été testées dans des travaux tels que [Salton et Buckley, 1988; Singhal *et al.*, 1996].

3.2.2. Autres formes de représentation

Les techniques citées dans la section précédente sont spécialement dédiées à la représentation vectorielle. Plusieurs autres techniques de représentation ont été proposées dans la littérature, nous citons : indexation par les phrases de **[**Fuhr et Buckley, 1991; Caropreso *et al.*, 2001; Boulaknadel, 2005**] ;** indexation par les racines lexicales [Aljlayl et Frieder, 2002], indexation par les n-grammes **[**Miller *et al.*, 1999; Keselj *et al.*, 2003], etc.

Il faut noter aussi que le principe sur lequel sont bâties certaines méthodes (ex. méthodes des modèles thématiques probabilistes) représente lui-même une forme de représentation des documents.

3.2.3. Représentation par la réduction de l'espace de représentation

Une autre vision pour la représentation des documents consiste carrément à transformer l'espace de représentation en un espace plus réduit. D'après [Sebastiani, 2002] il existe deux approches pour réduire la dimension de l'espace de représentation des textes :

a. Sélection des attributs *(Features selection)*

Ici un score est associé à chaque attribut pour déterminer son degré de pertinence pour un document donné, les attributs ayant les scores les plus faibles sont éliminés. Plusieurs techniques de sélection d'attributs ont été développées en vue de réduire la dimension de l'espace de représentation. Chacune de ces techniques utilise des critères lui permettant de rejeter les attributs jugés inutiles à la tâche de classification en générale. Parmi les critères qu'utilisent les techniques de sélection d'attributs, nous retrouvons [Yang et Pederson, 1997; Aggarwal et Zhai, 2013]: la fréquence dans le document *(document frequency)*, le gain d'information *(information gain)*, l'information mutuelle *(mutual information)*, la statistique du *chi-2* (χ^2), la force du terme *(term strength)*, etc.

b. Extraction de termes

Les techniques d'extraction de termes, construisent un sous ensemble à partir d'une combinaison algébrique linéaire des attributs de l'espace de représentation, pour maximiser la performance de la classification et éliminer les problèmes liés aux synonymies, polysémies et homonymies.

L'indexation par la sémantique latente (*Latent Semantic Indexing*, LSI) [Deerwester *et al.*, 1990] qui a essayé de répondre au problème de la synonymie, a été la première méthode dans ce domaine, puis des méthodes conçues au départ à la détection thématique se sont imposées, notamment, l'analyse sémantique latente probabiliste (*Probabilistic Latent Semantic Analysis*, PLSA) [Hofmann, 1999a; Hofmann, 1999b], la Factorisation en Matrice Non-négative (*Non-negative Matrix Factorisation*, NMF) [Lee et Seung, 1999; Lee et Seung, 2001], l'allocation latente de *Dirichlet* (*Latent Dirichlet Allocation*, LDA) [Blei *et al.*, 2003]. Dans la section 4 nous allons revenir sur la représentation des documents dans cette dernière méthode.

3.3. Mesures de similarité en classification non supervisée textuelle classique

Dans une CNST utilisant l'une des techniques classiques, la similarité entre deux documents peut être mesurée par plusieurs métriques telles que la distance euclidienne, la distance du cosinus la distance de *Kullback Leibler*, etc.

Dans l'espace vectoriel, par exemple, la similarité entre deux documents représentés par leurs deux vecteurs respectifs est calculée à l'aide d'une corrélation quelconque entre les deux vecteurs. Une telle corrélation peut être, par exemple, celle du cosinus de l'angle formé entre ces deux vecteurs. La mesure du cosinus est une technique qui découle du faite que, si deux vecteurs ont approximativement, les mêmes attributs alors ils pointent vers la même direction dans l'espace de représentation (figure 2.1). Donc pour calculer la similarité entre deux documents représentés par leurs deux vecteurs d_i et d_j en utilisant le cosinus, on a :

$$sim(d_i, d_j) = \cos(\alpha) = \frac{d_i . d_j}{|d_i \| d_j|} = \frac{\sum_k d_{ik} . d_{jk}}{\sqrt{\sum_k d_{ik}^2 . \sum_k d_{jk}^2}} \quad (2.5)$$

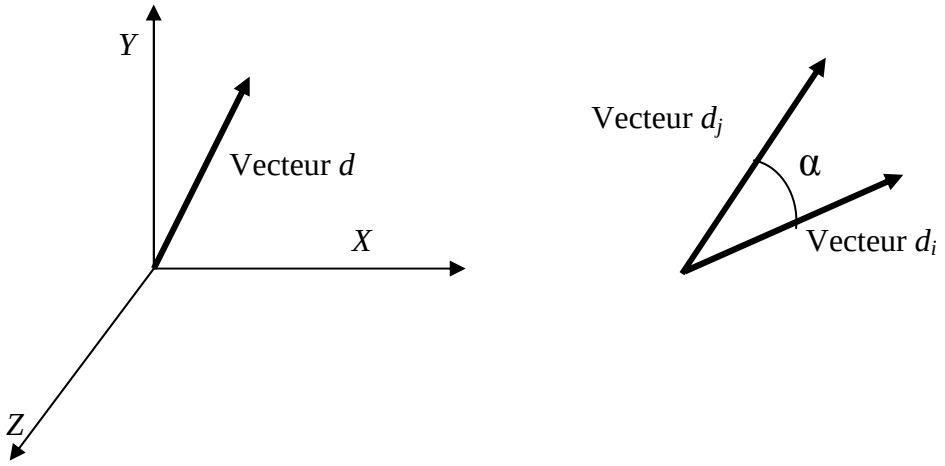


Figure 2.1 : Représentation et Similarité entre deux documents dans l'espace vectoriel.

4. Les modèles thématiques et la classification non supervisée textuelle

Les modèles thématiques probabilistes (*Topics models*) ont investi, ces dernières années plusieurs domaines du *textmining* comme la recherche d'information, le résumé automatique la CNST, etc. [Lu *et al.*, 2011]. Ils sont basés sur le principe de création d'un modèle probabiliste pour les documents de la collection en considérant que :

- un document comme un mélange probabiliste de thématiques latentes ;
- une thématique est définie par une distribution de probabilités sur les mots.

Les modèles thématiques probabilistes proposés dans la littérature partagent globalement le principe exposé ci-dessus, mais diffèrent principalement dans la manière de choisir les distributions de probabilités des thématiques sur les documents et les distributions de probabilités des mots sur les thématiques.

Avec les deux modèles les plus connus, à savoir, PLSA (*Probabilistic Latent Semantic Analysis*) [Hofmann, 1999a] et LDA (*Latent Dirichlet Allocation*) [Blei *et al.*, 2003], PLSA n'impose aucune hypothèse sur la distribution des thématiques sur les documents (à ce stade ce modèle n'est pas probabiliste); c'est-à-dire que chaque document est traité à part, d'ailleurs c'est le défaut majeur qui la rend non générative pour les documents n'appartenant pas à la collection. Dans LDA, par contre, chaque thématique est caractérisée par une distribution multinomiale sur les mots qui lui sont associés. La loi de *Dirichlet* est utilisée pour permettre

un choix judicieux des paramètres des distributions multinomiales, ce qui permet de pallier aux limites de PLSA. Dans ce qui suit nous nous intéresserons à cette méthode.

4.1. LDA, le modèle

Comme il a été mentionné auparavant, LDA repose sur l'idée qu'une collection textuelle de *D* documents est modélisée sur une mixture de *K* thèmes (ou sujets), chaque thème est une distribution probabiliste de *V* mots (figure 2.2) où *V* est le vocabulaire de toute la collection. Le modèle LDA initial décrit dans [Blei *et al.*, 2003] est complexe et a été sujet de plusieurs recherches. Nous présentons dans ce qui suit le processus génératif le plus répondu (figure 2.2) [Blei *et al.*, 2003; Wei et Croft, 2006 ; Steyvers et Griffiths, 2007]:

1. pour chaque thème *z*, une distribution multinomiale $P(w|z)$ spécifiée par ϕ_z qui définie les occurrences des mots du vocabulaire *V* est échantillonnée avec la distribution de *Dirichlet* de paramètre β. Cette distribution peut être interprétée comme la probabilité de l'occurrence d'un mot *w* dans un document du thème *z*.
2. pour chaque document *d*, une distribution multinomiale $P(z)$ spécifiée par θ_d sur les thèmes est échantillonnée avec la distribution de *Dirichlet* de paramètre α.
3. pour chaque mot *w* dans le document *d*, un seul thème est choisi selon la distribution multinomiale θ_d.
4. chaque mot est échantillonné avec une distribution multinomiale ϕ_z.

Ce processus est représenté dans la figure suivante :

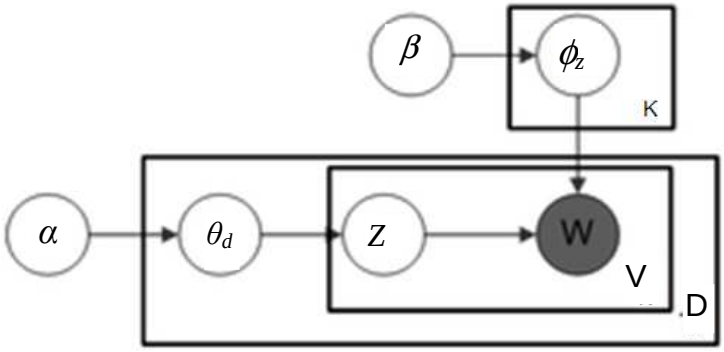


Figure 2.2 : Modèle graphique de LDA [Steyvers et Griffiths, 2007].

Ici β peut être perçu comme étant l'observation a priori du nombre de fois où un mot est échantillonné à partir d'un thème *z*, et α comme étant l'observation a priori du nombre de fois où un thème est échantillonné dans un document *d*, avant d'avoir observé les mots réels de ce document.

Les paramètres ϕ et θ indiquent, respectivement, quels sont les mots importants pour un thème donné et quels sont les thèmes importants pour un document donné, ils sont définis comme suit :

4.1.1. Loi de *Dirichlet*

La distribution de *Dirichlet* de dimension *K* permet de définir une distribution multinomiale $\theta=(\theta_1, \theta_2, \ldots, \theta_K)$ sur *K-1* simplex régulier, telle que $\forall i, \theta_i \geq 0$ et $\sum_{i=1}^{k} \theta_i = 1$. Sa densité est de la forme:

$$P(\theta_d|\alpha) = Dirichlet(\theta_d|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \left(\theta_{d\,i}\right)^{\alpha_i - 1} \quad (2.6)$$

L'équation 2.6 permet d'obtenir une distribution multinomiale de paramètre θ pour le mélange de thèmes dans le document *d*, avec $\alpha \in R^k$, $\alpha_i > 0$ et $\Gamma(x)$ est la fonction *Gamma* (l'extension de la fonction factorielle). Selon [Blei *et al.*, 2003; Steyvers et Griffiths, 2007 ; Lu *et al.*, 2011], il est plus commode mathématiquement, pour optimiser les inférences, d'utiliser une loi de *Dirichlet* symétrique avec un seul hyper-paramètre α en prenant $\alpha = \alpha_1 = \alpha_2 = \ldots = \alpha_k$. L'équation 2.6 devient alors :

$$P(\theta_d|\alpha) = Dirichlet(\theta_d|\alpha) = \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k} \prod_{i=1}^{k} \left(\theta_{d\,i}\right)^{\alpha - 1} \quad (2.7)$$

4.1.2. Génération de la collection

Pour générer la collection de documents un placement d'une *Dirichlet* symétrique a priori sur la distribution des thèmes θ, donne une distribution lissée (*smoothed distribution*) avec un degré de lissage déterminé par la valeur de α. De même, un placement d'une loi de *Dirichlet* symétrique a priori avec l'hyper-paramètre β sur la distribution ϕ permet d'obtenir une distribution lissée de mots sur les thèmes . Donc étant donnés les hyper-paramètres α et β, la probabilité de la collection entière est obtenue en intégrant sur θ et ϕ [Blei *et al.*, 2003; Steyvers et Griffiths, 2007 ; Lu *et al.*, 2011]:

$$P(D_{1..N}|\alpha,\beta) = \int_{\phi_1} \ldots \int_{\phi_k} \prod_{i=1}^{K} P(\phi_i/\beta) \prod_{d=1}^{N} \int P(\theta_d/\alpha) \prod_{w \in N_d} \left(\sum_{j=1}^{k} \theta_{dj} \phi_{jw} \right)^{c(w,d)} d\,\theta_d d\phi_1 \ldots d\phi_k \quad (2.8)$$

où $c(w,d)$ est le nombre de fois où le mot w apparait dans le document d, K est le nombre de thèmes et N est le nombre total des documents dans la collection.

Pour bien comprendre le modèle LDA reprenons-le sous la forme linéaire algébrique (figure 2.3). Selon [Steyvers et Griffiths, 2007] nous pouvons dire que le produit de ϕ (matrice à colonnes stochastiques mot par thème) et θ (matrice à colonnes stochastiques thème par document) est la représentation originale de la matrice originale mot par document.

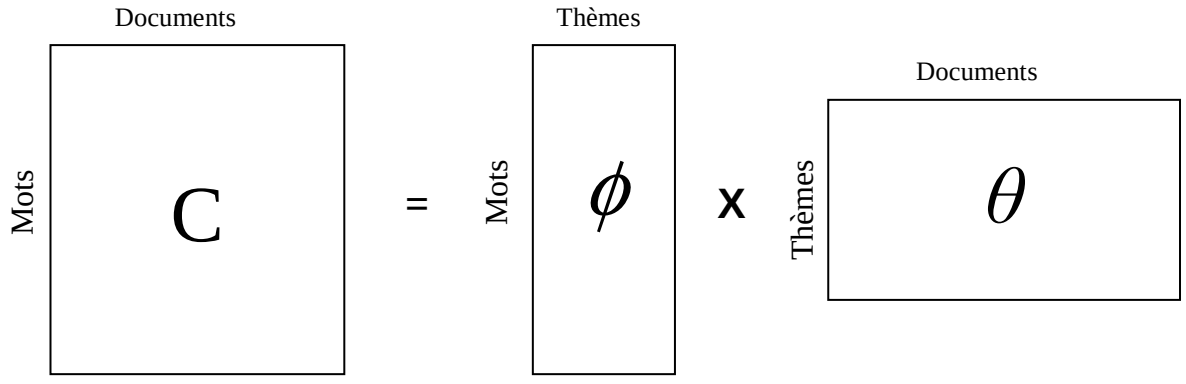


Figure 2.3: Modèle algébrique de LDA [Steyvers et Griffiths, 2007].

4.2. Inférence et estimation des paramètres

Les variables et les paramètres du modèle décrit plus haut ne sont pas connus initialement, il faut essayer de les apprendre à partir des données observables, c'est-à-dire les mots des documents. Etant donnés les hyper-paramètres α et β, le rôle de l'inférence est de déterminer les variables cachées θ et ϕ d'un document *d* à partir des mots de ce document.

Le modèle LDA décrit dans [Blei *et al.*, 2003] est complexe et ne peut être résolu par une inférence exacte. Plusieurs techniques d'inférence approximatives ont été proposées dans la littérature, notamment : les méthodes variationnelles (*variational methods*) [Blei *et al.*, 2003], l'espérance – propagation (*expectation propagation)* de [Griffiths et Steyvers, 2004], et l'échantillonnage de *Gibbs* (*Gibbs sampling*) de [Griffiths et Steyvers, 2004] basé sur les méthodes *Chaines de Markov* (*Markov Chain Monte Carlo*, MCMC). Dans le cinquième chapitre nous allons revenir en détail sur l'estimation des paramètres en utilisant l'échantillonnage de *Gibbs.*

4.3. La classification non supervisée textuelle et LDA

Selon [Lu et al, 2011] il y a deux approches pour l'utilisation des modèles thématiques en CNST. La première approche utilise ces modèles pour transformer la représentation originale de grande dimension des documents (basée mots) en une représentation de dimension réduite (basée thèmes), puis applique un algorithme standard comme les K-Moyennes sur cette nouvelle représentation.

L'autre approche repose sur l'hypothèse où chaque thème *z*, après estimation des paramètres θ et ϕ, devient un nouveau cluster et les documents sont assignés à ce cluster selon l'équation suivante :

$$arg\,max_{Z=1..K}\,\theta_d \tag{2.9}$$

4.4. Mesures de similarité

Dans les techniques classiques de la CNST, la similarité entre deux documents est calculée en mesurant la distance entre ces documents. Dans les modèles thématiques le principe est totalement différent, la similarité est calculée entre les mots d'une part, et d'autre part entre les thèmes.

4.4.1. Similarité entre les documents

Deux documents d_1 et d_2 sont similaires si leurs distributions thématiques θ_{d1} et θ_{d2} sont proches. La divergence *Kullback Leibler* (*KL*) [Kullback et Leibler, 1951] est la mesure par excellence de la divergence ou dissimilarité entre deux distributions. Si *p* et *q* représentent deux distributions de deux documents, alors la divergence de *Kullback Leibler* est mesurée comme suit:

$$KL(p,q) = \sum_{j=1}^{K} p_j log_2 \frac{p_j}{q_j} \tag{2.10}$$

Ici il faut noter que la fonction *KL(p,q)* est asymétrique et dans la majorité des applications on applique une fonction symétrique basée sur cette fonction :

$$KLS = \frac{1}{2}[KL(p,q) + KL(q,p)] \quad (2.11)$$

Une autre mesure est aussi utilisée, c'est la mesure de divergence *Jensen-Shannon (JS)* symétrique [Lin, 2002], elle est calculée selon :

$$JS = \frac{1}{2}[KL(p,(p+q)/2) + KL(q,(p+q)/2)] \quad (2.12)$$

L'équation 2.12 mesure la similitude entre p et q par l'intermédiaire de leur moyenne. Deux distributions p et q sont similaires si elles sont similaires à leur moyenne $(p + q) /2$.

Selon [Steyvers et Griffiths, 2007] il est également possible de considérer les distributions de thèmes en tant que vecteurs et appliquer les mesures classiques telles que la distance euclidienne ou le cosinus.

4.4.2. Similarité entre les mots

Deux mots m_1 et m_2 sont similaires s'ils apparaissent dans le même thème [Steyvers et Griffiths, 2007]. Cette similarité peut être calculée via la similarité entre les deux probabilités conditionnelles $\theta_1 = P(z|m_i=m_1)$ et $\theta_2 = P(z|m_i=m_2)$ respectives pour ces deux mots. Chacune de ces deux probabilités peut être calculée par l'une des deux mesures symétriques la divergence de *Kullback Leibler* ou la divergence de *Jensen-Shannon*.

5. Mesures d'évaluation de la classification non supervisée textuelle

De nombreuses mesures d'évaluation des résultats renvoyés par une classification non supervisée textuelle sont présentées dans la littérature. De même que dans la classification non supervisée classique (mentionnés dans le premier chapitre) deux types d'examens sont distingués :

5.1. Examen interne *(Internal quality)*

C'est le premier critère d'évaluation de la structure des clusters générée. Ici la structure de chaque cluster est évaluée à travers sa densité en calculant une mesure dite similarité globale *(overall similarity)*. Cette mesure est égale à la moyenne des similarités entre tous les documents du même cluster pris deux à deux.

5.2. Examen externe *(External quality)*

Consiste à comparer la structure obtenue par la classification non supervisée (ensemble de clusters) avec une structure élaborée a priori (ensemble de classes), en utilisant l'une des mesures qui existent dans la littérature. Dans ce qui suit nous allons exposer quatre des mesures les plus utilisées :

5.2.1. Indice de *Rand*

L'indice de *Rand* [Rand, 1971] calcule la similarité entre deux partitions dans le but d'évaluer l'exactitude ou l'accord de la classification obtenue par rapport à une

autre préétablie. Il est calculé en examinant l'appartenance des documents d'un cluster pris deux par deux à la même classe. On dit alors, que :

- Une paire de documents est une vraie positive (*VP*) si les deux documents sont de la même classe et sont placés dans le même cluster ;
- Une paire est vraie négative (*VN*) quand les deux documents sont de classes différentes et sont placés dans deux clusters différents ;
- Une paire fausse positive (*FP*) correspond à deux documents de classes différentes placés dans le même cluster.
- Une paire fausse négative (*FN*) correspond à deux documents de la même classe dans deux clusters différents.

L'indice de *Rand* est alors défini selon l'équation suivante:

$$RI = \frac{(VP+VN)}{(VP+FP+FN+VN)} \tag{2.13}$$

5.2.2. Indice de *Jaccard*

L'indice de *Jaccard* [Hamers *et al.*, 1989] est similaire à celui de *Rand* à l'exception qu'il ne tient pas compte des paires vraies négatives. Il est calculé selon l'équation:

$$JI = \frac{VP}{(VP+FP+FN)} \tag{2.14}$$

5.2.3. F-mesure (*F-measure*)

C'est la mesure la plus connue en recherche d'information. La F-mesure [Van Rijsbergen, 1979 ; Larsen et Aone, 1999] est une combinaison harmonique de deux métriques célèbres, la précision et le rappel. Selon le domaine d'utilisation, plusieurs définitions et formules de calcul ont été données à ces deux métriques. Selon [Manning *et al.*, 2008] elles sont calculées par les équations suivantes :

$$P = \frac{VP}{(VP+FP)} \tag{2.15}$$

$$R = \frac{VP}{(VP+FN)} \tag{2.16}$$

La F-mesure est alors calculée comme suit :

$$F - mesure = \frac{2*P*R}{(P+R)} \tag{2.17}$$

5.2.4. Entropie *(Entropy)*

L'entropie représente la distribution des classes sur les clusters produits [Zhao et Karypis, 2001]. Elle calcule le degré de désordre dans un cluster [Shannon, 1948]. Une entropie faible reflète un cluster homogène alors qu'une entropie élevée reflète le contraire. L'équation suivante représente l'entropie d'une partition de *k* clusters :

$$Entropie = \sum_{j=1}^{k} \frac{n_j}{n} E_j \tag{2.18}$$

n_j représente le nombre de documents dans le cluster *j*, *n* représente le nombre total des

documents dans la collection et E_j représente l'entropie du cluster *j*, elle est calculée selon l'équation :

$$E_j = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_{ij}}{n_j} log \frac{n_{ij}}{n_j} \quad (2.19)$$

n_{ij} est le nombre de documents de la classe *i* assignés au cluster *j*, *q* est le nombre total de classes dans la collection et n_j est le nombre de documents dans le cluster *j*.

6. Conclusion

Dans ce chapitre nous avons essayé de donner un aperçu des récents développements en classification non supervisée textuelle. Nous avons commencé par donner les domaines d'utilisation de cette dernière, nous avons ensuite passé en revu les différentes opérations relatives à une CNST.

La deuxième partie de ce chapitre a été consacrée à LDA, une technique très réputée parmi les modèles thématiques probabilistes et son emploi en CNST.

Nous avons ensuite enchainé par les mesures de similarité utilisées en classification non supervisée classique ainsi que celles utilisées dans les modèles thématiques probabilistes. Nous avons constaté que le principe de mesure de la similarité entre les deux familles (classiques et thématiques probabilistes) est totalement différent.

Nous avons terminé ce chapitre par la présentation d'une panoplie de mesures d'évaluation des résultats de la CNST notamment, l'indice de *Rand*, celui de *Jaccard*, la F-mesure et enfin l'Entropie.

Quelque soit le degré de performance des méthodes de CNST exposées plus haut, la compréhension des contenus des résultats obtenus (clusters) par ces méthodes reste plus au moins difficile par l'utilisateur, une description leurs sera indispensable. Dans le chapitre suivant nous aborderons les différentes techniques de la description ou encore de la labellisation des résultats de la CNST.

Chapitre III

Description ou labellisation

Chapitre III
Description ou labellisation

1. Introduction

Dans les différents domaines où la CNST est utilisée, la présentation des clusters “muets”, “nus” ou sans descriptions risque de faire perdre un temps précieux aux utilisateurs. En effet n’ayant aucune indication sur le contenu des clusters, les utilisateurs risquent de se désorienter entre les différents clusters, en quête de l’information recherchée. Nombreuses sont les recherches qui se sont intéressées à développer des méthodes de la CNST, mais relativement un nombre réduit est destiné à la description des résultats de cette dernière.

La description ou la labellisation ou encore l’étiquetage des clusters générés par un processus de classification non supervisée consiste à décrire ces clusters par des étiquettes (*labels*) qui peuvent être des mots ou des phrases. Deux questions sont alors naturellement posées :

1. Etant donné un ensemble de clusters, quelle technique employer pour décrire chacun de ces clusters ?
2. Commet évaluer la pertinence de ces descriptions?

Pour répondre à ces deux questions les recherches se sont multipliées pour donner une panoplie de techniques qui seront exposées dans la section suivante. La majorité de ces techniques de description sont destinées aux résultats des recherches web, qui sont généralement des “bouts” de textes connus sous l’appellation de *snippets*, et ne prennent pas en considération la totalité du texte original à la source de ces *snippets* [Navarro *et al.*, 2011].

2. Quelles conditions pour une bonne description

Quelque soit le type de la description adoptée, plusieurs recherches telles [Zhang *et al.*, 2009; Stefanowski et Weiss, 2007] se sont mis d’accord sur un ensemble d’exigences que cette description doit remplir :

Concision *(Conciseness)* qui signifie que la description du cluster doit être aussi courte que possible, mais suffisante pour reflété le contenu de ce cluster.

Compréhensibilité (*Comprehensibility*) qui indique la clarté de la description elle-même pour la bonne description du cluster ;

Précision *(Accuracy)* ou **transparence** (*transparency*) qui indique la capacité de refléter tout le contenu du cluster ;

Distinction *(Distinctiveness)* qui signifie que les mots descriptifs doivent fréquemment apparaitre dans les clusters qu'il représente et apparaissent rarement dans les autres clusters.

Pour mesurer le degré de satisfaction d'une description de ces exigences plusieurs heuristiques ont été employées. Par exemple la mesure la plus simple de concision des mots descriptifs est la longueur. Elle peut être aussi mesurée par le nombre de caractères ou de mots dans la phrase descriptive. Reste à savoir que le jugement humain présente un facteur incontournable et unanimement le meilleur moyen pour mesurer le degré de cette satisfaction.

3. Techniques de description

Pour décrire une CNST l'approche la plus intuitive est de faire recours à l'expertise humaine pour parcourir les différents clusters et donner des descriptions manuelles comme celles que l'on retrouve dans les travaux de [Glenisson *et al.*, 2005; Lai et Wu,2005]. Deux problèmes se présentent alors :

- Le nombre de documents dans la collection est très important ;
- L'objectivité des descriptions données par l'expertise humaine (chaque expert à sa propre vision sur les documents lus)[1].

Ces deux inconvénients ont poussés les chercheurs à se tourner vers la description automatique. Dans la littérature, il ya plusieurs techniques possibles pour construire une description. Un système parfait lit tous les documents et génère une description appropriée, incluant, si besoin, des mots qui ne figurent pas dans ces documents[2]. Ici nous n'étudions que les techniques qui s'articulent autour des contenus des documents.

3.1. Description par les mots clés (*Keys-words*)

C'est la technique la plus répondue en description. Elle consiste à choisir, selon des mesures de pondération, les mots les plus fréquents dans le cluster pour le décrire. Les précurseurs de cette technique sont [Cutting *et al.*, 1992 ; Cutting *et al.*, 1993 ; Hearst et Pedersen, 1996] qui utilisent une fréquence normalisée comme poids des mots les plus représentatifs. Selon [Manning *et al.*, 2008] deux catégories de cette techniques sont distinguées :

3.1.1. Description interne des clusters (*Cluster-internal labeling*)

Cette description sélectionne les mots clés en s'appuyant uniquement sur le contenu du cluster à décrire. Un calcul de la fréquence des mots *(term frequency*, TF) selon l'équation 3.1 est alors effectué[3], les mots avec les poids *P* les plus grands sont retenus.

[1] Ici mieux vaut donner à l'expert humain une description et le laisser juger sa pertinence que de le laisser lui même définir cette description.
[2] Ces techniques font intervenir des ressources autres que les documents telles que *WordNet* [Hotho et Stumme, 2002] et ne font pas partie de la présente étude.
[3] Un calcul simple des poids des mots clés peut être envisagé en ne prenant que la fréquence d'apparitions des mots dans la collection.

$$P_{m,d} = \begin{cases} 1 + \log(TF_{t,d}) & si \quad TF_{t,d} > 0 \\ 0 & sinon \end{cases} \tag{3.1}$$

Deux cas de figures sont envisageables, soit que seuls les mots du document le plus proche du centroïde sont pris en compte, soit que tous les mots du cluster le sont.

3.1.2. Description différentielle (*Differential cluster labeling*)

Dans cette description, les mots les plus fréquents sont ceux qui différencient un cluster par rapport à un autre. Dans la littérature nous retrouvons plusieurs pondérations pour ces mots. La pondération par IDF *(Inverse Document Frequency)* calculée par l'équation 3.2, représente l'une des plus connues.

$$IDF_m = log \frac{N}{1+df_m} \tag{3.2}$$

où N est le nombre de document dans la collection et df_m est le nombre de documents dans lesquels le mot m apparait.

D'autres pondérations comme la pondération par l'information mutuelle (*Mutuelle information*, MI), la pondération par le test Chi carré (χ^2) sont aussi utilisées. Il faut aussi noter qu'une troisième catégorie dite Hybride [Niu *et al.*, 2012] comprend des techniques de pondération telles que la fameuse TF-IDF (déjà présentée) dans les travaux de [Yang *et al.*, 2000].

3.2. Description par des phrases

Motivés par la nature humaine qui préfère décrire les choses par des phrases plutôt que par des mots clés, plusieurs chercheurs se sont investis dans la description par les phrases. Avec la structure et le principe sur lesquels est fondé STC, décrit dans le deuxième chapitre, [Zamir et Etzioni, 1998] prennent les phrases qui servaient à regrouper les documents comme étant des descriptions de clusters. [Lawrie *et al.*, 2001] eux, sélectionnent les phrases de description avec le plus grand poids sous TF-IDF. *Descriptive K-Means* [Stefanowski et Weiss, 2007] génère les phrases les plus porteuses de sens à partir du vecteur centroïde de chaque cluster formé par les K-Moyennes puis effectue une réaffectation des documents en conséquence. Avec Lingo [Osinski *et al.*, 2004], la décomposition en valeurs singulières (*Singular Value Decomposition*, SVD) est utilisée, au cours d'une phase dite *cluster label induction*, pour déterminer les phrases les plus descriptives à partir de celles sélectionnées dans une phase préalable dite *frequents phrases extraction*, c'est le principe de la DCF (*Description Comming First*) décrite dans le deuxième chapitre.

Une tentative de remédier au principal désavantage de cette approche qui est la dégradation de la lecture de la description des clusters *(the readability of clustering description)* selon [Zhang *et al.*, 2009], combine DCF et DCL *(Description Comming Last)* en sélectionnant les phrases avec les meilleurs scores parmi celles générées par les deux techniques DCF et DCL.

Dans la littérature nous retrouvons aussi les travaux de [Anton et Croft, 1996] qui extraient des phrases (composées d'un ou de plusieurs mots) à partir de la collection et limitant un seuil d'apparition de 50 % dans l'ensemble de la collection. [Glover *et al.*, 2002] utilisent un modèle statistique basé sur l'utilisation des informations contenues dans la

structure hiérarchique (cluster parent, cluster enfant et cluster actuel) pour définir une description du cluster actuel ou en cours de description.

3.3. Autres techniques de description

D'autres techniques de description sont exposées dans la littérature, nous mentionnons la description par le titre du document le plus proche du centroïde du cluster correspondant. Cette technique est considérée par [Manning *et al.*, 2008] comme une description interne des clusters.

Une autre technique de description est celle du résumé. En effet plusieurs recherches ont été introduites au cours de la DUC 2004 (*Document Understanding Conferences, 2004*) disponible sur le net[1]. Un travail très intéressant de [Wang *et al.*, 2011] qui consiste à utiliser une matrice dite de sens conjointement avec la matrice de mots pour générer un résumé multi-documents par cluster. [Muhr *et al.*, 2010] étudient l'apport de la relation entre les clusters d'une organisation hiérarchique aux techniques de descriptions traditionnelles basées sur les mots fréquentiels, la divergence de *Jensen-Shannon*, le test Chi carré (χ^2) et le gain d'information.

Il faut noter que [Pantel et Ravichandran, 2004 ; Niu *et al.*, 2012] présentent deux bonnes études comparatives entre les différentes techniques de description. Notons aussi que dans la présentation précédente des différentes techniques de description nous n'avons pas fait de distinction entre celles orientées résultats de la recherche web et celles orientées collections en hors line.

4. Mesures d'évaluation de la description

Bien que le jugement humain soit préférable pour l'évaluation de la performance de la description des résultats d'une CNST par rapport aux techniques automatiques, cette approche est très coûteuse et difficile à répéter avec des paramètres différents (ex. avant et après prétraitement, sur plusieurs techniques, sur plusieurs collections). En plus, dans la majorité des cas ce jugement diffère d'un expert à un autre sur une même technique. Dans ce qui suit nous donnons les différentes mesures généralement utilisées pour évaluer les résultats d'une technique de description.

Dans la littérature aucune méthodologie n'a gagné l'unanimité des recherches, néanmoins les mesures citées dans ce qui suit sont souvent utilisées pour évaluer le degré de satisfaction des conditions que doit remplir une description à savoir la concision, la compréhensibilité, la précision et la distinction.

4.1. Jugement humain *(Human Judge)*

Le jugement humain est généralement utilisé lorsqu'on ne dispose d'aucune indication sur les clusters à décrire. Cette mesure consiste alors à lire les documents de chaque cluster et donner un jugement sur sa description. Selon [Dostal *et al.*, 2013] dans un jugement humain une des trois valeurs suivante doit être affectée pour mesurer la description de chaque cluster de la CNST:

[1] http://www-nlpir.nist.gov/projects/duc/guidelines/2004.html

- **Correcte (*Correct*):** la description est jugée complètement correcte ;
- **Fausse (*Wrong*):** la description est jugée complètement fausse ;
- **Acceptable (*Acceptable*):** la description est jugée comme relativement correcte et peut être acceptée.

Ensuite, la qualité de la description globale est mesurée en comptant le nombre des descriptions correctes, fausses et acceptables.

Comme il a été mentionné auparavant il est difficile de tomber sur un même jugement entre plusieurs experts, une moyenne entre les différents jugements est alors faite pour évaluer une technique.

4.2. Mesures d'évaluation automatiques

Dans la littérature nous retrouvons plusieurs mesures pour évaluer la performance d'une technique de description automatique. La majorité de ces mesures se basent sur l'appréciation humaine pour être calculées, nous citons :

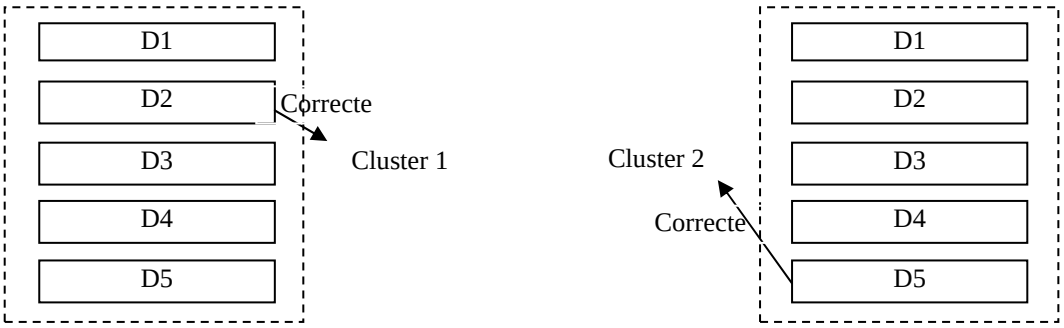
4.2.1. Match@N et MRR@N

Définie par [Treeratpituk et Callan, 2006**],** la mesure Match@N est calculée selon le procédé suivant :

Soit une liste de descriptions (mots ou phrases) retournée par une technique de description pour un cluster. Match@N indique l'existence d'au moins d'une description correcte parmi les *N* premières descriptions retournées, c'est-à-dire, mesurer la capacité de cette technique à retourner des descriptions correctes.

Match@N est un indicateur binaire (vrai ou faux) qui augmente d'une façon monotone lorsque *N* augmente. Ici le rang de la description dans les *N* descriptions n'est pas pris en compte. La valeur finale de cette mesure est calculée en prenant le rapport du nombre de descriptions correctes et le nombre total des descriptions (clusters).

L'exemple suivant illustre le fonctionnement de cette mesure sur deux clusters :



Description du cluster 1 **Description du cluster 2**

Match@2 = 1/2 = 0.5
Match@5 = 2/2 = 1

Le rang réciproque moyen MRR@N (*Mean Reciprocal Rank*) [Treeratpituk et Callan, 2006] est calculé en prenant l'inverse du rang de la première description correcte par les *N* premières, ou zéro si aucune description dans la liste retournée n'est correcte. La valeur finale de MRR@N est la moyenne des rangs réciproques de tous les clusters.

Le MRR@N détermine la capacité d'une technique de description à placer des descriptions justes dans les premiers rangs de l'ensemble des descriptions retournées.

Reprenons l'exemple précédant :
MRR@2 = (1/2 + 0)/2 = 0.25
MRR@5 = (1/2 + 1/5)/2 = 0.35

Il faut noter que généralement la valeur 5 est prise comme valeur de référence de *N* pour mesurer la performance d'une technique de description.

4.2.2. Chevauchement (*Overlap*) et précision (*Precision*)

Le chevauchement est généralement destiné à la mesure de la performance de la description par des phrases. Il mesure la similarité entre chaque phrase des *N* extraites et celle prédéfinie pour un cluster. La similitude est basée sur le nombre de mots partagés entre les deux phrases. Le chevauchement entre une phrase extraite p_i et la phrase prédéfinie p_c est défini comme suit:

$$Overlap(p_i, p_c) = \frac{|p_i \cap p_c|}{|p_i \cup p_c|} \quad (3.3)$$

La moyenne des *N* chevauchements représente la valeur du chevauchement des *N* phrases pour le cluster en question.

La deuxième mesure est la précision, elle consiste à donner une indication de la façon dont la meilleure phrase qui décrit le mieux le cluster est classée [Hammouda *et al.*, 2005]. La meilleure phrase est la première des *N* phrases de la description avec le meilleur score de chevauchement. La précision des *N* phrases de la description *pN* pour le cluster concerné est alors la suivante :

$$Précision\ (pN, p_t) = Chevauchement\ (p_{max}, p_t) * \left[1 - \frac{Rang(p_{max})-1}{N}\right] \quad (3.4)$$

4.2.3. Autres mesures d'évaluation des descriptions

D'autres mesures d'évaluation des descriptions peuvent être retrouvées dans la littérature, nous citons la Précision et le Rappel dans [Nanhong *et al.*, 2010 ; Kashireddy *et al.*, 2013] qui ont le même principe que les mesures plus haut, c'est-à-dire la comparaison avec une description prédéfinie humaine.

Une autre mesure est la $sim_{F\text{-measure}}$ qui réunie les mesures *Exact Match*, *Partial Match*, *Overlap Match* [Turel et Can, 2011]. Chacune de ces mesures est comparée avec une description prédéfinie humaine (*Ground Truth*) sous un aspect différent que les autres mesures.

5. Conclusion

La tâche de description des résultats d'une CNST est une tâche très délicate, ceci est confirmé par l'inexistence d'une technique d'évaluation qui à l'unanimité des recherches dans ce domaine. Nous avons commencé ce chapitre par donner les caractéristiques qu'une description doit remplir, ceci reste un peu subjectif étant donné que l'opinion humaine diffère sur une même description.

Nous avons ensuite cité les techniques de description largement répondues, ceci dit, il existe d'autres techniques dans la littérature mais qui se rapportent à des sujets différents du notre.

Enfin, nous avons terminé par donner les mesures de performance des descriptions les plus connues dans les deux approches de description, humaine et automatique, et décrit en détail leurs modes de fonctionnement.

Chapitre IV

La langue arabe et le traitement automatique

Chapitre IV
La langue arabe et le traitement automatique

1. Introduction

367 millions de personnes[1], entre le golf arabo-persique et l'océan atlantique, parlent la langue arabe. Le nombre d'utilisateurs d'internet parmi cette population ne cessent de grandir et par conséquent le volume d'information dans cette langue suit cette augmentation.

Par sa nature morphologique et sa syntaxique, la langue arabe, langue sémitique a été toujours une langue difficile à maitriser que se soit dans l'écrit ou le parlé. En traitement automatique le problème est encore amplifié [Aljlayl et Frieder, 2002 ; Larkey *et al.*, 2005]. Pour alléger cette langue, l'Arabe Modern Standard (*Modern Standard Arabic*, MSA) qui est une forme simplifiée de l'arabe classique [Khoja, 2001 ; Farghaly et Shaalan, 2009] est apparue. Le MSA n'utilise pas les formes compliquées que l'on retrouve dans l'arabe classique du *VII* ème siècle, mais il emploie plutôt, ce que nous qualifions de formes légères, ainsi il garde une même distance entre l'arabe classique et les différents dialectes d'origine arabe. Le MSA est utilisé dans les institutions académiques, les médias, les recherches, etc.

Même en allégeant l'arabe classique en MSA, cette langue reste morphologiquement très riche [Diab *et al.*, 2004], cette richesse a été la source d'une grande ambiguïté dans le traitement automatique de cette langue. A l'inverse des autres langues, beaucoup de recherches sont dédiées à l'enlèvement de cette ambiguïté. Même dans les autres axes de recherches telles que la recherche d'information, la traduction automatique et le résumé automatique, une grande partie est consacrée aux prétraitements qui visent à atténuer ce caractère morphologique difficile.

2. Particularités de la langue arabe

La langue arabe compte 28 lettres (Tableau 4.1) qui peuvent se raccorder entre elles (sauf les lettres ذ ,د ,ز ,ر ,و ,ا qui ne se joignent pas à gauche) et qui changent de forme et de présentation selon leurs positions (au début, au milieu ou à la fin du mot). Le tableau 4.2 montre les variations de l'écriture de la lettre ك (kef) selon la position dans le mot.

Pareille aux langues chinoise, japonaise et coréenne, et par opposition aux langues européennes, l'orientation de l'écriture est de droite à gauche et sans capitalisation. D'autres particularités qui vont être citées plus loin, comme la non vocalisation, l'agglutination, la

[1] Source : *Internet Word Stat* http://www.internetworldstats.com/stats7.htm

structure particulière combinant schème et radical distinguent la langue arabe des autres langues.

Lettre arabe	Correspondant français	Prononciation
ا	a	Alef
ب	b	Ba’
ت	t	Ta’
ث	Th	Tha’
ج	j	Jim
ح	h	Hha’
خ	Kh	Kha’
د	d	Dal
ذ	d	Thal
ر	r	Ra
ز	z	Zayn
س	s	Sin
ش	Sh	Shin
ص	S	Sad

Lettre arabe	Correspondant français	Prononciation
ض	D	Dad
ط	T	Tah
ظ	Z	Thah
ع	‘ ‘	Ayn
غ	Gh	Ghayn
ف	f	Fa
ق	q	Qaf
ك	k	Kaf
ل	l	Lam
م	m	Mim
ن	n	Nun
ه	h	Ha
و	W	Waw
ي	y	Ya

Tableau 4.1 : Les 28 lettres de l’alphabet arabe [Leclerc, 2000**].**

A la fin d’une lettre non joignable	A la fin	Au milieu	Au début
ك	ـك	ـكـ	كـ

Tableau 4.2 : Formes de la lettre ك kef.

2.1. Voyelles courtes et voyelles langues

La forme des voyelles est une particularité de la langue arabe. Elles sont de deux types. Les voyelles courtes ou brèves qui ont la forme d’une marque diacritique (ـَـِـُـْ) placée au dessus ou au dessous des lettres, tandis que les voyelles longues (ا،و،ي) collent aux consonnes et sont toujours écrites, même dans les formes non vocalisées. Le *tanwiin* (ـًـٍـٌ) est un autre genre de voyelles [El Kassas, 2005], il marque l’indéfini et est réalisé par un signe diacritique fusionné au signe de la voyelle courte.

Les voyelles sont nécessaires à la lecture et à la compréhension correcte d’un texte, elles permettent de différencier des mots ayant la même représentation (Tableau 4.3). Cependant, elles ne sont utilisées (ici on parle des voyelles courtes) que dans des contextes spéciaux tels que les livres didactiques, les dictionnaires ou le *Coran* [Larkey *et al.*, 2005], les textes courants rencontrés dans les journaux et les livres n’en comportent habituellement pas. En plus certaines lettres comme آ ,أ ou إ peuvent être normalisées par ا (Alef!) ; de même que pour les lettres ى et ه qui représentent respectivement ي et ة [Larkey et Connell, 2001; Farghaly et Shaalan, 2009].

Mot sans voyelle	1ère interprétation		2ème interprétation		3eme interprétation	
علم	عَلَمَ	Il a enseigné	عِلْمٌ عَلَمٌ	Science Drapeau	عُلِّمَ عُلِمَ	Il a été enseigné Il a été su

Tableau 4.3 : Ambiguïté causée par l’absence des voyelles dans le mot علم.

Ici il faut noter qu'un lecteur familier avec l'arabe ne trouvera pas vraiment de difficulté pour lire correctement un texte sans voyelles étant donné que l'ambiguïté causée par l'absence des voyelles courtes est atténuée par l'association de formes, de sens, de contexte, etc.

2.2. Notion de schème et morphologie

Trois catégories de mots sont à la base du lexique arabe; les verbes, les noms et les particules. Ces catégories sont dérivées de quelques milliers de racines (10000) [Darwish, 2002 ; Larkey *et al.*, 2005] voir 20000 par d'autres sources[1]. Les racines des verbes et des noms sont souvent à trois consonnes radicales ou trilitères (85 % du nombre total), et avec un degré moindre à quatre consonnes et rarement à cinq [Darwish, 2002; Tuerlinckx, 2004 ; Saad et Ashour, 2010].

La notion de schème est à l'origine des dérivations des mots en langue arabe. A partir d'une même racine et à l'aide de cette notion une famille de mots de différents schémas est générée [Baloul *et al.*, 2002]. Ce phénomène est une caractéristique spécifique aux langues sémitique. Une multitude de schèmes [De Roeck et Al Fares, 2000], qui sont essentiellement des adjonctions et des manipulations de la racine (Tableau 4.4), sont utilisés, certains sont plus complexes, tel que le redoublement d'une consonne (مسّك) ou l'allongement d'une consonne de la racine (voyelles longues, مسكا). Ainsi, nous pouvons dériver un grand nombre de noms, de formes et de temps verbaux.

Schèmes	RLK	غلق	Notion de fermer		MSK	مسك	Notion de tenir
$R_1aR_2aR_3a$	RaLaKa	غَلَقَ	Il a fermé		MaSaKa	مَسَكَ	Il a tenu
$R_1âR_2iR_3$	RâLiK	غَالِق	Fermant		MâSiK	مَاسِك	Teneur
$maR_1R_2uR_3$	maRLuK	مَغْلُوق	Fermé		maMSuK	مَمْسُو□	Tenu
$R_1uR_2iR_3a$	RuLiKa	غُلِقَ	Il a été fermé		MuSiKa	مُسِك	Il a été tenu
$taR_1R_2iR_3$	taRLiK	تَغْلِقُ	Elle ferme		taMSiK	تَمْسِكُ	Elle tient
…							

Tableau 4.4 : Exemple de schèmes pour les mots غلق et مسك.

R_i (Lettres en majuscule): désignent les consonnes de base de la racine ;
â, a, i,... : désignent les voyelles ;
m, t,... : sont des consonnes de dérivation utilisées dans les schèmes.

2.3. Agglutination

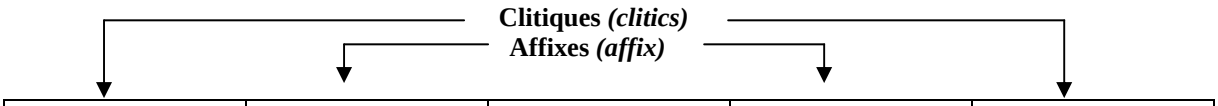
Le mot ou l'unité graphique, suite de graphèmes entre deux blancs, correspond le plus souvent en langue arabe non pas à une forme ou une unité susceptible de figurer sous une entrée lexicale ou lemme, mais à une suite de formes collées les unes aux autres [Tuerlinckx, 2004], ceci est dû au caractère agglutinant de cette langue. Une phrase en langue latine, par exemple, peut correspondre à un mot en langue arabe [Diab *et al.*, 2004].

Le collage du corps schématique, des antéfixes ou proclitiques (*proclitics*) qui sont des prépositions ou des conjonctions, des affixes (*affix*), qui représentent les préfixes et les suffixes et qui expriment les traits grammaticaux et indiquent les fonctions comme le cas du

[1] http://fr.wiktionary.org/wiki/Annexe:Schémes_arabes

nom, mode du verbe et les modalités (nombre, genre, personne,…) et des postfixes ou enclitiques (*enclitics*) qui sont des pronoms personnels, forment les mots agglutinés.

Par exemple le mot **أَتَأْكُلُونَهَا,** qui veut dire : « Est-ce que vous la mangez ?», se décompose selon le tableau 4.5 [Kelaiaia et Merouani, 2013] :



Postfixe	Suffixe	Corps schématique	Préfixe	Antéfixe
هَا	ونَ	أَكْلُ	تَ	أ
Pronom suffixe complément du nom	Suffixe verbal exprimant le pluriel	Verbe	Préfixe verbal du temps de l'inaccompli	Conjonction d'interrogation

Tableau 4.5 : Structure d'un mot arabe.

Le tableau 4.6 dresse l'ensemble des antéfixes, les préfixes, les suffixes et les postfixes de la langue arabe.

Postfixes		Suffixes		Préfixes		Antéfixes	
كما	Votre(s) en dual	ون	Marqueurs de dual, pluriel et le féminin pour les noms.	ا	Les lettres qui représentent la personne de conjugaison des verbes dans l'inaccompli (présent, futur)	وبالـ	et avec le [la, les]
هما	Leur (s) en dual	ين		نـ		والـ	et le [la, les]
كن	Votre(s) en pluriel féminin	ات		يـ		بالـ	avec le [la, les]
هن	Leur (s) en pluriel féminin	ان		تـ		فالـ	Ensuite le [la, les] (l'inaccompli)
هم	Leur (s)	تان				كالـ	comme le [la, les]
كم	Votre(s)	تين				ولـلـ	et pour le [la, les]
ها	Son, sa, ses du féminin	ة				الـ	le [la, les]
ه	Son, sa, ses du masculin	تما	Terminaisons de conjugaison pour les verbes.			وبـ	et avec
نا	Notre (s)	ون				ولـ	et pour
ك	Votre, to, ta	ين				للـ	pour le [la, les]
ي	Mon, ma, mes	وا				فسـ	ensuite (l'inaccompli)
		تا				فبـ	ensuite avec
		تم				فلـ	ensuite pour
		تن				و-□	et (l'inaccompli)
		نا				كـ	Comme
		ت				-□	l'inaccompli
		ن				فـ	Ensuite
		ا				و	et
		ي				بـ	Avec
		و				لـ	Pour

Tableau 4.6 : Antéfixes, préfixes, suffixes et postfixes de la langue arabe.

2.4. Catégories des mots

L'arabe considère trois catégories de mots [Khoja, 2001 ; Maamouri et Bies, 2004], verbe nom et particules ou lettres dans certains ouvrages. Certains grammairiens ajoutent une

catégorie instruments ou articles recoupant plus ou moins celle des particules [Tuerlinckx, 2004], alors que d'autres donnent une toute autre catégorisation [El Kassas, 2005].

2.4.1. Verbe

Le verbe est une entité exprimant un sens qui dépendra des facteurs suivants :

- Le temps (accompli, inaccompli, impératif) ;
- Le nombre de sujet (singulier, dual, pluriel) ;
- Le genre du sujet (masculin, féminin) ;
- La personne (première, deuxième, troisième) ;
- Le mode (actif, passif).

Pour conjuguer les verbes, on ajoute des préfixes et des suffixes (tableau 4.6). La langue arabe dispose de trois temps [El Kassas, 2005] :

a. L'accompli

Il exprime une action accomplie dans le temps par le rattachement de l'un des suffixes dans le tableau 4.6. Par exemple pour le pluriel féminin on a : غلقن RalaKna (elles ont fermé), pour le pluriel masculin on a : غلقوا RaLaKuu (Ils ont fermé).

b. L'inaccompli

Il exprime les deux temps, le présent et le futur :

Présent : exprimé par le rattachement d'un préfixe, il définit une action en cours d'accomplissement. Par exemple pour la troisième personne du singulier on a يغلق yaRLiKu (il ferme) pour le masculin et تغلق taRLiKu (elle ferme) pour le féminin.

Futur : marqué par le rattachement au début du verbe des antéfixes سـ (sa), فسـ (fasa) ou وسـ (wasa) ou le positionnement du mot outil سوف (sawfa) avant le verbe (en plus d'un préfixe de l'inaccompli présent), ce temps définit une action qui se déroulera dans le futur. Par exemple on a سيغلق sayaRliku (il fermera) ou سوف يغلق sawfa yaRLiKu (il va fermer).

c. L'impératif

Ce temps n'existe qu'à la deuxième personne, il définit une action directive pour l'accomplissement. Par exemple pour la deuxième personne du singulier on a : اغلق aRLiK (fermes) pour le masculin et اغلقي aRLiKi (fermes) pour le féminin.

En plus du rattachement d'un suffixe, le rattachement d'un préfixe est en fonction du verbe à conjuguer. Par exemple, pour le verbe اكل (manger) on peut ne pas rajouter le préfixe ا (a) ce qui donne كل (KuL), ou كلي (KuLi) pour le féminin.

2.4.2. Nom

Le nom est l'élément désignant un être ou un objet qui exprime un sens indépendant du temps. La catégorie des noms regroupe toutes les unités lexicales référant à un sens qui n'est pas lié au temps. Cette catégorie comprend l'adjectif et le substantif (اصفة والموصوف).

Les substantifs sont de deux catégories, ceux qui sont dérivés de la racine verbale et ceux qui ne le sont pas comme les noms propres et les noms communs. Les noms de la première catégorie sont tirés du rattachement des suffixes (tableau 4.6) comme suite :

Le masculin pluriel: un des deux suffixes ين ou و est rajouté selon la position du mot dans la phrase (sujet ou complément d'objet). Par exemple: جزائريين ou جزائريو (*Algériens*).

Le féminin singulier: le suffixe ة est rattaché au mot. Par exemple جزائرية (*Algérienne*).

Le féminin pluriel: le suffixe ات est rattaché au mot. Par exemple جزائريات (*Algériennes*).

Le Pluriel irrégulier: il ne suit pas une règle bien précise, il s'obtient par l'ajout ou la suppression des lettres au début, au milieu ou à la fin du mot en question. Par exemple : تركي qui veut dire *Turque* devient أتراك pour le pluriel masculin.

Il faut noter que cette situation est retrouvée en langue anglaise mais avec un nombre réduit de noms et un petit nombre de verbe très fréquents [Larkey *et al.*, 2002]. Aussi, à l'inverse du français, par exemple, les pronoms personnels représentant le sujet en langue arabe sont raccordés au verbe conjugué.

Exemple:

Pour le verbe غلق on aura :
غلقت pour le singulier ;
غلقتما pour le dual ;
غلقتم ou غلقتن pour le pluriel.

2.4.3. Particules (lettres الحروف ou outils الأدوات)

Les particules jouent trois rôles essentiels dans le texte arabe. Le premier est sémantique, les particules d'introduction comme كا (il était), d'explication comme أي (c'est-à-dire), de conséquence comme لعل (peut être), etc. contribuent fondamentalement à la compréhension du texte. Le deuxième rôle est la cohérence et l'enchainement dans le texte tel que و (et), ثم (ensuite), etc. Le troisième rôle, quant à lui, est la situation et le déroulement des évènements dans le temps ou l'endroit (المكا والزما) comme بعد (après), قبل (avant), ذنا (depuis), lieu حيث (où), etc.

Comme le cas des noms et des verbes, les particules peuvent aussi s'agglutiner avec des affixes et des clitiques ce qui rajoute une certaine complexité quant à leur identification, on a par exemple بعدك (après toi) et قبلك (avant toi).

3. Problèmes posés au traitement automatique de la langue arabe

Depuis les premières recherches en traitement automatique sur la langue arabe attribuées selon [Tuerlinckx, 2004] à *David Cohen* qui propose un essai d'analyse automatique dès 1961 (*Cohen*, 1961/1970), plusieurs chercheurs se sont intéressés à une multitude d'axes dans cette langue, notamment le résumé automatique, la catégorisation, la classification non supervisée, la détection des entités nommées, l'analyse du discours, etc.

Toutes ces recherches devaient surmonter certains problèmes générés par les particularités de la langue citées précédemment telles que, le sens des mots causé par

l'absence des voyelles, la catégorie lexicale des mots, la forme des mots qui change et suit la position de ces derniers dans la phrase, l'agglutination, la synonymie, etc. Certains de ces problèmes relèvent du TALN, alors que d'autres représentent l'axes de recherches des approches statistiques. Dans le cadre de notre étude nous nous limitons aux deux cas suivants:

3.1. Détection de racine

Le caractère flexionnel de la langue arabe influe d'une façon négative sur les résultats obtenus par l'application des méthodes statistiques sur les documents dans cette langue. Il faut donc ramener tous les mots ayant subis des flexions sur la forme (*surface forms*) [Aljlayl et Frieder, 2002] à leur racine. Par exemple le mot **كتب** (a écrit) qui indique l'idée d'écriture, a parmi ses formes flexionnelles les mots **مَكتب** (bureau), **مكتبة** (bibliothèque), **كاتب** (auteur), **كتا** (livre), **كتبتم** (vous avez écrit), **سيكتب** (il va écrire)**,** etc.

Pour ramener ces mots à leurs racines, il y a deux manières de faire et se présentent :

La première consiste à utiliser un dictionnaire de racines [Khoja, 2001] pour trouver celles qui correspondent aux flexions en question (comme décrit dans le deuxième chapitre), ce qui un peut difficile eu égard à la richesse de la langue arabe et la nécessité de la mise à jour de ce dictionnaire.

La deuxième technique consiste à utiliser le stemming léger (*light stemming*) ce qui permet de trouver la racine de quelques formes flexionnelles, dans notre exemple, les mots **مكتبة, كتبتم** et **سيكتب** (tableau 4.7).

Flexions	Préfixes/ antéfixes	suffixes/ postfixes	Résultats du stemming léger
مكتب	□	/	مكتب
مكتبة	□	ة	مكتب
كاتب	/	/	كاتب
كتا□	/	/	كتا□
كتبتم	/	تم	كتب
سيكتب	سي	/	كتب

Tableau 4.7 : Résultats du stemming léger sur les flexions du mot كتب.

3.2. Agglutination et ordre d'élimination des éléments flexionnels

Avec l'absence des voyelles, l'ordre de troncature des éléments flexionnels est très important. La troncature des préfixes avant les suffixes (et réciproquement) peut influencer les résultats.

Par exemple en prenant le mot ألمهم (leur douleur) et en faisant une normalisation agressive (c'est-à-dire أ → ا) nous aurons une ambiguïté au niveau du sens. Cela est expliqué dans le tableau suivant :

Préfixes/ antéfixes	Suffixes/ postfixes	Résultats du stemming léger	Signification
/	هم	الم	Douleur, Fait la douleur
ال	/	مهم	Important
ال	هم	م	La lettre M

Tableau 4.8 : ambiguïté généré par le stemming léger du mot ألمهم.

4. Travaux connexes

Dans la littérature nous retrouvons plusieurs travaux qui se sont intéressés à la classification supervisée ou catégorisation de textes arabes tels que [El-Kourdi *et al.*, 2004] qui ont utilisé l'algorithme *Naïve Bayes* pour la classification d'une collection de documents tirés du web, [El-Halees, 2006] qui a utilisé un système à base de règles d'association pour la classification d'un ensemble de documents, [Al-Harbi *et al.*, 2008] qui ont testé deux célèbres algorithmes à savoir SVM et C5.0 sur sept collections en langue arabe, alors que [Al-Shargabi *et al.*, 2011] ont effectué une étude comparative entre trois algorithmes de classification sur une collection de 2363 textes en langue arabe. D'autres travaux similaires peuvent êtres retrouvés dans la littérature.

En CNST sur la langue arabe nous avons recensé deux travaux, le premier est celui de [Sawaf *et al.*, 2001] qui ont employé une approche statistique (basée sur la technique de la maximisation de l'entropie) pour la CNST d'une base d'articles arabes couvrant plusieurs domaines tels que la politiques, l'économie, etc. Le deuxième est celui de [Huot et Coupet, 2005] qui ont développé un algorithme intégré dans le logiciel *Insight Discoverer Cluster* qui, à partir de descripteurs en arabe, regroupe les documents similaires dans des clusters en fonction de leurs ressemblances sémantiques et de leurs proximités thématiques.

En ce qui concerne l'utilisation des modèles thématiques probabilistes en langue arabe, nous avons recensé une seule recherche majeure celle de [Brahmi *et al.*, 2011] qui étudie l'influence du stemming sur la classification supervisée avec LDA.
De même, pour la description des résultats de la CNST et mis à part les travaux décrits dans le troisième chapitre, destinés dans leur majorité à la langue anglaise, la présente étude est la première en la matière sur la langue arabe.

5. Conclusion

Dans ce chapitre, nous avons mis en valeur les particularités de la langue arabe. Particularités à caractère spécialement productif, dérivationnel et flexionnel. Un mot dans cette langue peut avoir un grand nombre de formes et de sens générés par le caractère flexionnel de cette langue. Mais la non voyellation et l'agglutination des mots par rapport à d'autres langues comme le français ou l'anglais rend le traitement automatique de cette langue très compliqué.

Dans ce chapitre, nous avons décrit la morphologie des mots de cette langue, suivie de la structure de ses mots ainsi que leurs catégories qui sont essentiellement, le verbe, le nom et les particules.

Nous avons par la suite cité les différents problèmes qui pourraient être engendrés par les particularités de cette langue en traitement automatique, et nous nous sommes attardés sur le problème de la détection de la racine, ainsi que les ambiguïtés qui peuvent être posées par l'ordre de suppression des éléments flexionnels. Nous avons argumenté cela par des exemples bien précis. Nous avons terminé ce chapitre par citer les travaux que nous avons jugés majeurs sur la langue arabe et qui se rapportent à la présente étude. Nous reviendrons sur d'autres problèmes qui peuvent être posés par cette langue dans son traitement automatique dans le sixième chapitre.

Chapitre V

LDK-Means (Latent Descriptive K-Means): Nouvelle approche

Chapitre V
LDK-Means (Latent Descriptive K-Means): Nouvelle approche

1. Introduction

La classification non supervisée textuelle descriptive vise à donner une description (labellisation) automatique claire et compréhensible par l'utilisateur aux contenus des résultats (clusters) de cette classification, reliés thématiquement, sans l'intervention de l'expertise humaine. Comme nous avons vu au cours du troisième chapitre la majorité des travaux de description se sont focalisés sur les résultats rendus par les moteurs de recherches sur le web. Plusieurs facteurs ont contribué à la mise à l'écart de la CNST sur les collections textuelles ou bases documentaires en hors ligne (*offline*). Ces facteurs sont de plusieurs natures, nous citons:

- **Facteurs qui relèvent de la nature de la langue utilisée comme :**
  - La technique de représentation des documents[1] tels que les sacs de mots, les n-grammes, les phrases, etc. ;
  - Les prétraitements appliqués en post-classification qui varient en complexité selon la langue dans laquelle les textes sont rédigés ;

- **Facteurs qui relèvent de la classification proprement dite comme :**
  - La grande dimension des textes qui implique la manipulation des vecteurs de représentation à l'ordre de dizaine de milliers de mots et ce quel que soit la technique de représentation employée;
  - Le grand nombre de textes qu'une collection peut contenir et qui est généralement de l'ordre d'une dizaine de milliers de textes voir plus;

  Ces deux points exigent des techniques de classification non supervisée relativement acceptables en complexité de calcul ;

- **Facteurs qui relèvent de la manière de description des clusters comme :**
  - Quelle est la technique la plus compréhensible par l'utilisateur, les mots seuls ou mots clés, les phrases à mots fréquentiels successifs (ou non) ou phrases proprement dites (phrases grammaticales) et dans ce cas quel type utiliser (la phrase nominale ou la phrase verbale)?

[1] Non comprises les techniques de représentation relatives à la classification non supervisée des résultats retournés par les moteurs de recherches.

- Envisager l'emploi des techniques qui relèvent du TALN (traitement automatique du langage naturel) dans le cas où on opte pour une description par des phrases grammaticales ;

- **Facteurs qui relèvent de la validation des solutions obtenues comme :**
  - Quelles mesures d'évaluation utiliser ?
  - Coût élevé des moyens de validation des solutions proposées (généralement la validation se fait à travers l'expertise humaine); etc.

Tous ces facteurs nous ont obligés à réfléchir sur la méthodologie avec laquelle nous allons aborder la problématique visée. En fait, s'ajouter à cela la nature de la langue arabe et l'absence d'une collection homologuée disponible immédiatement, étant donné que les collections comme celles de LDC *(Linguistic Data Consortium)* sont difficiles à obtenir dans l'immédiat, ce qui rendait notre tâche encore plus ardue.

Ainsi, notre premier objectif était de trouver des collections sur lesquelles nos recherches seront conduites, nous avons opté alors pour quatre collections qui seront décrites dans le prochain chapitre. Puis, nous nous somme attaqués au développement et à la validation d'une nouvelle approche de la classification non supervisée textuelle descriptive intitulée LDK-Means (*Latent Descriptive K-Means*). Ce développement a été précédé par deux études qui consistaient à déterminer quelle technique de classification utiliser et d'autre part comment décrire les résultats obtenus.

Dans ce qui suit nous allons exposer les différentes étapes qui nous ont menées aux résultats exposés dans le prochain chapitre.

2. Stratégie de l'étude menée

La stratégie que nous avons adoptée est très intuitive et facile à implémenter, elle s'articule autour de trois phases essentielles conduites via notre approche intitulée LDK-Means et décrites comme suit :

2.1. Phase 1: Préparation des quatre collections

Au cours de cette phase, nous avons procédé à la préparation des quatre collections pour les traitements ultérieures, c'est-à-dire, passé de la forme brute des textes *(raw text)* à la forme souhaitable en post-classification. Nous avons alors suivi les étapes suivantes:

1. Transformation automatique des textes rédigés initialement sous plusieurs formats d'encodage vers un encodage standard (CP1256 ou UTF-8) ;

2. Filtrage des caractères de ponctuation, Translittération et Tokenisation ;

3. Nettoyage des mots vides de sens ou mots outils (*Stop words*) ;

4. Stemming[1].

Ici il faut noter que le choix de travailler avec l'encodage CP1256 et la translittération

[1] Ici nous utiliserons le stemming léger (chapitre 2, section 3.1.3).

est lié à la nature morphologique des caractères arabes qui peut parfois provoquer des défaillances des outils avec lesquels nous avons travaillé[1]. Le même problème a été soulevé dans les recherches de [Saad et Achour, 2010].

2.2. Phase 2: Classification non supervisée textuelle

Au cours de cette phase il fallait décider soit à développer une nouvelle technique de CNST basée sur l'amélioration de l'une des techniques déjà existantes, soit à utiliser la plus répandue.

Ici, mise à part les techniques traditionnelles (telles que les K-Moyennes ou les méthodes hiérarchiques), méthodes généralement appréciées dans ce domaine, nous avons remarqué l'absence d'une technique de CNST qui a l'unanimité des recherches sur sa performance du point de vue classification textuelle et surtout à grande échelle; donc, essayer de développer une technique meilleure (avec des améliorations minimes généralement sur la vitesse d'exécution) serait totalement absurde.

En plus les méthodes de complexité plus que linéaires sont quasiment prohibées. Ainsi des méthodes telles que les K-Moyennes, proposée en 1967 par *MacQueen* sont largement souhaitables, d'ailleurs cette dernière, souvent prise comment méthode de référence *(baseline method)*, a été à la base de plusieurs technique de CNST descriptive [Stefanowski et Weiss, 2007]. Tout ceci nous a encouragés à opter pour ladite K-Moyennes, surtout après l'analyse des résultats obtenus au cours de la phase intermédiaire décrite dans la section 4.1.

2.2.1. Pourquoi K-Moyennes et non pas LDA

Ce choix est motivé, d'une part, par son faible degré de complexité qui avoisine les *O(knt)* où *k* est le nombre de clusters, *n* est le nombre de textes, et *t* est le nombre d'itérations (en général $k << t << n$), et d'autre part, la facilité de remédier à ses principales inconvénients tels que :

- Choix de *k ;*
- Sensibilité aux exceptions et aux données bruitées ;

Plusieurs recherches se sont consacrées à la résolution de ses inconvénients et peuvent être retrouvées dans la littérature. En plus ce choix a été réconforté par les résultats obtenus à l'issue de l'étude comparative effectuée initialement entre cette méthode et la méthode LDA fréquemment lancée dans le domaine de la CNST [Kelaiaia et Merouani, 2016].

Un autre facteur qui a été décisif dans ce choix est le principe de fonctionnement de la méthode LDA en CNST (décrit en deuxième chapitre). Dans cette dernière les clusters sont formés à base du thème dominant, donc même c'est l'évaluation de la CNST par LDA est relativement meilleure que les K-Moyennes [Kelaiaia et Merouani, 2013b; Kelaiaia et Merouani, 2014], cette dernière ne prend en charge qu'un seul thème, à l'inverse des K-Moyennes qui elle, regroupe les textes en prenant en compte tous les thèmes et ceci selon le principe de calcul de la similarité entre deux textes. D'autre part l'existante de plusieurs thèmes dans un texte suppose l'existence d'une relation entre ces derniers.

La figure 5.1 montre comment les deux méthodes regroupent les textes. Avec la

[1] D'ailleurs c'est le cas de la majorité des travaux sur la langue arabe.

méthode LDA le premier cluster sera formé par les textes ayant pour thème dominant celui avec les mots en rouge, le deuxième sera formé des textes ayant pour thème dominant celui avec les mots en vert. Avec les K-Moyennes les documents sont regroupés en prenant en compte la similarité entre l'ensemble de mots formants les textes.

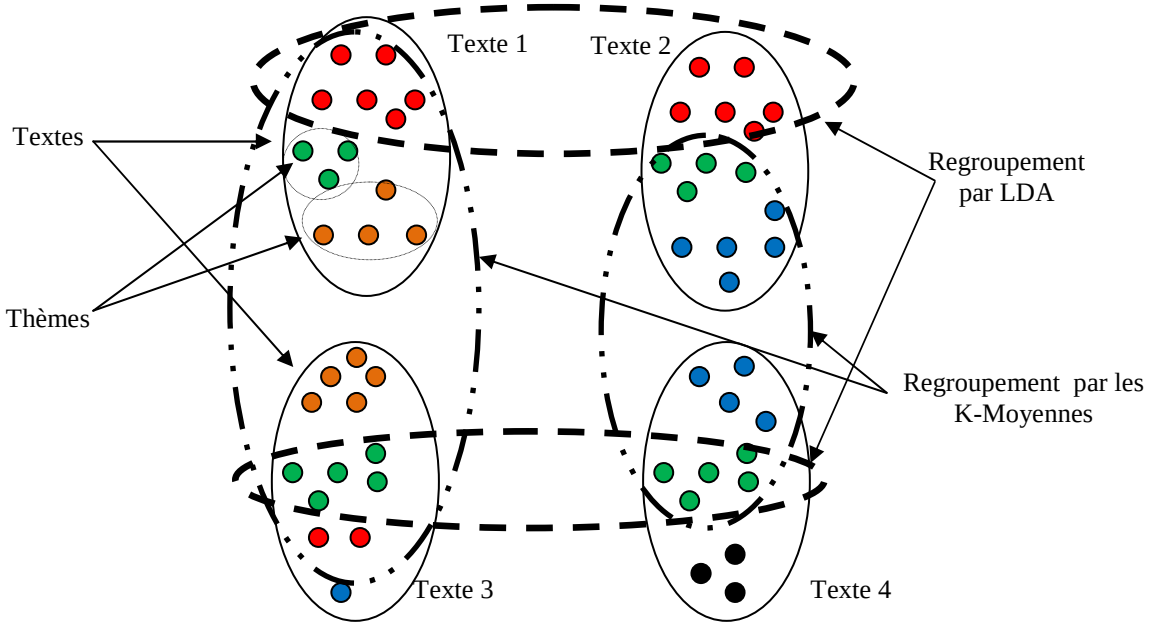


Figure 5.1 : Regroupement des textes par les K-Moyennes et LDA.

2.2.2. Classification non supervisée textuelle

Une fois le choix de la méthode avec laquelle la classification sera faite, cette dernière est lancée sur chaque collection. Ici, le choix du nombre de clusters *k* est égale au nombre des classes prédéfinies de chaque collection, ce choix n'est pas fixe puisque l'approche proposée est flexible est configurable selon la volonté de l'utilisateur. Notons qu'un nombre de clusters élevé donne des clusters contenant des textes très semblables dans le cas contraire, les clusters seront diversifiés.

Ici l'approche proposée reste ouverte à tout changement de technique de classification puisque les phases classification et description sont totalement indépendantes.

2.2.3. Présentation des méthodes utilisées

Dans les deux sections *a* et *b* suivantes, nous allons décrire les fondements des deux fonctions utilisées par notre approche LDK-Means incorporées dans deux environnements, MALLET (*MAchine Learning for LanguagE Toolkit*) [Mccallum, 2002] et Text Garden [Mladenic et Grobelnik, 2003] et qui représentent l'implémentation des deux méthodes K-Moyennes et LDA. Pour cette dernière nous avons évité de rentrer dans les considérations et les fondements mathématiques, qui eux peuvent être retrouvés en détail dans les travaux à l'origine de cette méthode.

a. La méthode LDA

a.1. Le modèle

Le modèle LDA est celui décrit dans [Blei *et al.*, 2003; Wei et Croft, 2006 ; Steyvers et

Griffiths, 2007] avec le processus génératif déjà décrit dans la section 4.1 du deuxième chapitre.

a.2. Estimation des paramètres avec l'échantionnage de Gibbs (Gibbs sampling)

Avant de présenter comment se fait l'estimation des paramètres par l'échantionnage de *Gibbs*, commençant par donner un aperçu sur son fondement qui est basée sur les méthodes de chaines de *Markov Monte Carlo* (*Markov chain Monte Carlo,* MCMC).

Ces méthodes consistent à construire une chaîne de *Markov* sur les variables cachées dont la loi stationnaire est la distribution *a posteriori* cherchée. Nous rappelons qu'une chaîne de *Markov* est définie par une loi de transition $p(\theta^{t+1},z^{t+1}|\theta^t,z^t)$ d'un état (θ^t,z^t) vers un état suivant (θ^{t+1},z^{t+1}), et peut converger vers une loi stationnaire qui est laissée invariante par la transition. Une fois l'état stationnaire atteint, les échantillons donnés par la chaîne de *Markov* suivent cette loi stationnaire qui est la distribution voulue.

L'échantionnage de *Gibbs*, est une méthode de chaines de *Markov* où la transition de la chaîne est donnée par la loi conditionnelle d'une variable cachée étant donnés les observations et l'état courant des autres variables cachées. Pour LDA, on peut successivement obtenir des échantillons $\theta|w,z$ puis $z_n|\theta,w$.

b. La méthode des K-Moyennes

b.1. Fonctionnement

Le fonctionnement de la méthode des K-Moyennes utilisée suit l'algorithme de *McQueen* décrit par l'algorithme 1.1 dans le premier chapitre. Chaque cluster est représenté par un centroïde qui est calculé selon l'équation suivante :

$$C_S = \frac{1}{|S|}\sum_{X_i \in S} X_i \qquad (5.1)$$

où C_S est le centroïde du cluster S et X_i un élément du cluster.

b.2. Mesure de similarité

La mesure de similarité utilisée est celle du cosinus déjà décrite dans la section 3.3 du deuxième chapitre.

2.3. Phase 3: Description des partitions générées

La dernière phase de l'approche LDK-Means est la description du contenu des partitions générées par la CNST. Comme nous avons vu dans le troisième chapitre, la description des résultats d'une CNST locale ou en hors ligne (*offline*) est une tâche très délicate, surtout que la majorité des recherches précédentes se sont focalisées sur la classification des résultats retournés par les moteurs de recherches telles que STC et Lingo.

A première vue, nous pouvons prendre ces recherches et les appliquer sur notre cas, mais ici le problème et que les algorithmes dans ces recherches sont conçus pour fonctionner sur les fichiers entêtes des pages web (*snippets*) et ne prennent pas en considération la totalité du texte de la page web et donc ne prennent pas les relations intra-textes (dans le texte lui-même) et extra-textes (entre les documents) [Turel et Can, 2011].

Au début, nous avons investigué la possibilité de prendre en charge l'aspect

grammatical pour déterminer quelles sont les phrases les plus pertinentes et les plus représentatives, les phrases nominales ou les phrases verbales ?. En langue anglaise plusieurs recherches se sont mise d'accord sur les phrases nominale [Weiss, 2006]. En langue arabe le cas est très délicat. Le caractère flexionnel et agglutinant rend la tâche plus difficile, Il fallait faire appel à des experts en linguistique arabe, d'une part et d'autre part faire un étiquetage grammatical des quatre collections, ce qui était quasiment impossible. Ceci nous a poussé à laisser tomber cet axe là et se tournés vers les mots fréquentiels latents et les phrases fréquentielles latentes[1]. Ces deux techniques, sont inspirées du principe de la détection thématique et sont décrites comme suit :

2.3.1. Mots latents ou Mots fréquentiels latents

L'une des vocations des modèles thématiques probabilistes étant l'extraction des mots sémantiquement reliées entre eux [Wei et Croft, 2006; Lu *et al.*, 2011] pour décrire un thème, le modèle de la distribution multinomiale des mots définit, ainsi, d'une part le poids d'un mot pour définir un thème et d'autre part l'association d'apparition conjointe avec les autres mots qui définissent le même thème (quatrième règle du processus génératif, deuxième chapitre, section 4.1). Ceci est traduit par :

- L'inférence thématique attribue, selon une probabilité d'apparition, chaque occurrence d'un mot à un thème.
- Un classement selon le nombre d'occurrences de chaque mot par thème est effectué.
- Chaque thème est alors décrit par une suite de mots pondérés par leurs poids de présence dans ce thème.

Formellement ceci est traduit comme suit :

Soit $T_1, T_2, \ldots, T_K$ les thèmes qui forment les documents de la collection. Chaque thème T_k est représenté par la représentation 5.2 :

$$T_k=(na_{1k}, pa_{1k}), (na_{2k},pa_{2k}), \ldots, (na_{nk,} pa_{nk}) \quad (5.2)$$

où na_{ik} représente le nombre d'apparition du *ième* mot dans le thème T_k et pa_{ik} représente la probabilité conditionnelle d'apparition du *ième* mot dans le thème T_k par rapport au vocabulaire de représentation (les mots thématiquement liés).

Principe de description

La description par les mots fréquentiels latents est inspirée du principe décrit plus haut. En supposant qu'un cluster représente un thème et en lançant le processus de détection thématique sur les textes formant ce cluster, nous pourrons obtenir une description par des mots sémantiquement reliés.

2.3.2. Phrases latentes

Linguistiquement, les utilisateurs préfèrent décrire les choses avec des phrases plutôt qu'avec des mots séparés. Le problème majeur avec la description par les phrases est que ces dernières sont omises avec la majorité des techniques de pondération, d'où est venue la motivation de décrire les résultats de notre CNST avec les phrases fréquentielles latentes.

[1] Ici, le principe de la latence défini la relation cachée entre les descripteurs (mots ou phrases) latents.

L'idée de description par les phrases fréquentielles latentes dans notre cas est simple, son fonctionnement est le même que celui des mots fréquentiels latents sauf qu'ici deux (ou plus) mots latents successifs sont connectés est pris comme étant une phrase fréquentielle latente. Ceci est traduit comme suit :

Soit le même schéma de description du thème T_k cité dans la section précédente. Une phrase fréquentielle latente est prise comme étant les mots contigus (deux ou plus) qui possèdent les mêmes probabilités conditionnelles d'apparition et les mêmes nombres d'apparitions.

2.3.3. Mots fréquentiels : la méthode de référence (*Baseline method*)

Pour mesurer la performance de la description par les mots et les phrases latentes, nous avons pris comme méthode de référence *(baseline method)* la description par les mots fréquentiels, cette technique est la technique la plus connue dans ce domaine, elle consiste à prendre uniquement les mots qui apparaissent le plus souvent dans un nombre donné de textes, ce qui la rend une technique naturelle pour la description.

3. Expérimentations intermédiaires

Avant d'entamer les expérimentations avec l'approche que nous avons proposée nous avons effectué deux tests préliminaires.

3.1. Etude comparative

Le but de cette étude est de comparer les performances de la méthode des K-Moyennes avec celles de LDA. Au départ nous avons testé les deux méthodes sur la collection OSAc [Kelaiaia et Merouani, 2013b] puis l'étude a été élargie aux trois autres collections (El Sulaiti, BBC, AlWatan) [Kelaiaia et Merouani, 2014 ; Kelaiaia et Merouani, 2016].

Comme il a été mentionné précédemment, le but de cette étude est de décider quelle méthode utiliser dans notre approche étant donné que LDA, une technique relativement récente, a prouvé ses performances dans le domaine de la segmentation thématique.

3.2. Description des classes prédéfinies

La deuxième étape que nous avons accomplie, toujours dans l'optique des tests avant les expérimentations de l'approche proposée, consiste à mesurer la fiabilité de description par les mots fréquentiels et les phrases fréquentielles latentes des classes prédéfinies d'une collection. Notre choix de cette dernière s'est porté sur la collection CCA de [El Sulaiti, 2003].

Le choix de la collection CCA est justifié par le nombre relativement petit des textes contenus dans cette dernière par rapport aux trois autres collections (406 textes après nettoyage), ce nombre est relativement acceptable pour une lecture par des experts humains. En faite, on ne pouvait pas imposer à une personne de lire plus d'un millier de textes étant donné que cette expertise est la mieux placée pour trancher si une telle ou telle technique est meilleure.

4. Mesures d'évaluation

Comment il a été mentionné auparavant, le choix d'une mesure d'évaluation que se soit en CNST ou sa description est très délicat, néanmoins au cours de notre étude nous avons utilisé les plus répondues.

4.1. Qualité de la CNST

Afin d'évaluer la qualité de la CNST par les deux techniques nous avons utilisé quatre mesures reconnues performantes dans ce domaine, l'indice de *Rand*, l'indice de *Jaccard*, la F-mesure et l'entropie.

Le calcul de ces quatre mesures se fait en comparant la partition générée par une CNST (ensemble de clusters) à une autre prédéfinie (ensemble de classes).

4.2. Qualité de la description des classes prédéfinies

Mise à part l'expertise humaine, il très difficile de trouver une mesure qui a l'unanimité des recherches pour évaluer la description automatique d'une partition en clusters générée par une technique de CNST [Dostal *et al.*, 2013] ou une partition de classes prédéfinies. Même les mesures les plus répondues font appel à cette expertise.

Pour déterminer la qualité de la description des classes prédéfinies, nous avons utilisé deux mesures déjà décrites dans le troisième chapitre, qui sont Match@N et MRR@N avec la valeur de N égale à 5.

Dans la littérature ces deux mesures sont surtout utilisées pour mesurer la qualité de la description automatique des classes déjà décrite manuellement.

4.3. Qualité de la description des clusters générés par la CNST

Pour mesurer la qualité de l'approche proposée nous avons fait appel à l'expertise humaine à travers la mesure jugement humain *(human judge)* [Dostal *et al.*, 2013] déjà décrite dans le troisième chapitre section 4.1. Ce choix est motivé par l'absence de toute information sur la partition de clusters générée.

5. Traitement des nouveaux documents

Une des questions les plus difficiles et les plus pertinentes en classification est : comment traiter ou affecter les nouveaux documents ? Pour répondre à cette question deux solutions se présentent :

- **Affectation directe :** Consiste à affecter les nouveaux documents au cluster le plus proche, c'est-à-dire calculer la distance (ex. en utilisant une mesure telle que le cosinus) entre le centroïde de chaque cluster et le document à classer. Le centroïde avec la plus petite distance l'emporte.
  Après l'affectation d'un certains nombre de documents (proportionnel au nombre total des documents de la collection) un relancement de tout le processus de classification et de description doit être effectué.

- **Lancement du processus de la classification non supervisée et de la description :** Ce choix est motivé par deux paramètres :
  - Le nombre réduit de documents dans la collection ;
  - La puissance de calcul des machines.

Ici il faut noter que même avec un nombre de documents relativement grand dans une collection, la puissance de calcul des machines peut jouer à la faveur de ce choix. Par exemple une machine avec les caractéristiques de : *Microprocesseur Intel Core I3 2.4 GHZ, 4 GO de RAM*, LDK-Means a mis à peu prés quinze (15) minutes pour classer et décrire la collection OSAc sous la forme nettoyée décrite dans le chapitre suivant.

6. Conclusion

Ce chapitre se voulait être une présentation de la démarche que nous avons suivie afin de tester une nouvelle approche de description des résultats de la classification non supervisée textuelle en langue arabe. LDK-Means rassemble deux méthodes très célèbres en classification et en détection thématique probabiliste. Par ce rassemblement nous avons voulu exploiter les points positifs de ces deux méthodes en simplicité de classification (les K-Moyennes) et en performances de description (LDA).

Même si on peut reprocher à la présente approche d'être basée sur les K-Moyennes qui a plusieurs défauts, LDK-Means est flexible est la technique sur laquelle elle est basée peut être remplacée par une autre méthode.

Nous avons essayé d'apporter des solutions à plusieurs questions posées actuellement en CNST telles que : le traitement de langue arabe en classification, la grande dimension et la description des résultats de la CNST en hors ligne. Nous avons argumenté les décisions prises que ce soit pour la CNST que pour la description de ses résultats. Nos arguments sont fondés sur une étude comparative entre deux techniques de CNST et une étude de l'application de la description proposée sur les classes prédéfinies de la collection CCA de [El Sulaiti, 2003].

Nous nous sommes tournés ensuite, vers la présentation du fonctionnement de notre approche en détail.

Pour l'évaluation des résultats obtenus par notre approche, nous nous sommes alignés aux recherches conduites dans ce domaine, on optant pour les mesures réputées efficaces dans ce genre d'évaluation, à savoir, l'expertise humaine.

Enfin, le déroulement et les résultats obtenus par cette démarche sont exposés dans le chapitre suivant.

Chapitre VI

Expérimentations, résultats et discussions

Chapitre VI
Expérimentations, résultats et discussions

1. Introduction

Ce chapitre traite les résultats obtenus au fur et à mesure du déroulement des différentes expérimentations menées sur les quatre collections via l'approche que nous avons développée intitulée LDK-Means. Nous allons commencer par présenter les quatre collections textuelles sur lesquelles ont été menés nos travaux ainsi que les différents outils utilisés.

2. Collections textuelles utilisées

La majorité des expérimentations de la présente étude ont été conduites sur les quatre collections suivantes :

- *Corpus of Contemporary Arabic* (CCA): compilé par *Latifa El Sulaiti* [El Sulaiti, 2003] de la radio du Qatar.
- *Al watan on-line newspaper* : collecté d'*Alwatan on-line newspaper* pendant l'année 2004 par *Mourad Abbas* [Abbas *et al.*, 2004].
- *BBC Arabic corpus*: membre de OSAC (*Open Source Arabic Corpora*), collectés de *bbcarabic.com* par *Motaz K. Saad* [Saad et Achour, 2010].
- *Open Source Arabic Corpus* (OSAc): membre de OSAC (*Open Source Arabic Corpora*), collectée à partir de multiples sites par *Motaz K. Saad* [Saad et Achour, 2010].

Les caractéristiques de ces collections sont décrites dans le tableau 6.1 ci-après :

Collections	Nbr de classes (prédéfini)	Nbr de documents	Nbr de mots (Million)	Taille (MOctet)	Format
CCA[1]	15	432	0,82	10	XML + TXT (UTF8)
Al Watan on-line newspaper	6	20291	10	118	TXT (UTF8)
BBC Arabic corpus	7	4763	1,86	29	TXT (UTF8)
OSAc	11	22429	18,18	182	TXT (UTF8)

Tableau 6.1: Les quatre collections utilisées en expérimentations.

Le tableau 6.2 décrit en détail les classes prédéfinies des quatre collections textuelles utilisées.

[1] En enlevant les documents corrompus le nombre total des documents pris dans les expérimentations est revenu à 406 documents.

Collections	Classes	Nbr de documents
CCA	*Short stories*	25
	Education	8
	Religion	19
	Autobiography	70
	Sociology	30
	Tourist/travel	60
	Recipes	9
	Science	57
	Sports	4
	Economics	29
	Children' stories	26
	Health and medicine	32
	Interviews	20
	Politics	10
	Spoken	7
	Total	**402**
Al Watan on-line newspaper	*Culture*	2782
	Religion	3860
	Economy	3468
	Local News	3596
	International News	2035
	Sports	4550
	Total	**20291**
BBC Arabic corpus	*Middle east news*	2356
	World news	1489
	Economy	296
	Sport	219
	Word press	49
	Science and technology	232
	Art and culture	122
	Total	**4763**
OSAc[1]	*Astronomy*	557
	Law	944
	Economy	3102
	Education	2626
	Entertainment	982
	History	3233
	Recipe	2373
	Story	726
	Religion	3171
	Health	2296
	Sport	2419
	Total	**22429**

Tableau 6.2 : Les classes prédéfinies des quatre collections textuelles.

[1] Les classes initiales au nombre de onze (11) ont été subdivisées en cinquante (50) classes.

3. Présentation des outils utilisés

L'outil LDK-Means fait appel à des fonctions incorporées dans deux environnements, MALLET et *Text Garden.*

3.1. Environnement MALLET (*MAchine Learning for LanguagE Toolkit*)

Nous avons fait appel à deux fonctions du paquetage (*package*) MALLET[1] (vollet *Topic Models* implémenté selon [Yao *et al.*, 2009]), un *Java open source* destiné aux traitements statistiques du langage naturel tels que la classification supervisée et non supervisée textuelle, la modalisation thématique, extraction d'information, et d'autres applications d'apprentissage automatique relatives au texte.

3.1.1. Importation des documents

La fonction *"Import-dir"* consiste à importer des fichiers en format texte au format interne supporté par MALLET.

Exemple d'utilisation

import-dir --input e:\sulaiti\trans --output Trans.mallet --preserve-case true --keep-sequence true

Cette instruction transforme les documents du dossier *"e:\sulaiti\trans"* dans le fichier *"Trans.mallet"* sous le format supporté par les fonctions MALLET.

- *"--preserve-case"* préserve la casse des caractères des mots.
- *"--keep-sequence"* garde la séquence des mots dans le document, elle est nécessaire pour la détection thématique.

3.1.2. Construction des modèles thématiques (utilisée comme outil de la CNST)

La deuxième fonction que nous avons utilisée est *"train-topics"*, cette fonction construit un modèle thématique à partir d'un fichier de document généré par la fonction *import-dir*.

Exemple d'utilisation

Train-topics --input Trans.mallet --num-topics 15 --output-doc-topics TransDocTop
Cette instruction affecte les documents aux thèmes générés (ici 15 thèmes) dans le fichier *TransDocTop*.

Train-topics --input Trans.mallet --num-topics 15 --output-topic-keys TransKeyTop
Cette instruction génère les mots descriptifs de chaque thème dans le fichier *TransKeyTop*.

Train-topics --input Trans.mallet --num-topics 15 --xml-topic-phrase-report TransPhraseTop
Cette instruction génère les phrases descriptives de chaque thème dans le fichier *TransKeyTop*.

[1] http://mallet.cs.umass.edu

3.2. Environnement *Text Garden (Text-Mining Software Tools)*

Ici aussi nous avons fait appel à deux fonctions de l'environnement *Text Garden*[1], un ensemble d'outils redirigés en C++, destinés aux tâches du *Text-Mining* telles que la classification supervisée et non supervisée textuelle et d'autres applications d'apprentissage automatique relatives au texte.

3.2.1. Importation des documents

La fonction *"Txt2Bow"* consiste à importer des fichiers en format texte au format BOW (*Bag Of Words*) ou sac de mots.

Exemple d'utilisation

```
Txt2bow -idir: e:\sulaiti\trans -o:trans.bow
```

Cette instruction transforme les documents du dossier *"e:\sulaiti\trans"* dans le fichier *"Trans.bow"* sous le format supporté par les fonctions *de Text Garden*.

3.2.2. Fonction des K-Moyennes

La deuxième fonction que nous avons utilisée est *"bowkmeans"*, cette fonction construit une partition avec le nombre de clusters spécifié avec la méthode des K-Moyennes à partir du fichier de documents généré par la fonction *"Txt2Bow"*.

Exemple d'utilisation

```
Bowkmeans -i:Trans.bow -ctrials:7 -clusts:50 -ot:TransKMeans.txt  -sdnm:T
```

Cette instruction applique les K-Moyennes sur le fichier *"Trans.bow"*

- *"-ctrials"* indique le nombre d'itérations à effectuer.
- *"-clusts"* indique le nombre de clusters à générer.
- *"-ot"* indique le fichier dans lequel sera généré la description des clusters ainsi que les mots fréquents.
- *"-sdnm"* préserve les noms de documents dans le fichier généré par l'option *"-ot"*.

3.3. Configuration des paramètres

Les fonctions utilisées citées plus haut sont configurables via des paramètres qui prennent des valeurs par défauts dans le cas de leurs omissions. Les paramètres cités dans l'exemple ne sont pas les seuls, d'autres paramètres sont proposés par les fonctions utilisées.

4. Déroulement des différentes étapes de l'étude menée

L'étude menée est passée par plusieurs étapes importantes avant d'arriver aux objectifs fixés. Nous décrivons ci-après les dites étapes.

[1] http://ailab.ijs.si/dunja/TextGarden/

4.1. Prétraitement et préparation des documents

Comme il a été mentionné dans le deuxième chapitre, le processus de préparation des collections est une phase cruciale dans n'importe quel processus de traitement automatique du langage naturel. Dans ce qui suit, nous allons voir en détail tout le processus de prétraitement exposé dans la phase 1 de la section 3 du cinquième chapitre.

4.1.1. Normalisation du texte

C'est la première étape avec laquelle nous avons entamé notre travail. Elle consiste à transformer l'ensemble des documents des quatre collections sous un format facilement manipulable par les traitements suivants dans le processus. L'arabe étant encodé sous plusieurs formats d'encodage, LDK-Means permet de passer de n'importe quel format reconnaissable par *Microsoft Word*[1] vers le format texte avec un encodage UTF-8 *(Unicode Text Format-8)* ou un encodage Windows-1256 (ou cp1256).

4.1.2. Nettoyage de textes, Translittération et Tokenisation

Nettoyage : Le nettoyage de textes enlève toutes les séquences de caractères qui peuvent engendrer un bruit dans le traitement (Balises XML, chiffres, caractères latins, symboles, …) et ne garder que les séquences de lettres arabes.

Translittération : Les séquences de lettres arabes sont translittérées en des séquences de lettres latines (format standard ASCII). En raison d'incorporation des outils cités dans la section 2 plus haut, nous avons utilisé deux translittérations (tableau 6.3).

Caractère arabe	Caractère de latin		Caractère arabe	Caractère latin	
	1ère forme	2ème forme		1ère forme	2ème forme
ا	A	A	س	s	S
ئ	F	u	ش	P	5
إ	c	6	ص	S	0
ؤ	e	9	ض	D	1
أ	a	8	ط	T	2
آ	B	7	ظ	Z	3
ء	C	C	ع	E	E
ب	b	B	غ	g	g
ت	t	T	ف	f	f
ة	p	P	ق	q	q
ث	v	V	ك	k	k
ج	j	J	ل	l	l
ح	H	H	م	m	m
خ	x	X	ن	n	n
د	d	D	ه	H	4
ذ	O	O	و	w	w
ر	r	R	ي	y	y
ز	z	Z	ى	Y	i

Tableau 6.3: Caractères arabes et leurs correspondants en caractères latins (translitération).

[1] Intégration de la bibliothèque *Microsoft Word XX Object Library*.

Il faut noter qu'au cours de cette étape, les marques diacritiques sont aussi éliminées et une normalisation dite légère (*Light normalization*)[1] est appliquée pour les lettres ا, أ, إ, آ, ئ, ؤ, ء et ى, ي . Aussi le التطويل (Tatweel) ou "el kasheeda", qui est une forme de rallongement des lettres arabes, est aussi éliminé.

Tokenisation : Les séquences en caractères translitérés sont transformées en *Tokens* (des mots séparés par des retours chariot), ce qui permet de faciliter l'élimination des mots outils ou mots vides de sens (contenus dans le fichier Stoplist.txt qui sera présenté plus loin) et les traitements suivants.

La figure 6.1 donne le résultat de l'application des trois opérations citées plus haut pour le document *CHD01*.txt de la collection El-Sulaiti:

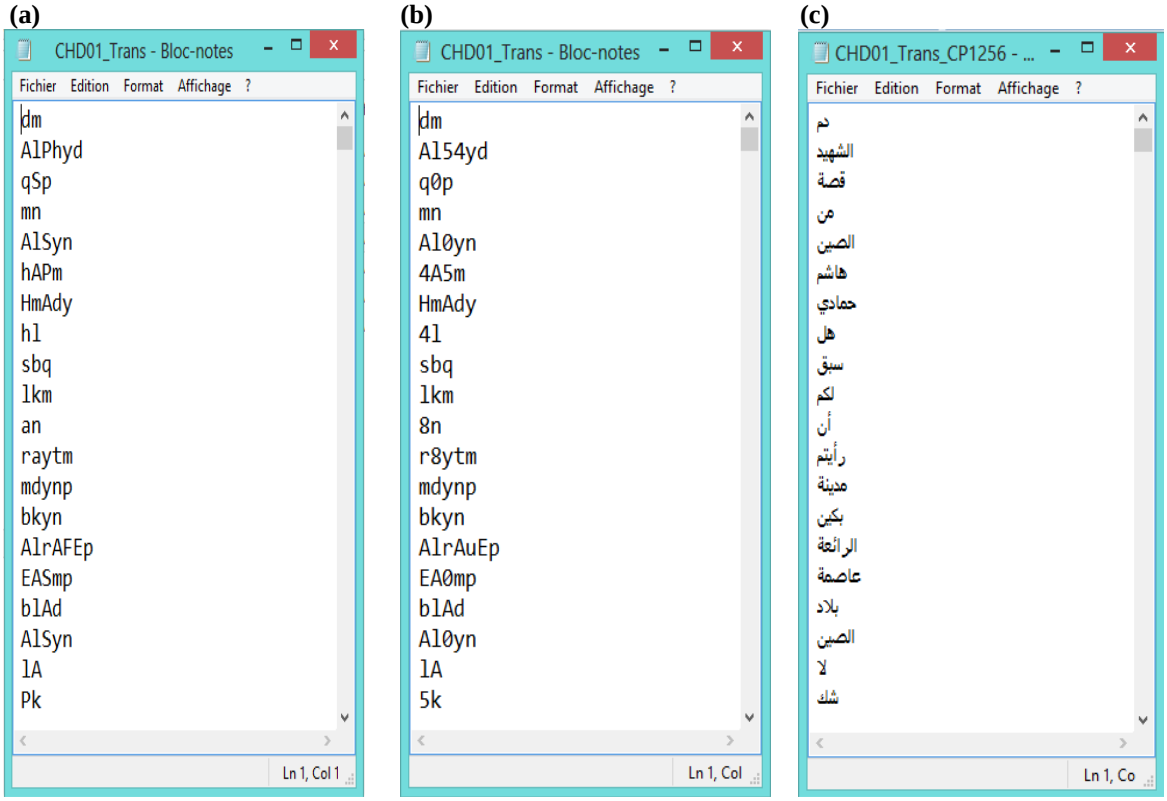


Figure 6.1 : (a) Première forme de translittération, (b) Deuxième forme de translittération, (c) Correspondant en lettres arabes.

4.1.3. Elimination des mots outils (*Stop words*)

Comme les autres langues, l'arabe contient aussi des mots outils (ou mots vides de sens) qui ne véhiculent pas un sens particulier utile pour la sémantique du texte. Plusieurs auteurs ont mis en place leurs propres listes de mots outils en plus des mots fréquemment utilisés pour les employer dans leurs recherches.

Les deux listes les plus répandues, et largement reprises par d'autres travaux dans ce domaine, sont celles de *Darwish K.* [Darwish et Oard, 2002] renferment 131 mots, et celle de

[1] Il existe une autre normalisation dite normalisation agressive (*aggressive normalization*) dans laquelle deux normalisations sont faites: ا, أ, إ, آ, ئ, ؤ, ء → A et ى, ي → y .

de *Khoja S.* [Khoja et Garside, 1999] renfermant 168 mots. La liste que nous avons construite (en annexe) comprend 875 mots différents englobant presque toutes les prépositions et particules de la langue arabe. La raison pour laquelle la taille de notre liste est plus large par rapport à celle de *Darwish* et *Khoja* est due à deux facteurs :

- Combinaison de plusieurs listes à partir de plusieurs sources (ex. les listes accompagnant les quatre collections) ;
- La prise en compte de l'agglutination que nous avons expliquée dans le quatrième chapitre. Le fait qu'une conjonction de coordination, par exemple, colle à un mot outil génère une multitude de formes pour ce même mot outil. Ainsi, nous pouvons trouver dans notre liste plusieurs formes pour ce mot. Par exemple, pour le mot هذا (ceci) nous avons comme extensions وهذا (et ceci), فهذا (ensuite ceci), etc.

Ici, il faut noter que les fichiers obtenus après l'opération de translitération sont appelés fichiers translittérés (Figure 6.2: *CHD01_Trans.txt*). L'opération de suppression des mots outils donne naissance à des fichiers appelés fichiers translitérés nettoyés (*ex CHD01_ClTrans.txt*).

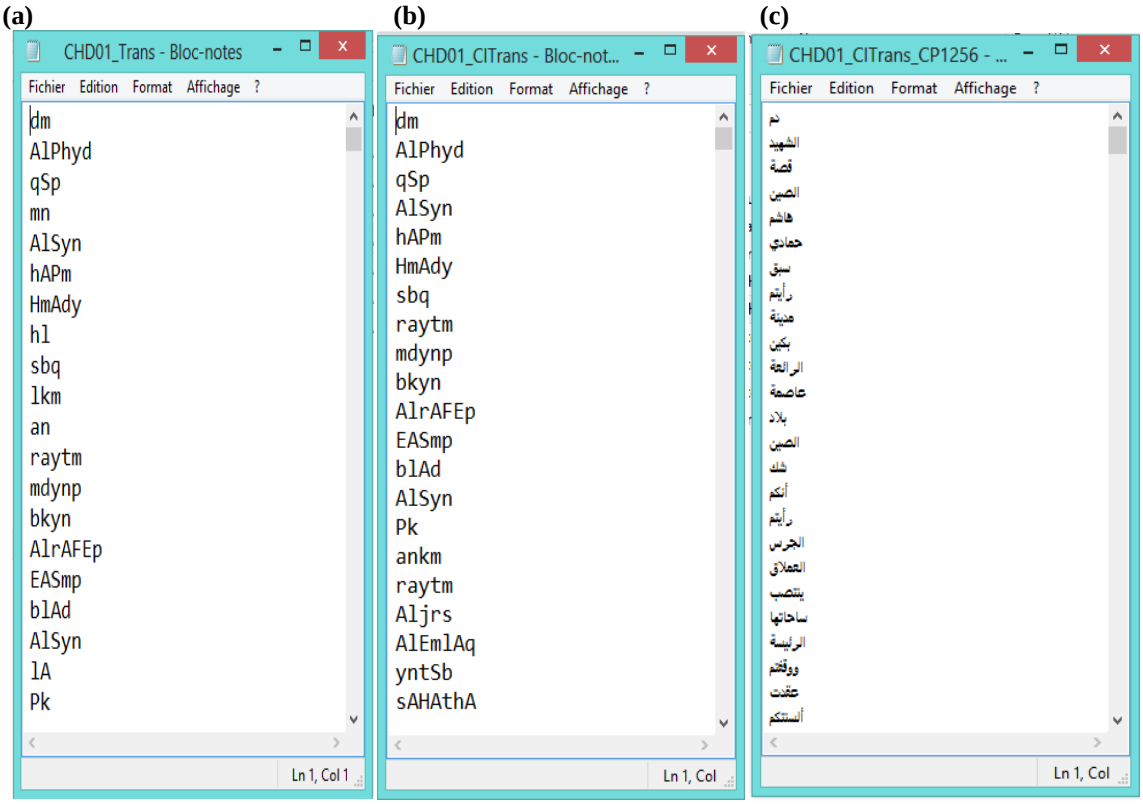


Figure 6.2 : (a) Forme translittérée (b) Forme nettoyée (c) Correspondant en lettres arabes.

4.1.4. Stemming

Comme il a été mentionné dans le deuxième chapitre les algorithmes de stemming en langue arabe les plus répandus sont *Al-Stem* **[**Darwish et Oard, 2002**]** et *StemmerLight10* [Larkey *et al.*, 2005].

Ce genre d'algorithmes opère une légère troncature sur le début et la fin des mots. Pour

ce faire, des listes de préfixes et de suffixes à une lettre, à deux lettres et à trois lettres sont établies. Le choix de ces listes est déterminé généralement selon des statistiques. Ces statistiques analysent les fréquences d'occurrence des préfixes et des suffixes sur les mots d'une grande collection de textes. La décision de tronquer un préfixe ou un suffixe d'un mot est faite selon de simples règles comme la longueur de mots. Par exemple, on ne peut pas tronquer un préfixe à trois lettres d'un mot de longueur quatre.

Dans notre étude nous avons travaillé avec la liste exposée dans Tableau 6.4, qui est à la base de l'algorithme *Al-Stem* [Darwish et Oard, 2002].

Préfixes							
فا	ويـ	الـ	ومـ	نتـ	متـ	بتـ	والـ
لا	فيـ	للـ	كمـ	بمـ	و-	يتـ	فالـ
با	وا	ليـ	فمـ	لمـ	ستـ	لتـ	بالـ
Suffixes							
ا	ـة	ين	ية	هم	ه	وه	ات
	ـه	يه	ك	هن	م	ـان	وا
	ي	ية	نا	ها	كم	ي	ون

Tableau 6.4 : Liste des préfixes et suffixes les plus fréquents *(Al-stem)*.

Pour obtenir des textes « stemmés » nous procédons comme suit :

1. Au début de chaque mot enlevé (remplacement par un vide) :
   Pour [wfb] Al:
   le préfixe Al (ال)
   les lettres w,f,b (و ف ب) si elles sont suivies par le préfixe Al (ال)
   de même pour : [bylmtwsn]t,[blwkf]m,[Al]l, [wlsf]y,[wflb]A

2. A la fin de chaque mot enlevé (remplacement par un vide) :
   At,wA,tA,wn,wh,An,ty,th,tm,km,hA,tk,yp,nA,p,h,y,A
   Pour h[nm]:
   le suffixe h (ه)
   les lettres n, m (ن م) si elles sont précédées par le suffixe h (ه)
   de même pour y[nh]

Le fragment[1] de code suivant (rédigé en PERL), figure 6.3 permet l'obtention des fichiers stems :

```
foreach $line (@lines) {
      if ($line =~ /^([wfb]Al|[bylmtwsn]t|[blwkf]m|[Al]l|[wlsf]y|[wflb]A|)(.*?)
            (At|wA|tA|wn|wh|An|ty|th|tm|km|h[nm]|hA|yp|tk|nA|y[nh]|[phyA]|)$/)
                  {
```

Figure 6.3 : Extrait du code PERL de l'algorithme de stemming.

La figure 6.4 montre le résultat de l'opération du stemming sur le document *CHD01_ClTrans.txt.*

[1] Ce fragment est tiré de l'algorithme *Al-stem* (rédigé en langage *PERL*) [Darwish *et al.*, 2005]

Inconvénient de l'algorithme de stemming

L'application de l'algorithme de stemming peut engendrer parfois des défaillances telles que, par exemple, `bArzA` (بارزا) devient `rz` (رز) ce qui est absurde. L'effet de ces aberrations est, heureusement, atténué par leurs raretés de présence dans les textes stemmés générés.

4.1.5. Résultats de la phase de prétraitement et de préparation des documents

A la fin de cette phase nous aurons pour chaque collection six dossiers qui contiennent respectivement les fichiers translitérés, les fichiers translitérés nettoyés et les fichiers des stems pour la première forme de translittération et même chose pour la deuxième forme.

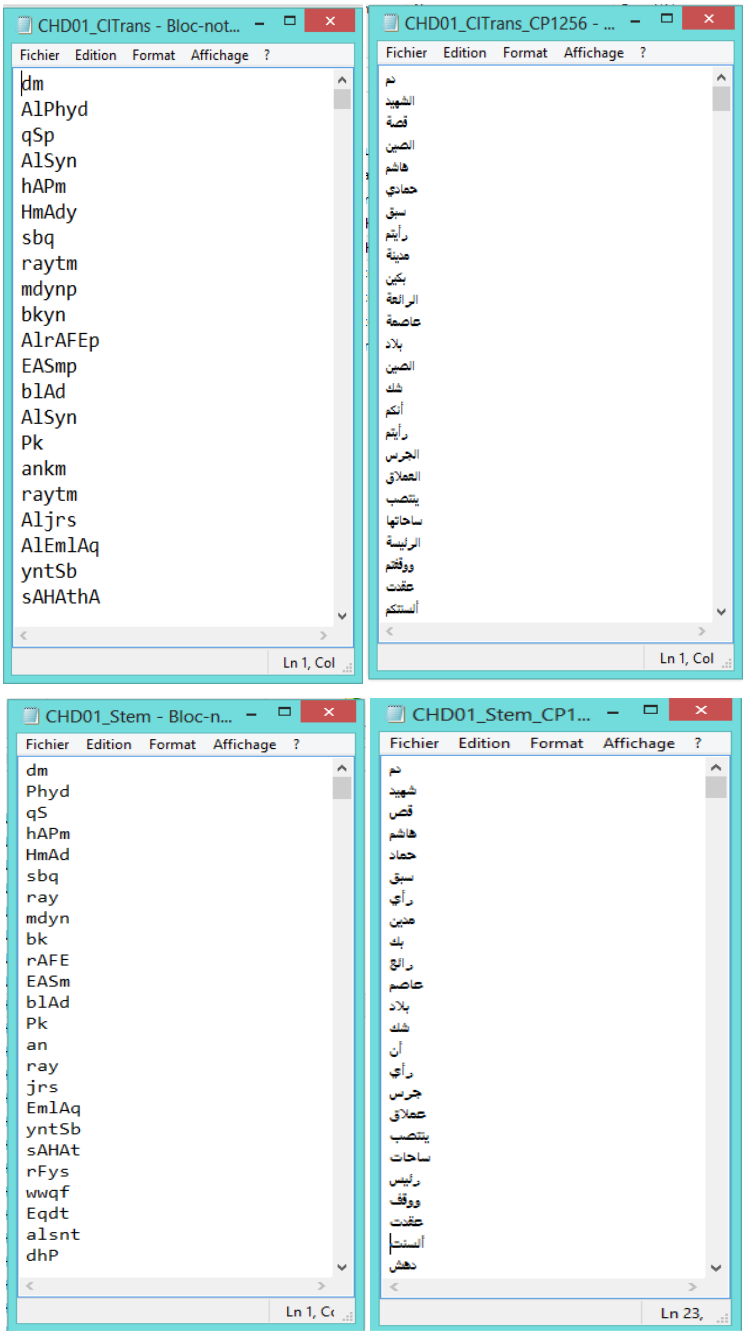


Figure 6.4 : Résultats de l'opération du stemming sur le document *CHD1_ClTrans.txt.*

4.2. Classification non supervisée textuelle: LDA versus K-Moyennes

Une fois les prétraitements et la préparation des documents sont terminés, nous avons soumis les quatre collections textuelles sous les trois formes (translittérée, nettoyée et stemmée) au processus de classification avec les deux méthodes LDA et K-Moyennes.

Il est à noter qu'au début de cette étape l'étude a été entamée sur la collection *OSAc* [Kelaiaia et Merouani, 2013] avec la F-Mesure et l'Entropie comme mesures de performance. Puis cette étude a été élargie aux trois autres collections [Kelaiaia et Merouani, 2014] avec quatre mesures de performances à savoir l'indice de *Rand*, l'indice de *Jaccard*, la F-Mesure et l'Entropie.

4.2.1. Réglage des paramètres des deux fonctions

a. Méthode LDA

- *"--num-topics"* : est pris égale au nombre de classes prédéfinies pour chacune des quatre collections (respectivement 15, 6, 7, 50) ;

Les valeurs des autres paramètres sont prises par défauts, nous citons les plus importants :

- *"--num-iterations"* : 1000 itérations (valeur prise par défaut);
- *"--Alpha"* : 50.0 (valeur prise par défaut) ;
- *"--Beta"* : 0.01 (valeur prise par défaut) ;
- *"--Gamma"* : 0.01 (valeur prise par défaut).

Les valeurs de *Alpha et Beta* sont des valeurs expérimentales définis par [Steyvers et Griffiths, 2004] à l'issue de plusieurs expérimentations.

b. Méthode K-Moyennes

- *"-clusts"* : est pris égale au nombre de classes prédéfinies pour chacune des quatre collections (respectivement 15, 6, 7, 50) ;
- *"-ctrials"* : 7 itérations (stabilisation).

4.2.2. Résultats et discussions

Commençons par présenter les résultats de chaque collection à part :

a. Résultats par collection

a.1. Collection CCA : Pour cette collection nous avons obtenus les résultats suivants :

Mesures / Méthodes / Collection	Indice de *Rand*	Indice de *Jaccard*	F-Mesure	Entropie
K-Moyennes Trans	0,859	0,136	0,240	0,548
K-Moyennes ClTrans	0,869	0,163	0,280	0,576
K- Moyennes Stem	0,866	0,168	0,287	0,529
LDA Trans	0,899	0,287	0,446	0,353
LDA ClTrans	0,898	0,308	0,471	0,354
LDA Stem	0,914	0,363	0,532	0,307

Tableau 6.5 : Evaluation de la CNST sur la collection CCA.

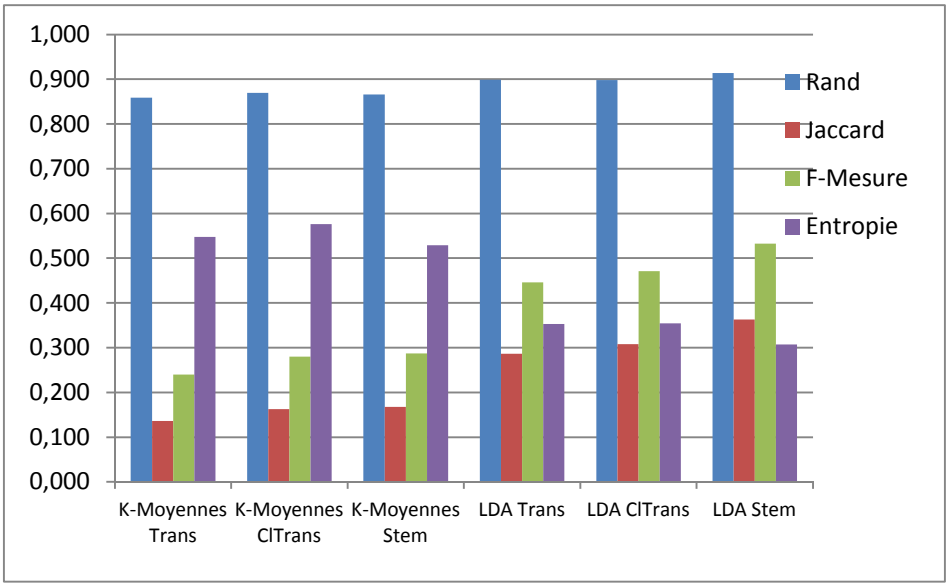


Figure 6.5 : Evaluation de la CNST sur la collection CCA.

Du tableau 6.5 et la figure 6.5 nous remarquons que :

- Les quatre mesures semblent, en général, indiquer la performance de LDA sur les trois formes par rapport aux K-Moyennes ;
- L'indice de *Jaccard* et la F-mesure indiquent l'amélioration de la qualité de la classification avec l'avancement du processus du prétraitement (translitération, nettoyage et stemming) pour les deux méthodes, ceci est à l'encontre des deux autres indices qui, eux, indiquent le contraire. Ceci est expliqué par la non prise en considération des documents mal classés dans les deux premiers indices.

a.2. Collection BBC :

Mesures / Méthodes / Collection	Indice de *Rand*	Indice de *Jaccard*	F-Mesure	Entropie
K-Moyennes Trans	0,627	0,143	0,250	0,595
K-Moyennes ClTrans	0,644	0,165	0,284	0,527
K-Moyennes Stem	0,645	0,173	0,295	0,541
LDA Trans	0,638	0,167	0,286	0,517
LDA ClTrans	0,630	0,154	0,266	0,521
LDA Stem	0,641	0,174	0,297	0,500

Tableau 6.6 : Evaluation de la CNST sur la collection BBC.

Pour la collection BBC, le tableau 6.6 et la figure 6.6 nous donnent :

- Sur la forme nettoyée les résultats de LDA sont moins bons que ceux obtenus pour les deux autres formes ;
- A l'inverse des résultats obtenus avec LDA, les résultats obtenus avec les K-Moyennes s'améliorent au fur et à mesure du processus du prétraitement ;
- Sur la deuxième forme (nettoyée) la méthode des K-Moyennes donne des résultats bien meilleurs que ceux obtenus avec LDA.

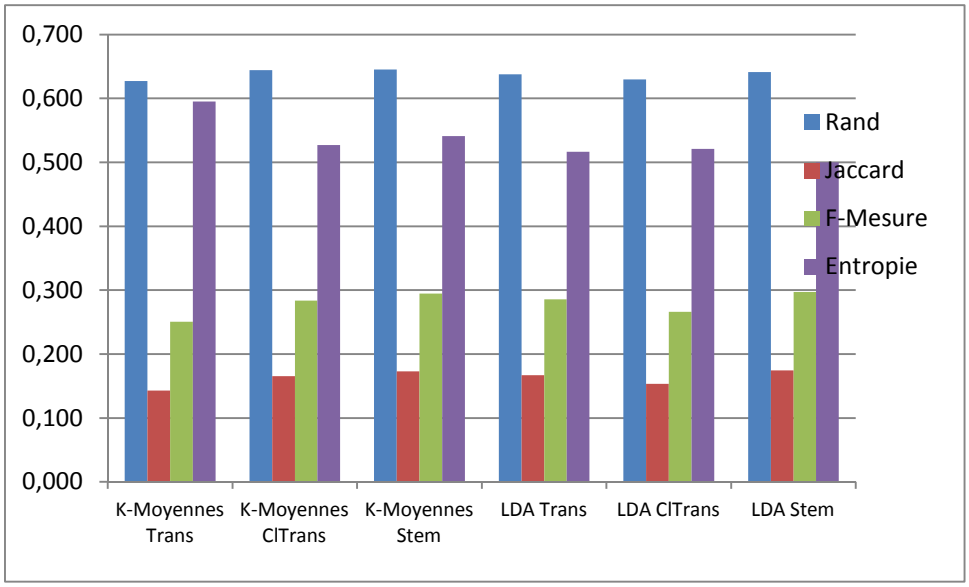


Figure 6.6 : Evaluation de la CNST sur la collection BBC.

a.3. Collection OSAc : Les résultats sont exposés dans le tableau 6.7:

Mesures / Méthodes / Collection	Indice de *Rand*	Indice de *Jaccard*	F-Mesure	Entropie
K-Moyennes Trans	0,953	0,343	0,511	0,199
K-Moyennes ClTrans	0,955	0,359	0,528	0,197
K-Moyennes Stem	0,954	0,330	0,496	0,198
LDA Trans	0,971	0,576	0,731	0,155
LDA ClTrans	0,972	0,586	0,739	0,152
LDA Stem	0,968	0,558	0,716	0,165

Tableau 6.7 : Evaluation de la CNST sur la collection OSAc.

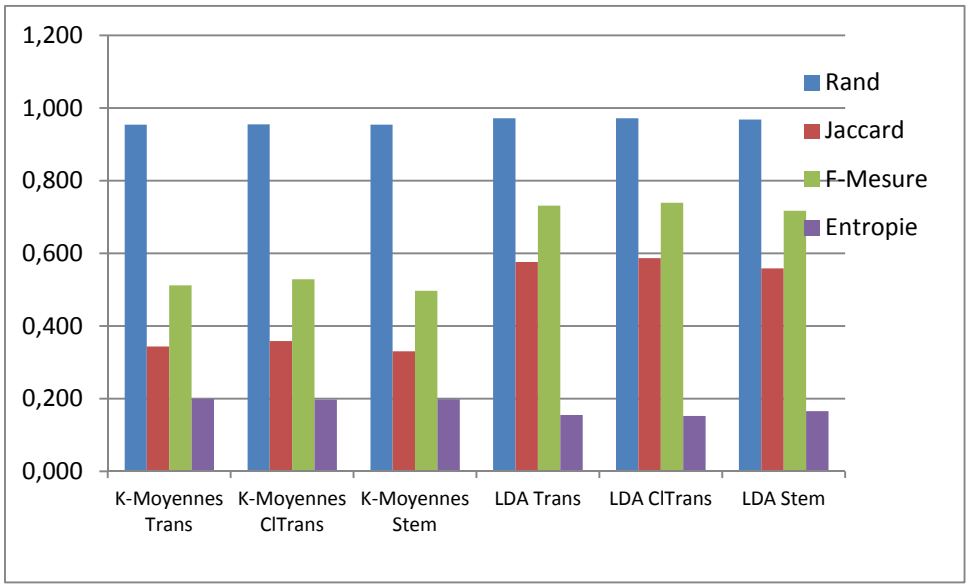


Figure 6.7 : Evaluation de la CNST sur la collection OSAc.

A partir du tableau 6.7 et la figure 6.7 nous remarquons que :

- De même que pour la collection CCA, les quatre mesures indiquent la performance de LDA sur les trois formes par rapport aux K-Moyennes ;
- Sur la forme nettoyée les deux méthodes ont un rendement meilleur que sur les autres formes ;
- Le processus de prétraitement agit bien sur la qualité de la CNST par la méthode LDA par rapport à la méthode des K-Moyennes ;
- Avec l'augmentation du nombre de cluster, la qualité de la CNST avec LDA s'améliore.

a.4. Collection Al-Watan :

Mesures Méthodes / Collection	Indice de *Rand*	Indice de *Jaccard*	F-Mesure	Entropie
K-Moyennes Trans	0,835	0,401	0,573	0,494
K-Moyennes ClTrans	0,838	0,385	0,556	0,479
K-Moyennes Stem	0,877	0,493	0,660	0,395
LDA Trans	0,861	0,452	0,622	0,383
LDA ClTrans	0,860	0,454	0,625	0,387
LDA Stem	0,877	0,480	0,648	0,371

Tableau 6.8 : Evaluation de la CNST sur la collection Al Watan.

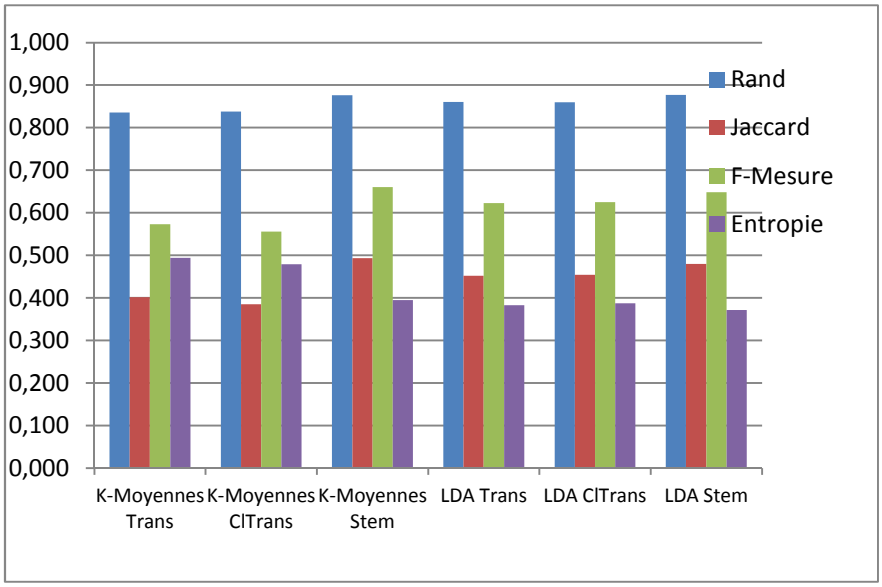


Figure 6.8 : Evaluation de la CNST sur la collection Al Watan.

Enfin, pour la collection Al Watan et à partir du tableau 6.8 et la figure 6.8 nous constatons que:

- Les quatre mesures indiquent la performance de LDA sur les deux formes translitérée et nettoyée par rapport aux K-Moyennes ;

- En général, les quatre indices indiquent une bonne réaction des deux méthodes au processus de prétraitement.

b. Synthèse des résultats sur les quatre collections

Les figures 6.9, 6.10, 6.11 et 6.12 et le tableau 6.9 synthétisent les résultats obtenus. Ces résultats, à première vue, montrent la performance de LDA par rapport aux K-Moyennes.

Mesure / Collection	Indice de *Rand*	Indice de *Jaccard*	F-Mesure	Entropie
Forme translittérée	2,35%	11,43%	12,77%	10,74%
Forme nettoyée	1,33%	10,76%	11,34%	9,12%
Forme stemmée	1,49%	10,29%	11,39%	7,97%

Tableau 6.9 : Moyennes de performances obtenues avec LDA par rapport aux K-Moyennes.

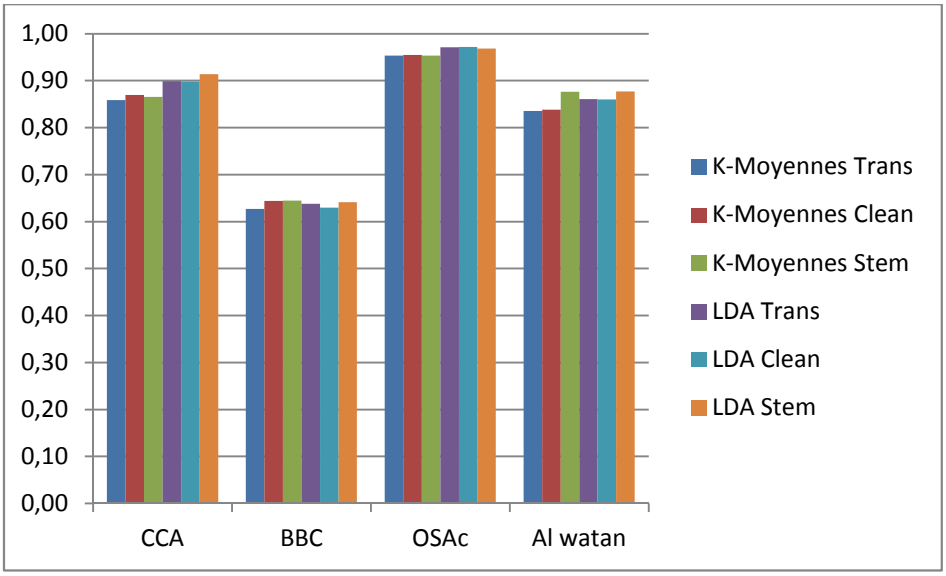


Figure 6.9 : Performances de LDA et K-Moyennes avec l'Indice de *Rand*.

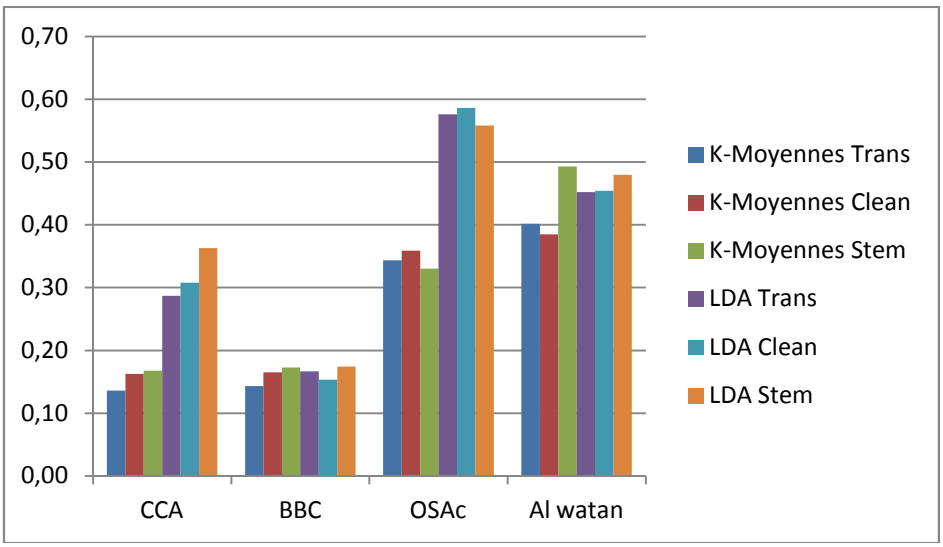


Figure 6.10 : Performances de LDA et K-Moyennes avec l'Indice de *Jaccard*.

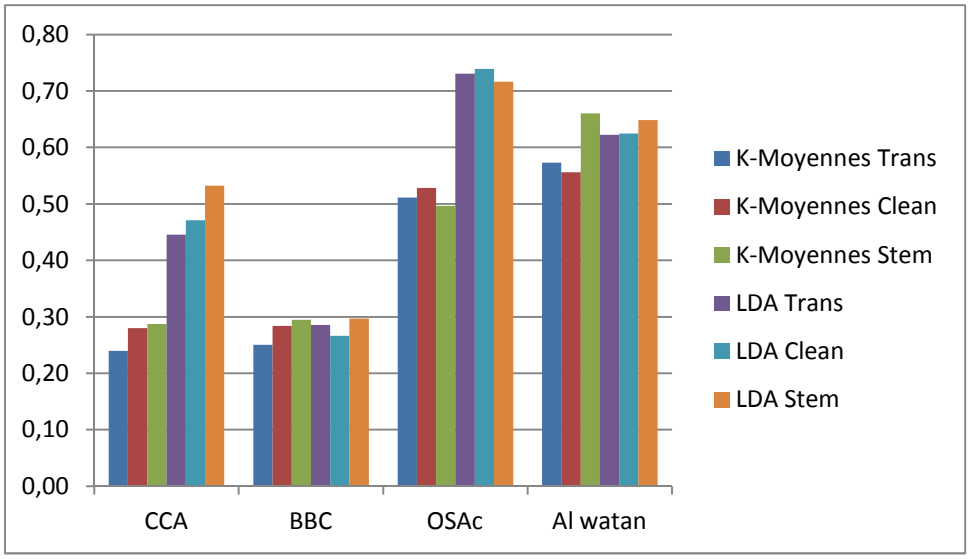


Figure 6.11 : Performances de LDA et K-Moyennes avec la F-Mesure.

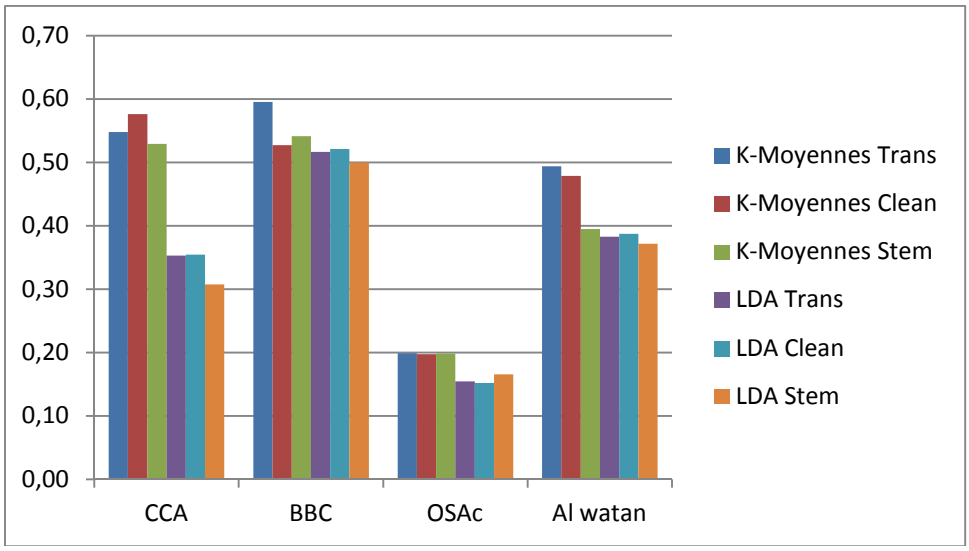


Figure 6.12 : Performances de LDA et K-Moyennes avec l'Entropie.

b.1. Performances de LDA par rapport aux K-Moyennes :

D'après le tableau 6.9, nous observons que LDA apporte une amélioration par rapport aux K-Moyennes sur les trois formes des quatre collections textuelles. Cette performance est en opposition avec les résultats obtenus dans [Lu *et al.*, 2011] qui stipule que les résultats obtenus avec les modèles thématiques probabilistes dans la CNST sont moins performants par rapport à ceux obtenus par les méthodes de classification traditionnelles telles que les K-Moyennes. Plus loin nous allons revenir sur ce point.

b.2. Influence de l'élimination des mots outils sur la qualité de la CNST :

Le tableau 6.10 compare les résultats de la CNST effectuée sur les formes translitérée et nettoyée avec les deux méthodes. Selon ces résultats la méthode des K-Moyennes réalise de meilleurs résultats que ceux obtenus avec LDA, ce qui veut dire qu'avec LDA la suppression des mots outils dans le texte arabe diminue la qualité de la classification obtenue. Nous allons revenir sur ce phénomène plus loin.

Ces résultats sont en accord avec ceux obtenus dans [Lu *et al.*, 2011] et confirme leurs contradiction avec les recherches de [Blei *et al.*, 2003 ; Griffiths et Steyvers, 2004] qui stipule qu'avec LDA la suppression des mots outils est une étape de prétraitement nécessaire.

Mesure / Méthode	Indice de *Rand*	Indice de *Jaccard*	F-Mesure	Entropie
K-Moyennes	0,80%	1,19%	1,84%	1,42%
LDA	-0,22%	0,53%	0,41%	-0,19%

Tableau 6.10 : Moyennes des performances obtenues avec LDA et K-Moyennes avec l'élimination des mots outils. (Comparaison entre la forme translittérée et la forme nettoyée).

b.3. Influence du stemming sur la qualité de la CNST :

Le tableau 6.11 compare les résultats de la CNST effectuée sur les formes translitérées et stemmées obtenues avec les deux méthodes. Nous remarquons que le stemming a amélioré la qualité de la classification obtenue, ce qui est peut-être dû principalement à l'effet du stemming, qui a contribué à supprimer les flexions des mots qui ont le même stem, ceci augmente la probabilité d'avoir les documents similaires dans le même cluster. En outre, nous remarquons aussi que, comme pour l'élimination des mots outils, la méthode des K-Moyennes semble réaliser de meilleurs résultats que ceux obtenus avec LDA.

Mesure / Méthode	Indice de *Rand*	Indice de *Jaccard*	F-Mesure	Entropie
K-Moyennes	1,66%	3,49%	4,12%	4,33%
LDA	0,80%	2,35%	2,74%	1,56%

Tableau 6.11 : Moyennes des performances obtenues avec LDA et K-Moyennes avec stemming. (Comparaison entre la forme translittérée et la forme stemmée).

c. LDA ou K-Moyennes pour la classification ?

Les recherches dans [Lu *et al.*, 2011] citées dans la section *b.1* plus haut ont été menées sur les deux collections *Reuters 21578* et *TDT2*, deux collections qui sont en langue anglaise. Au début et en comparant ces résultats à ceux obtenus au cours de nos recherches nous avons pensé que cela est peut être à l'origine de la baisse de performance des modèles thématiques par rapport aux K-Moyennes, c'est-à-dire que les caractéristiques morphosyntaxiques de la langue arabe (caractéristiques flexionnelles, dérivationnelle et agglutinant dans notre cas) ont influencé la performance de ces modèles dans la classification non supervisée, mais en affinant l'analyse de nos recherches nous avons constaté que cela est peut être faux.

Reprenant ce qui a été dit plus haut. En récapitulant les résultats rapportés par les quatre mesures de performance et à première vue LDA parait faire de meilleures performances que celles accomplies par la méthode des K-Moyennes. Revenant maintenant sur la répartition des textes de chaque classe sur les clusters formés par les deux méthodes pour la collection CCA par exemple. En examinant cette répartition nous avons remarqué que LDA a regroupé les textes selon des passages contenus dans ces derniers (des thèmes) sans prendre en compte la totalité des textes. Pour bien comprendre ce qui a été dit, prenant l'exemple des deux fichiers *Aut02_Trans* et *Aut06_Trans* qui décrivent respectivement l'autobiographie de deux grands écrivains arabes Mustapha ElSubai (مصطفى السباعي) et *Mohamed Moufdi Zakaria* (محمد مفدي

زكريا). La première méthode ne tient compte que des passages qui décrivent l'apport des deux hommes aux révolutions arabes contre l'occupation française (même cluster) à l'inverse de la méthode des K-Moyennes qui quant à elle, tient en compte d'autres passages qui décrivent d'autres cotés qui ont entouré la vie des deux hommes (deux clusters différents).

En conclusion à cette étude comparative, les résultats obtenus décrits plus haut ne nous ont pas convaincu pour l'utilisation de LDA dans notre processus LDK-Means eu égard à la complexité de cette dernière et le principe sur lequel elle se base en CNST.

d. Remarques

Au cours de cette étude comparative nous avons relevé plusieurs facteurs qui ont fortement influé les quatre mesures de performance (peut être qu'il fallait prendre d'autres mesures), tels que :

- le nombre de documents par collection ainsi que le nombre de clusters (c'est-à-dire que les clusters sont surpeuplés) ;
- l'origine des textes (rédacteur), le type des textes ;
- peut être qu'il fallait étudier chaque collection à part avec ses propres mots outils et son propre processus de prétraitement ; etc.

Une autre remarque importante est que nous n'avons pas pu confirmer que le caractère morphosyntaxique de la langue arabe a fortement influencé la classification avec les deux techniques, c'est-à-dire que les recherches dans le même axe sur les autres langues telles l'anglais ont montré des résultats similaires tels que l'amélioration des performances avec l'illimitation des mots outils et le stemming. L'utilisation du MSA (*Modern Standar Arabic*) dans les médias, principales sources des quatre collections, et le style de rédaction des textes formant ces collections qui n'utilise pas les formes complexes de cette langue (l'agglutination forte, exemple cité dans la section 2.3 dans le quatrième chapitre, n'est pas fréquemment utilisée) sont à l'origine de ces résultats.

4.3. Description des classes prédéfinies de la collection CCA en utilisant les deux techniques de l'approche proposée

Le but de cette étape est de tester la capacité des deux techniques de l'approche proposée intitulées mots fréquentielles latents et phrases fréquentielles latentes à décrire les classes prédéfinies de la collection CCA.

4.3.1. Processus de description

Pour atteindre cet objectif, nous avons généré les trois descriptions (mots fréquentiels latents, phrases fréquentielles latentes et mots fréquentiels) des quinze (15) classes de la collection CCA[1] avec les trois formes translittérée, nettoyée et stemmée.

a. Résultats obtenus

Les figures 6.13,…, 6.21 donnent un exemple de description par les trois techniques, mots fréquentiels, mots fréquentiels latents et phrases fréquentielles latentes, de la classe *Autobiography* sous les trois formes, translitérée, nettoyée et stemmée.

[1] L'utilisation de cette collection est argumentée dans le cinquième chapitre section 2.2.

De même, les tableaux 6.12,…, 6.20 donnent les descriptions générées par les trois techniques sous les trois formes de la collection CCA et leurs évaluations par les deux mesures Match@5 et MRR@5.

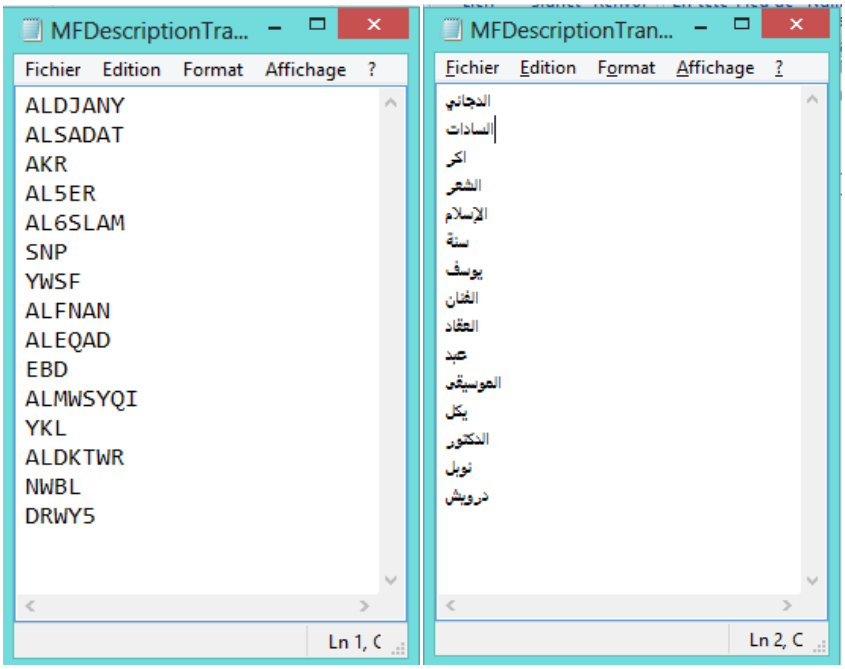


Figure 6.13 : Description par les mots fréquentiels de la classe *Autobiography* translitérée.

Classe	Décription manuelle	Mots fréquentiels			
		Rang mot fréquentiel	Mot fréquentiel	Match@5	MRR@5
0	Autobiographys	4	الشعر	1	0,25
1	Children' stories	3	الأرانب	1	0,33
2	Economics	2	الذهب	1	0,50
3	Education	2	المعلمة	1	0,50
4	Health and medicine	1	العين	1	1,00
5	Interviews	0		0	0,00
6	Politics	1	قطر	1	1,00
7	Recipes	0		0	0,00
8	Religion	1	الحجاب	1	1,00
9	Science	4	الكمبيوتر	1	0,25
10	Short stories	0		0	0,00
11	Sociology	4	الحضارة	1	0,25
12	Spoken	2	التعليم	1	0,50
13	Sports	5	البطولة	1	0,20
14	Tourist/travel	4	السياحة	1	0,25
Match@5				0,80	
MRR@5					0,40

Tableau 6.12 : Evaluation de la description par les mots fréquentiels de la collection CCA sous la forme translitérée.

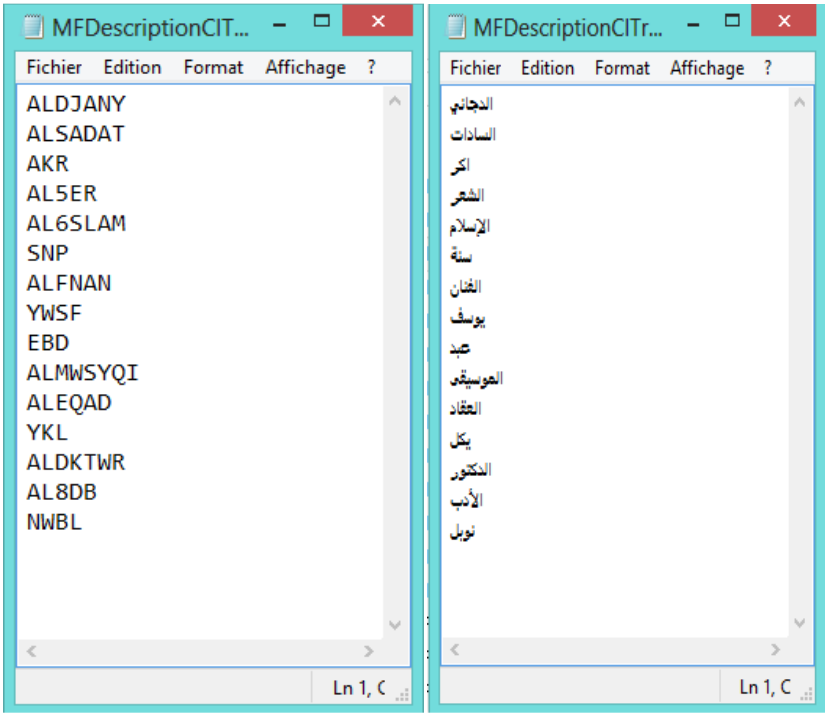


Figure 6.14 : Description par les mots fréquentiels de la classe *Autobiography* nettoyée.

Classe	Classe Label	Mots fréquentiels			
		Rang mot fréquentiel	Mot fréquentiel	Match@5	MRR@5
0	Autobiographys	4	الشعر	1	0,25
1	Children' stories	3	الأرانب	1	0,33
2	Economics	2	الذهب	1	0,50
3	Education	2	المعلمة	1	0,50
4	Health and medicine	1	العين	1	1,00
5	Interviews	5	الرواية	1	0,20
6	Politics	1	قطر	1	1,00
7	Recipes			0	0,00
8	Religion	1	الحجاب	1	1,00
9	Science	2	الكمبيوتر	1	0,50
10	Short stories	0		0	0,00
11	Sociology	3	الحضارة	1	0,33
12	Spoken	1	التعليم	1	1,00
13	Sports	2	البطولة	1	0,50
14	Tourist/travel	1	السياحة	1	1,00
Match@5				0,87	
MRR@5					0,54

Tableau 6.13 : Evaluation de la description par les mots fréquentiels de la collection CCA sous la forme nettoyée.

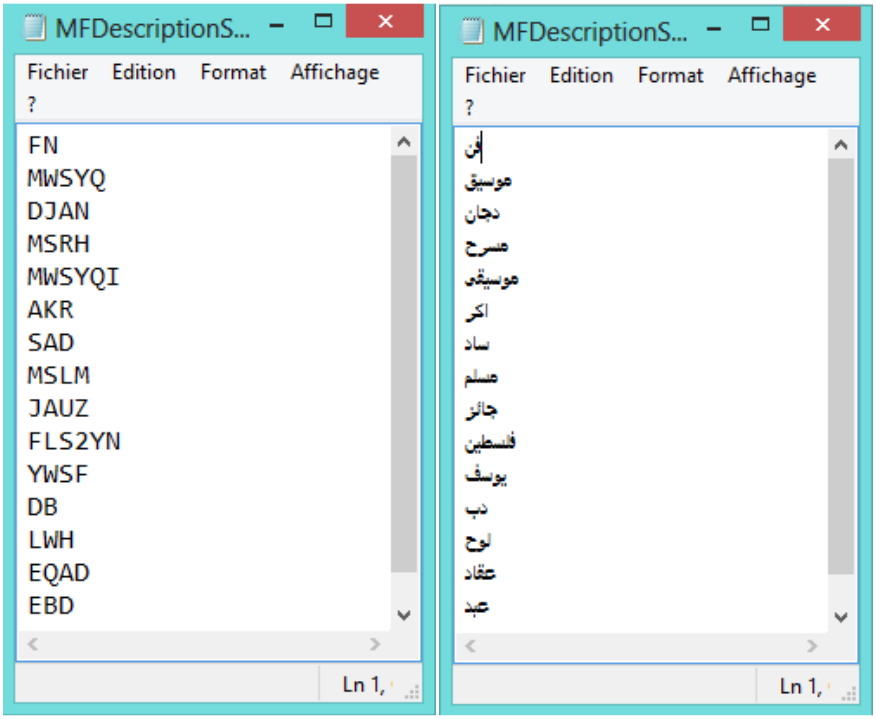


Figure 6.15 : Description par les mots fréquentiels de la classe *Autobiography* stemmée.

Classe	Classe Label	Mots fréquentiels			
		Rang mot fréquentiel	Mot fréquentiel	Match@5	MRR@5
0	Autobiographys	1	فن	1	1,00
1	Children' stories	2	عصفور	1	0,50
2	Economics	4	ذهب	1	0,25
3	Education			0	0,00
4	Health and medicine	1	اكتئاب	1	1,00
5	Interviews	0		0	0,00
6	Politics	1	فرنس	1	1,00
7	Recipes	2	سكر	1	0,50
8	Religion	1	حجاب	1	1,00
9	Science	2	جين	1	0,50
10	Short stories	0		0	0,00
11	Sociology	1	حضار	1	1,00
12	Spoken	1	جامع	1	1,00
13	Sports	1	مبارا	1	1,00
14	Tourist/travel	1	سياح	1	1,00
Match@5				0,80	
MRR@5					0,65

Tableau 6.14 : Evaluation de la description par les mots fréquentiels de la collection CCA sous la forme stemmée.

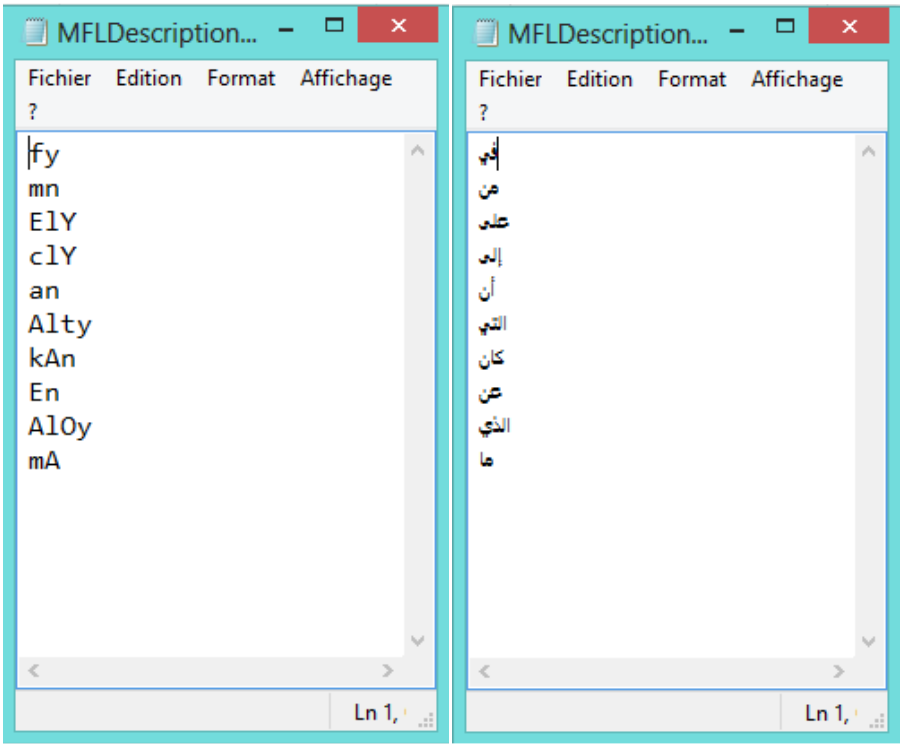


Figure 6.16 : Description par les mots fréquentiels latents de la classe *Autobiography* translitérée.

Classe	Décription manuelle	Mots fréquentiels latents			
		Rang mot fréquentiel latent	Mot fréquentiel latent	Match@5	MRR@5
0	Autobiographys	0		0	0,00
1	Children' stories	0		0	0,00
2	Economics	0		0	0,00
3	Education	0		0	0,00
4	Health and medicine	0		0	0,00
5	Interviews	0		0	0,00
6	Politics	0		0	0,00
7	Recipes	4	ملعقة	1	0,25
8	Religion	0		0	0,00
9	Science	0		0	0,00
10	Short stories	0		0	0,00
11	Sociology	0		0	0,00
12	Spoken	5	التعليم	1	0,20
13	Sports	0		0	0,00
14	Tourist/travel	0		0	0,00
Match@5				0,13	
MRR@5					0,03

Tableau 6.15 : Evaluation de la description par les mots fréquentiels latents de la collection CCA sous la forme translittérée.

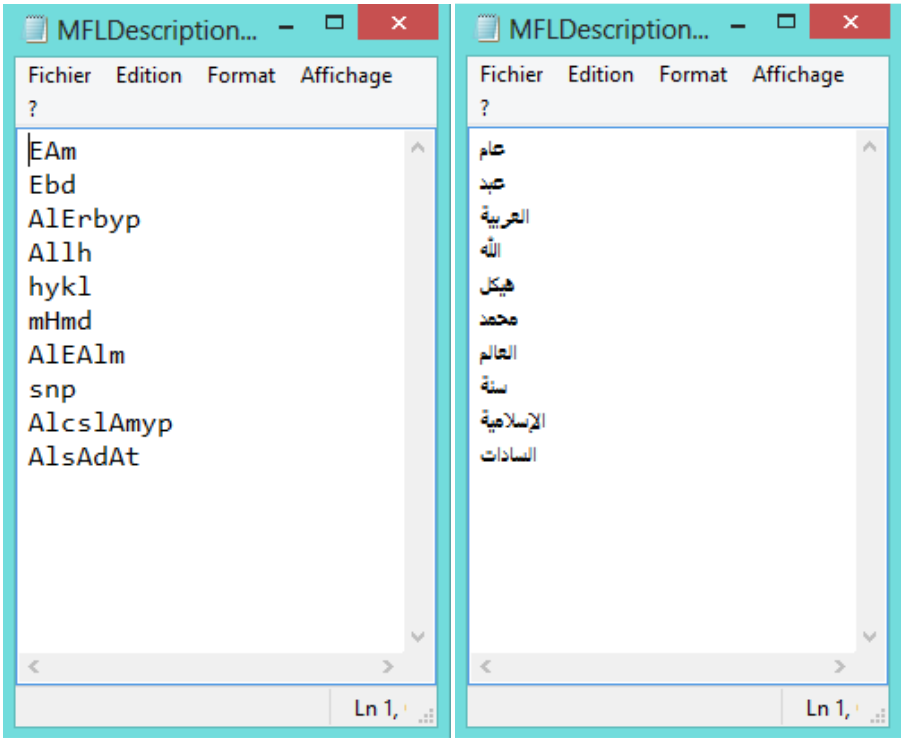


Figure 6.17 : Description par les mots fréquentiels latents de la classe *Autobiography* nettoyée.

Classe	Classe Label	Mots fréquentiels latents			
		Rang mot fréquentiel latent	Mot fréquentiel latent	Match@5	MRR@5
0	Autobiographys	3	العربية	1	0,33
1	Children' stories	3	دبدوب	1	0,33
2	Economics	1	السعودية	1	1,00
3	Education	1	التعليم	1	1,00
4	Health and medicine	1	الدم	1	1,00
5	Interviews	1	العربي	1	1,00
6	Politics	1	العراق	1	1,00
7	Recipes	1	ملعقة	1	1,00
8	Religion	1	الله	1	1,00
9	Science	2	العالم	1	0,50
10	Short stories	0		0	0,00
11	Sociology	2	العربية	1	0,50
12	Spoken	1	التعليم	1	1,00
13	Sports	1	البطولة	1	1,00
14	Tourist/travel	2	السياحة	1	0,50
Match@5				0,93	
MRR@5					0,74

Tableau 6.16 : Evaluation de la description par les mots fréquentiels latents de la collection CCA sous la forme nettoyée.

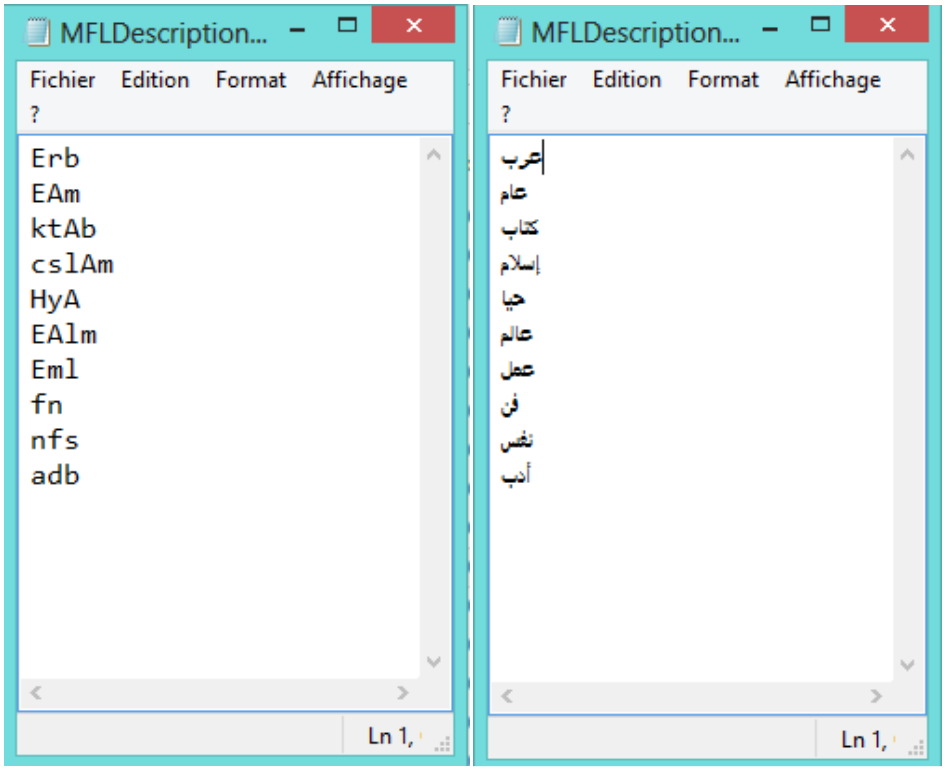


Figure 6.18 : Description par les mots fréquentiels latents de la classe *Autobiography* stemmée.

Classe	Classe Label	Mots fréquentiels latents			
		Rang mot fréquentiel latent	Mot fréquentiel latent	Match@5	MRR@5
0	Autobiographys	1	عرب	1	1,00
1	Children' stories	3	ديدوب	1	0,33
2	Economics	1	شرك	1	1,00
3	Education	1	تعليم	1	1,00
4	Health and medicine	2	جراح	1	0,50
5	Interviews	1	عرب	1	1,00
6	Politics	1	عراق	1	1,00
7	Recipes	1	زيت	1	1,00
8	Religion	1	إسلام	1	1,00
9	Science	1	عالم	1	1,00
10	Short stories			0	0,00
11	Sociology	1	عرب	1	1,00
12	Spoken	1	تعليم	1	1,00
13	Sports	1	بطول	1	1,00
14	Tourist/travel	1	سياح	1	1,00
Match@5				0,93	
MRR@5					0,86

Tableau 6.17 : Evaluation de la description par les mots fréquentiels latents de la collection CCA sous la forme stemmée.

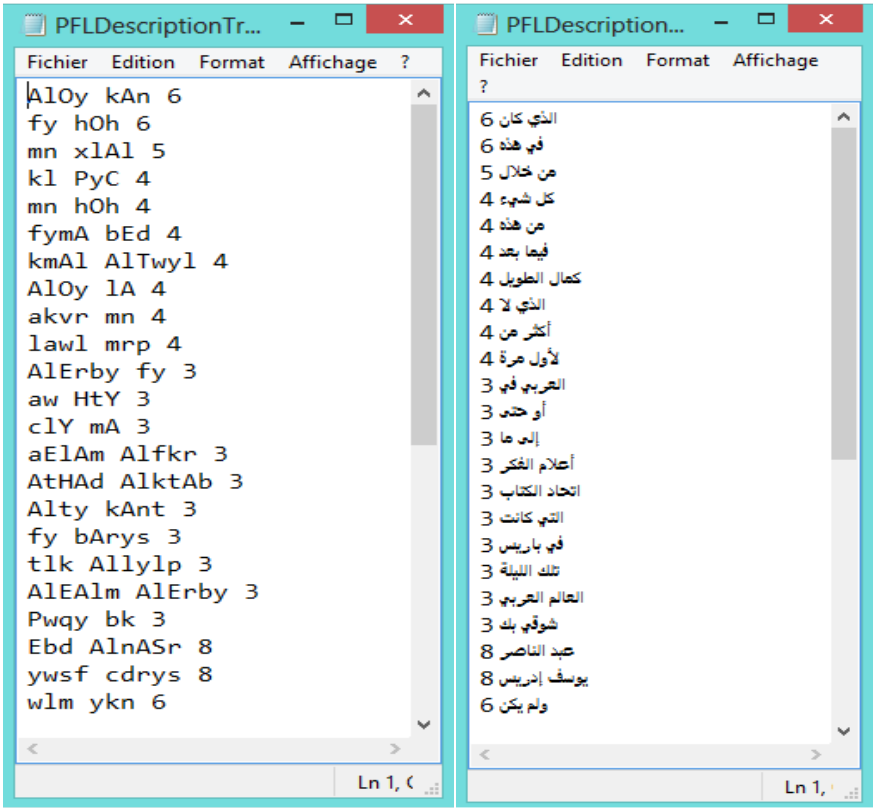


Figure 6.19 : Description par les phrases fréquentielles latentes de la classe *Autobiography* translitérée.

Classe	Décription manuelle	Phrases fréquentielles latentes			
		Rang phrase fréquentielle lattente	phrase fréquentielle lattente	Match@5	MRR@5
0	Autobiographys			0	0,00
1	Children' stories	2	الحذاء الخشبي	1	0,50
2	Economics	1	المملكة العربية السعودية	1	1,00
3	Education	2	البحث العلمي	1	0,50
4	Health and medicine	3	القدرة البصرية	1	0,33
5	Interviews	1	العالم العربي	1	1,00
6	Politics	1	الولايات المتحدة	1	1,00
7	Recipes	2	ملعقة صغيرة	1	0,50
8	Religion	1	أهل السنة	1	1,00
9	Science	3	الكائنات الحية	1	0,33
10	Short stories	0		0	0,00
11	Sociology	0		0	0,00
12	Spoken	0		0	0,00
13	Sports	3	دوري ابطال	1	0,33
14	Tourist/travel	2	لدول الخليج	1	0,50
Match@5				0,73	
MRR@5					0,47

Tableau 6.18 : Evaluation de la description par les phrases fréquentielles latentes de la collection CCA sous la forme translittérée.

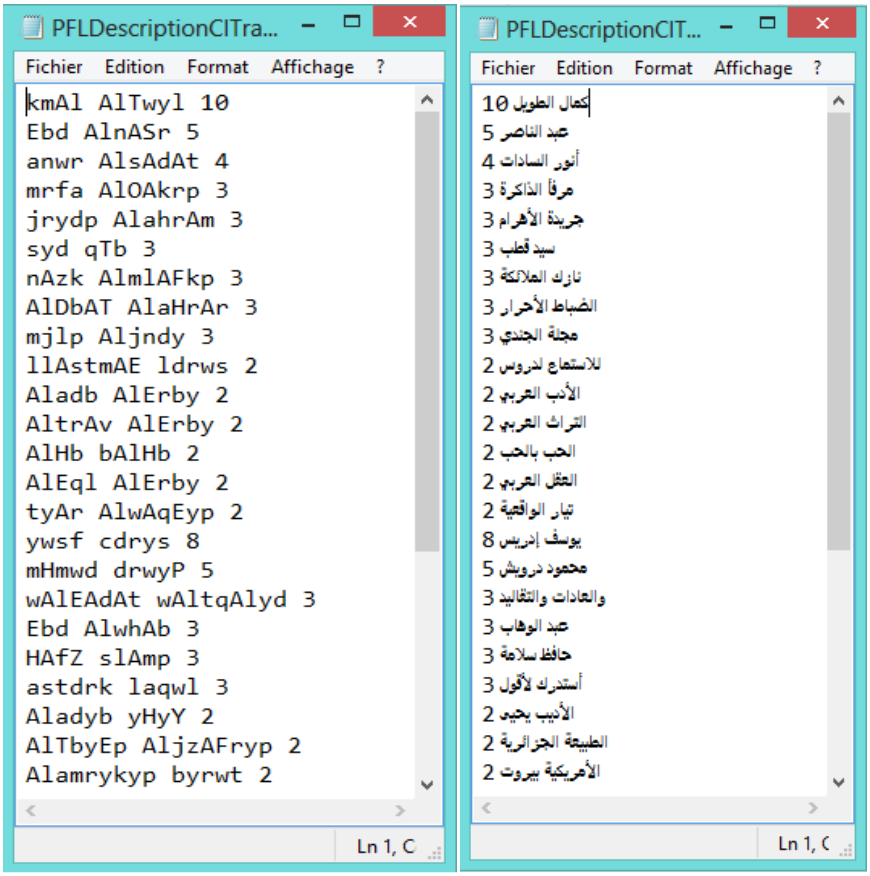


Figure 6.20 : Description par les phrases fréquentielles latentes de la classe *Autobiography* nettoyée.

Classe	Classe Label	Phrases fréquentielles latentes			
		Rang phrase fréquentielle lattente	phrase fréquentielle lattente	Match@5	MRR@5
0	Autobiographys			0	0,00
1	Children' stories	3	الأطفال والابتسام	1	0,33
2	Economics	1	العربية السعودية	1	1,00
3	Education	1	البحث العلمي	1	1,00
4	Health and medicine	1	القدرة البصرية	1	1,00
5	Interviews	0		0	0,00
6	Politics	1	الضفة الغربية	1	1,00
7	Recipes	1	ملعقة طعام	1	1,00
8	Religion	1	الله تعالى	1	1,00
9	Science	2	الكائنات البحرية	1	0,50
10	Short stories	0		0	0,00
11	Sociology	1	الحضارة الحديثة	1	1,00
12	Spoken	1	كرسي الأستاذية	1	1,00
13	Sports	1	كرة القدم	1	1,00
14	Tourist/travel	1	الشرق الاوسط	1	1,00
Match@5				0,80	
MRR@5					0,72

Tableau 6.19 : Evaluation de la description par les phrases fréquentielles latentes de la collection CCA sous la forme nettoyée.

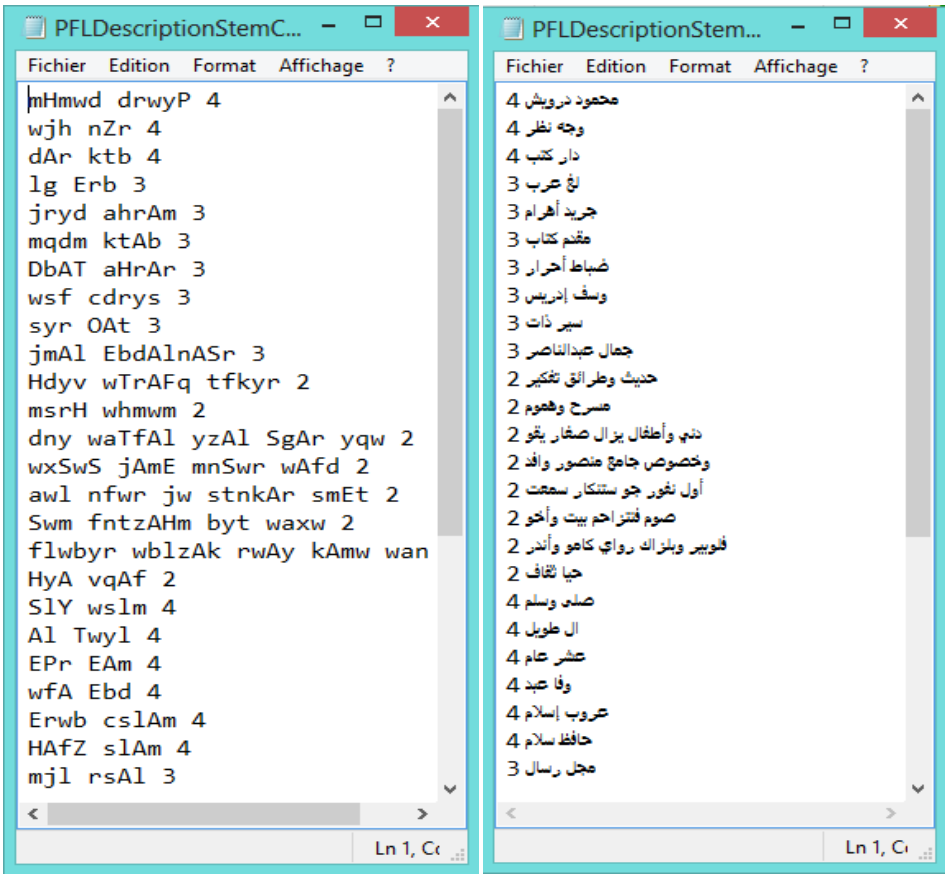


Figure 6.21 : Description par les phrases fréquentielles latentes de la classe *Autobiography* stemmée.

Classe	Classe Label	Phrases fréquentielles latentes			
		Rang phrase fréquentielle lattente	phrase fréquentielle lattente	Match@5	MRR@5
0	Autobiographys	0		0	0,00
1	Children' stories	4	حذاء خشب	1	0,25
2	Economics	1	عرب سعود	1	1,00
3	Education	1	تعليم مدرس منزل	1	1,00
4	Health and medicine	2	تنفس أثناء	1	0,50
5	Interviews	1	عالم عرب	1	1,00
6	Politics	1	ولاي متحد	1	1,00
7	Recipes	2	قدر نار	1	0,50
8	Religion	2	أهل سن	1	0,50
9	Science	1	ولاي متحد	1	1,00
10	Short stories	0		0	0,00
11	Sociology	1	لغ عرب	1	1,00
12	Spoken	1	كرس أستاذ	1	1,00
13	Sports	1	لكر لقدم	1	1,00
14	Tourist/travel	1	مملك عرب سعود	1	1,00
Match@5				0,87	
MRR@5					0,72

Tableau 6.20 : Evaluation de la description par les phrases fréquentielles latentes de la collection CCA sous la forme stemmée.

b. Synthèse des résultats

Les figures 6.22,…, 6.27 donnent les résultats de la description des trois techniques par forme.

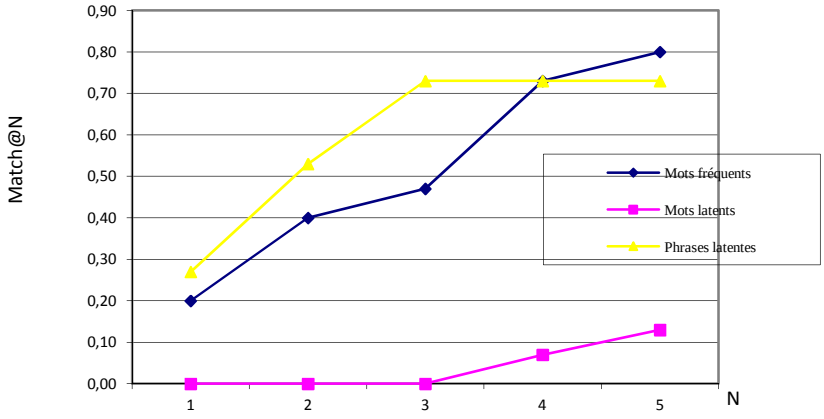


Figure 6.22 : Evaluation de la description des classes prédéfinies sous la forme translittérée de la collection CCA avec la mesure Match@N.

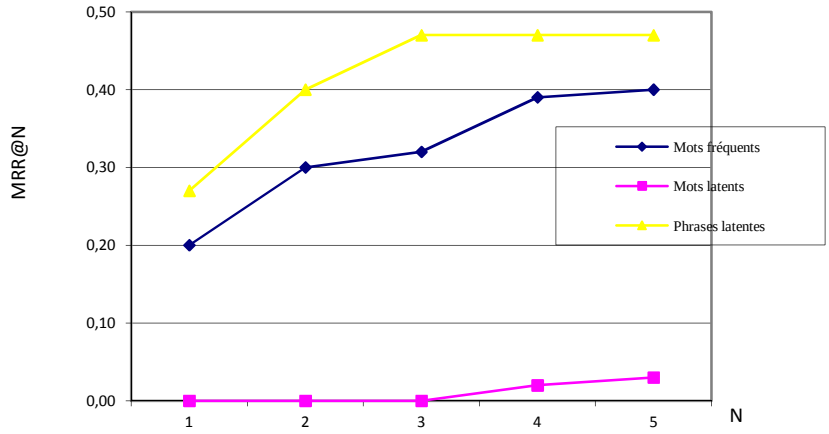


Figure 6.23 : Evaluation de la description des classes prédéfinies sous la forme translittérée de la collection CCA par la mesure MRR@N.

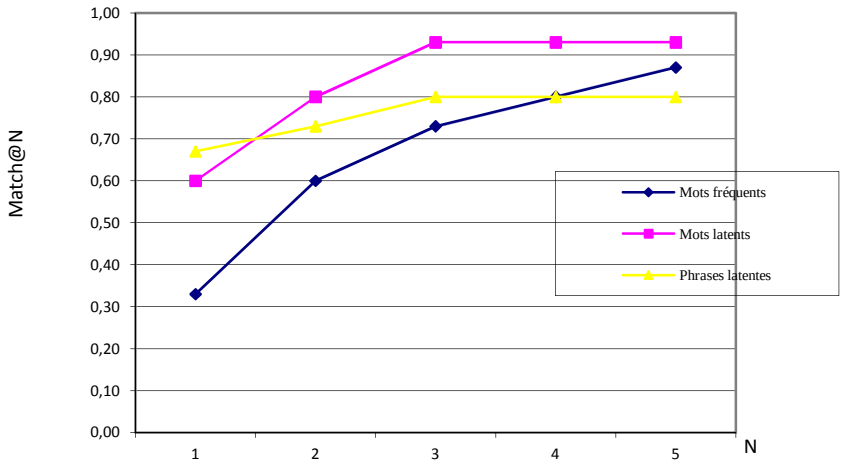


Figure 6.24 : Evaluation de la description des classes prédéfinies sous la forme nettoyée de la collection CCA par la mesure Match@N.

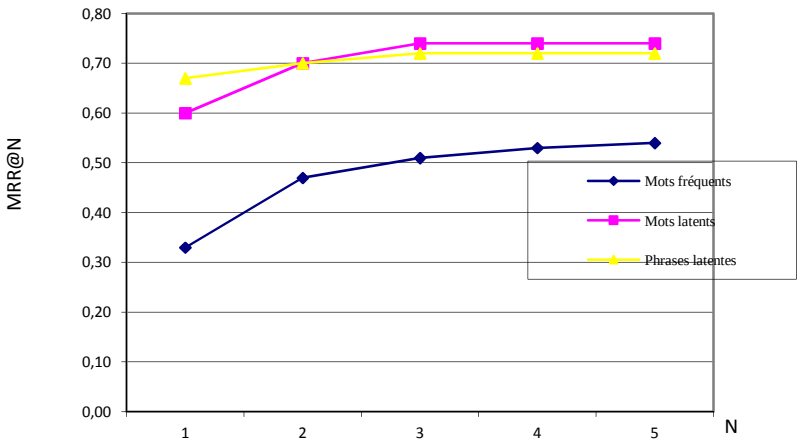


Figure 6.25 : Evaluation de la description des classes prédéfinies sous la forme nettoyée de la collection CCA par la mesure MRR@N.

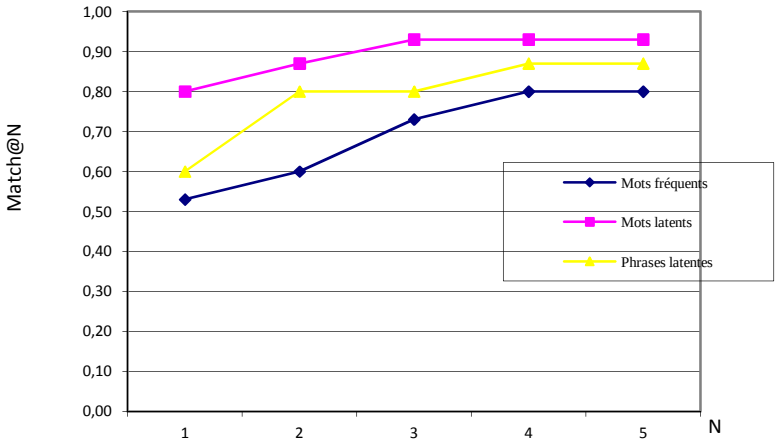


Figure 6.26 : Evaluation de la description des classes prédéfinies sous la forme stemmée de la collection CCA par la mesure Match@N.

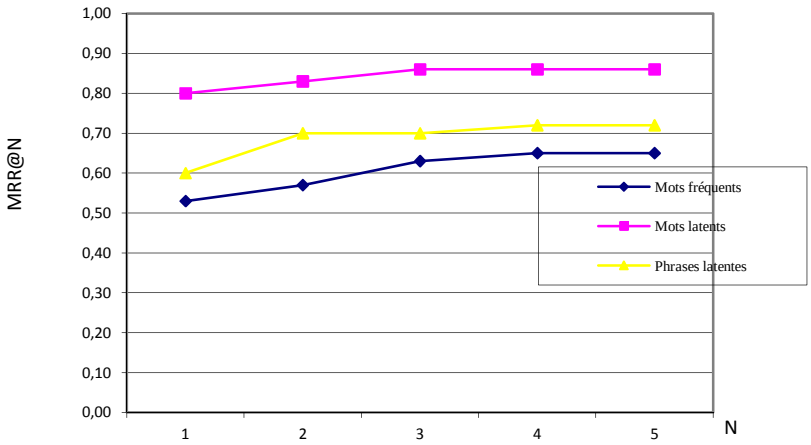


Figure 6.27 : Evaluation de la description des classes prédéfinies sous la forme stemmée de la collection CCA par la mesure MRR@N.

4.3.2. Description des classes sous la forme translittérée

A première vue sur les deux figures 6.22 et 6.23, la description avec les mots fréquentiels latents est très faible pour les deux mesures d'évaluation. En examinant de prés les descriptions rendues par cette technique des classes prédéfinies, nous avons constaté que la quasi-totalité des mots latents sont des mots outils.

Par exemple la description de la première classe « *Autobiography* » est :

"ما الذي عن كان التي أن إلى على من في"

Le raisonnement sur lequel est basée la détermination des mots fréquentiels latents est à l'origine de ce résultat (représentation 5.2, cinquième chapitre): les mots latents les plus fréquents qui apparaissent souvent ensemble représentent la description et c'est le cas des mots outils. Nous avons pensé au début que ce phénomène était relatif à la collection en cours d'étude (*CCA*), mais en lançant le même processus de description sur les trois autres collections nous nous sommes aperçus que le même phénomène s'est répété. Dans la littérature nous n'avons pas trouvé une référence à ce sujet sur les autres langues, ceci ne nous laisse pas dire que c'est une particularité stricte à la langue arabe, par contre nous pouvons dire que les mots outils sont d'une utilisation relativement fréquentes dans les textes arabes rédigés en *MSA*.

Par contre dans le cas des phrases latentes ce phénomène n'apparait plus puisque par définition les mots constituant une phrase latente se succèdent ce qui n'est pas le cas pour les mots outils sauf quelques exceptions pré (ex. "من عليها" qui veut dire "qui est sur elle").

Nous remarquons aussi la capacité des phrases fréquentielles latentes de description par rapport aux mots fréquentiels : La description avec les phases fréquentielles latentes performe de bons résultats avec la mesure Match@N pour N<5 (des phrases qui remplissent en majorité les trois conditions de concision, de compréhensibilité et de précision[1]), en réussissant à placer de bonnes descriptions dans les premiers rangs, alors que la technique de référence n'a réussi son dépassement qu'avec N>=5. Ceci est confirmé avec la mesure MRR@N.

Ici il faut noter que le principal défaut de la description par les phrases est celui de la génération, parfois, des phrases non significatives. Un exemple concret est illustré dans la figure 6.19 pour décrire la classe *Autobiography*.

4.3.3. Description des classes sous la forme nettoyée

Les figures 6.24 et 6.25 décrivent les résultats obtenus avec la forme nettoyée des textes de la collection CCA. Nous remarquons la nette amélioration et la performance des mots fréquentiels latents effectués par rapport aux phrases fréquentielles latentes et les mots fréquentiels. La performance des phrases fréquentielles latentes par rapport aux mots fréquentiels est la même que sur la forme translittérée.

Notons ici que les phrases insignifiantes retrouvées pour décrire la forme translittérée tendent à disparaitre, ceci est dû, essentiellement, à la suppression des mots outils. Aussi, la domination des phrases nominales dans la description générée qui sont plus apparentes que dans la description de la première forme (translitérée). Ces phrases sont généralement des

[1] Ici nous avons donné un jugement un peu abusif puisque le jugement de satisfaction des phrases candidates à la description de ces conditions demande des techniques plus rigoureuses et une analyse plus approfondie.

entités nommées, des groupes nominales, etc. Ce phénomène est expliqué, toujours, par le principe sur lequel est basé le fonctionnement de cette description.

4.3.4. Description des classes sous la forme stemmée

Dans les figures 6.26 et 6.27 les résultats obtenus avec les mots fréquentiels latents s'améliorent encore plus par rapport à ceux obtenus sur la forme nettoyée. La même chose s'est produite pour les phrases fréquentielles latentes.

Ici le majeur inconvénient, malgré les bonnes performances, reste la lisibilité des mots et des phrases de la description. Le processus de description des formes stemmées engendre parfois des descriptions qui peuvent être incompréhensibles par les lecteurs qui ne maitrisent pas totalement la langue arabe. Ceci nous a obligé à se poser des questions sur la performance de traitement du stemming que nous avons utilisé et à réfléchir sur son amélioration dans les travaux futurs.

4.3.5. Influence du processus de prétraitement sur la qualité de la description

Les figures 6.22,…, 6.27 ne nous permettent pas de voir en claire l'influence du processus de prétraitement sur la qualité de la description par les trois techniques, prise une par une. Les figures 6.28,…, 6.33 comparent les performances de description, par technique, dudit processus et illustre l'incidence de ce dernier sur les performances de ces techniques.

a. Mots fréquentiels

Les deux figures 6.28 et 6.29 montrent une amélioration de la description sur les deux formes, nettoyée et stemmée. Un léger avancement sur la forme stemmée est remarqué au début de la description et qui sera rattrapé aux rangs suivants.

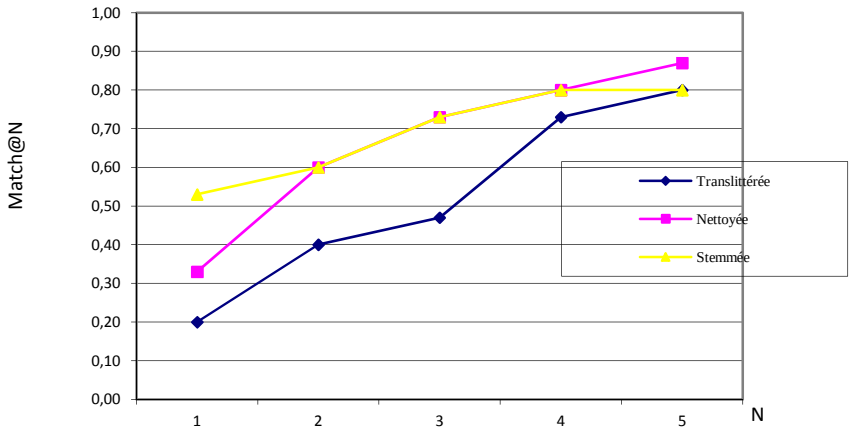


Figure 6.28 : Evaluation de la description des trois formes (translittérée, nettoyée et stemmée) des classes prédéfinies de la collection CCA par les mots fréquentielles avec la mesure Match@N.

La figure 6.29 montre une nette performance de la technique des mots fréquentiels sur la forme stemmée puis sur la forme nettoyée par rapport à la forme translittérée. Ceci est expliqué par le fait que les mots outils ont été pris comme étant des descriptions sur la forme translitérée, avec l'avancement du prétraitement, ces mots disparaissent et les descriptions s'améliorent. Avec le stemming plusieurs mots vont se rapportés aux mêmes stems et donc leurs fréquences va augmenter.

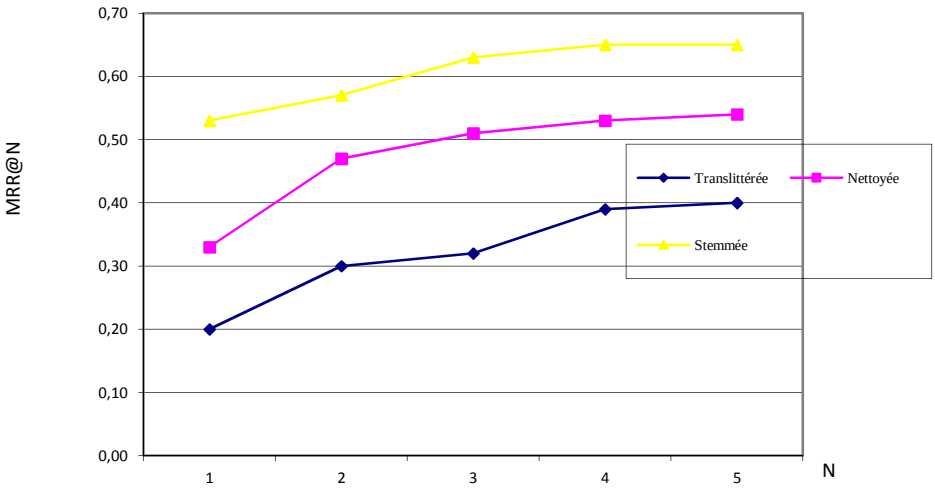


Figure 6.29 : Evaluation de la description des trois formes (translittérée, nettoyée et stemmée) des classes prédéfinies de la collection CCA par les mots fréquentielles avec la mesure MRR@N.

b. Most fréquentiels latents

Les deux figures 6.30 et 6.31 montrent les mauvais résultats obtenus sur la forme translitérée, ceci est dû, comme il a été mentionné dans la section 4.3.2, au principe de fonctionnement de la technique en question et la présence des mots outils. Sur les formes nettoyées et stemmées les résultats sont bien meilleurs.

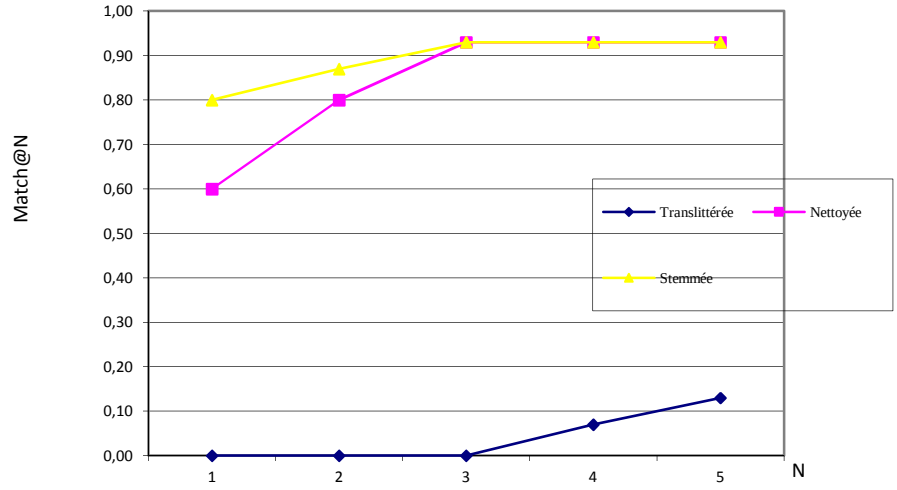


Figure 6.30 : Evaluation de la description des trois formes (translittérée, nettoyée et stemmée) des classes prédéfinies de la collection CCA par les mots fréquentiels latents avec la mesure Match@N.

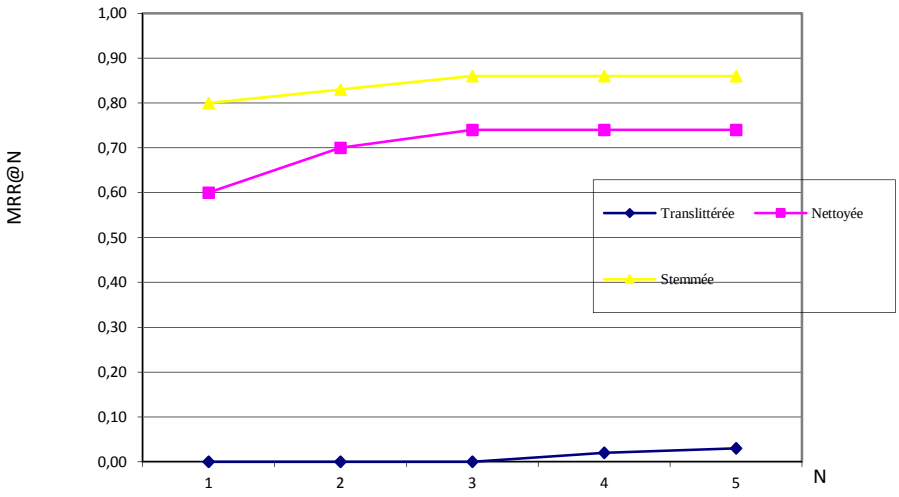


Figure 6.31 : Evaluation de la description des trois formes (translittérée, nettoyée et stemmée) des classes prédéfinies de la collection CCA par les mots fréquentiels latents avec la mesure MRR@N.

c. Phrases fréquentielles latentes

Pour les phrases fréquentielles latentes nous remarquons sur la figure 6.32 qu'il y a une certaine non domination des performances de la technique sur l'une des deux formes nettoyée et stemmée. Ceci est peut être causé par la non lisibilité des phrases sur la forme stemmée lors de l'évaluation de la technique par l'expert humain.

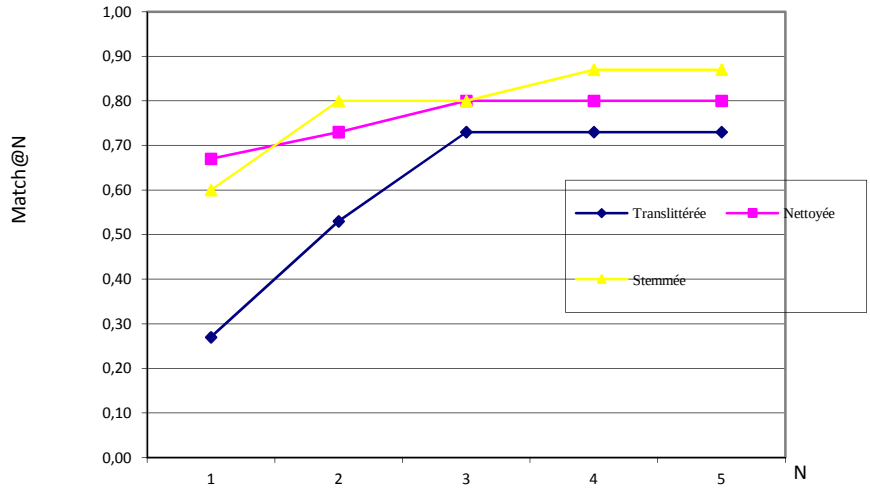


Figure 6.32 : Evaluation de la description des trois formes (translittérée, nettoyée et stemmée) des classes prédéfinies de la collection CCA par les phrases fréquentielles latentes avec la mesure Match@N.

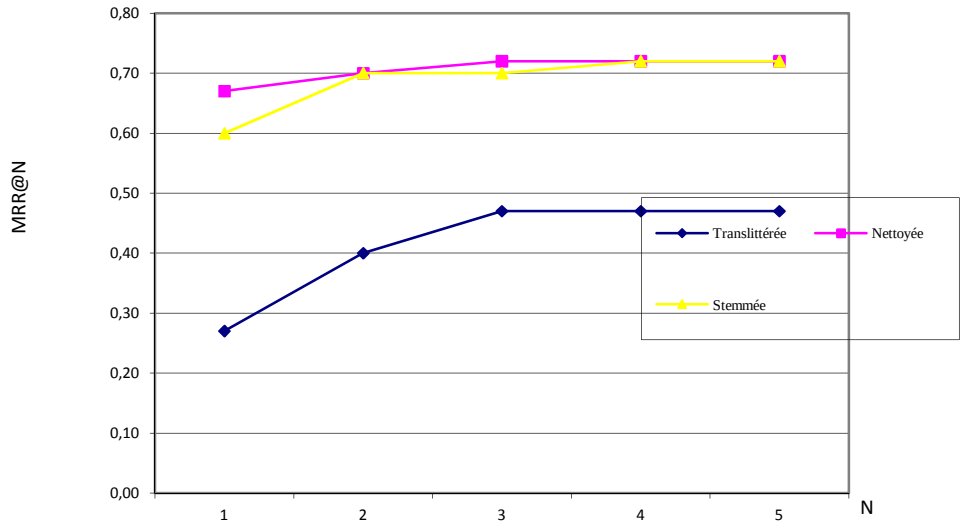


Figure 6.33 : Evaluation de la description des trois formes (translittérée, nettoyée et stemmée) des classes prédéfinies de la collection CCA par les phrases fréquentielles latentes avec la mesure MRR@N.

4.3.6. Performances des deux techniques de description proposées

En conclusion à cette phase nous pouvons dire que les deux techniques de description de l'approche LDK-Means ont réussi une performance relativement bonne et nous pouvons maintenant passer aux tests de l'approche entière.

4.4. Application de LDK-Means (Approche proposée)

En résumé de la présentation de l'approche LDK-Means dans le cinquième chapitre, le processus passe par deux étapes. La première est la classification non supervisée textuelle qui consiste à générer une partition cohérente de clusters de la collection CCA (partition décrite en annexe). La deuxième étape, quant à elle, consiste à décrire chaque cluster par les mots fréquentiels latents et les phrases fréquentielles latentes.

4.4.1. Exemple de descriptions obtenues sur le premier cluster

Dans ce qui suit nous présentons un exemple de description du premier cluster de la partition générée par la première étape de LDK-Means par les mots fréquentiels latents et les phrases fréquentielles latentes. L'exemple comporte aussi une description générée par les mots fréquentielles.

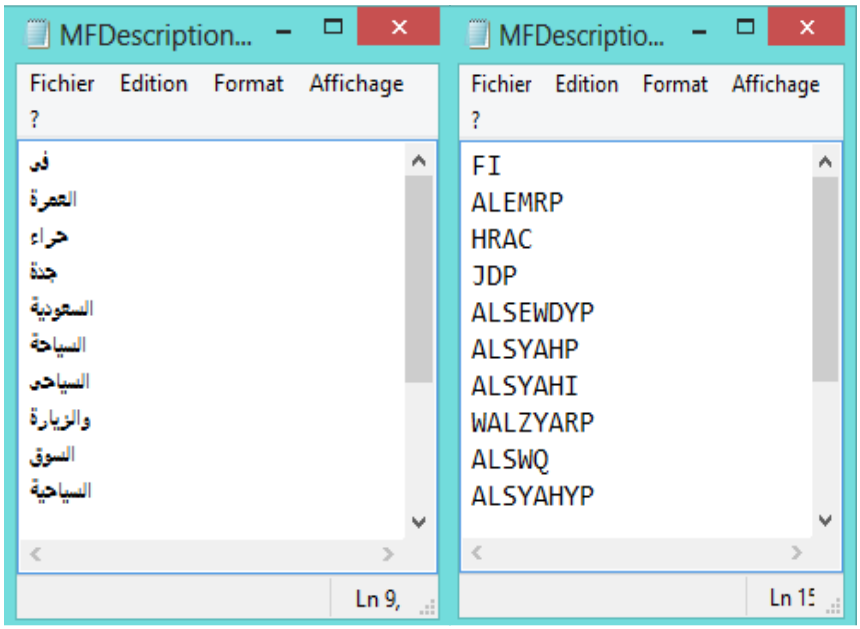


Figure 6.34 : Description par les mots fréquentiels du cluster 1 sous la forme translitérée.

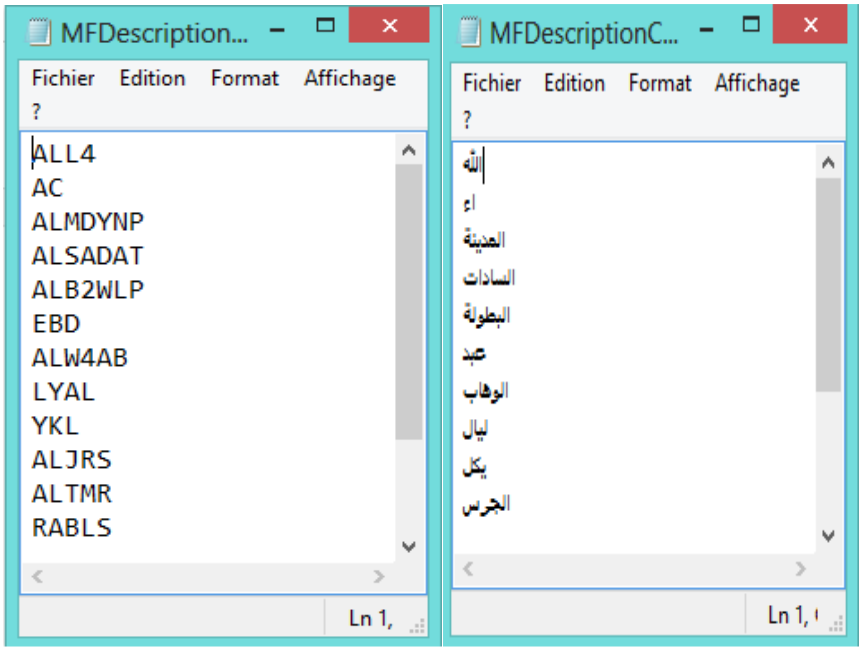


Figure 6.35 : Description par les mots fréquentiels du cluster 1 sous la forme nettoyée.

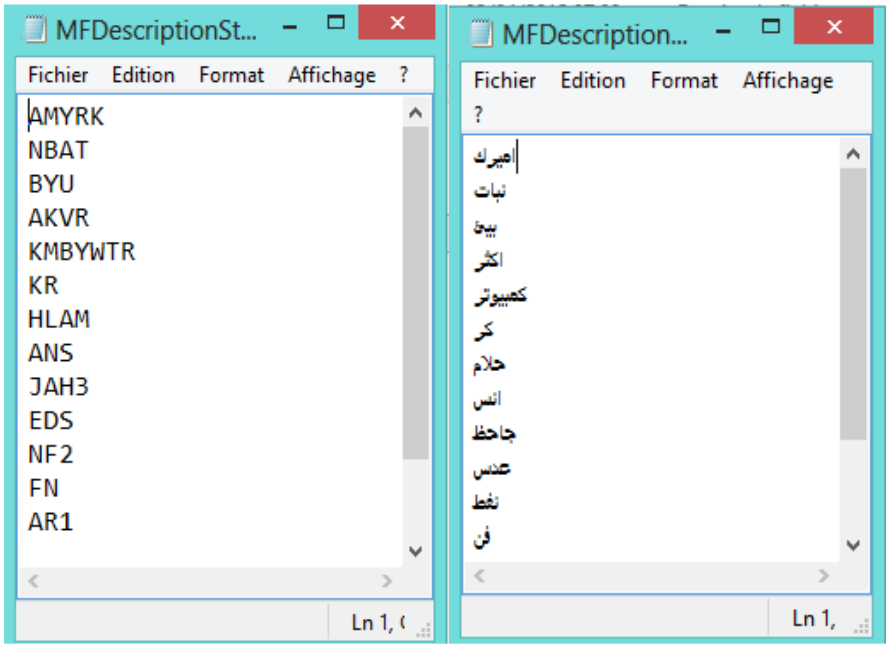


Figure 6.36 : Description par les mots fréquentiels du cluster 1 sous la forme stemmée.

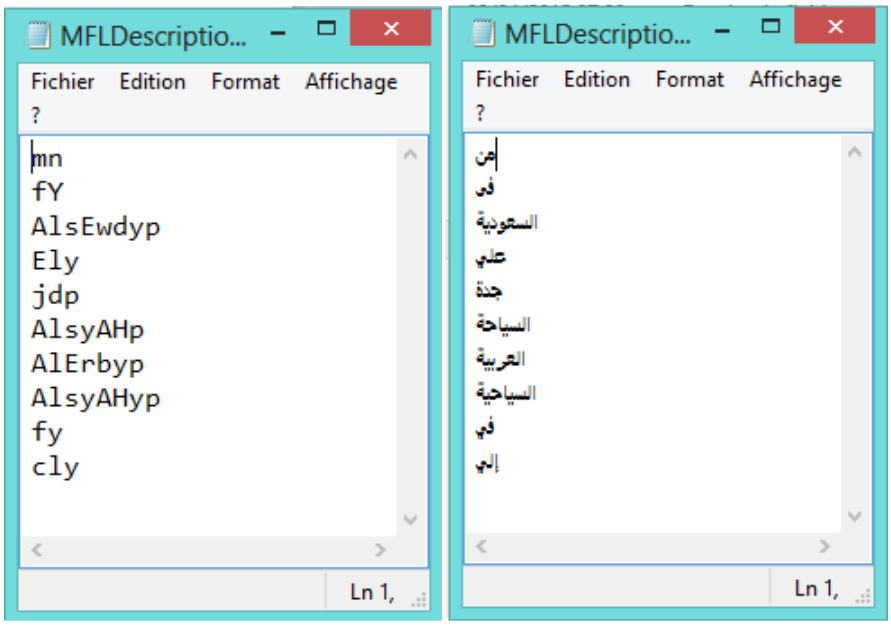


Figure 6.37 : Description par les mots fréquentiels latents du cluster 1 sous la forme translitérée.

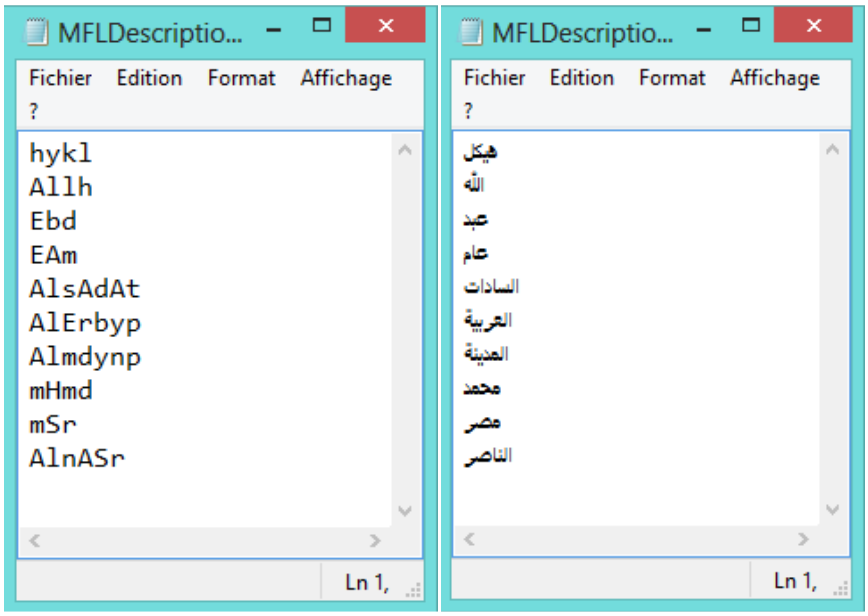


Figure 6.38 : Description par les mots fréquentiels latents du cluster 1 sous la forme nettoyée.

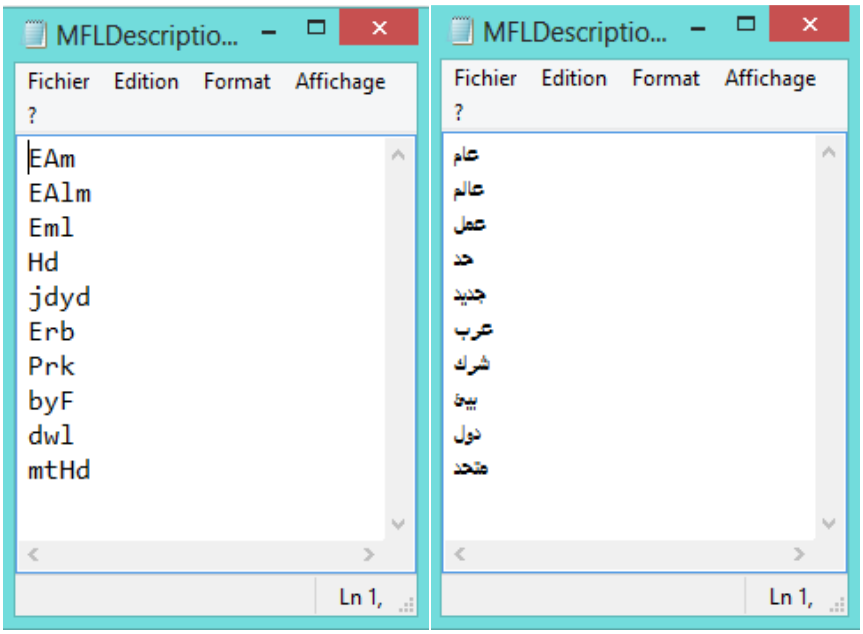


Figure 6.39 : Description par les mots fréquentiels latents du cluster 1 sous la forme stemmée.

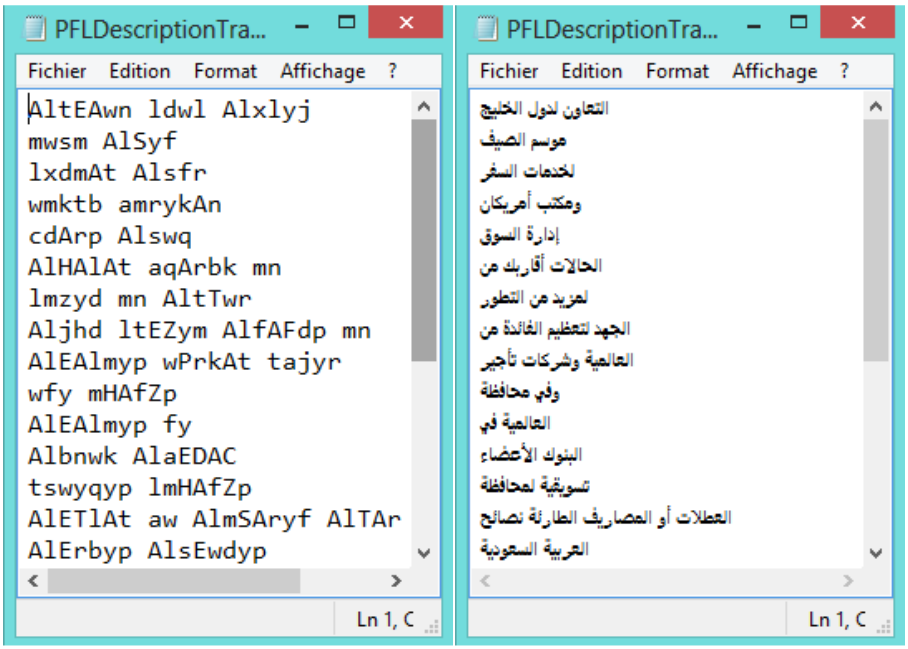


Figure 6.40 : Description par les phrases fréquentielles latentes du cluster 1 sous la forme translitérée.

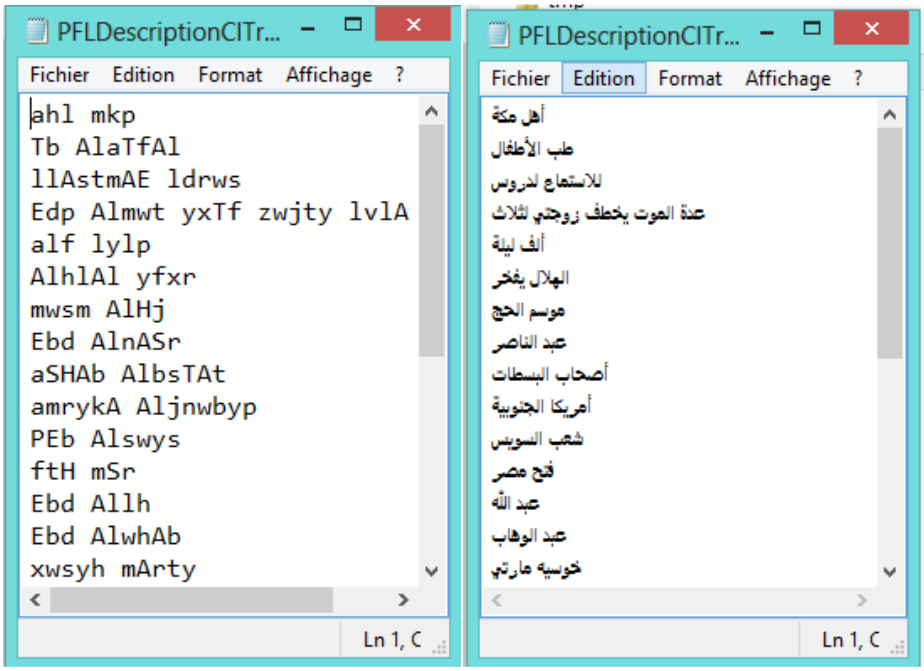


Figure 6.41 : Description par les phrases fréquentielles latentes du cluster 1 sous la forme nettoyée.

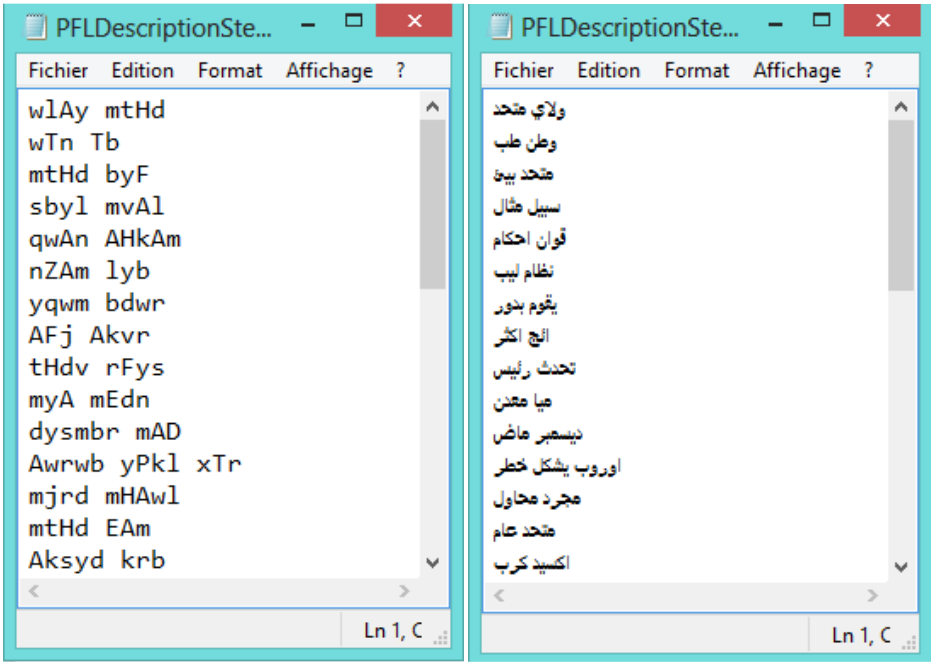


Figure 6.42 : Description par les phrases fréquentielles latentes du cluster 1 sous la forme stemmée.

4.4.2. Résultats de l'évaluation manuelle

Pour évaluer ces descriptions nous avons fait recours à l'expertise humaine. Le travail de l'expert humain consistait, pour chaque cluster, à lire touts les textes de ce dernier, lire les trois descriptions générées qui lui sont associées à savoir, les mots fréquentiels, les mots fréquentiels latents et les phrases fréquentielles latentes et d'affecter un des trois jugements (Correcte, Acceptable, Fausse) à chacune de ces descriptions. Ici nous avons pris en compte deux évaluations différentes. Les résultats de cette évaluation ont été alors les suivants :

a. Forme translitérée

Sur la forme translitérée les deux évaluations, qui apparaissent dans tableau 6.21 et synthétisées dans le tableau 6.22, se sont pratiquement mises d'accord sur les performances des mots fréquentielles latents un peu plus meilleures que les mots fréquentiels. Les phrases fréquentielles latentes, quant à elles, performent des résultats moins bons. Ces résultats vont à l'encontre des résultats obtenus sur les classes prédéfinies de la collection CCA dans la section 4.3.

Remarquant ici la différence flagrante entre les deux évaluations des phrases fréquentielles latentes pour le jugement des descriptions fausses (0 et 4). Ceci est dû au fait de la nature du jugement humain qui diverge d'une personne à une autre, d'ailleurs, et comme il a été mentionné au troisième chapitre, c'est le défaut majeur avec ce type d'évaluation.

Technique / Cluster	Mots fréquentiels		Mots fréquentiels latents		Phrases fréquentielles latentes	
	Evaluation 1	**Evaluation 2**	**Evaluation 1**	**Evaluation 2**	**Evaluation 1**	**Evaluation 2**
Cluster 0	Acceptable	Acceptable	Acceptable	Correcte	Acceptable	Acceptable
Cluster 1	Correcte	Correcte	Correcte	Acceptable	Acceptable	Correcte
Cluster 2	Correcte	Acceptable	Acceptable	Acceptable	Fausse	Fausse
Cluster 3	Correcte	Correcte	Correcte	Correcte	Fausse	Acceptable
Cluster 4	Acceptable	Acceptable	Correcte	Correcte	Acceptable	Acceptable
Cluster 5	Correcte	Correcte	Correcte	Acceptable	Acceptable	Acceptable
Cluster 6	Fausse	Fausse	Acceptable	Acceptable	Acceptable	Correct
Cluster 7	Acceptable	Correcte	Correcte	Correcte	Acceptable	Acceptable
Cluster 8	Correcte	Acceptable	Acceptable	Acceptable	Fausse	Fausse
Cluster 9	Acceptable	Acceptable	Acceptable	Acceptable	Fausse	Fausse
Cluster 10	Acceptable	Correcte	Correcte	Acceptable	Fausse	Fausse
Cluster 11	Acceptable	Acceptable	Correcte	Correcte	Acceptable	Acceptable
Cluster 12	Correcte	Correcte	Correcte	Correcte	Acceptable	Correcte
Cluster 13	Correcte	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable
Cluster 14	Correcte	Correcte	Correcte	Correcte	Acceptable	Acceptable

Tableau 6.21 : Evaluation manuelle des descriptions de la CNST sur la collection CCA avec les trois techniques (forme translittérée).

Technique / Evaluation	Mots fréquentiels		Mots fréquentiels latents		Phrases fréquentielles latentes	
	Evaluation 1	**Evaluation 2**	**Evaluation 1**	**Evaluation 2**	**Evaluation 1**	**Evaluation 2**
Correcte	8	7	9	7	5	3
Acceptable	6	7	6	8	10	8
Fausse	1	1	0	0	0	4

Tableau 6.22 : Performances des descriptions de la CNST sur la collection CCA avec les trois techniques (forme translittérée).

b. Forme nettoyée

Sur la forme nettoyée (tableaux 6.23 et 6.24) la performance des mots fréquentiels latents est nettement mieux que les deux autres techniques. En éliminant les mots outils cette performance a sensiblement augmenté ce qui a influencé positivement sur la compréhensibilité et la précision de la description avec cette technique. Notons seulement le fait que les phrases fréquentielles latentes marquent cinq descriptions fausses (dans les deux évaluations). Il faut remarquer aussi qu'ici il y a une grande concordance entre les deux évaluations.

Technique / Cluster	Mots fréquentiels		Mots fréquentiels latents		Phrases fréquentielles latentes	
	Evaluation 1	**Evaluation 2**	**Evaluation 1**	**Evaluation 2**	**Evaluation 1**	**Evaluation 2**
Cluster 0	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable
Cluster 1	Acceptable	Acceptable	Correcte	Correcte	Acceptable	Acceptable
Cluster 2	Correcte	Correcte	Acceptable	Correcte	Acceptable	Correcte
Cluster 3	Acceptable	Acceptable	Correcte	Correcte	Fausse	Fausse
Cluster 4	Acceptable	Correcte	Correcte	Correcte	Fausse	Fausse
Cluster 5	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable
Cluster 6	Correcte	Correcte	Acceptable	Acceptable	Acceptable	Acceptable
Cluster 7	Correcte	Correcte	Correcte	Correcte	Fausse	Fausse
Cluster 8	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable
Cluster 9	Acceptable	Acceptable	Correcte	Correcte	Fausse	Fausse
Cluster 10	Acceptable	Acceptable	Correcte	Acceptable	Acceptable	Acceptable
Cluster 11	Acceptable	Correcte	Correcte	Correcte	Acceptable	Correcte
Cluster 12	Acceptable	Acceptable	Correcte	Correcte	Fausse	Fausse
Cluster 13	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable
Cluster 14	Acceptable	Acceptable	Correcte	Correcte	Acceptable	Acceptable

Tableau 6.23 : Evaluation manuelle des descriptions de la CNST de la collection CCA avec les trois techniques (forme nettoyée).

Technique / Evaluation	Mots fréquentiels		Mots fréquentiels latents		Phrases fréquentielles latentes	
	Evaluation 1	**Evaluation 2**	**Evaluation 1**	**Evaluation 2**	**Evaluation 1**	**Evaluation 2**
Correcte	3	5	9	9	0	2
Acceptable	12	10	6	6	10	8
Fausse	0	0	0	0	5	5

Tableau 6.24 : Performances des descriptions de la CNST sur la collection CCA avec les trois techniques (forme nettoyée).

c. Forme stemmée

Sur la forme stemmée (tableaux 6.25 et 6.26) les mots fréquentiels latents, donnent des meilleurs résultats par rapport aux deux autres techniques. Ces résultats sont moins bons que sur la forme nettoyée. Ici nous pensons que la lisibilité (compréhensibilité) des mots stemmés est la cause principale de cette baisse de performance.

Pour les phrases fréquentielles latentes, les deux évaluations se sont mises d'accords sur quatre descriptions fausses, par contre il y a une certaine divergence dans la détermination des descriptions correctes et acceptables. Ici aussi nous pensons que la lisibilité est à l'origine de ces résultats.

Technique / Cluster	Mots fréquentiels		Mots fréquentiels latents		Phrases fréquentielles latentes	
	Evaluation 1	**Evaluation 2**	**Evaluation 1**	**Evaluation 2**	**Evaluation 1**	**Evaluation 2**
Cluster 0	Acceptable	Acceptable	Correcte	Acceptable	Acceptable	Acceptable
Cluster 1	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable
Cluster 2	Correcte	Acceptable	Correcte	Correcte	Acceptable	Correcte
Cluster 3	Correcte	Correcte	Correcte	Correcte	Fausse	Fausse
Cluster 4	Correcte	Correcte	Acceptable	Acceptable	Acceptable	Acceptable
Cluster 5	Acceptable	Acceptable	Acceptable	Acceptable	Correcte	Correcte
Cluster 6	Acceptable	Acceptable	Correcte	Correcte	Acceptable	Acceptable
Cluster 7	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable
Cluster 8	Acceptable	Correcte	Acceptable	Acceptable	Fausse	Fausse
Cluster 9	Correcte	Correcte	Correcte	Correcte	Fausse	Fausse
Cluster 10	Acceptable	Acceptable	Correcte	Correcte	Fausse	Fausse
Cluster 11	Acceptable	Acceptable	Correcte	Correcte	Acceptable	Correcte
Cluster 12	Acceptable	Correcte	Acceptable	Acceptable	Acceptable	Acceptable
Cluster 13	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable
Cluster 14	Acceptable	Acceptable	Correcte	Correcte	Acceptable	Correcte

Tableau 6.25 : Evaluation manuelle des descriptions de la CNST de la collection CCA avec les trois techniques (forme stemmée).

Technique / Evaluation	Mots fréquentiels		Mots fréquentiels latents		Phrases fréquentielles latentes	
	Evaluation 1	**Evaluation 2**	**Evaluation 1**	**Evaluation 2**	**Evaluation 1**	**Evaluation 2**
Correcte	4	5	8	7	1	4
Acceptable	11	10	7	8	10	7
Fausse	0	0	0	0	4	4

Tableau 6.26 : Performances des descriptions de la CNST de la collection CCA avec les trois techniques (forme stemmée).

4.4.3. Discussions

En comparant les descriptions obtenues sur une structure établie manuellement ou préétablie de la collection *CCA* et celles qui résultent de l'application des K-Moyennes dans l'approche LDK-Means, nous remarquons que l'apparition des mots outils dans la description avec les mots fréquentiels latents sur la forme translittérée, n'est pas la même. Le nombre de ces mots qui a été dominant dans la description des classes prédéfinies a sensiblement diminué. Le premier facteur qui a contribué au retournement de cette situation est la structure des clusters formés par les K-Moyennes. En examinant cette structure (exposée en annexe) nous avons constaté qu'elle est formée de textes de plusieurs classes. Ceci veut dire qu'il y a une très forte relation entre le style de rédaction d'un type de texte et l'emploi et la fréquence des mots outils, c'est-à-dire que la rédaction d'un texte d'autobiographie ne se fait pas de la même manière qu'un texte qui décrit un thème qui relève de l'éducation.

La deuxième grande constatation est qu'avec les deux formes nettoyée et stemmée les phrases fréquentielles ont marqué plusieurs fausses descriptions (dégradation de la compréhensibilité et de la précision). On revenant toujours sur la structure générée par les K-Moyennes dans LDK-Means nous avons constaté que beaucoup de textes ont changé d'emplacement, ceci est lié directement au processus de nettoyage et de stemming, s'ajoute à ceci ce même processus de stemming qui rend les descriptions générées pour la forme stemmée parfois incompréhensibles ce qui induit des erreurs sur la lisibilité de la description.

Il faut noter ici que si on juge que le stemming améliore la qualité des clusters nous pouvons imaginer un mécanisme de reconstitution qui restitue les mots des descriptions à l'état initial avant stemming, ainsi nous pourrons remédier aux problèmes de l'incompréhension et de la lisibilité.

Enfin, la description avec l'approche LDK-Means par les mots fréquentiels latents a montré une bonne performance. La tentative d'injection d'une certaine sémantique dans le calcul des représentants d'une thématique, en ne prenant en compte que les mots reliés qui apparaissent ensemble a visiblement réussi. Par contre ce même principe n'a visiblement pas marché avec la deuxième description qui est les phrases fréquentielles latentes. Ceci est dû à la fréquence d'apparition de ces phrases qui est relativement petit par rapport aux mots fréquentiels latents, ce qui est à l'origine de la dégradation des performances de cette technique.

5. Conclusion

Dans le présent chapitre nous avons exposé en détails les différentes phases de l'étude menée pour l'aboutissement aux résultats obtenus par l'approche LDK-Means.

Nous avons constaté en premier lieu que la nature des collections influe considérablement sur la qualité des résultats d'une classification non supervisée. En effet, la richesse de la langue arabe peut jouer un grand rôle dans la détermination de la nature d'une collection. Nous avons aussi constaté parfois la contradiction entre les mesures de performance de la classification, ceci est peut être dû à un choix qui n'est pas adéquat à la langue arabe.

En second lieu nous avons vu que la valeur élevée d'une mesure de qualité d'une classification ne veut pas dire que c'est une bonne classification. En effet, la présence d'un grand nombre de mots outils peut influencer sur cette mesure d'évaluation surtout pour les méthodes thématiques probabilistes telles que LDA.

Enfin nous avons testé l'approche de description par les mots fréquentiels latents et les phrases fréquentielles latentes avant de passer à la description des résultats de la CNST. Nous avons constaté une performance relativement bonne de la description par la première technique et moins bonne pour la deuxième. Ces performances ont été soutenues par la suppression des mots outils.

En conclusion à ce chapitre nous avons conçu et testé une approche basée sur deux grandes techniques de renommée. Cette approche a montré plusieurs points positifs que nous pouvons synthétiser dans ce qui suit :

- Une souplesse dans la préparation et la classification des collections que se soit de petite ou de grande dimension.
- Une description des clusters générés par deux techniques à savoir les mots fréquentiels latents et les phrases fréquentielles latentes, qui ont donné des bons résultats pour la première et relativement bons pour la deuxième. Néanmoins la deuxième technique peut être utilisée et donner de très bon résultats sur les collections avec des classes préétablies.

Conclusion générale et perspectives

L'augmentation phénoménale du volume de l'information sur le support textuel a intensifié l'intérêt porté à la classification automatique textuelle. Que se soit pour l'amélioration des résultats des recherches, l'exploration ou l'organisation des collections, cette classification est un passage incontournable. Deux grandes famille de la classification se distinguent, supervisée et non supervisée.

La classification supervisée est une technique très couteuse du point de vue ressources mobilisées par rapport à la classification non supervisée. Dans la littérature beaucoup de recherches se sont intéressées à la classification non supervisée orientée web (*online*), laissant de côté la classification des grandes collections (*offline*).

La description ou la labellisation des regroupements de documents ou textes générés par une classification s'est vue dédiée par les chercheurs au domaine de la recherche sur le web. Les techniques employées dans la classification descriptive orientée web sont spécifiques aux textes courts (ou comme on à vue des fichiers entêtes). L'application de ces techniques sur des collections à grandes échelles avec des documents relativement longs nécessite des modifications majeures voir même un changement radicale sur le principe de fonctionnement, ceci avec des résultats qui ne sont pas sures.

De même que pour la classification non supervisée textuelle et pour la description, les recherches sur la langue arabe n'ont pas suivi ceux des autres langues, malgré la richesse des morphosyntaxiques que cette langue possède et le nombre de ses utilisateurs sur le web qui est en augmentation constante.

L'apparition, ces dernières années, des nouvelles techniques telles que les méthodes thématiques probabilistes *(topic models)* ces dernières années et leurs performances intéressantes dans certains domaines du *textminning* a poussé les chercheurs à les tester dans le domaine de la classification.

L'objectif de la présente étude a été de débattre tout ceci en essayant de répondre aux questions posées au début de cette thèse et reformulées dans les trois questions majeures suivantes:

- Quel est l'‘apport des techniques relativement nouvelles dites méthodes thématiques probabilistes à la classification non supervisée textuelle par rapport aux techniques classiques ?
- Quelle est la réaction de ces techniques aux caractéristiques de la langue arabe ?
- Comment décrire automatiquement les partitions générées par une classification non supervisée en langue arabe ?

Pour ce faire, il fallait passer en revu et en premier lieu, les majeurs techniques de classification non supervisée et celles qui sont réorientées vers la classification non supervisée textuelle. Nous avons, par l'occasion, exposé les différentes étapes visant à préparer et normaliser les textes dans les formats facilement manipulables par ces techniques.

Nous avons ensuite présenté un état de l'art des différentes techniques de descriptions des résultats de la classification automatique et les techniques de l'évaluation de leurs performances. Le dernier chapitre de l'état de l'art a été consacré à la langue arabe et ses caractéristiques les plus importantes.

Nous avons entamé la deuxième partie de cette thèse par l'élaboration de l'approche proposée LDK-Means et la description des différentes étapes franchies. Les expérimentations reflétant ces étapes et leurs résultats et discussions respectifs ont étaient décrites dans le dernier chapitre.

L'étude que nous avons menée et l'approche que nous avons montée intitulée LDK-Means nous a permis de mettre la lumière sur plusieurs points. La difficulté rencontrée dans la description des résultats de la classification non supervisée textuelle est une tâche très ardue, encore plus quand la langue traitée est une langue aussi complexe que la langue arabe.

La première étape que nous avons franchie a été une étude comparative entre l'une des méthodes les plus connue et les plus ancienne dans la classification non supervisée à savoir les K-Moyennes et une autre méthode, très en vogue, conçue initialement pour la détection thématique et réorienter ensuite vers divers domaines du *textminning*. Cette méthode qui est LDA a apporté une légère amélioration par rapport à la première méthode en classification. L'étude a été menée sur quatre collections en langue arabe CCA, OSAc, BBC et Al Watan.

Après l'étude de différents paramètres tels que le mode de fonctionnement, le facteur temps, nous avons suivi notre intuition et opté pour les K-Moyennes pour la classification et réserver LDA pour la description des clusters générés selon deux techniques. La première est dite mots fréquentiels latents, la deuxième est dite phrases fréquentielles latentes.

La phase de description a été commencée par les tests des deux dites techniques ainsi qu'une technique très connue dans ce domaine, sur les classes prédéfinies de la collection CCA [El Sulaiti, 2003]. Le choix de cette collection est dû essentiellement au facteur humain qui représente un élément incontournable pour l'évaluation des descriptions. Les résultats de cette phase ont placé les mots fréquentiels latents en premier ordre de performance sur les formes prétraitées. En second ordre, les phrases fréquentielles latentes ont montré une certaine performance par rapport à la technique de référence.

La deuxième partie de cette phase a été la description des clusters générés par la méthode des K-Moyennes, ici aussi la technique des mots fréquentiels latents à montrer une certaine dominance sur les deux autres techniques. Nous avons espéré avoir de bon résultats avec la technique des phrases fréquentielles latentes mais le prétraitement, et à l'inverse de la description des classes prédéfinies, a considérable dégrader les performances de cette dernière.

Nous estimons que nous avons apporté par l'approche LDK-Means une contribution dans les recherches sur la classification non supervisée en langue arabe et sa description. Ici

n’oublions pas que plusieurs facteurs restent à prendre en considération, ces facteurs peuvent influer considérablement sur l’étude en question tels que :

- Type des collections textuelles ;
- Mesures d’évaluation adéquates ;
- Certains mots outils pour un domaine peuvent ne pas l’être pour un autre ;
- Stemming employé ; etc.

S’ajoute à ces facteurs les erreurs d’ordre grammatical et syntaxique qu’une collection peut contenir. Ces erreurs sont imprévisibles et peuvent perturber les résultats de n’importe quelle approche. Un control en amont ou à la source doit être effectué pour contenir ces erreurs lors de la construction des collections textuelles.

La présente étude ayant été achevée plusieurs horizons sont ouverts comme une continuation sur cet axe. En premier lieu il serait très intéressant de prolonger notre étude et l’extension de LDK-Means aux opérations de prétraitement plus poussées telles que la lemmatisation, une opération de prétraitement très importante en langue arabe. Aussi une implication des chercheurs en linguistique de la langue serait très souhaitable afin de déterminer l’incidence des mots outils sur la sémantique dans les collections textuelles et de déterminer une liste unifiée de ces mots.

Il serait très intéressant aussi de réfléchir sur la prise en compte des variantes de la méthode LDA telles que la *Dynamic Topic Model* (DTM) [Blei et Lafferty, 2003], *Correlated Topic Model* (CTM) [Blei et Lafferty, 2007], et étudier l’apport de ces variantes.

Enfin, essayer d’adapter les algorithmes tels que Lingo [Osinski *et al.*, 2004] et STC [Zamir et Etzioni, 1998] pour la classification et la description de la classification à grande échelle serait très intéressante eu égard aux performances qu’ils ont accompli dans la classification non supervisée des résultats retournés par les moteurs de recherches.

Bibliographie

1. [Aggarwal et Zhai, 2013] C. C. Aggarwal and C. X. Zhai, "**A survey of text clustering algorithms**", in Mining Text Data (2012), pp 77-128, Springer.
2. [Al-Harbi *et al.*, 2008] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M.S. Khorsheed, and A. Al-Rajeh, "**Automatic Arabic Text Classification**", JADT 2008: 9es Journées Internationales d'Analyse Statistique des Données Textuelles. (pp. 77-83), France, 2008.
3. [Aljlayl et Frieder, 2002] M. Aljlayl and O. Frieder, "**On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach**", the International Conference on Information and Knowledge Management (CIKM), November, Virginia, USA, 2002.
4. [Andrews et Fox, 2007] N.O. Andrews and A. Edward, "**Recent Developments in Document Clustering**", Tech. rept. TR-07-35. Department of Computer Science, Virginia Tech, USA, 2007.
5. [Anick et Vaithyanathan, 1997] P. G. Anick and S. Vaithyanathan, "**Exploiting Clustering and Phrases for Context-Based Information Retrieval**", in Proceedings of ACM/SIGIR'97, pp. 314-323, 1997.
6. [Anton et Croft, 1996] V.L. Anton and W.B. Croft, "**An Evaluation of Techniques for Clustering Search Results**", Technical Report, Department of Computer Science, University of Massachusetts, Amherst, USA, 1996.
7. [Baloul et *al.*, 2002] S. Baloul, M. Alissali, M. Baudry et P. Boula, "**Interface syntaxe-prosodie dans un système de synthèse de la parole à partir du texte en arabe**", 24es Journées d'Étude sur la Parole, 24-27 juin 2002 Nancy, pp.329-332. France, 2002.
8. [Basu et *al.*, 2002] S. Basu, A. Banerjee, and R. Mooney, "**Semi-supervised clustering by sending**", Proceedings of the 19th International Conference on Machine Learning, (ICML-2002), pp. 19-26, Sydney, Australia, 2002.
9. [Bekkerman *et al.*, 2001] R. Bekkerman, R. El-Yaniv, Y. Winter, and N. Tishby, "**On Feature Distributional Clustering for Text Categorization**", ACM SIGIR Conference, 2001.
10. [Berkhin, 2002] P. Berkhin, "**Survey of Clustering Data Mining Techniques**", Accrue Software, CA, USA, 2002.
11. [Bishop, 2006] C.M. Bishop, "**Pattern Recognition and Machine Learning**". Springer, 2006.
12. [Blei *et al.*, 2003] D. M. Blei, A. Y. Ng, and M. I. Jordan, "**Latent dirichlet allocation**". Journal of Machine Learning Research, 3:993–1022, 2003.
13. [Blei et Lafferty, 2003] D. Blei and J. Lafferty, "**Dynamic topic models**", In Proceedings of the 23rd International Conference on Machine Learning, 2006.
14. [Blei et Lafferty, 2007] D. Blei and J. Lafferty, "A correlated topic model of science", Annals of Applied Statistics, 1(1):17–35, 2007.
15. [Boulaknadel, 2005] S. Boulaknadel, "**Utilisation des syntagmes nominaux dans un système de recherche d'information en langue arabe**", LINA FRE CNRS 2729- Université de Nantes 2, France, 2005.
16. [Cutting *et al.*, 1992] D.R. Cutting, J. O. Pedersen, D. Karger, and J. W. Tukey, "**Scatter/Gather: A cluster-based approach to browsing large document collections**", In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, 318–329. 1992.
17. [Caropreso *et al.*, 2001] M. F. Caropreso, S. Matwin, and F. Sebastiani, "**A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization**", Text databases and document management: theory and practice, pages 78–102, 2001.
18. [Carpineto *et al.*, 2009] C. Carpineto, S. Osinski, G. Romano, and D. Weiss, "**A survey of Web clustering engines**", In: ACM Computing Surveys, Volume 41 , Issue 3, pp. 1–38. ACM, New York, USA, 2009.

19. [Carpineto et Romano, 2010] C. Carpineto and G. Romano, **"Optimal Meta Search Results Clustering",** *SIGIR'10*, Geneva, Switzerland, 2010.

20. [Chapelle *et al.*, 2006] O. Chapelle, B. Schölkopf, and A. Zien, ***"Semi-Supervised Learning"***. MIT Press, 2006.

21. [Chavent et *al.*, 1999] M. Chavent, C. Guinot, Y. Lechevallier et M. Tenenhaus, **"Méthodes divisives de classification et segmentation non supervisée : recherche d'une typologie de la peau humaine saine"**, Revue de Statistique Appliquée, Vol. 47, pp. 87-99. 1999.

22. [Chowdhury *et al.*, 2002] A. Chowdhury, M. Aljlayl, E. Jensen, S. Beitzel, D. Grossman, and O. Frieder, "**Linear combination based on document structure and varied stemming for Arabic retrieval**", IIT at TREC 2002 : In Proceedings of the Text Retrieval Conference (TREC-11), pages 299_310, 2002.

23. [Cleuzieu, 2004] G. Cleuzieu, "**Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information**", Thèse présentée pour l'obtention du grade de Docteur de l'Université d'Orléans, France, 2004

24. [Cutting et al., 1992] D. R. Cutting, J. O. Pedersen, D. Karger, and J.W. Tukey, "**Scatter/Gather: A cluster-based approach to browsing large document collections**", In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, 318–329. 1992.

25. [Cutting *et al.*, 1993] D. R. Cutting, D. Karger, and J. O. Pedersen, "**Constant interaction-time Scatter/Gather rowsing of large document collections**", In Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval., pages 126-35, 1993.

26. [Darwish et Oard, 2002] K. Darwish and D. W. Oard, "**CLIR Experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval"**, TREC, Gaithersburg: NIST, pp 703-710, USA, 2002.

27. [Darwish, 2002] K. Darwish, "**Building a Shallow Morphological Analyzer in One Day**". In Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02),Philadelphia, PA, USA. 2002.

28. [De Roeck et Al Fares, 2000] A. N. De Roeck and W. Al-Fares, "**A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots**", in the 38th Annual Meeting of the ACL. Hong Kong. 2000.

29. [Deerwester *et al.*, 1990] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, **"Indexing by latent semantic analysis**", Journal of the American Society of Information science, (JASIS) 416(6): 391–407, USA, 1990.

30. [Diab *et al.*, 2004] M. Diab, K. Hacioglu, and D. Jurafsky, "**Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks**", Stanford University, USA, 2004.

31. [Diday, 1971] E. Diday, "**Une nouvelle méthode de classification et reconnaissance des formes la méthode des nuées dynamiques**", Revue de statistique appliquée, tome 19 n°:2, p. 19-33. Société française des statistiques, France, 1971.

32. [Dostal *et al.*, 2013] M. Dostal, M. Nykl, and K. Ježek, "**Cluster labeling with linked data**", Journal of Theoretical and Applied Information Technology, Vol. 53 No.3, 2013.

33. [Dubin, 2004] D. Dubin, "**The most influential paper Gerard Salton never wrote**", Library Trends, 52(4):748–764, 2004.

34. [Duda *et al.*, 2001] R. Duda, P. Hart, and D. Stork, "**Pattern Classification**", John Wiley and Sons, Inc, 2001.

35. [El-Halees, 2006] A. El-Halees, "**Mining Arabic Association Rules for Text Classification**", in the proceedings of the first international conference on Mathematical Sciences, *Al-Azhar University of Gaza, Palestine*, 2006.

36. [El-Kourdi *et al.*, 2004] M. El-Kourdi, A. Bensaid, and T. Rachidi, "**Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm**", 20th International Conference on Computational Linguistics, Geneva, 2004.

37. [El Kassas, 2005] D. El Kassas, "**Une étude contrastive de l'arabe et du français dans une perspective de génération multilingue", ,** UFR Linguistique, Université Paris 7 – Denis Diderot, France, 2005.

38. [El Sulaiti, 2004] L. El Sulaiti, "**L'arabe contemporain**", Radio Qatar, Qatar, 2003.

39. [Eldesouki *et al.*, 2009] M. I. Eldesouki, W. M. Arafa, and K. M. Darwish, **"Stemming techniques of Arabic Language: Comparative Study from the Information Retrieval Perspective"** The Egyptian Computer Journal , Vol. 36 No. 1, June, 2009.

40. [Evans *et al.*, 1998] D. A. Evans, A. Huettner, X. Tong, P. Jansen, and P. Subasic, "**Notes on the Effectiveness of Clustering in Ad-Hoc Retrieval**", in Proceedings of TREC-7, NIST special publication, 1998.

41. [Everitt *et al.*, 2001] B. S. Everitt, S. Landau, And M. Leese, "Cluster Analysis", 4th Ed. Oxford University Press, 2001.

42. [Faber, 1994] V. Faber, "**Clustering and the Continuous k-Means Algorithm"**, Los Alamos Science Number 22, USA, 1994.

43. [Farghaly et Shaalan, 2009] A. Farghaly and K. Shaalan, "**Arabic Natural Language Processing: Challenges and Solutions**", ACM Transactions on Asian Language Information Processing, Vol. 8, No. 4, Article 14, Pub. date: December, 2009.

44. [Fraley et Raftery, 1999] C. Fraley and A. Raftery, "**Mclust : Software for model-based cluster and discriminant analysis**", In Tech Report 342. Dept. Statistics, Univ. of Washington, 1999.

45. [Fuhr et Buckley, 1991] N. Fuhr and C. Buckley, "**A probabilistic learning approach for document indexing"**, In ACM Transactions on Information Systems, volume 9, pages 223–248, France, 1991.

46. [Glenisson *et al.*, 2005] P. Glenisson, W. Glänzel, F. Janssens, and B. De Moor, "**Combining Full Text and Bibliometric Information in Mapping Scientific Disciplines**", Information Processing & Management 41(6), 1548–1572, 2005.

47. [Glover *et al.*, 2002] E. Glover, D. M. Pennock, S. Lawrence, and R. Krovetz, "**Inferring Hierarchical Descriptions**", In: Proceedings of the 11th International Conference on Information and Knowledge Management (CKIM2002), McLean, VA, 2002: 4-9.

48. [Griffiths et Steyvers, 2004] T. L. Griffiths and M. Steyvers, "**Finding scientific topics**", Proceedings of the National Academy of Science, 101, 5228-5235, 2004.

49. [Hamers *et al.*, 1989] L. Hamers, Y. Hemeryck, G. Herweyers, M. Janssen, H. Keters, R. Rousseau, and A. Vanhoutte, "**Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula**", Inf Process Manag 25(3):315–318, 1989.

50. [Hammouda *et al.*, 2005] K. Hammouda, D. Matute, and M. Kamel, "**Corephrase: Keyphrase extraction for document clustering**", Machine Learning and Data Mining in Pattern Recognition, pages 634–634, 2005.

51. [Han et Kamber, 2006] J. Han and M. Kamber, "**Data Mining: Concepts and Techniques**", second edition: Diane Cerra, 2006.

52. [Harman et Voorhees, 1997] D. Harman and E. Voorhees (Eds), **The Fifth Text REtrieval Conference (TREC-5)**, NIST, 1997.

53. [Harman et Voorhees, 1998] D. Harman and E. Voorhees (Eds), **The Sixth Text REtrieval Conference (TREC-6)**, NIST, 1998.

54. [Hearst et Pederson, 1996] M. A. Hearst and J. O. Pedersen, "**Reexamining the cluster hypothesis: Scatter/gather on retrieval results**", In Proceedings of ACM SIGIR, pp. 76–84, 1996.

55. [Hofmann, 1999a] T. Hofmann, "**Probabilistic latent semantic analysis**", In K. B. Laskey & H. Prade (Eds.), UAI (pp.289–296), 1999.

56. [Hofmann, 1999b] T. Hofmann, "**Probabilistic latent semantic indexing**", In SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, 1999.

57. [Huot et Coupet, 2005] C. Huot et P. Coupet, "**Le Text Mining sur la langue Arabe : application au traitement des sources ouvertes**", TEMIS SA, Paris, France, 2005.

58. [Jain *et al.*, 1999] A.K. Jain, M.N. Murty, and P.J. Flynn, "**Data Clustering: A Review",** ACM Computing Surveys, Vol. 31, No. 3, September 1999.

59. [Jain, 2009] A. Jain, "**Data Clustering: 50 Years Beyond KMeans**", Pattern Recognition Letters, vol. 31, pp. 651-666, 2009.

60. [Jain et Dubes, 1988] A. K Jain, and R. C. Dubes, "**Algorithms for clustering data**", Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ. 1988.

61. [Kashireddy *et al.*, 2013] S. D. Kashireddy, S. Gauch, and S. M. Billah, "**Automatic class labeling for CiteSeerX**", In Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on (Vol. 1, pp. 241-245). IEEE, 2013.

62. [Kaufman et Rousseeuw, 1990a] L. Kaufman and P. J. Rousseeuw, "**Partitioning Around Medoids (Program PAM)**", in "**Finding Groups in Data: An Introduction to Cluster Analysis**", John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9780470316801.ch2, 1990.

63. [Kaufman et Rousseeuw, 1990b] L. Kaufman and P. Rousseeuw, "**Finding Groups in Data. An Introduction to Cluster Analysis**", Wiley-Interscience, New York, 1990.

64. [Kelaiaia et Merouani, 2013a] A. Kelaiaia and H. F. Merouani, "**Influence of stemming on Clustering of Arabic texts: Comparative Study in Document Retrieval**", Digital Library
URI: http://www.ijcaonline.org/archives/volume63/number14/10536-5529 Issue ISBN: 973-93-80872-94-0 DOI: 10.5120/10536-5529, 2013.

65. [Kelaiaia et Merouani, 2013b] A. Kelaiaia and H. F. Merouani, "**Clustering with Probabilistic Topic Models on Arabic Texts**", Conference CIIA 2013**,** Studies in Computational Intelligence, Springer-Verlag, Volume 488, 2013, pp 65-74. 2013.

66. [Kelaiaia et Merouani, 2014] A. Kelaiaia et H. F. Merouani, "**Etude comparative entre LDA et K-Moyennes en classification non supervisée sur la langue arabe**", Conférence Internationale sur l'Intelligence Artificielle et les Technologies de l'Information (ICA2IT'14) du 10 au 12 Mars 2014, Ouargla, Algérie, 2014.

67. [Kelaiaia et Merouani, 2016] A. Kelaiaia and H. F. Merouani, "**Clustering with probabilistic topic models on arabic texts: A comparative study of LDA and K-means**", The International Arab Journal of Information Technology (IAJIT), Volume 13, num 02, March, 2016.

68. [Keselj *et al.*, 2003] V. Keselj, F. Peng, N. Cercone, and C. Thomas, "**N-gram-based ˇ author profiles for authorship attribution**", In PACLING'03, pages 255–264, August, 2003.

69. [Khoja et Garside, 1999] S. Khoja and S. Garside, "Stemming Arabic Text", Technical report, Computing department, Lancaster University, Lancaster, U.K., 1999.

70. [Khoja, 2001] S. Khoja, "**APT: Arabic Part-of-speech Tagger**", Actes de l'atelier des étudiants de Second Meeting of the North American Chapter of the Association for Computational Linguistics, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, 2001.

71. [Kullback et Leibler, 1951] S. Kullback and R. A. Leibler, "**On Information and Sufficiency**", Annuals of Mathematical Statistics, 22:49–86, 1951.

72. [Lai et Wu, 2005] K.K. Lai and S.J. Wun, "**Using the Patent Co-citation Approach to Establish a New Patent Classification System**", Information Processing & Management 41(2), 313–330, 2005.

73. [Larsen et Aone, 1999] B. Larsen and C. Aone, "**Fast and effective text mining using linear time document clustering**", in Proceedings of the conference on knowledge discovery and data mining, pp 16–22, 1999.

74. [Law *et al.*, 2005] M. H. C. Law, A. Topchy, and A. K. Jain, "**Model-based clustering with probabilistic constraints**", In Proceedings of the 2005 SIAM International Conference on Data Mining (SDM-05), 2005.

75. [Lawrie *et al.*, 2001] D. Lawrie, W. B. Croft, A. L. Rosenberg, "**Finding Topic Words for Hierarchical Summarization**", In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information retrieval (SIGIR'01) : 249-357, New Orlean, LA, USA, 2001.

76. [Larkey et Connell, 2001] L. S. Larkey and M. E. Connell, "**Arabic information retrieval**" at UMASS in TREC-10. In Proceedings of the 10th Text Retrieval Conference (TREC'01), 2001.

77. [Larkey *et al.*, 2005] L. S. Larkey, L. Ballesteros, and M. E. Connell, "**Light Stemming for Arabic Information Retrieval**", Univ. of Massachusetts, Dept. of Computer Science, USA, 2005.

78. [Leclerc, 2000] J. Leclerc, " **L'aménagement linguistique dans le monde**",

http://www.tlfq.ulaval.ca/axl/monde/famarabe.htm, 2000.

79. [Lee et Seung, 1999] D. D. Lee and H. S. Seung, "**Learning the parts of objects with nonnegative matrix factorization**", Nature, Volume 401: pp. 391–407, October 1999, www.nature.com, 1999.

80. [Lee et Seung, 2001] D. D. Lee and H. S. Seung, "**Algorithms for non-negative matrix factorization**", In NIPS 13, pages 556–562, 2001.

81. [Lemur project, 2014], "**The Lemur Toolkit for Language Modeling and Information Retrieval"**, http://www.lemurproject.org

82. [Lin, 2002] J. Lin, "Divergence Measures based on the Shannon Entropy", IEEE Transactions on Information Theory, 37(1):145–151, 2002.

83. [Lloyd, 1982] S. Lloyd, "Least squares quantization in PCM", IEEE Transactions on Information Theory, 28, 129–137. Originally as an unpublished Bell laboratories Technical Note (1957), 1982.

84. [Lu *et al.*, 2011] Y. Lu, Q. Mei, and C. Zhai, "**Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA**", Information Retrieval (2011) 14:178–203, 2011.

85. [Maamouri et Bies, 2004] M. Maamouri and A. Bies, "**Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools"**, LDC, University of Pennsylvania, Philadelphia, PA 19104, USA, 2004.

86. [Maarek et Wecker, 1994] Y. S. Maarek and A. J. Wecker, "**The Librarian's Assistant: automatically organizing on-line books into dynamic bookshelves**", In Proceedings of the International Conference on Intelligent Multimedia Information Retrieval Systems and Management (RIAO'94), 1994.

87. [MacQueen, 1967] J. MacQueen, "Some methods for classification and analysis of multivariate observations", Pages 281–297 of: Fifth Berkeley Symposium on Mathematics, Statistics and Probability. University of California Press, 1967.

88. [Manning *et al.*, 2008] C. D. Manning, P. Raghavan, and H. Schtze, "**An Introduction to Information Retrieval**", Cambridge University Press, 2008.

89. [McCallum, 1996] A. K. McCallum, "**Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering**", http://www.cs.cmu.edu/~mccallum/bow,1996.

90. [McCallum, 2002] A. K. McCallum, "**MALLET: A Machine Learning for Language Toolkit**", http://mallet.cs.umass.edu, 2002.

91. [McLachlan et Basford, 1988] G. J. McLachlan and K. E. Basford, "**Mixture Models, Inference and Applications to Clustering**", Marcel Dekker, New York, 1988.

92. [Miller *et al.*, 1999] E. Miller, D. Shen, J. Liu, and C. Nicholas, "**Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System**", Journal of Digital Information, 1(5), 1999.

93. [Mladenic et Grobelnik, 2003] D. MLADENIĆ and M. GROBELNIK, "**Text and web mining**", Kluwer Academic Publishers, USA, 2003.

94. [Muhr *et al.*, 2010] M. Markus, K. Roman Kern, and M. Granitzer, "**Analysis of Structural Relationships for Hierarchical Cluster Labeling**", SIGIR'10, July 19–23, 2010, Geneva, Switzerland, 2010.

95. [Nanhong *et al.*, 2010] Y. Nanhong, S. Gauch, Q. Wang, and H. Luong, "**An Adaptive Ontology based Hierarchical Browsing System for CiteSeerx**", The Second International Conference on Knowledge and Systems Engineering, Hanoi, Vietnam, October 7-9, 2010, pp. 203-208, 2010.

96. [Navarro et al., 2011] E. Navarro, Y. Chudy, B. Gaume, G. Cabanac et K. Pinel-Sauvagnat, "Kodex ou comment organiser les résultats d'une recherche d'information par détection de communautés sur un graphe biparti ? ", In CORIA'11, Avignon, ARIA, pp. 25-40, 2011.

97. [Ng et Han, 1994] R.T. Ng and J. Han, "**CLARANS: a method for clustering objects for spatial data mining"** Knowledge and Data Engineering, IEEE Transactions on , vol.14, no.5, pp.1003,1016, Sep/Oct, 2002.

98. [Niu *et al.*, 2012] N. Niu, S. M. A. Reddivari, S., Mahmoud, T. Bhowmik, and S. Xu, "**Automatic labeling of software requirements clusters**", In Search-Driven Development-Users, Infrastructure, Tools and Evaluation (SUITE), 2012 ICSE Workshop on (pp. 17-20). IEEE, 2012.

99. [Osinski et al., 2004] S. Osinski, J. Stefanowski, and D. Weiss, "**Lingo: Search results clustering algorithm based on singular value decomposition**". In Proceedings of the International IIS: Intelligent Information Processing and Web Mining Conference, Advances in Soft Computing, pages 359– 368, Zakopane, Poland, 2004.

100. [Pasquier, 2003] N. Pasquier, "**Fouille de Données: Classification non supervisée**", Université de Nice Sophia-Antipolis Laboratoire I3S, France, 2003.

101. [Porter, 1980] M. F. Porter, "**An Algorithm for Suffix Stripping**", http://tartarus.org/~martin/PorterStemmer/def.txt, 1980.

102. [Rand, 1971] M. Rand, "**Objective criteria for the evaluation of clustering methods**", Journal of the American Statistical Association, Vol. 66, 1971, pp. 846–850, 1971.

103. [Řehůřek et Sojka, 2011] R. Řehůřek and P. Sojka,"**Gensim – Python Framework for Vector Space Modelling**", NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, http://radimrehurek.com/gensim/, 2011.

104. [Rokach, 2010] L. Rokach, "**A survey of Clustering Algorithms**", O. Maimon, L. Rokach (eds.), Data Mining and Knowledge Discovery Handbook, 2nd ed., 2010.

105. [Rosen-zvi *et al.*, 2004] M. Rosen-zvi, T. Griffiths, M. Steyvers, and P. Smyth, "**The author-topic model for authors and documents**", in Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Banff, Alberta, Canada, 2004.

106. [Salton, 1989] G. Salton, "**Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer**", Addison-Wesley, 1989.

107. [Salton *et al.*, 1975] G. Salton, A. Wong, and C. S. Yang, "**A vector space model for automatic indexing**", Communications of the ACM, v.18 n.11, p.613-620, Nov. 1975.

108. [Salton et Buckley, 1988] G. Salton and C. Buckley, "**Term-weighting approaches in automatic text retrieval"**, Information Processing and Management, 24(5), pp. 513–523, 1988.

109. [Sawaf *et al.*, 2001] H. Sawaf, J. Zaplo, and H. Ney, "**Statistical Classification Methods for Arabic News Articles"**, AIXPLAIN AG Monnetstrasse 18 D-52146 Würselen, Germany, 2001.

110. [Schmid, 1994] H. Schmid, "**Probabilistic part-of-speech tagging using decision trees**", International Conference on New Methods in Language Processing, Manchester, UK, 1994.

111. [Schütze et Silverstein, 1997] H. Schütze and C. Silverstein, "**Projections for efficient document clustering**", In ACM SIGIR Forum (Vol. 31, No. SI, pp. 74-81). ACM, 1997

112. [Sebastiani, 2002] F. Sebastiani, "**Machine Learning in Automated Text Categorization"**, Consiglio Nazionale delle Ricerche, Italy, 2002.

113. [Shannon, 1948] C. E. Shannon, "**A mathematical theory of communication**", Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, 1948.

114. [Singhal *et al.*, 1996] A. Singhal, C. Buckley, and M. Mitra, "**Pivoted Document Length Normalization**", ACM SIGIR Conference, pp. 21–29, 1996.

115. [Sokal et Sneath, 1963] R. R. Sokal and P. H. Sneath, "**Principles of numerical taxonomy**", Principles of numerical taxonomy, 1963.

116. [Stefanowski et Weiss, 2007] J. Stefanowski and D. Weiss, "**Comprehensible and Accurate Cluster Labels in Text Clustering**", Conference RIAO2007, Pittsburgh PA, U.S.A. May 30-June 1, 2007 - Copyright C.I.D. Paris, France, 2007.

117. [Steinbach *et al.*, 2000] M. Steinbach, G. Karypis, and V. Kumar, "**A comparison of document clustering techniques**". In *KDD workshop on text mining* (Vol. 400, No. 1, pp. 525-526), 2000.

118. [Steyvers et Griffiths, 2007] M. Steyvers and T. Griffiths, "**Probabilistic topic models**", Handbook of Latent Semantic Analysis, pages 424–440, 2007.

119. [Tuerlinckx, 2004] L. Tuerlinckx, "**La lemmatisation de l'arabe non classique**", JADT: 7es Journées internationales d'Analyse statistique des Données Textuelle, France, 2004.

120. [Turel et Can, 2011] A. Turel and F. Can, "**A new approach to search result clustering and labeling**", In Information Retrieval Technology (pp. 283-292). Springer Berlin Heidelberg, 2011.

121. [Turenne, 2000] N. Turenne, "**Apprentissage statistique pour l'extraction de concepts à partir du textes. Application au filtrage d'information textuelle**", Université Louis Pasteur, Starsbourg, France, 2000.

122. [Treeratpituk et Callan, 2006] P. Treeratpituk and J. Callan, "**Automatically labeling hierarchical clusters**". In: Proceedings of the 2006 International Conference on Digital Government Research, pp. 167–176. ACM, 2006.

123. [Tyron et Bailey, 1970] R. C. Tryon and D. Bailey, "**Cluster analysis**", McGraw-Hill, New York, 1970.

124. [Van Rijsbergen, 1979] C. J. van Rijsbergen, "**Information Retrieval**", Buttersworth, London, second edition, 1979.

125. [Wang et McCallum, 2006] X. Wang and A. McCallum, "**Topics over time: A non-markov continuous-time model of topical trends**", In KDD '06: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 424–433). New York, NY, USA: ACM, 2006.

126. [Wang et al., 2011] D. Wang, S. Zhu, T. Li, Y. Chi, and Y. Gong, "**Integrating document clustering and multidocument summarization**", ACM Transactions on Knowledge Discovery from Data (TKDD), 5(3), 14, 2011.

127. [Wei et Croft, 2006] X. Wei, and W. B. Croft, "**LDA-based document models for ad-hoc retrieval**", In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 178-185). ACM, 2006.

128. [Weiss, 2006] D. Weiss, "**Descriptive Clustering as a Method for Exploring Text Collections**", PhD Thesis, Poznan University of technology Institute of Computing Science, Poland, 2006.

129. [Weiss *et al.*, 1996] R. Weiss, B. Vélez, and M. A. Sheldon, "**HyPursuit: A hierarchical network search engine that exploits content-link hypertext clustering**", In Proceedings of the the seventh ACM conference on Hypertext(pp. 180-193). ACM, 1996.

130. [Xiao, 2010] Y. Xiao, "**A Survey of Document Clustering Techniques & Comparison of LDA and moVMF**", In: CS 229 Machine Learning Final Projects, 2010.

131. [Xu et Wunsch, 2005] R. Xu and D. Wunsch, "**Survey of clustering algorithms**", Neural Networks, IEEE Transactions on, 16(3), 645-678, 2005.

132. [Yang *et al.*, 2000] Y. Yang, T. Ault, T. Pierce, and C. W. Lattimer, "**Improving text categorization methods for event tracking**", In Proceedings of the 23rd ACMSIGIR conference (pp. 65–72), 2000.

133. [Yang et Pederson, 1997] Y. Yang, and J. O. Pedersen, "**A comparative study on feature selection in text categorization**", In ICML (Vol. 97, pp. 412-420), 1997.

134. [Yao *et al.*, 2009] M. Yao and A. K. McCallum, "**Efficient Methods for Topic Model Inference on Streaming Document Collections**", KDD'09, June 28–July 1, 2009, Paris, France. Copyright ACM 978-1-60558-495-9/09/06, 2009.

135. [Zamir et Etzioni, 1998] O. Zamir, and O. Etzioni, "**Web document clustering: A feasibility demonstration**", In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, 46–54, 1998.

136. [Zhang *et al.*, 2009] C. Zhang, H. Wang, Y. Liu, and H. Xu, "**Document clustering description extraction and its application**", In Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy (pp. 370-377). Springer Berlin Heidelberg, 2009.

137. [Zhang et Dong, 2004] D. Zhang and Y. Dong, "**Semantic, hierarchical, online clustering of web search results**", In Advanced Web Technologies and Applications (pp. 69-78). Springer Berlin Heidelberg, 2004.

138. [Zhao et Karypis, 2001] Y. Zhao and G. Karypis, "**Criterion functions for document clustering: Experiments and analysis**", Technical Report #01-40, University of Minnesota, 2001.

139. [Zhong et Ghosh, 2003] S. Zhong and J. Ghosh, "**A unified framework for model-based clustering**", The Journal of Machine Learning Research, 4, 1001-1037, 2003.

140. [Zhong et Ghosh, 2005] S. Zhong and J. Ghosh, "**Generative model-based document clustering: a comparative study**", Knowledge and Information Systems, 8(3), 374-384, 2005.

Productions personnelles

[Kelaiaia et Merouani, 2013a] A. Kelaiaia and H. F. Merouani, "**Influence of stemming on Clustering of Arabic texts: Comparative Study in Document Retrieval**", Digital Library
URI: http://www.ijcaonline.org/archives/volume63/number14/10536-5529 Issue ISBN: 973-93-80872-94-0 DOI: 10.5120/10536-5529, 2013.

[Kelaiaia et Merouani, 2013b] A. Kelaiaia and H. F. Merouani, "**Clustering with Probabilistic Topic Models on Arabic Texts**", Conference CIIA 2013**,** Studies in Computational Intelligence, Springer-Verlag, Volume 488, 2013, pp 65-74. 2013.

[Kelaiaia et Merouani, 2014] A. Kelaiaia et H. F. Merouani, "**Etude comparative entre LDA et K-Moyennes en classification non supervisée sur la langue arabe**", Conférence Internationale sur l'Intelligence Artificielle et les Technologies de l'Information (ICA2IT'14) du 10 au 12 Mars 2014, Ouargla, Algérie, 2014.

[Kelaiaia et Merouani, 2016] A. Kelaiaia and H. F. Merouani, "**Clustering with probabilistic topic models on arabic texts: A comparative study of LDA and K-means**", The International Arab Journal of Information Technology (IAJIT), Volume 13, num 02, March, 2016.

Annexes

1. Structure d’un fichier texte de la collection CCA


```
<?xml version="1.0" encoding="UTF-8"?>
- <tei.2>
  - <teiHeader id="AUT02">
    - <fileDesc>
      - <titleStmt>
          <title> !إدوارد سعيد لو كان مسلماً لترحمنا عليه</title>
          <author> د. محمد الأحمري </author>
        - <respStmt>
            <resp>compiled by</resp>
            <name>Latifa Al-Sulaiti</name>
          </respStmt>
        </titleStmt>
      - <publicationStmt>
          <publisher> مؤسسة الوقف الأسلامي , Saudi Arabia</publisher>
          <pubPlace> Saudi Arabia </pubPlace>
          <date>2003</date>
        </publicationStmt>
      - <sourceDesc>
          <p>created in machine-readable form in http://www.lahaonline.com </p>
        </sourceDesc>
      </fileDesc>
    - <encodingDesc>
      - <projectDesc>
          <p>Texts collected for use in the Corpus of Contemporary Arabic project, June, 2003</p>
        </projectDesc>
      - <samplingDecl>
          <p>Whole text of 1330 words copied from the site</p>
        </samplingDecl>
      </encodingDesc>
    - <profileDesc>
      - <creation>
          <date value="2003-10">October 2, 2003</date>
          <rs type="city">Riyadh, Saudi Arabia</rs>
        </creation>
        <langUsage>Arabic</langUsage>
      - <textClass>
        - <textDesc n="Autobiography">
            <channel mode="w">print; written</channel>
            <constitution type="single"/>
            <derivation type="original"/>
            <domain type="Arts"/>
            <factuality type="fact"/>
            <interaction type="none"/>
            <preparedness type="prepared"/>
          </textDesc>
        - <particDesc>
          - <person id="P1" age="Unknown" sex="M">
            - <birth date="Unknown">
                <date> Unknown </date>
                <name type="place"> Unknown </name>
                <nationality>= Unknown </nationality>
              <date> Unknown </date>
              <name type="place"> Unknown </name>
              <nationality>= Unknown </nationality>
            </birth>
            <firstLang>Arabic</firstLang>
            <langKnown>Unknown</langKnown>
            <residence> Unknown </residence>
            <education> Unknown </education>
            <occupation>Unknown</occupation>
          </person>
        </particDesc>
      </textClass>
    </profileDesc>
  </teiHeader>
- <text>
  - <body>
      <p>!إدوارد سعيد شغل المثقفين في العالم منذ ربع قرن</p>
      <p> على الجهتين إعلان عن محاضرة لـ"إدوارد سعيد"، فتأكدت منه عن حقيقة قدوم المحاضر فعلا إن اللقاء هو فقط عرض لفيلم مسجل عن محاضرة له سابقة، وأنهم يتوقعون جمهورا كبيرا لها</p>
      <p> ديار، وقد ربطت السياسة بين موقفيهما، فهو يناصر بحكم المعايشة والاختلاف الموقف قضايا العرب في أمريكا، وهو كاثوليكي، ومن أول من سمعت منه وصف المتعصبين اليروشلمائت الإسرائيل</p>
      <p> طائفة تسخر بلاده لمصلحة إسرائيل، ويهز أريحيته للحديث ذكر رجلين فقط في تاريخ العرب الثقافي الحديث، هما: إدوارد سعيد ومحمد حسين فضل الله؛ لأنه يراهما "متطرفين جدا في الذكاء</p>
      <p> ية تاريخية بين إدوارد سعيد وبرنارد لويس، بعد ضجة كتاب الاستشراق لإدوارد سعيد، وقد أدار المواجهة مؤرخ شهير هو "وليام مكنيل"، مؤلف كتب مهمة في التاريخ، مثل: "صعود الغرب"</p>
      <p> رأيه وحجته، ثم وقت للرد، ثم يجيبان عن أسئلة قليلة من القاعة، والمناقشة في وقت محدد.. يضاء قرب نهاية كل سباق نور أصفر موح بنهاية الوقت، ثم ضوء أحمر يوقف المتحدث. قال:</p>
      <p> الذي استولى على المجال دون منافس قادر على المواجهة. قال: لقد كانت لدينا جميعا أن يستمع نخبة المثقفين اليساريين الذين شغا معسكر الثقافة المتعلقة بالعالم العربي والإسلامي،</p>
      <p> حياته الباقية الطويلة في الإفساد الثقافي إلى اليوم.. إفساد الرؤية الغربية لتعالم الإسلامي، وإفساد الحكم على الثقافات، وتشويهها قصر طاقته، والتهوين منها، ونشر الرعب في قلوب الغربيين مما</p>
      <p> ثقافة اليمين المتطرف الصهيوني، وقد فاد الموقف اليساري في هذه القضايا إدوارد سعيد، المثقف البقظ، وحف به قوم لا يقل بعضهم شهرة عنه، رأت فيه معادا لتوجهها، ومدافعا عن المظلومين</p>
      <p> مرة واحدة في قاعة "ويستمنستر" في لندن بقدم مقابلة أو محاورة عن المنطقة، وعن جوانب من حياته، وكانت استاذت القاعة قبل بدء اللقاء وأثناء المحاضرة.. كان بجانبي أستاذ في أحد</p>
      <p> تنهى اللقاء، قلت: كانت المحاضرة "جيدة"، فغضب جاري وقال: ماذا تقول؟! إنها "عظيمة" وليست مجرد جيدة.. أدركت من استكثار جاري أني أنني هضمت مستوى اللقاء، وشعرت أنني قصرت</p>
      <p> في حق المحاضر. </p>
      <p> ويبدأ حديثه مع كاتب استكتدي يتحادثان في كتاب أقه الأخير عن الموسيقى، ولأن إدوارد حجة في الموسيقى كما يكتبون عنه، فعندما يقوّمون عمله وإنتاجه يضعون إلى جانب مكانته في</p>
      <p> ه للناس من مجرد ناقد أدبي أو أديب، أو مثقف سياسي ذي صوت عال للمسألة الفلسطينية إلى مفكر عالمي مضاد للاستعمار ومثقافته، شق الكتاب للناس دربا جديدا في التعرف على ظاهرة</p>
      <p> تابا مهما، وقد اهتمت بموضوع لغة الاستعمار، وكتبت رسالة علمية لها أعجبها في هذا الجانب، مما أغرى المثقفين الإسرائيليين بدعوتها وتكريمها، تخفيفا من حملتها عليهم، أو حرقا لها عن</p>
      <p> ياسيين؛ ليصوغ منها حملة شرسة على الاستعمار وأربابه، وربما رآها آخرون حملة ناقصة، عياء، دل عليها كتابان تاليان هما كتاب "تغطية الإسلام" و"ثقافة الإمبريالية"، لقد جار عليه</p>
      <p> إعجاب إدوارد بفوكو، كبير، حتى إنه اهتم بحضور محاضراته، وربما حضر درسه الأفتتاحي. وكان كتاب إدوارد سعيد قوية لكتابات فوكو، وتطبيقا للنظرية تجاوز بالتطبيق والتفريع فكرة "صاحب</p>
      <p> نظرية المعرفة سلطة" أو "المعرفة تستتبع السلطة" وأسلوبه العالي نفخ الحياة في جفاف التنظير</p>
      <p> ريف الغرب واستغلاله للمعرفة، وسيلة للاستعمار، وكان شجاعا بتجاوز عقدة خنوع المثقف وحرصه وتبعيته ونهافته، وخالف نهج مثقفي العرب في المهجر الذين يلوذون بالصمت خوفا من</p>
      <p> نفوذ اليهود، وحرصا على مواقعهم الوظيفية، الضم لمنظمة التحرير بقناعة</p>
      <p> الدنيا عليه في جامعة يتقذ فيها يهود نيويورك، ولم يبال بعرائضهم المطالبة بطرده، يقول: لم أملك إلا أن أقف على الحدود الثيالية وأرمي المحتلين بحجر، والتقطت الصورة له وهو يرجمهم</p>
      <p> بية اليهودية ملفا يذكر، فزلت هذه المطاردة من ثبوع قضيته. وخالف عرفات وصلحه بشجاعة، وخالف آراء كثير من مثقفي العرب في الموقف من مذابح هتلر لليهود، فهو يصدق حدوثها،</p>
      <p> كرين، ويتعاطف مع ضحاياها، وله في المسألة الفلسطينية رأي جريء حيث يطالب بدولة ديموقراطية واحدة في فلسطين لجميع، تحكمها الأغلبية ويرعى حق الأقلية، ويخالف من يقول بدولتين</p>
      <p> أحد الكتب القرنسية المثيرة التي كتبت مطولا عن خيانة المثقفين وتبعيتهم. وهنا تلاحظ ذلك الجانب المكروه للوصوليين، وسعيه المرفوع الذي يجد به ظهورهم، لم يكن يملك ما يخالف عليهم،</p>
      <p> مهاتفا بالخير؛ لأنا سنقضي ساعة من المتعة والفكر، ويراعة المواجهة، فشجعان العقول قليل.. وهو من المتة التي تستطيع أن تكشف حدود الحرية الفكرية في أمريكا، وما أصعب أن ترى</p>
      <p> حدود الحرية. </p>
      <p> الفلسطينية في الغرب، وكان مطلعا ومتابعا للأحداث ومعتقا قطنا، ومتحدثا آسرا، يفوق أسلوب حديثه أسلوب كتابته، كتبه القديمة والحديثة دائما معروضة في طبعات جديدة، لا يتنهي حولها</p>
      <p> الجدل، كان صيادا وعارضا للفكرة، مجيدا ومبدعا في اعتراضه ومؤثرا. عمقه في أدب الإنجليزية لا يبارى، وأجاد الفرنسية، ثم عاد لبيروت وتمكن من العربية</p>
      <p> صور المثقف" أي كتاب المقابلة الطويلة معه التي أجراها بارسيان؛ لرأيت فرق الطريقتين. لقد كان لسان العرب الحر، محاضرا ومحاورا، أما بعد حسم معركته مع نويس فقد قل من فكر في</p>
      <p> مواجهته.</p>
      <p> !..ودّع الناس مفكرا ومناضلا ثقافيا لا يديل له، ولا مقارب، ويغيب آثاره مدرسة في النزاهة ومكافحة الظلم، كان يعتقد الشك . كما وصف نفسه.. ولو كان مسلما لترحمنا عليه</p>
    </body>
  </text>
```

2. Liste des mots outils

طالما شيء شبه شرع شتان هؤلاء هي هيهات هيا ها هاهنا هاك هاته هاتي هاتين هاتان هب هج هكذا هل هلا هلم هم هما هن هنا هناك هنالك وهنا وهناك وهنالك فهنا فهناك فهنالك هذه هذي هذين هذا هذان هو يا يفعل يكون يلي يمكن يناقش يقدر يتضمن يوم يومئذ آه آها آي آيف آأن آالذي آالتي آان آانت آض آكل آل آلا آلما آم آما آناء آنفا آذا آذلك آخر إيه إياه إياها إياهم إياهما إياهن إياي إياك إياكم إياكما إياكن إيانا إحدى إلي إليه إليك إليكم إليكما إليكن إلا إلى إما إن وإن إنما وإنما فإنما إذ إذا إذما إذن إى إزاء أطعم أشياء أهلا أي أيا أياك أيان أيضا أين أآثر أبدا أضحى أصبح أصلا أعطى أعلم أف أفأمامك أحد أجل أجمع أل ألا ألفى أم أما أمام أمامك أمد أمس أمسى أن أنشأ أنه أنها أنا أنبأ أنت أنتم أنتما أنتن أنى أقبل أرى أسكن أو أوشك أوه أول أولئك أولاء أولالك أولات أولو أخبر أخذ أخرى أخو أى اشياء اهلا أي اي ايا اياك ايضا اين آآثر ابان ابدا ابتدأ اصبح اف افضل افضله افضلها افضلهم افضلهما احد احده احدى احسن احسنه احسنها احسنهم احسنهما اضحى الي اليه اليها اليهم اليهما اليوم الآن الآخر الألاء الألى الأخرى الا الاخر الاخري الفعل الحالي اللائي اللاتي اللاتى اللاواتي اللذي اللذين اللذان اللتي اللتيا اللتين اللتان اللواتي اللواتى المؤثر المؤثره المقال المقالات المتعلق المتعلقه الموضوع الذي الذين الذى القطع التي الوثائق الخ الى اما امسى ان انه انها انهم انهما انا انبرى انقلب انت انتم انتما انتن اذ اذا اذما اذن ارتد استحال اتجاه اتخذ اثناء او اوجد اول اولاء اولات اولو اخلولق اخر اخري اخرون اخرى ازاء بطآن بؤسا به بها بهم بهما بهن بهذا وبهذا فبهذا بي بيد بين بينه بينها بينهم بينهما بينما بئس بإن بأن بالفعل بان بات بد بدلا بعد وبعد فبعد بعدا بعدما بعض بفعل بفعله بفعلها بفعلهم بغتة بضع بك بكم بكما بكن بل بله بلا بلى بما بماذا بمن بن بنا برح بس بخ بخصوص صه صار صباح صبر صبرا صدقا صراحة دائما درى دواليك دون دونك عيانا عين عاد عامة عد عده عدا عجبا عل علي عليه عليها عليهم عليك علم علق على عما عن عنه عنها عنهم عنهما عند عندما عسى عوض فهذه فهذا فهو في فيه فيها فيهم فيهما فيكون فيما فإن فاعل فاعله فالذي فالتي فان فانه فبه فبين فعل فعله فعلها فعلا فعلى فعندما ففي ففيما فحتى فضلا فكان فكانت فكذلك فلهذه فلهذا فلان فلم ولم فلما فمن فمنه فمنها فمنهم فمنهما فمنذ فذلك فقط فقد فرادى فربما فتلك فو فوق فخلال فى غير غادر غالبا غدا غداة حي حين فحين وحين حيث وحيث فحيث حيثما حاشا حاي حبذا حدث حجا حمدا حمو حذار حقا حرى حسب حتي حتى حول ضد ضمن جير جدا جعل جل جلل جميع كي كيف كيفما كيت كأي كأين كأن كاد كان كانت كل كلها كلا كلاهما كلاكما كلما كلتا كم كما كذا كذلك كرب كسا كثيرا كخ له لها لهم لهما لهن لهذه لهذا لي ليس ليست ليت ليته ليتها ليتهم لئن لأن لا ولا فلا لايزال لان لاسيما لات لازال لبيك لدي لديه لديها لديهم لدن لدى لعل لعمر لفي لفعل لفعله لفعلها لفعلهم لك لكي لكيلا لكم لكما لكن لكنه لكنها لكنهم لكنهما للذي للتي لم لما لماذا لمن لن لنا لذلك لقد لو لولا لوما مؤثر مؤثره مؤثرا مه مهما ما مايزال ماانفك مابعد مابرح مادام ماعدا مافتئ ماذا ماقبل مازال مع معه معاذ مكانك مكانكم مكانكما مكانكن مم مما ممن من منه منها منهم منهما منذ مذ مقال مقالا مره مرارا مرة مساء متي متيقن متيقنه متعلق متعلقه متحقق متحققه متى مثل مثله مثلها مثلهم مثلا موضوع نبا نعم نعما نفس نحن نحو نخ ذه ذهب ذي ذين ذينك ذيت ذا ذاك ذان ذانك ذات ذاته ذاتها ذاتهم ذلك وذلك فذلك ذلكم ذلكما ذلكن ذو ذوي وا ذوات ذواتي ذواتا ذوو ذى قط قطع قطعا قاطبة قادر قادره قادرا قام قبل قبلما قد له قليل قلما قلة ريث ريثما رأى راح رب ربما رجع رويدك رزق سهلا سيكون سيما ساء ساءما سبت سحقا سمعا سقى سرا سرعان ستكون سواء سواءا سوف سوى ته تي تين تينك تان تانك تارة تبدل تعلم تعسا تفعل تحت تحول تضمن تضمنت تجاه تكون تلك تلكما تلقاء تمكن تناقش تقدر ترك تتضمن تخذ ثم ثمة و وشكان وهي وهب وهذه وهذا وهو وي ويكأن ويكون وآذلك وأبدا وأن فأن وا واها والذي والتي وان وبه وبين وبالفعل وبفعل وعلى وعن وعندما وفي ففي وفيما ففيما وفعل وحيد وحيده وحيدة وحتى وجد وكان وكانت وكذلك وله ولهذه ولهذا وليس ولايزال ولفعل ولكن ولذلك وما ومع ومن ومنه ومنها ومنهم ومنهما ومنذ وقد وراء وراءه وراءها وراءهم وراءهما وراءك وربما ورد وتلك وثائق وخلال خاصه خاصة خال خبر خصوص خصوصا خلا خلافا خلال خلف زعم زود كثير كثيرة أقل أكثر أكبر أصغر أعلى أدنى الكثير الكثيرة الأقل الأكثر الأكبر الأصغر الأعلى الأدنى رغم ورغم فرغم كنت وكنت فكنت قال فقال وقال قالت فقالت وقالت وأنا فأنا إنه إنها وإنه وإنها فإنه فإنها أنني وأنني ولا قليلا تم تمت ولكنه ولكنها فكل وكل لكل كله كلها داخل خارج بالنسبة ء ؤ ئ آ أ أ ب ت ة ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ى ي يجب عدد كبير صغير كبيرة صغيرة العدد الكبير الصغير الكبيرة الصغيرة بهذه وبهذه فبهذه

3. Partitions générées par LDK-Means

3.1. Forme translittérée

Cluster-0:
'To49_Trans.txt' : 0.681862 'To40_Trans.txt' : 0.630023 'To42_Trans.txt' : 0.625675 'To37_Trans.txt' : 0.553176 'To27_Trans.txt' : 0.514405 'To24_Trans.txt' : 0.441017 'To20_Trans.txt' : 0.436613 'To33_Trans.txt' : 0.280994

Cluster-1:
'Ec03_Trans.txt' : 0.38708 'ec07_Trans.txt' : 0.38251 'S19_Trans.txt' : 0.377092 'AUT64_Trans.txt' : 0.334918 'AUT54_Trans.txt' : 0.316788 'S11_Trans.txt' : 0.308229 'AUT69_Trans.txt' : 0.302101 'Edu10_Trans.txt' : 0.296947 'S09_Trans.txt' : 0.292304 'S22_Trans.txt' : 0.280438 'S18_Trans.txt' : 0.274402 'CHD01_Trans.txt' : 0.265908 'CHD06_Trans.txt' : 0.247841 'S24_Trans.txt' : 0.24553 'HM27_Trans.txt' : 0.232355 'CHD17_Trans.txt' : 0.226221

Cluster-2:
'Rec04_Trans.txt' : 0.549097 'Rec05_Trans.txt' : 0.540611 'Rec03_Trans.txt' : 0.513667 'Rec06_Trans.txt' : 0.507595 'Rec07_Trans.txt' : 0.491306 'Rec08_Trans.txt' : 0.468977 'Rec02_Trans.txt' : 0.421658 'Rec01_Trans.txt' : 0.333028 'HM15_Trans.txt' : 0.295043 'HM24_Trans.txt' : 0.274942 'Sc52_Trans.txt' : 0.255884 'Rec09_Trans.txt': 0.244193 'Ec18_Trans.txt' : 0.240991 'Sc56_Trans.txt' : 0.23012 'HM23_Trans.txt' : 0.222865 'HM22_Trans.txt' : 0.214405 'Sc55_Trans.txt' : 0.208103 'Sc62_Trans.txt' : 0.203624 'HM16_Trans.txt' : 0.198655 'HM21_Trans.txt' : 0.19471 'HM26_Trans.txt' : 0.193442 'Sc44_Trans.txt' : 0.178697 'HM12_Trans.txt' : 0.177114 'HM20_Trans.txt' : 0.159736 'Soc29_Trans.txt': 0.159439 'HM25_Trans.txt' : 0.156242 'CHD12_Trans.txt': 0.145848 'S14_Trans.txt' : 0.13244 'To25_Trans.txt' : 0.129961 'Int14_Trans.txt' : 0.126985

Cluster-3:
'Soc20_Trans.txt' : 0.334618 'AUT63_Trans.txt' : 0.332911 'AUT61_Trans.txt' : 0.317959 'HM18_Trans.txt' : 0.300707 'AUT53_Trans.txt' : 0.297968 'Edu07_Trans.txt' : 0.29479 'AUT57_Trans.txt' : 0.294209 'AUT62_Trans.txt' : 0.277897 'AUT71_Trans.txt' : 0.274226 'AUT31_Trans.txt' : 0.272166 'Edu09_Trans.txt' : 0.271093 'Soc28_Trans.txt' : 0.265916 'AUT29_Trans.txt' : 0.264046 'Edu08_Trans.txt' : 0.261914 'HM19_Trans.txt' : 0.258493 'AUT68_Trans.txt' : 0.254831 'Sc29_Trans.txt' : 0.254482 'Soc03_Trans.txt' : 0.228364 'AUT34_Trans.txt' : 0.222141 'Int21_Trans.txt' : 0.220411 'HM11_Trans.txt' : 0.216158 'Edu05_Trans.txt' : 0.21517 'HM28_Trans.txt' : 0.214114 'AUT72_Trans.txt' : 0.214078 'Edu12_Trans.txt' : 0.213626 'Sc48_Trans.txt' : 0.212034 'Sc50_Trans.txt' : 0.200652 'Pol02_Trans.txt' : 0.197792 'Spo05_Trans.txt' : 0.196971 'CHD08_Trans.txt' : 0.183633 'Int13_Trans.txt' : 0.171239

Cluster-4:
'AUT56_Trans.txt' : 0.39667 'AUT70_Trans.txt' : 0.374671 'Int20_Trans.txt' : 0.364104 'Int23_Trans.txt' : 0.349755 'AUT66_Trans.txt' : 0.342473 'Int16_Trans.txt' : 0.341705 'Int06_Trans.txt' : 0.339014 'AUT55_Trans.txt' : 0.336135 'Soc16_Trans.txt' : 0.330804 'AUT25_Trans.txt' : 0.312807 'AUT67_Trans.txt' : 0.287712 'Int22_Trans.txt' : 0.283632 'Int10_Trans.txt' : 0.249059 'Soc22_Trans.txt' : 0.239722 'S12_Trans.txt' : 0.239476 'S13_Trans.txt' : 0.232677 'AUT30_Trans.txt' : 0.229549 'AUT41_Trans.txt' : 0.229173 'S20_Trans.txt' : 0.228579 'Soc06_Trans.txt' : 0.224716 'AUT35_Trans.txt' : 0.210421 'AUT36_Trans.txt' : 0.209855 'AUT28_Trans.txt' : 0.208422 'AUT43_Trans.txt' : 0.200879 'S10_Trans.txt' : 0.176888 'S15_Trans.txt' : 0.171943 'CHD07_Trans.txt' : 0.16566 'CHD04_Trans.txt' : 0.157773

Cluster-5:
'Pol03_Trans.txt' : 0.441481 'Pol10_Trans.txt' : 0.375337 'Soc07_Trans.txt' : 0.368282 'Pol07_Trans.txt' : 0.348823 'Pol08_Trans.txt' : 0.348711 'Soc01_Trans.txt' : 0.347031 'Soc09_Trans.txt' : 0.34239 'Int17_Trans.txt' : 0.324472 'Soc08_Trans.txt' : 0.309406 'Soc25_Trans.txt' : 0.305957 'Soc19_Trans.txt' : 0.299197 'AUT49_Trans.txt' : 0.290268 'Soc26_Trans.txt' : 0.279777 'Soc02_Trans.txt' : 0.278674 'Soc23_Trans.txt' : 0.273933 'AUT40_Trans.txt' : 0.265375 'Rel09_Trans.txt' : 0.261417 'Ec19_Trans.txt' : 0.254709 'Ec08_Trans.txt' : 0.250272 'Soc12_Trans.txt' : 0.25022 'Soc21_Trans.txt' : 0.246013 'Soc30_Trans.txt' : 0.238082 'Sc31_Trans.txt' : 0.226585 'AUT50_Trans.txt' : 0.213222 'AUT47_Trans.txt' : 0.187076

Cluster-6:
'Sc66_Trans.txt' : 0.699462 'Pol05_Trans.txt' : 0.689953 'Pol04_Trans.txt' : 0.680822 'Sc70_Trans.txt' : 0.654536 'Sc69_Trans.txt' : 0.645328 'Ec28_Trans.txt' : 0.640893 'Ec27_Trans.txt' : 0.63648 'Sc67_Trans.txt' : 0.63429 'Sc68_Trans.txt' : 0.568936 'Pol06_Trans.txt' : 0.566751 'Ec20_Trans.txt' : 0.319019 'SC60_Trans.txt' : 0.198371 'AUT33_Trans.txt' : 0.181823 'AUT32_Trans.txt' : 0.161164

Cluster-7:
'Int01_Trans.txt' : 0.523282 'S07_Trans.txt' : 0.511621 'To60_Trans.txt' : 0.494458 'Rel02_Trans.txt' : 0.491195 'Int05_Trans.txt' : 0.490353 'AUT16_Trans.txt' : 0.483079 'To61_Trans.txt' : 0.478993 'Int02_Trans.txt' : 0.47754 'AUT07_Trans.txt' : 0.475199 'To58_Trans.txt' : 0.470023 'AUT21_Trans.txt' : 0.464648 'S05_Trans.txt' : 0.464558 'Sc06_Trans.txt' : 0.454386 'Int03_Trans.txt' : 0.447727 'S04_Trans.txt' : 0.44638 'Sc01_Trans.txt' : 0.443274 'AUT10_Trans.txt' : 0.43722 'Rel01_Trans.txt' : 0.434809 'CHD26_Trans.txt' : 0.434098 'Int04_Trans.txt' : 0.427506 'CHD27_Trans.txt' : 0.426989 'Pol01_Trans.txt' : 0.409729 'Rel04_Trans.txt' : 0.408358 'AUT17_Trans.txt' : 0.400735 'To59_Trans.txt' : 0.395346 'CHD25_Trans.txt' : 0.387973 'S01_Trans.txt' : 0.385201 'S03_Trans.txt' : 0.37172 'AUT08_Trans.txt' : 0.3686 'HM05_Trans.txt' : 0.363489 'CHD20_Trans.txt' : 0.361943 'AUT11_Trans.txt' : 0.358905 'AUT06_Trans.txt' : 0.35387 'AUT04_Trans.txt' : 0.353162 'AUT23_Trans.txt' : 0.347737 'S08_Trans.txt' : 0.333839 'Edu03_Trans.txt' : 0.333838 'S06_Trans.txt' : 0.328084 'AUT20_Trans.txt' : 0.312013 'AUT14_Trans.txt' : 0.309813 'AUT22_Trans.txt' : 0.308246 'AUT19_Trans.txt' : 0.293647 'Ec29_Trans.txt' : 0.285711 'AUT03_Trans.txt' : 0.271064 'CHD21_Trans.txt' : 0.254551 'AUT05_Trans.txt' : 0.250701 'AUT13_Trans.txt' : 0.240058 'AUT12_Trans.txt' : 0.237959 'Edu02_Trans.txt' : 0.230199 'CHD22_Trans.txt' : 0.205985 'CHD24_Trans.txt' : 0.204411 'Edu01_Trans.txt' : 0.195133 'CHD23_Trans.txt' : 0.192218 'To50_Trans.txt' : 0.190476 'S02_Trans.txt' : 0.107931 'CHD19_Trans.txt' : 0.0965736 'S23_Trans.txt' : 0.0913832 'CHD14_Trans.txt' : 0.0715248 'CHD13_Trans.txt' : 0.0637937 'CHD11_Trans.txt' : 0.0568793

Cluster-8:
'To48_Trans.txt' : 0.808261 'To38_Trans.txt' : 0.736657 'To47_Trans.txt' : 0.692323 'Int15_Trans.txt' : 0.688037 'Soc04_Trans.txt' : 0.682952 'Sc20_Trans.txt' : 0.64881 'To39_Trans.txt' : 0.646169 'To36_Trans.txt' : 0.582468 'To44_Trans.txt' : 0.500793

'To41_Trans.txt' : 0.495374 'To46_Trans.txt' : 0.448575 'Sc22_Trans.txt' : 0.411343 'To45_Trans.txt' : 0.409029 'To14_Trans.txt' : 0.269861 'To09_Trans.txt' : 0.256172 'To18_Trans.txt' : 0.254902 'To16_Trans.txt' : 0.248326 'To06_Trans.txt' : 0.229607 'To52_Trans.txt' : 0.21626 'To19_Trans.txt' : 0.206406

Cluster-9:
'Sc57_Trans.txt' : 0.37539 'Sc42_Trans.txt' : 0.361596 'Sc30_Trans.txt' : 0.359329 'Sc58_Trans.txt' : 0.356647 'Sc25_Trans.txt' : 0.355674 'Sc33_Trans.txt' : 0.35551 'Sc21_Trans.txt' : 0.353485 'Sc38_Trans.txt' : 0.33624 'Sc32_Trans.txt' : 0.32682 'Sc23_Trans.txt' : 0.325876 'Sc26_Trans.txt' : 0.324679 'Sc37_Trans.txt' : 0.32207 'Sc40_Trans.txt' : 0.312522 'Sc54_Trans.txt' : 0.309167 'Sc41_Trans.txt' : 0.305871 'Sc27_Trans.txt' : 0.301254 'Sc43_Trans.txt' : 0.290813 'Sc28_Trans.txt' : 0.29005 'Sc53_Trans.txt' : 0.279917 'HM13_Trans.txt' : 0.271805 'Sc24_Trans.txt' : 0.270727 'Soc14_Trans.txt' : 0.268022 'Sc45_Trans.txt' : 0.241258 'HM08_Trans.txt' : 0.239132 'Sc46_Trans.txt' : 0.238665 'HM09_Trans.txt' : 0.228711 'S31_Trans.txt' : 0.217442 'Sc51_Trans.txt' : 0.206775 'HM14_Trans.txt' : 0.20024 'HM10_Trans.txt' : 0.197374 'Sc49_Trans.txt' : 0.192699 'HM32_Trans.txt' : 0.177441 'HM30_Trans.txt' : 0.170016 'CHD18_Trans.txt' : 0.162329 'S21_Trans.txt' : 0.156934 'To02_Trans.txt' : 0.140081 'CHD05_Trans.txt' : 0.126299 'CHD16_Trans.txt' : 0.119921 ‘Edu06_Trans.txt' : 0.114015

Cluster-10:
'Rel10_Trans.txt' : 0.540071 'Rel14_Trans.txt' : 0.486277 'Rel16_Trans.txt' : 0.475654 'Rel15_Trans.txt' : 0.455645 'Rel18_Trans.txt' : 0.441076 'Int11_Trans.txt' : 0.435698 'Rel11_Trans.txt' : 0.430832 'Int18_Trans.txt' : 0.372114 'Rel17_Trans.txt' : 0.371066 'AUT46_Trans.txt' : 0.358929 'Rel06_Trans.txt' : 0.330005 'AUT51_Trans.txt' : 0.324439 'Rel12_Trans.txt' : 0.310193 'Rel07_Trans.txt' : 0.300498 'Rel05_Trans.txt' : 0.290677 'Int19_Trans.txt' : 0.288018 'AUT24_Trans.txt' : 0.283611 'AUT45_Trans.txt' : 0.281791 'AUT52_Trans.txt' : 0.276458 'Edu11_Trans.txt' : 0.275288 'Int12_Trans.txt' : 0.272809 'AUT59_Trans.txt' : 0.261226 'AUT60_Trans.txt' : 0.260969 'Rel08_Trans.txt' : 0.252872 'Spo01_Trans.txt' : 0.252228 'Soc15_Trans.txt' : 0.251532 'Rel19_Trans.txt' : 0.250531 'AUT38_Trans.txt' : 0.24322 ‘Soc18_Trans.txt' : 0.238534 'AUT58_Trans.txt' : 0.233304 'Spo02_Trans.txt' : 0.231291 'HM17_Trans.txt' : 0.229114 'AUT26_Trans.txt' : 0.224939 'Soc24_Trans.txt' : 0.222615 'AUT73_Trans.txt' : 0.222116 'Soc10_Trans.txt' : 0.218578 'Spo06_Trans.txt' : 0.214442 'Soc27_Trans.txt' : 0.213985 'AUT48_Trans.txt' : 0.213483 'AUT39_Trans.txt' : 0.20436 'AUT27_Trans.txt' : 0.186554 'Rel13_Trans.txt' : 0.185317 'AUT65_Trans.txt' : 0.18374 'Entr02_Trans.txt' : 0.179441 'Soc05_Trans.txt' : 0.179123 'AUT42_Trans.txt' : 0.164757 'S16_Trans.txt' : 0.161255 'S17_Trans.txt' : 0.142964 'AUT37_Trans.txt' : 0.136698 'CHD09_Trans.txt' : 0.131146 'Soc13_Trans.txt' : 0.125982 'CHD03_Trans.txt' : 0.109492 'CHD02_Trans.txt' : 0.10419

Cluster-11:
'Ec13_Trans.txt' : 0.630999 'Ec09_Trans.txt' : 0.594384 'Ec10_Trans.txt' : 0.588533 'Ec05_Trans.txt' : 0.549531 'Ec16_Trans.txt' : 0.380268 'Sc59_Trans.txt' : 0.369309 'AUT44_Trans.txt' : 0.322258 'Sc47_Trans.txt' : 0.291266

Cluster-12:
'To34_Trans.txt' : 0.633541 'Ec14_Trans.txt' : 0.581093 'To23_Trans.txt' : 0.555174 'To11_Trans.txt' : 0.542754 'To13_Trans.txt' : 0.520445 'To29_Trans.txt' : 0.460636 'To30_Trans.txt' : 0.455117 'To28_Trans.txt' : 0.440623 'Ec26_Trans.txt' : 0.414392 'To26_Trans.txt' : 0.396927 'Ec01_Trans.txt' : 0.392307 'Ec12_Trans.txt' : 0.390143 'To54_Trans.txt' : 0.386876 'To05_Trans.txt' : 0.359242 'To04_Trans.txt' : 0.357466 'To10_Trans.txt' : 0.335491 'Ec22_Trans.txt' : 0.335411 'Ec25_Trans.txt' : 0.329079 'To56_Trans.txt' : 0.325673 'To07_Trans.txt' : 0.325072 'To22_Trans.txt' : 0.321392 'Ec17_Trans.txt' : 0.313917 'Ec11_Trans.txt' : 0.307896 'To12_Trans.txt' : 0.29989 'To31_Trans.txt' : 0.299144 'To57_Trans.txt' : 0.298198 'Soc11_Trans.txt' : 0.269677 'Ec24_Trans.txt' : 0.266205 'To21_Trans.txt' : 0.233743 'Ec04_Trans.txt' : 0.220764 'Ec15_Trans.txt' : 0.212633 'To53_Trans.txt' : 0.211712 'To01_Trans.txt' : 0.208995 'Ec06_Trans.txt' : 0.205368 'Ec02_Trans.txt' : 0.202615 'Soc17_Trans.txt' : 0.180178 'To03_Trans.txt' : 0.149615 ‘To17_Trans.txt' : 0.139324

Cluster-13:
'To08_Trans.txt' : 0.448491 'Spo04_Trans.txt' : 0.426476 'To35_Trans.txt' : 0.423494 'Spo03_Trans.txt' : 0.419812 'To43_Trans.txt' : 0.418985 'To15_Trans.txt' : 0.409977 'To32_Trans.txt' : 0.39246 'CHD15_Trans.txt' : 0.323984

Cluster-14:
'Sc07_Trans.txt' : 0.444617 'Sc08_Trans.txt' : 0.424191 'HM04_Trans.txt' : 0.410439 'Sc63_Trans.txt' : 0.410001 'Sc05_Trans.txt' : 0.409016 'HM02_Trans.txt' : 0.407569 'Sc10_Trans.txt' : 0.406454 'SC61_Trans.txt' : 0.391222 'Sc64_Trans.txt' : 0.384988 'HM06_Trans.txt' : 0.37238 'HM31_Trans.txt' : 0.362765 'Sc65_Trans.txt' : 0.360261 'HM03_Trans.txt' : 0.351757 'HM07_Trans.txt' : 0.301024 'Edu04_Trans.txt' : 0.296067 'Rel03_Trans.txt' : 0.280983 'Ec23_Trans.txt' : 0.280175 'Sc09_Trans.txt' : 0.279057 'HM29_Trans.txt' : 0.276304 'Ec21_Trans.txt' : 0.274958 'AUT09_Trans.txt' : 0.271586 'Sc02_Trans.txt' : 0.269737 'Sc03_Trans.txt' : 0.262359 ‘Pol09_Trans.txt' : 0.260172 'AUT02_Trans.txt' : 0.246532 'Sc04_Trans.txt' : 0.238143 'HM01_Trans.txt' : 0.221603 'To55_Trans.txt' : 0.117531

3.2. Forme nettoyée

Cluster-0:
'AUT59_ClTrans.txt' : 0.360264 'AUT60_ClTrans.txt' : 0.360235 'Spo01_ClTrans.txt' : 0.315099 'Spo02_ClTrans.txt' : 0.2939
'Soc24_ClTrans.txt' : 0.292153 'Spo06_ClTrans.txt' : 0.271642 'Soc01_ClTrans.txt' : 0.248769 'Sc29_ClTrans.txt' : 0.238735
'Soc28_ClTrans.txt' : 0.233403 'AUT43_ClTrans.txt' : 0.229241 'AUT47_ClTrans.txt' : 0.227961 'AUT26_ClTrans.txt' : 0.226364
'AUT40_ClTrans.txt' : 0.224711 'Rel19_ClTrans.txt' : 0.21962 'Spo03_ClTrans.txt' : 0.217506 'To06_ClTrans.txt' : 0.215943
'Sc22_ClTrans.txt' : 0.214715 'AUT27_ClTrans.txt' : 0.212705 'To18_ClTrans.txt' : 0.210513 'AUT58_ClTrans.txt' : 0.203498
'Sc28_ClTrans.txt' : 0.203334 'Soc20_ClTrans.txt' : 0.197824 'AUT42_ClTrans.txt' : 0.194861 'To19_ClTrans.txt' : 0.1914
'S11_ClTrans.txt' : 0.17855 'S19_ClTrans.txt' : 0.173852 'Int01_ClTrans.txt' : 0.17057 'CHD07_ClTrans.txt' : 0.164802
'AUT69_ClTrans.txt' : 0.164359 'CHD01_ClTrans.txt' : 0.161365 'To02_ClTrans.txt' : 0.158749 'CHD17_ClTrans.txt' : 0.154835
'HM14_ClTrans.txt' : 0.154257 'S24_ClTrans.txt' : 0.150904 'HM06_ClTrans.txt' : 0.140325

Cluster-1: 'Int12_ClTrans.txt' : 0.456516 'Rel11_ClTrans.txt' : 0.388638 'Soc08_ClTrans.txt' : 0.379576 'Rel17_ClTrans.txt' : 0.360241
'AUT46_ClTrans.txt' : 0.343657 'Int10_ClTrans.txt' : 0.327908 'Rel06_ClTrans.txt' : 0.321592 'Soc27_ClTrans.txt' : 0.31199
'AUT45_ClTrans.txt' : 0.304364 'Rel08_ClTrans.txt' : 0.300065 'Soc14_ClTrans.txt' : 0.299554 'Rel02_ClTrans.txt' : 0.291767
'AUT41_ClTrans.txt' : 0.281071 'AUT73_ClTrans.txt' : 0.266692 'Int22_ClTrans.txt' : 0.22584 'AUT36_ClTrans.txt' : 0.212721
'HM18_ClTrans.txt' : 0.201498 'HM07_ClTrans.txt' : 0.19544

Cluster-2:
'AUT49_ClTrans.txt' : 0.400722 'Int17_ClTrans.txt' : 0.376334 'Soc07_ClTrans.txt' : 0.35864 'Soc09_ClTrans.txt' : 0.35077
'Soc21_ClTrans.txt' : 0.308266 'Rel09_ClTrans.txt' : 0.304193 'Soc19_ClTrans.txt' : 0.3037 'Soc22_ClTrans.txt' : 0.300097
'Edu09_ClTrans.txt' : 0.292771 'Soc23_ClTrans.txt' : 0.281245 'AUT39_ClTrans.txt' : 0.268674 'Sc07_ClTrans.txt' : 0.252923
'Soc05_ClTrans.txt' : 0.252661 'AUT32_ClTrans.txt' : 0.250956 'Edu11_ClTrans.txt' : 0.249389 'Sc47_ClTrans.txt' : 0.216198
'CHD06_ClTrans.txt' : 0.215639

Cluster-3:
'To43_ClTrans.txt' : 0.421609 'To14_ClTrans.txt' : 0.405931 'To47_ClTrans.txt' : 0.395397 'To46_ClTrans.txt' : 0.365041
'To32_ClTrans.txt' : 0.354447 'Sc30_ClTrans.txt' : 0.35414 'To16_ClTrans.txt' : 0.350867 'To15_ClTrans.txt' : 0.344495
'Sc23_ClTrans.txt' : 0.336465 'Sc33_ClTrans.txt' : 0.322667 'S31_ClTrans.txt' : 0.291802 'S14_ClTrans.txt' : 0.279594
'Sc08_ClTrans.txt' : 0.245203 'To08_ClTrans.txt' : 0.235888

Cluster-4:
'SC61_ClTrans.txt' : 0.361095 'Sc65_ClTrans.txt' : 0.340121 'Sc59_ClTrans.txt' : 0.339923 'Sc25_ClTrans.txt' : 0.314611
'Sc42_ClTrans.txt' : 0.30081 'Sc57_ClTrans.txt' : 0.300417 'Sc56_ClTrans.txt' : 0.297935 'Sc27_ClTrans.txt' : 0.297765 'Sc21_ClTrans.txt'
: 0.295881 'Sc63_ClTrans.txt' : 0.285815 'Sc62_ClTrans.txt' : 0.27791 'Sc26_ClTrans.txt' : 0.276898 'Sc40_ClTrans.txt' : 0.267254
'Sc38_ClTrans.txt' : 0.266048 'Sc37_ClTrans.txt' : 0.262215 'Sc54_ClTrans.txt' : 0.25638 'Sc32_ClTrans.txt' : 0.255477 'Sc41_ClTrans.txt'
: 0.254071 'SC60_ClTrans.txt' : 0.250193 'Sc46_ClTrans.txt' : 0.246684 'Sc70_ClTrans.txt' : 0.243197 'Sc69_ClTrans.txt' : 0.242754
'HM15_ClTrans.txt' : 0.235557 'Sc64_ClTrans.txt' : 0.232952 'Ec18_ClTrans.txt' : 0.231621 'HM09_ClTrans.txt' : 0.230734
'Sc68_ClTrans.txt' : 0.229981 'Sc53_ClTrans.txt' : 0.228428 'Sc43_ClTrans.txt' : 0.222151 'Sc31_ClTrans.txt' : 0.213811
'Sc44_ClTrans.txt' : 0.212138 'Ec15_ClTrans.txt' : 0.209072 'Sc45_ClTrans.txt' : 0.206666 'Ec16_ClTrans.txt' : 0.206567
'Sc55_ClTrans.txt' : 0.205382 'Sc10_ClTrans.txt' : 0.191495 'HM24_ClTrans.txt' : 0.187503 'Sc24_ClTrans.txt' : 0.184092
'Sc51_ClTrans.txt' : 0.181614 'HM23_ClTrans.txt' : 0.181203 'Sc48_ClTrans.txt' : 0.178546 'Sc49_ClTrans.txt' : 0.171648
'HM21_ClTrans.txt' : 0.170546 'HM32_ClTrans.txt' : 0.163172 'HM20_ClTrans.txt' : 0.15891 'HM26_ClTrans.txt' : 0.158186
'HM19_ClTrans.txt' : 0.145326 'HM28_ClTrans.txt' : 0.140429 'HM12_ClTrans.txt' : 0.139209 'HM05_ClTrans.txt' : 0.137522
'Soc13_ClTrans.txt' : 0.103407 'Int14_ClTrans.txt' : 0.0974398 'Rec09_ClTrans.txt' : 0.0968561 'CHD04_ClTrans.txt' : 0.0939981

Cluster-5:
'AUT33_ClTrans.txt' : 0.367861 'AUT31_ClTrans.txt' : 0.364232 'Int19_ClTrans.txt' : 0.352692 'AUT63_ClTrans.txt' : 0.351767
'AUT29_ClTrans.txt' : 0.322863 'Soc18_ClTrans.txt' : 0.311937 'AUT66_ClTrans.txt' : 0.30335 'AUT71_ClTrans.txt' : 0.286621
'Edu05_ClTrans.txt' : 0.272479 'AUT37_ClTrans.txt' : 0.27192 'Soc03_ClTrans.txt' : 0.269955 'AUT35_ClTrans.txt' : 0.25417
'Soc30_ClTrans.txt' : 0.242185 'Sc01_ClTrans.txt' : 0.241106 'Sc05_ClTrans.txt' : 0.23555 'AUT44_ClTrans.txt' : 0.228064
'CHD08_ClTrans.txt' : 0.213421

Cluster-6:
'AUT13_ClTrans.txt' : 0.386527 'AUT10_ClTrans.txt' : 0.335244 'AUT21_ClTrans.txt' : 0.312137 'Rel01_ClTrans.txt' : 0.303494
'AUT04_ClTrans.txt' : 0.301343 'AUT08_ClTrans.txt' : 0.288069 'AUT14_ClTrans.txt' : 0.282766 'AUT67_ClTrans.txt' : 0.271218
'AUT65_ClTrans.txt' : 0.261895 'AUT64_ClTrans.txt' : 0.261825 'AUT51_ClTrans.txt' : 0.245763 'Soc11_ClTrans.txt' : 0.245309
'AUT16_ClTrans.txt' : 0.244228 'To59_ClTrans.txt' : 0.23528 'Pol05_ClTrans.txt' : 0.233645 'To61_ClTrans.txt' : 0.232488
'To58_ClTrans.txt' : 0.232487 'AUT09_ClTrans.txt' : 0.230449 'Int06_ClTrans.txt' : 0.220748 'Spo05_ClTrans.txt' : 0.213528
'Soc10_ClTrans.txt' : 0.213223 'Spo04_ClTrans.txt' : 0.212214 'AUT17_ClTrans.txt' : 0.210499 'Rel03_ClTrans.txt' : 0.206899
'AUT23_ClTrans.txt' : 0.205295 'Int03_ClTrans.txt' : 0.204541 'AUT19_ClTrans.txt' : 0.191529 'AUT50_ClTrans.txt' : 0.182499
'AUT53_ClTrans.txt' : 0.171822 'HM03_ClTrans.txt' : 0.167441 'AUT11_ClTrans.txt' : 0.164387 'HM04_ClTrans.txt' : 0.160887
'AUT28_ClTrans.txt' : 0.157957 'Int13_ClTrans.txt' : 0.15353 'Soc29_ClTrans.txt' : 0.147605 'Sc02_ClTrans.txt' : 0.142195
'Sc09_ClTrans.txt' : 0.137874 'To52_ClTrans.txt' : 0.131701 'S02_ClTrans.txt' : 0.129298

Cluster-7:
'Soc15_ClTrans.txt' : 0.450265 'Rel05_ClTrans.txt' : 0.3573 'Int20_ClTrans.txt' : 0.306925 'AUT56_ClTrans.txt' : 0.301558
'AUT55_ClTrans.txt' : 0.298276 'Soc16_ClTrans.txt' : 0.282229 'AUT70_ClTrans.txt' : 0.279043 'AUT61_ClTrans.txt' : 0.273481
'AUT54_ClTrans.txt' : 0.272539 'Int23_ClTrans.txt' : 0.265826 'Rel12_ClTrans.txt' : 0.264461 'AUT57_ClTrans.txt' : 0.261355
'Int16_ClTrans.txt' : 0.257704 'Rel07_ClTrans.txt' : 0.250889 'S12_ClTrans.txt' : 0.248023 'AUT24_ClTrans.txt' : 0.24449
'AUT25_ClTrans.txt' : 0.239274 'Ec19_ClTrans.txt' : 0.231914 'AUT62_ClTrans.txt' : 0.227528 'HM29_ClTrans.txt' : 0.223037
'HM30_ClTrans.txt' : 0.221409 'Soc25_ClTrans.txt' : 0.208319 'AUT68_ClTrans.txt' : 0.204528 'HM11_ClTrans.txt' : 0.198767
'AUT72_ClTrans.txt' : 0.196946 'S20_ClTrans.txt' : 0.193942 'S15_ClTrans.txt' : 0.1875 'S13_ClTrans.txt' : 0.181073

'HM31_ClTrans.txt' : 0.17739 'S10_ClTrans.txt' : 0.157043 'CHD02_ClTrans.txt' : 0.155604 'Sc50_ClTrans.txt' : 0.15554
'Pol01_ClTrans.txt' : 0.153674 'S06_ClTrans.txt' : 0.146088 'CHD13_ClTrans.txt' : 0.139425 'To55_ClTrans.txt' : 0.120489

Cluster-8:
'S23_ClTrans.txt' : 0.379169 'Sc58_ClTrans.txt' : 0.314411 'S22_ClTrans.txt' : 0.307258 'S16_ClTrans.txt' : 0.304333
'HM27_ClTrans.txt' : 0.297944 'HM08_ClTrans.txt' : 0.296789 'HM13_ClTrans.txt' : 0.295168 'CHD14_ClTrans.txt' : 0.288536
'S21_ClTrans.txt' : 0.287498 'CHD19_ClTrans.txt' : 0.276465 'CHD15_ClTrans.txt' : 0.273928 'S09_ClTrans.txt' : 0.263427
'CHD03_ClTrans.txt' : 0.263359 'S17_ClTrans.txt' : 0.262432 'Int02_ClTrans.txt' : 0.238564

Cluster-9:
'Ec27_ClTrans.txt' : 0.428851 'Ec24_ClTrans.txt' : 0.412405 'Ec28_ClTrans.txt' : 0.389492 'Ec23_ClTrans.txt' : 0.358307
'Ec20_ClTrans.txt' : 0.315798 'Sc04_ClTrans.txt' : 0.310075 'AUT22_ClTrans.txt' : 0.30959 'AUT03_ClTrans.txt' : 0.299726
'Sc67_ClTrans.txt' : 0.295734 'AUT07_ClTrans.txt' : 0.289559 'Int05_ClTrans.txt' : 0.283884 'AUT20_ClTrans.txt' : 0.277949
'AUT12_ClTrans.txt' : 0.275205 'AUT06_ClTrans.txt' : 0.274209 'Ec29_ClTrans.txt' : 0.273643 'Sc52_ClTrans.txt' : 0.213876
'HM02_ClTrans.txt' : 0.204994

Cluster-10:
'Rel15_ClTrans.txt' : 0.487238 'Rel10_ClTrans.txt' : 0.483445 'Rel16_ClTrans.txt' : 0.450558 'Rel14_ClTrans.txt' : 0.446074
'Rel18_ClTrans.txt' : 0.424142 'Int18_ClTrans.txt' : 0.409857 'AUT38_ClTrans.txt' : 0.322838 'Soc26_ClTrans.txt' : 0.303669
'Ec03_ClTrans.txt' : 0.272514 'ec07_ClTrans.txt' : 0.268449 'HM17_ClTrans.txt' : 0.254623 'AUT48_ClTrans.txt' : 0.253372
'Rel13_ClTrans.txt' : 0.212612 'Entr02_ClTrans.txt' : 0.198735 'AUT34_ClTrans.txt' : 0.198245 'CHD18_ClTrans.txt' : 0.196288
'To50_ClTrans.txt' : 0.194787 'To35_ClTrans.txt' : 0.193215 'Rel04_ClTrans.txt' : 0.189928 'Soc02_ClTrans.txt' : 0.186655
'CHD16_ClTrans.txt' : 0.177266 'AUT02_ClTrans.txt' : 0.167184 'Soc06_ClTrans.txt' : 0.166021 'To03_ClTrans.txt' : 0.158321
'To25_ClTrans.txt' : 0.158091 'CHD11_ClTrans.txt' : 0.149068

Cluster-11
'To42_ClTrans.txt' : 0.695065 'To34_ClTrans.txt' : 0.653615 'To23_ClTrans.txt' : 0.574877 'To13_ClTrans.txt' : 0.537594
'Ec14_ClTrans.txt' : 0.530584 'To11_ClTrans.txt' : 0.5301 'To38_ClTrans.txt' : 0.518562 'To29_ClTrans.txt' : 0.487116 'To28_ClTrans.txt'
: 0.453932 'To40_ClTrans.txt' : 0.449099 'To49_ClTrans.txt' : 0.432396 'To26_ClTrans.txt' : 0.40631 'To30_ClTrans.txt' : 0.399271
'Ec01_ClTrans.txt' : 0.373264 'To05_ClTrans.txt' : 0.372899 'Ec12_ClTrans.txt' : 0.367938 'Ec26_ClTrans.txt' : 0.367617
'To04_ClTrans.txt' : 0.354383 'To54_ClTrans.txt' : 0.350987 'To39_ClTrans.txt' : 0.329491 'To31_ClTrans.txt' : 0.32381
'Ec25_ClTrans.txt' : 0.312806 'To10_ClTrans.txt' : 0.30808 'Ec17_ClTrans.txt' : 0.304066 'To56_ClTrans.txt' : 0.302201
'To07_ClTrans.txt' : 0.293501 'To12_ClTrans.txt' : 0.293138 'To22_ClTrans.txt' : 0.284248 'Ec11_ClTrans.txt' : 0.272316
'To57_ClTrans.txt' : 0.271899 'Ec22_ClTrans.txt' : 0.271451 'Int15_ClTrans.txt' : 0.269126 'To48_ClTrans.txt' : 0.264525
'Ec10_ClTrans.txt' : 0.246824 'To20_ClTrans.txt' : 0.232799 'Ec05_ClTrans.txt' : 0.226006 'To27_ClTrans.txt' : 0.221028
'To21_ClTrans.txt' : 0.213132 'To53_ClTrans.txt' : 0.207388 'Ec13_ClTrans.txt' : 0.202437 'To24_ClTrans.txt' : 0.199938
'To41_ClTrans.txt' : 0.197153 'To01_ClTrans.txt' : 0.196029 'Ec04_ClTrans.txt' : 0.192351 'Ec02_ClTrans.txt' : 0.186806
'Ec21_ClTrans.txt' : 0.170841 'Ec06_ClTrans.txt' : 0.165696 'To45_ClTrans.txt' : 0.162051 'To33_ClTrans.txt' : 0.158438
'Soc17_ClTrans.txt' : 0.156228 'To17_ClTrans.txt' : 0.120618 'To09_ClTrans.txt' : 0.111452 'To36_ClTrans.txt' : 0.0973765

Cluster-12:
'Pol06_ClTrans.txt' : 0.528757 'Edu03_ClTrans.txt' : 0.514137 'Edu07_ClTrans.txt' : 0.509605 'Pol07_ClTrans.txt' : 0.42994
'Edu04_ClTrans.txt' : 0.42421 'Pol09_ClTrans.txt' : 0.384811 'Edu06_ClTrans.txt' : 0.383472 'Pol10_ClTrans.txt' : 0.377957
'Int21_ClTrans.txt' : 0.371795 'HM16_ClTrans.txt' : 0.247911 'HM25_ClTrans.txt' : 0.232399

Cluster-13:
'Rec04_ClTrans.txt' : 0.500117 'Rec05_ClTrans.txt' : 0.498232 'Rec03_ClTrans.txt' : 0.481939 'Rec06_ClTrans.txt' : 0.466543
'Rec07_ClTrans.txt' : 0.461571 'Rec08_ClTrans.txt' : 0.417251 'Rec02_ClTrans.txt' : 0.412686 'Rec01_ClTrans.txt' : 0.341082
'CHD25_ClTrans.txt' : 0.289464 'CHD26_ClTrans.txt' : 0.285966 'CHD27_ClTrans.txt' : 0.275352 'CHD22_ClTrans.txt' : 0.23956
'S05_ClTrans.txt' : 0.235484 'CHD21_ClTrans.txt' : 0.223579 'S07_ClTrans.txt' : 0.216524 'S08_ClTrans.txt' : 0.213927
'CHD23_ClTrans.txt' : 0.211677 'CHD24_ClTrans.txt' : 0.206472 'S03_ClTrans.txt' : 0.199574 'S01_ClTrans.txt' : 0.191842
'S04_ClTrans.txt' : 0.189801 'CHD20_ClTrans.txt' : 0.181458 'Sc06_ClTrans.txt' : 0.180814 'Sc66_ClTrans.txt' : 0.177209
'To60_ClTrans.txt' : 0.174035 'Int04_ClTrans.txt' : 0.146417 'CHD12_ClTrans.txt' : 0.14519 'Ec09_ClTrans.txt' : 0.143793
'Sc03_ClTrans.txt' : 0.138647 'HM01_ClTrans.txt' : 0.13834 'S18_ClTrans.txt' : 0.137851 'Edu10_ClTrans.txt' : 0.13695
'CHD05_ClTrans.txt' : 0.134731 'HM22_ClTrans.txt' : 0.133555

Cluster-14:
'Soc04_ClTrans.txt' : 0.457185 'Pol03_ClTrans.txt' : 0.38107 'Pol08_ClTrans.txt' : 0.3564 'Edu08_ClTrans.txt' : 0.319917
'Pol04_ClTrans.txt' : 0.303237 'Edu12_ClTrans.txt' : 0.299192 'Int11_ClTrans.txt' : 0.287673 'Edu02_ClTrans.txt' : 0.284841
'AUT52_ClTrans.txt' : 0.273395 'Pol02_ClTrans.txt' : 0.265614 'Edu01_ClTrans.txt' : 0.260353 'Sc20_ClTrans.txt' : 0.259116
'AUT05_ClTrans.txt' : 0.248312 'Soc12_ClTrans.txt' : 0.237035 'To44_ClTrans.txt' : 0.228312 'Ec08_ClTrans.txt' : 0.225186
'AUT30_ClTrans.txt' : 0.223664 'To37_ClTrans.txt' : 0.217173 'HM10_ClTrans.txt' : 0.198868 'CHD09_ClTrans.txt' : 0.190198

3.3. Forme stemmée

Cluster-0:
'Sc65_Stem.txt' : 0.433327 'Sc59_Stem.txt' : 0.406784 'Sc63_Stem.txt' : 0.404782 'Sc66_Stem.txt' : 0.361625 'Pol09_Stem.txt' : 0.352251 'Sc67_Stem.txt' : 0.349286 'Sc64_Stem.txt' : 0.342161 'Sc62_Stem.txt' : 0.342012 'SC60_Stem.txt' : 0.303955 'Sc69_Stem.txt' : 0.295966 'Pol04_Stem.txt' : 0.293252 'Sc25_Stem.txt' : 0.288832 'Int01_Stem.txt' : 0.238215 'Soc05_Stem.txt' : 0.235504 'AUT54_Stem.txt' : 0.230443 'AUT67_Stem.txt' : 0.227365 'S13_Stem.txt' : 0.220052 'S11_Stem.txt' : 0.211672 'S10_Stem.txt' : 0.197248 'HM17_Stem.txt' : 0.193548 'AUT13_Stem.txt' : 0.191008 'HM28_Stem.txt' : 0.187852 'CHD26_Stem.txt' : 0.187511 'S06_Stem.txt' : 0.183308 'Spo05_Stem.txt' : 0.182172 'Int22_Stem.txt' : 0.177924 'AUT69_Stem.txt' : 0.170853 'CHD14_Stem.txt' : 0.163708 'Soc13_Stem.txt' : 0.156142

Cluster-1:
'CHD20_Stem.txt' : 0.457316 'CHD19_Stem.txt' : 0.457316 'CHD02_Stem.txt' : 0.355541 'S17_Stem.txt' : 0.325098 'S23_Stem.txt' : 0.324702 'Soc29_Stem.txt' : 0.276152 'S20_Stem.txt' : 0.275357 'S22_Stem.txt' : 0.265568 'S08_Stem.txt' : 0.262511 'CHD08_Stem.txt' : 0.250244 'S09_Stem.txt' : 0.248558 'CHD24_Stem.txt' : 0.244306 'Soc02_Stem.txt' : 0.242914 'CHD17_Stem.txt' : 0.241592 'Edu08_Stem.txt' : 0.235645 'AUT40_Stem.txt' : 0.234917 'AUT42_Stem.txt' : 0.226077

Cluster-2:
'Pol03_Stem.txt' : 0.509932 'Pol07_Stem.txt' : 0.502281 'Pol06_Stem.txt' : 0.441894 'Soc08_Stem.txt' : 0.419356 'Pol08_Stem.txt' : 0.416144 'Pol10_Stem.txt' : 0.405169 'Int10_Stem.txt' : 0.38527 'Edu07_Stem.txt' : 0.345917 'AUT20_Stem.txt' : 0.339752 'Pol02_Stem.txt' : 0.339344 'Soc27_Stem.txt' : 0.336478 'AUT09_Stem.txt' : 0.336044 'Edu03_Stem.txt' : 0.333305 'Edu04_Stem.txt' : 0.324287 'Int06_Stem.txt' : 0.29486 'AUT12_Stem.txt' : 0.287089 'Pol05_Stem.txt' : 0.285568 'AUT41_Stem.txt' : 0.280767 'Int21_Stem.txt' : 0.259433 'Edu05_Stem.txt' : 0.241854 'AUT16_Stem.txt' : 0.238043 'AUT03_Stem.txt' : 0.23179 'AUT02_Stem.txt' : 0.228965 'Spo04_Stem.txt' : 0.220427 'Spo01_Stem.txt' : 0.220051 'AUT43_Stem.txt' : 0.203583 'Edu02_Stem.txt' : 0.202248 'Spo06_Stem.txt' : 0.199824 'AUT47_Stem.txt' : 0.187886 'AUT05_Stem.txt' : 0.180782 'To61_Stem.txt' : 0.173568 'CHD27_Stem.txt' : 0.130978 'CHD11_Stem.txt' : 0.124624 'CHD13_Stem.txt' : 0.117839 'CHD15_Stem.txt' : 0.11262

Cluster-3:
'To14_Stem.txt' : 0.506586 'To48_Stem.txt' : 0.486825 'To09_Stem.txt' : 0.476583 'To44_Stem.txt' : 0.441096 'To52_Stem.txt' : 0.427445 'To32_Stem.txt' : 0.425039 'To06_Stem.txt' : 0.415277 'To18_Stem.txt' : 0.400152 'To55_Stem.txt' : 0.396046 'To04_Stem.txt' : 0.394334 'To16_Stem.txt' : 0.390391 'To15_Stem.txt' : 0.364338 'To47_Stem.txt' : 0.363534 'To19_Stem.txt' : 0.35486 'To46_Stem.txt' : 0.318147 'To43_Stem.txt' : 0.285896 'S01_Stem.txt' : 0.284271 'To33_Stem.txt' : 0.245077 'To25_Stem.txt' : 0.24342 'Spo03_Stem.txt' : 0.199749 'To22_Stem.txt' : 0.185968 'Spo02_Stem.txt' : 0.18189 'S12_Stem.txt' : 0.156903 'To24_Stem.txt' : 0.151069 'Sc04_Stem.txt' : 0.126923

Cluster-4:
'HM13_Stem.txt' : 0.504036 'Sc53_Stem.txt' : 0.393766 'Sc07_Stem.txt' : 0.39362 'HM14_Stem.txt' : 0.366243 'Ec18_Stem.txt' : 0.354624 'HM08_Stem.txt' : 0.342953 'HM10_Stem.txt' : 0.319198 'HM31_Stem.txt' : 0.313853 'HM15_Stem.txt' : 0.30655 'HM18_Stem.txt' : 0.299513 'HM29_Stem.txt' : 0.290041 'HM16_Stem.txt' : 0.288497 'HM01_Stem.txt' : 0.285407 'HM32_Stem.txt' : 0.282566 'HM22_Stem.txt' : 0.274117 'Sc09_Stem.txt' : 0.271487 'Rel03_Stem.txt' : 0.270045 'HM07_Stem.txt' : 0.268633 'HM04_Stem.txt' : 0.255127 'Sc26_Stem.txt' : 0.253956 'HM02_Stem.txt' : 0.251091 'HM12_Stem.txt' : 0.248854 'Sc56_Stem.txt' : 0.245802 'HM23_Stem.txt' : 0.2378 'HM24_Stem.txt' : 0.236511 'HM09_Stem.txt' : 0.228141 'Edu10_Stem.txt' : 0.219261 'Sc33_Stem.txt' : 0.218037 'HM26_Stem.txt' : 0.21657 'HM06_Stem.txt' : 0.194049 'HM21_Stem.txt' : 0.188952 'S07_Stem.txt' : 0.149484 'CHD09_Stem.txt' : 0.115126

Cluster-5:
'Rel15_Stem.txt' : 0.49865 'Int18_Stem.txt' : 0.485737 'Int11_Stem.txt' : 0.392728 'Rel14_Stem.txt' : 0.387864 'Rel16_Stem.txt' : 0.368644 'Rel18_Stem.txt' : 0.355674 'Rel11_Stem.txt' : 0.348294 'AUT46_Stem.txt' : 0.343895 'Rel10_Stem.txt' : 0.335342 'Int12_Stem.txt' : 0.33381 'Soc09_Stem.txt' : 0.329397 'Rel17_Stem.txt' : 0.328716 'Int16_Stem.txt' : 0.325759 'Rel07_Stem.txt' : 0.321769 'Int19_Stem.txt' : 0.320964 'AUT52_Stem.txt' : 0.314287 'Rel06_Stem.txt' : 0.313882 'AUT10_Stem.txt' : 0.306239 'AUT07_Stem.txt' : 0.305055 'AUT24_Stem.txt' : 0.297583 'Rel09_Stem.txt' : 0.284166 'Soc21_Stem.txt' : 0.283276 'Rel08_Stem.txt' : 0.281506 'Int04_Stem.txt' : 0.279346 'AUT21_Stem.txt' : 0.278533 'Rel02_Stem.txt' : 0.270629 'Soc26_Stem.txt' : 0.270186 'AUT38_Stem.txt' : 0.268238 'AUT48_Stem.txt' : 0.26724 'Rel12_Stem.txt' : 0.264981 'Soc07_Stem.txt' : 0.264649 'AUT49_Stem.txt' : 0.259535 'Rel01_Stem.txt' : 0.256231 'Soc15_Stem.txt' : 0.254067 'Rel05_Stem.txt' : 0.246445 'Soc18_Stem.txt' : 0.244075 'AUT23_Stem.txt' : 0.243672 'AUT60_Stem.txt' : 0.243223 'Edu11_Stem.txt' : 0.243178 'AUT59_Stem.txt' : 0.242891 'AUT22_Stem.txt' : 0.241003 'Soc04_Stem.txt' : 0.239807 'AUT31_Stem.txt' : 0.239456 'AUT27_Stem.txt' : 0.2342 'Rel19_Stem.txt' : 0.233398 'AUT45_Stem.txt' : 0.228364 'AUT25_Stem.txt' : 0.222712 'AUT51_Stem.txt' : 0.221033 'AUT29_Stem.txt' : 0.213368 'AUT39_Stem.txt' : 0.210111 'Soc19_Stem.txt' : 0.2027 'AUT66_Stem.txt' : 0.195001 'Soc23_Stem.txt' : 0.190765 'Edu09_Stem.txt' : 0.190079 'Soc28_Stem.txt' : 0.188718 'Rel04_Stem.txt' : 0.184796 'To60_Stem.txt' : 0.184027 'AUT04_Stem.txt' : 0.182875 'AUT26_Stem.txt' : 0.182395 'AUT30_Stem.txt' : 0.177351 'Rel13_Stem.txt' : 0.175508 'AUT11_Stem.txt' : 0.169196 'Int23_Stem.txt' : 0.168915 'AUT73_Stem.txt' : 0.164604 'Soc22_Stem.txt' : 0.163748 'AUT08_Stem.txt' : 0.160693 'AUT33_Stem.txt' : 0.156157 'Edu06_Stem.txt' : 0.151812 'AUT06_Stem.txt' : 0.14973 'Soc25_Stem.txt' : 0.148327 'Edu12_Stem.txt' : 0.148087 'AUT34_Stem.txt' : 0.143168 'Soc12_Stem.txt' : 0.138733 'AUT64_Stem.txt' : 0.133722 'Soc10_Stem.txt' : 0.131089 'S16_Stem.txt' : 0.127306 'CHD18_Stem.txt' : 0.123515 'Entr02_Stem.txt' : 0.112899 'S02_Stem.txt' : 0.0986412

Cluster-6:
'ec07_Stem.txt' : 0.578968 'Ec03_Stem.txt' : 0.52511 'To36_Stem.txt' : 0.358796 'To03_Stem.txt' : 0.352127 'Sc02_Stem.txt' : 0.347263 'AUT50_Stem.txt' : 0.346281 'Ec08_Stem.txt' : 0.330775 'Soc30_Stem.txt' : 0.330413 'CHD01_Stem.txt' : 0.317991

Cluster-7:
'Int02_Stem.txt' : 0.450077 'AUT65_Stem.txt' : 0.417304 'AUT53_Stem.txt' : 0.404894 'AUT17_Stem.txt' : 0.391767 'AUT71_Stem.txt' : 0.385002 'AUT68_Stem.txt' : 0.344842 'Int17_Stem.txt' : 0.324023 'Pol01_Stem.txt' : 0.309047 'Int05_Stem.txt' : 0.307363 'AUT70_Stem.txt' : 0.300948 'Int13_Stem.txt' : 0.270486 'CHD03_Stem.txt' : 0.245366 'CHD05_Stem.txt' : 0.238056 'HM27_Stem.txt' : 0.228076 'AUT44_Stem.txt' : 0.220325 'CHD04_Stem.txt' : 0.203447

Cluster-8:
'Ec21_Stem.txt' : 0.512869 'Ec15_Stem.txt' : 0.463106 'Sc57_Stem.txt' : 0.422458 'Sc42_Stem.txt' : 0.422422 'SC61_Stem.txt' : 0.410031 'Ec20_Stem.txt' : 0.3953 'Sc27_Stem.txt' : 0.387395 'Sc30_Stem.txt' : 0.387041 'Ec11_Stem.txt' : 0.381323 'Sc70_Stem.txt' : 0.373845 'Sc21_Stem.txt' : 0.3681 'Sc31_Stem.txt' : 0.351045 'Sc44_Stem.txt' : 0.339063 'Sc41_Stem.txt' : 0.326609 'Sc40_Stem.txt' : 0.299119 'Sc51_Stem.txt' : 0.298374 'Sc46_Stem.txt' : 0.281024 'Sc43_Stem.txt' : 0.280933 'Soc14_Stem.txt' : 0.280321 'Sc68_Stem.txt' : 0.270112 'Sc49_Stem.txt' : 0.263979 'Sc55_Stem.txt' : 0.216762 'HM03_Stem.txt' : 0.184053 'Ec04_Stem.txt' : 0.176286 'Int14_Stem.txt' : 0.15581 'Edu01_Stem.txt' : 0.147919 'HM25_Stem.txt' : 0.139965 'To08_Stem.txt' : 0.13003

Cluster-9:
'Rec04_Stem.txt' : 0.778953 'Rec05_Stem.txt' : 0.746446 'Rec06_Stem.txt' : 0.736692 'Rec03_Stem.txt' : 0.727367 'Rec07_Stem.txt' : 0.719876 'Rec02_Stem.txt' : 0.683566 'Rec01_Stem.txt' : 0.625853 'Rec08_Stem.txt' : 0.621752 'Soc01_Stem.txt' : 0.355378 'Rec09_Stem.txt' : 0.306531 'Sc37_Stem.txt' : 0.184621 'HM30_Stem.txt' : 0.159632 'CHD21_Stem.txt' : 0.154796 'S24_Stem.txt' : 0.151853

Cluster-10:
'To58_Stem.txt' : 0.390396 'Soc20_Stem.txt' : 0.390396 'Sc20_Stem.txt' : 0.377711 'Sc23_Stem.txt' : 0.354895 'Sc48_Stem.txt' : 0.34343 'Sc28_Stem.txt' : 0.308168 'Sc38_Stem.txt' : 0.288898 'Sc22_Stem.txt' : 0.285012 'Sc32_Stem.txt' : 0.280754 'AUT57_Stem.txt' : 0.278224 'CHD25_Stem.txt' : 0.27635 'S31_Stem.txt' : 0.272889 'Sc08_Stem.txt' : 0.271092 'Sc10_Stem.txt' : 0.255178 'Sc03_Stem.txt' : 0.249521 'HM05_Stem.txt' : 0.243224 'Sc24_Stem.txt' : 0.243178 'Sc05_Stem.txt' : 0.240295 'Ec16_Stem.txt' : 0.231787 'AUT36_Stem.txt' : 0.230532 'CHD23_Stem.txt' : 0.221816 'Sc50_Stem.txt' : 0.210168 'Soc24_Stem.txt' : 0.19536 'HM11_Stem.txt' : 0.190744 'CHD22_Stem.txt' : 0.189474 'Sc47_Stem.txt' : 0.167035

Cluster-11:
'Soc16_Stem.txt' : 0.590164 'Int20_Stem.txt' : 0.509011 'AUT63_Stem.txt' : 0.486533 'Soc03_Stem.txt' : 0.480382 'AUT56_Stem.txt' : 0.462257 'Int03_Stem.txt' : 0.454687 'Ec29_Stem.txt' : 0.37105 'AUT72_Stem.txt' : 0.357567 'AUT14_Stem.txt' : 0.330733 'AUT62_Stem.txt' : 0.317615 'To37_Stem.txt' : 0.261777

Cluster-12:
'Ec01_Stem.txt' : 0.547001 'Ec26_Stem.txt' : 0.507177 'Ec12_Stem.txt' : 0.46825 'Ec10_Stem.txt' : 0.448309 'Ec13_Stem.txt' : 0.441022 'Ec22_Stem.txt' : 0.431898 'Ec05_Stem.txt' : 0.424466 'Ec25_Stem.txt' : 0.397981 'Ec27_Stem.txt' : 0.388303 'Ec24_Stem.txt' : 0.361357 'Ec06_Stem.txt' : 0.323541 'Ec28_Stem.txt' : 0.322355 'Ec09_Stem.txt' : 0.314546 'To27_Stem.txt' : 0.304111 'To56_Stem.txt' : 0.295158 'Ec19_Stem.txt' : 0.285725 'Ec02_Stem.txt' : 0.274169 'Sc52_Stem.txt' : 0.239282 'Ec23_Stem.txt' : 0.2342 'To02_Stem.txt' : 0.199158 'To50_Stem.txt' : 0.156476 'AUT32_Stem.txt' : 0.149882 'AUT28_Stem.txt' : 0.147307 'AUT37_Stem.txt' : 0.146059 'CHD12_Stem.txt' : 0.145362 'Sc45_Stem.txt' : 0.137112 'CHD06_Stem.txt' : 0.135826 'CHD07_Stem.txt' : 0.132858

Cluster-13:
'Soc06_Stem.txt' : 0.414124 'AUT19_Stem.txt' : 0.408137 'AUT55_Stem.txt' : 0.381374 'Sc01_Stem.txt' : 0.334722 'Sc58_Stem.txt' : 0.28736 'Sc54_Stem.txt' : 0.286214 'AUT61_Stem.txt' : 0.285792 'AUT58_Stem.txt' : 0.25353 'S21_Stem.txt' : 0.247773 'Sc06_Stem.txt' : 0.246395 'S14_Stem.txt' : 0.245041 'HM20_Stem.txt' : 0.242349 'S05_Stem.txt' : 0.241772 'S19_Stem.txt' : 0.237979 'S04_Stem.txt' : 0.232446 'S18_Stem.txt' : 0.219683 'AUT35_Stem.txt' : 0.216755 'CHD16_Stem.txt' : 0.211702 'To35_Stem.txt' : 0.206182 'HM19_Stem.txt' : 0.197082

Cluster-14:
'To34_Stem.txt' : 0.811782 'To42_Stem.txt' : 0.794599 'To23_Stem.txt' : 0.7785 'To28_Stem.txt' : 0.717196 'To11_Stem.txt' : 0.707892 'To13_Stem.txt' : 0.657382 'To29_Stem.txt' : 0.64853 'To49_Stem.txt' : 0.637668 'To26_Stem.txt' : 0.612155 'To05_Stem.txt' : 0.59451 'To31_Stem.txt' : 0.591564 'To38_Stem.txt' : 0.589864 'To40_Stem.txt' : 0.584689 'Ec17_Stem.txt' : 0.541487 'To01_Stem.txt' : 0.520314 'Ec14_Stem.txt' : 0.469405 'To41_Stem.txt' : 0.394689 'To39_Stem.txt' : 0.377818 'To30_Stem.txt' : 0.372858 'To54_Stem.txt' : 0.368585 'To53_Stem.txt' : 0.334295 'To20_Stem.txt' : 0.326236 'Int15_Stem.txt' : 0.311436 'To21_Stem.txt' : 0.297239 'To10_Stem.txt' : 0.285181 'S03_Stem.txt' : 0.281557 'Soc17_Stem.txt' : 0.279942 'To12_Stem.txt' : 0.268531 'To57_Stem.txt' : 0.26293 'To07_Stem.txt' : 0.25104 'To45_Stem.txt' : 0.250548 'Soc11_Stem.txt' : 0.217915 'Sc29_Stem.txt' : 0.179809 'To17_Stem.txt' : 0.155369 'To59_Stem.txt' : 0.14856 'S15_Stem.txt' : 0.0703666 .