République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université de Mohamed El Bachir El Ibrahimi de Borj Bou Arréridj
Faculté des Mathématiques et d'Informatique
Département d'informatique

MEMOIRE

Présenté en vue de l'obtention du diplôme

Master en Informatique

Spécialité : Technologie de l'information et de la communication

THEME

La modélisation thématique pour le texte arabe

Présenté par :

BOUHALI Wiam

AMMARA Bachra

Soutenu publiquement le : jj/mm/aaaa

Devant le jury composé de:

Présidente : SAIFI Lynda

Examinateur : NOUIOUA Mourad

Encadrante : MOHDEB Djamila

2021/2022

Dédicace

Au nom de Dieu le Miséricordieux

Je dédie ce travail

À ma famille, qui m'a apporté tout le soutien et l'amour qui ont fait de moi ce que je suis aujourd'hui :

Surtout

A ma mère pour son amour, ses encouragements et son sacrifice

A mon père pour le soutien, l'affection et la confiance qu'il m'a accordé

A mes frères (Ridha, Shams El Dinne et zine El Abidine)
Et mes sœurs (Nassima et Imane)

A tous mes amis qui m'ont toujours encouragé, et à qui je souhaite plus de succès

A tous ceux que j'aime

Bachra

Dédicace

Je dédie ce travail :

A ma chère mère

A mon cher père

Qui n'ont jamais cessé de formuler des prières à mon égard, de me soutenir et de m'épauler pour que je puisse atteindre mes objectifs.

A mes frères Chamseddine et Saloh et Iyad

A mes sœurs Aya et Douâa

A ma tante Sarah

Pour leurs soutiens moral tout au long de mes études, et qui m'ont toujours encouragé, et à qui je souhaite plus de succès.

À tous mes amis de promotion de 2eme année Master TIC

A toute personne qui occupe une place dans mon cœur

Wiam

Remerciements

On remercie dieu le tout puissant de nous avoir donné la santé et la volonté d'entamer et de terminer ce mémoire.

Tout d'abord, ce travail ne serait pas aussi riche et n'aurait pas pu avoir le jour sans l'aide et l'encadrement de madame MOHDEB Djamila, on la remercie pour la qualité de son encadrement exceptionnel, pour sa patience, sa rigueur et sa disponibilité durant notre préparation de ce mémoire.

Nos vifs remerciements vont aux membres de jury pour avoir accepté de juger notre présent travail.

En fin, nous remercions toute personne qui a participé de prés ou de loin à l'accomplissement de ce mémoire.

Résumé

La modélisation thématique est un type d'analyse quantitative non-supervisée qui vise à découvrir des structures sémantiques latentes (thématiques ou sujets) qui apparaissent dans un ensemble de textes non structurés.

Le domaine de la modélisation thématique compte un certain nombre de méthodes et de techniques simples ou avancées qui permettent l'extraction des thématiques nécessaires pour résumer un contenu textuel donné. Cependant, la performance de ces méthodes est contestable dans les langues autres que la langue anglaise ou bien les langues qui sont basées sur les lettres latins.

Dans ce projet, nous avons implémenté une application qui permet d'appliquer les méthodes de modélisation thématique les plus utilisées, sur un jeu de données textuelles en langue arabe. Les résultats obtenus notent la performance du modèle LSA par rapport aux autres modèles concurrents, à savoir LDA et NMF.

Mots-clés : fouille de textes, TALN, modélisation thématique, LSA, LDA, NMF.

Abstract

Topic modeling is a type of unsupervised quantitative analysis that aims to discover latent semantic structures (themes or topics) that appear in a set of unstructured texts.

The field of topic modeling has a number of simple and advanced techniques that allow the extraction of topics that are necessary to summarize a given textual content. However, the performance of these methods is questionable in languages other than English or languages that are based on Latin letters.

In this project, we have implemented an application to exploit the most used topic modeling methods on a textual data set in Arabic. The results note the performance of the LSA model compared to other competing models, namely LDA and NMF.

Keywords: text mining, NLP, topic modeling, LSA, LDA, NMF.

ملخص

نمذجة الموضوع هي نوع من التحليل الكمي غير الخاضع للرقابة الذي يهدف إلى اكتشاف الهياكل الدلالية الكامنة (الموضوعات) التي تظهر في مجموعة من النصوص غير المنظمة.

يحتوي مجال النمذجة الموضوعية على عدد من الأساليب والتقنيات البسيطة أو المتقدمة التي تسمح باستخراج الموضوعات الضرورية لتلخيص محتوى نصي معين. ومع ذلك ، فإن أداء هذه الأساليب مشكوك فيه في لغات أخرى غير الإنجليزية أو تلك اللغات التي تعتمد على غير الحروف اللاتينية.

في هذا المشروع ، قمنا بتنفيذ تطبيق يسمح بتطبيق أساليب النمذجة المواضيعية الأكثر استخدامًا على مجموعة بيانات نصية باللغة العربية. النتائج التي تم الحصول عليها تشير إلى الأداء الجيد لنموذج LSA مقارنة بالنماذج المنافسة الأخرى، وبالتحديد LDA و NMF.

كلمات مفتاحية :التنقيب في النصوص, المعالجة الآلية للغات الطبيعية, النمذجة الموضوعية, LSA, LDA, NMF.

Table des matières

Liste des abréviations

BOW: Bag Of Word

IDF: Inverse Document Frequency

LDA : Latent Dirichlet Allocation

LSA : Latent Semantic Analysis

MT: Modélisation Thématique

NMF: Non-negative Matrix Factorization

PCA: Principal Component Analysis

PLSA: Probabilistic Latent Semantic Analysis

TF: Term Frenquency

Liste des figures

Liste des tableaux

Introduction générale

1. Contexte

Le monde assiste à une révolution rapide et continuée d'Internet et des plates-formes sociales en ligne. Parmi les conséquences de cette révolution la quantité gigantesque d'informations et de contenu textuel qui se créent à chaque seconde de manière qu'ona souvent du mal à en comprendre et à en gérer. Cette question et d'autres ne font que mettre en évidence la nécessité d'organiser et de résumer ces énormes corpus textuels pour faciliter le traitement et la découverte d'informations utiles. Ainsi, plusieurs techniques et applications utilisant les systèmes automatisés et le traitement automatique du langage naturel ont vu le jour dans le but d'analyser les données textuelles et faciliter sa compréhension à l'utilisateur ordinaire.

2. Problématique

La modélisation thématique (*en anglais*, topic modeling) est un type d'analyse quantitative non-supervisée qui vise à extraire des thèmes latents d'une quantité massive de documents textes. Son objectif principal est de rendre les données textuelles produites par les internautes plus interprétables et plus compréhensibles pour les systèmes automatiques.

Le domaine de la modélisation thématique compte un certain nombre de méthodes et de techniques simples ou avancées qui permettent l'extraction des thématiques nécessaires pour résumer un contenu textuel donné. Cependant, la performance de ces méthodes est contestable dans les langues autres que la langue anglaise ou bien les langues qui sont basées sur les lettres latins.

3. Objectif et contribution

Le but de ce projet d'étude est de mettre en œuvre un projet de modélisation thématique pour le contenu en langue arabe qui est souvent marginalisée dans ce domaine en raison de sa complexité morphologique, même s'il s'agit de l'un des langages les plus utilisés sur le web.

4. Structure du rapport

La méthodologie de ce projet est expliquée en détails dans ce mémoire. Ce dernier est composé de quatre chapitres :

- Le premier chapitre apporte des précisions sur la définition de la modélisation thématique, son importance, son processus, ses outils et ses domaines d'application.
- Le deuxième chapitre liste les approches et les techniques de modélisation thématique avec ses points forts et ses points faibles.
- Le troisième chapitre décrit la méthodologie du projet avec sa conception et son architecture algorithmique.
- Le dernier chapitre présente l'environnement matériel et logiciel qui nous a permis de mettre en place puis d'évaluer les techniques de modélisation thématique pour un corpus de textes en langue arabe.

Chapitre 01: Concepts de base

1.1. Introduction

Dans ce chapitre nous présentons un ensemble de concepts de base qui sont reliés au sujet de la modélisation thématique grâce auxquels nous pouvons comprendre tout ce qui sera mentionné dans les chapitres suivants.

D'abord nous définissons la modélisation thématique en se focalisant sur ses objectifs. Ensuite, nous expliquons son processus, ses applications et ses outils.

1.2. Définition de la modélisation thématique

La modélisation thématique est un type des modèles et des méthodes d'exploration de texte qui sont utilisés pour la découverte de structures sémantiques latentes (cachées) appelées « thématiques » (***topics,*** en anglais) qui se produisent dans un corpus des textes non structurés. Ces « thématiques » peuvent être représentées par un seul mot ou un groupe de mots qui résument bien et reflètent le contenu textuel du corpus [1].

1.3. Objectifs et importance de la modélisation thématique

La modélisation thématique (MT) joue un rôle très important dans le domaine d'analyse des textes. Ses principaux objectifs incluent :

- Trouver les formes d'utilisation des mots dans les documents et les textes qui partagent les mêmes structures.
- Modéliser les relations cachées qui régissent les données textuelles non structurées et complexes.
- Présenter de nouvelles approches de modélisation thématique différentes qui sont capables de traiter la corrélation entre les thématiques des données textuelles.
- Comprendre et gérer des textes courts tels que ceux rencontrés sur les réseaux sociaux.

La modélisation thématique peut être appliquée à de nombreux domaines tels que le traitement automatique du langage naturel, la recherche d'informations, la classification et le clustering de texte, l'apprentissage automatique et les systèmes de recommandation [2].

Les méthodes de MT peuvent être supervisées, non supervisées ou semi-supervisées ; peuvent utiliser des données structurées ou non structurées ; et peuvent être appliquées dans plusieurs domaines tels que la santé, l'agriculture, l'éducation, le commerce électronique, l'analyse d'opinion sur les réseaux sociaux. La MT peut être utilisée pour découvrir des sujets abstraits latents dans une collection de textes tels que des documents, des textes courts, des discussions, des publications Twitter et Facebook, des commentaires d'utilisateurs sur les pages web, des blogs et des e-mails [3].

1.4. Processus de la modélisation thématique

La première étape du processus d'exploration de texte consiste à collecter des données textuelles non structurées ou semi-structurées provenant de plusieurs sources de données telles les microblogs, les réseaux sociaux, les documents et les pages Web.

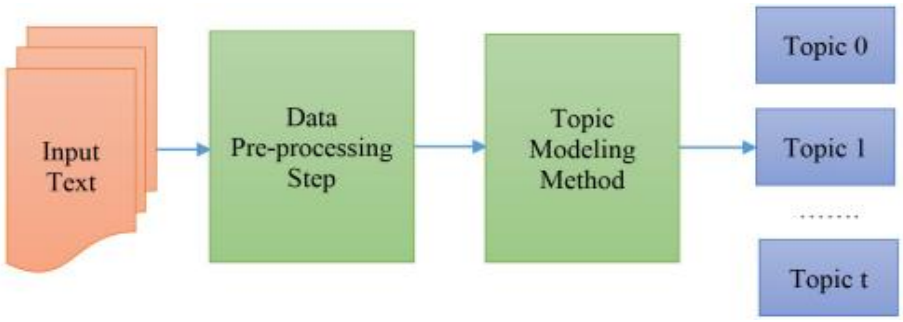

Figure 1. Processus de la modélisation thématique [2]

Par la suite, ces textes collectés vont subir un processus de nettoyage et de prétraitement de données qui permet entre autres la suppression des mots vides, la racinisation, la lemmatisation, la tokenisation, et l'identification des n-grammes :

- L’élimination des mots vides : les mots vides sont les mots les plus courants dans un langage donné qui ne sont généralement pas utiles pour l’objectif de l’extraction du texte, tels que les prépositions, les pronoms, les nombres...etc.
- La racinisation : la conversion des mots dans leur racine, en utilisant des algorithmes de *Stemming*.
- La lemmatisation : est utilisée pour améliorer la précision du modèle en retournant la forme de base ou de dictionnaire d’un mot.
- La segmentation (tokenisation) : diviser un texte en *tokens* ; des paragraphes en phrases des phrases en mots, des mots en lettres ou à d’autres éléments significatifs.
- L’identification des structures n-grammes telles que le bi-gramme (phrases contenant deux mots consécutifs) et trigramme (phrases contenant trois mots consécutifs).
- …etc.

Les données prétraitées doivent être ensuite converties dans un format structuré convenable afin de les utiliser comme des entrées pour les méthodes automatiques de la modélisation thématique.

Le modèle thématique appliqué analyse les entrées et donne comme résultat les motifs (thématiques) visibles ou cachés qui représentent en mieux le contenu textuel étudié.

1.5. Applications de la modélisation thématique

Avec l’émergence des médias sociaux, des modèles thématiques ont été utilisés dans plusieurs domaines. Nous citons entre autres:

La caractérisation et la recommandation de contenu : les sites de microblogging sont utilisés comme plateformes pour la création et la consommation du contenu textuel et multimédia. Une des applications populaires qui sont reliées à ces sites sont les systèmes de recommandation en ligne qui peuvent, par exemple, recommander un travail approprié à des

candidats intéressés en fonction de leurs informations, de leur histoire et sociologie, localisation, et autres contextes [4].

La classification de texte : les modèles thématiques sont utilisés pour l'ingénierie ou l'extraction des caractéristiques qui aident à la classification des textes. Ils peuvent être également utiles pour la réduction de la dimensionnalité du texte afin d'obtenir une faible représentation dimensionnelle qui simplifie la tâche de la classification [5].

L'analyse financière: dans de nombreuses activités commerciales telles que la structuration de la bourse, les modèles thématiques sont exploités pour l'utilisation de la valeur des actions afin d'inciter les sujets à diverses transactions sur le marché organisé et d'autres activités [6].

La bio-informatique: la modélisation thématique aide à consolider plus de connaissances bio-informatiques à travers les études liés aux textes qui sont générés par les dossiers cliniques des patients [7].

L'analyse des réseaux sociaux (SNA): l'analyse des données sociales en ligne est une application importante pour l'exploration et l'extraction des motifs qui redessinent et expliquent les profiles réels des utilisateurs web et des services [8].

1.6. Les outils de la modélisation thématique

Il existe de nombreux outils standards de modélisation thématique pour les tests et l'évaluation, parmi ces outils :

Stanford TMT : Stanford Topic Modeling Toolbox (TMT) est une ressource développée par The Stanford Natural Language Processing Group. TMT est conçu pour aider les chercheurs qui souhaitent analyser de grands documents textuels en offrant la possibilité d'importer et de manipuler des textes, de former des modèles thématiques pour créer des résumés textuels et de générer des sorties compatibles en suivant l'utilisation des mots sur les thématiques, le temps et d'autres groupements de données. TMT a été écrit en 2009-2010 et utilise une ancienne version de Scala [9][10].

VISTopic : est un outil de structure hiérarchique pour l'analyse visuelle qui vise à aider les utilisateurs à donner un sens à de grandes collections de documents en se basant sur de nombreux algorithmes de la modélisation thématique [11].

KEA : est un outil open source pour extraire automatiquement des phrases clés à partir du texte. Kea identifie les phrases clés candidates à l'aide de méthodes lexicales, calcule les valeurs des caractéristiques de chaque candidat texte et utilise l'apprentissage automatique pour prédire quels candidats représentent de bonnes phrases clés. KEA est programmé en Java et disponible sous la licence publique GNU [12].

MALLET : MALLET est un package basé sur Java pour le traitement statistique du langage naturel, la classification de documents, le clustering, la modélisation thématique, l'extraction d'informations et d'autres applications d'apprentissage automatique pour le texte. La boîte à outils de modélisation thématique MALLET contient des implémentations efficaces basées sur l'Allocation de Dirichlet Latente (LDA), l'Allocation Pachinko et de la LDA hiérarchique [13].

FiveFilters : est un outil logiciel gratuit pour obtenir des termes à partir de texte via un service Web. Cet outil créera une liste des termes pertinents d'un texte donné au format JSON [9].

Gensim : est une boîte à outils open source, implémentée en Python, pour la modélisation de sujets, l'indexation de documents et la recherche de similarités dans de grands corpus de données textuelles. Gensim est considéré être plus rapide que les autres outils de modélisation thématique [14].

R : l'environnement logiciel libre R pour le calcul statistique et les graphiques comprend trois packages de modélisation thématique, à savoir MALLET, topic models et LDA [15].

1.7. Conclusion

La modélisation thématique est très importante pour comprendre les données textuelles. Dans ce chapitre nous avons présenté les objectifs et le processus de la modélisation thématique ainsi que quelques pistes de recherche de ses domaines d'applications. Nous aborderons dans le chapitre suivant les modèles thématiques les plus connus et utilisés dans les études qui sont liées aux analyses textuelles.

Chapitre 02: Les techniques de la modélisation thématique

2.1. Introduction

Les quantités sans précédent d'informations textuelles qui circulent sur le web et sur les différentes plates-formes de médias sociaux ainsi que leurs diversités, a créé un besoin urgent à développer des méthodes et des systèmes automatisés pour en comprendre et analyser.

Dans ce chapitre, nous allons identifier les principaux fondements théoriques des modèles thématiques que nous avons appliqués et implémentés dans notre projet de fin d'études.

2.2. Approches de la modélisation thématique

Les algorithmes de modélisation thématique les plus populaires qui ont contribué au domaine de l'analyse de texte dans plusieurs applications incluent LSA (Latent Semantic Analysis), NMF (Non-Negative Matrix Factorization), PLSA (Probabilistic Latent Semantic Analysis), LDA (Latent Dirichlet Allocation) et PCA (Principal Component Analysis).

Ces méthodes tombent sous deux catégories: méthodes probabilistes et méthodes non-probabilistes (voir Figure 1). Les méthodes probabilistes comme PLSA et LDA sont principalement des modèles non supervisées. Ils peuvent être cependant utilisées dans des configurations supervisées ou semi-supervisées. Les approches non probabilistes sont des approches algébriques qui ont vu le jour au début des années 1990 avec le concept d'analyse sémantique latente et de la factorisation matricielle non négative en utilisant des modèles génératifs.

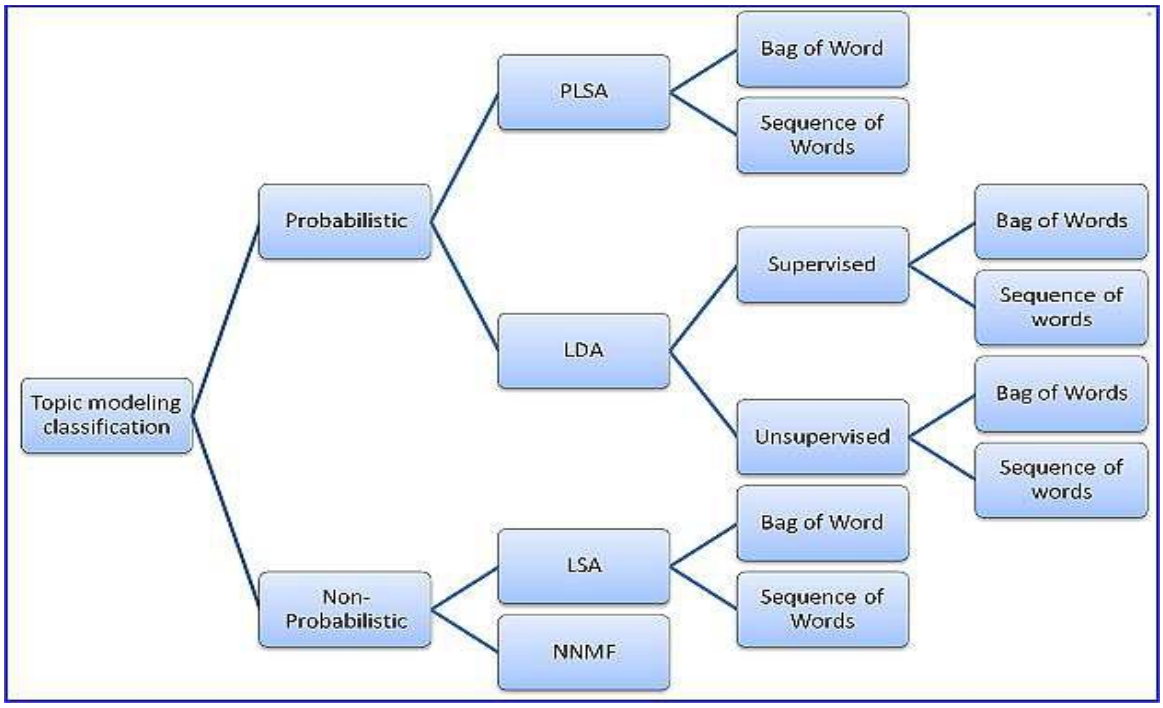

Figure 2. Classification des méthodes de modélisation thématique [16]

2.2.1. LSA: Latent Semantic Analysis

C'est une méthode en sémantique distributionnelle, qui peut être utilisée dans plusieurs domaines, comme la détection de sujets. Elle est devenue une référence de performance pour de nombreuses méthodes avancées. Les hypothèses distributionnelles constituent le fondement théorique de la méthode LSA, qui stipule que les termes ayant une signification similaire sont plus proches en termes d'utilisation contextuelle, en supposant que les mots qui sont proches dans leur signification apparaissent dans les parties liées de textes [2].

La LSA utilise une matrice Termes/Documents qui décrit l'occurrence de certains termes dans les documents à traiter. C'est une matrice creuse dont les lignes correspondent aux « termes » et dont les colonnes correspondent aux « documents ».

La LSA est une technique mathématique/statistique entièrement automatique pour établir des relations entre les termes contenus dans des documents, les documents eux-mêmes et des «

concepts » associés aux termes. En utilisant des opérations matricielles, nous pouvons obtenir ces approximations statistiques en décomposant essentiellement une matrice en trois facteurs[2].

$$A_{m,n} = U_{m,m} S_{m,n} V_{n,m}$$

- ✓ $A_{m,n}$: la matrice d'origine Termes/Documents avec m termes et n documents.
- ✓ $U_{m,m}$: une matrice orthogonale de vecteurs de termes.
- ✓ $S_{m,n}$: une matrice diagonale contenant des valeurs propres.
- ✓ $V_{n,n}$: une matrice orthogonale de vecteurs de documents.

Les avantages de LSA

- Elle résout le problème de rareté des données et capture les synonymes de mots.
- Elle ne nécessite pas une solide base en statistiques ou en théorie des probabilités.
- Elle exploite une structure unique en tant que facteurs.

Les inconvénients de LSA

- La difficulté d'étiqueter un sujet dans certains cas et d'établir un certain nombre des sujets.
- Le nombre de thématiques (sujets) qui doit être déterminé à l'avance, dépend du jugement humain et non pas du calcul statistique.
- La LSA ne capture pas la corrélation entre les thématiques si elles sont nombreuses.

2.2.2. PLSA (Probabilistic Latent Semantic Analysis)

PLSA est une technique de réduction de dimension dans le text mining qui utilise le concept du sac-de-mots (Bag of Words) pour détecter la co-occurrence sémantique de termes dans un corpus en suivant un cadre probabiliste. Elle est basée sur le principe que chaque mot généré à partir d'un seul sujet et un mot différent dans un document peut être généré à partir de différents sujets. PLSA se repose sur un modèle statistique appelé le modèle d'aspect. Le modèle

d'aspect est un modèle de variable latente pour la co-occurrence de données, qui associe des variables de classe non observées à chaque observation. La figure ci-dessous résume le fonctionnement de cette méthode:

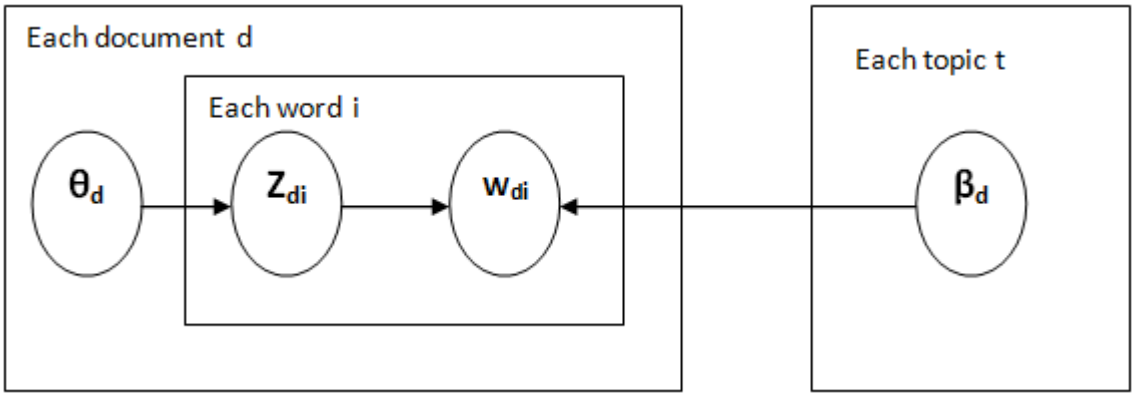

Figure 3. Représentation mathématique du modèle PLSA

Où :

- $\mathbf{W_{di}}$: le i^{ème} terme dans un document d.
- $\mathbf{Z_{di}}$: la thématique de terme i de document d.
- $\boldsymbol{\theta_d}$: la distribution de thématique d'un document d.
- $\boldsymbol{\beta_d}$: la distribution des termes du thématique t.

On a alors la probabilité d'observer l'ensemble du corpus D:

$$\mathbf{P(d)} = \prod \mathbf{d} \prod \mathbf{w_{di}} \sum_{\mathbf{t}} \boldsymbol{\theta_{d,t}} \boldsymbol{\beta_{t,Wdi}}$$

Cela suppose une indépendance conditionnelle étant donné une thématique non observée t:

$$\mathbf{P(w,d)} = \mathbf{P(d)} \sum_{\mathbf{t}} \mathbf{P(w/t)P(t/d)}$$

Les avantages de PLSA

- Elle s'appuie sur un formalisme mathématique solide.

Les inconvénients de PLSA

- La difficulté d'associer un thème à un nouveau document.
- Le nombre de paramètres du modèle est limité par rapport à la taille du vocabulaire qui peut être traité.

2.2.3. LDA : Latent Dirichlet Allocation

La LDA est un modèle probabiliste qui est considéré comme l'algorithme de modélisation thématique le plus populaire dans des applications réelles pour extraire des sujets à partir d'un corpus de documents car il fournit des résultats précis et peut être pré-entrainé en préalable.

L'allocation latente de Dirichlet (LDA) classifie le texte dans un document et les mots par sujet (thématique), ceux-ci sont modélisés sur la base des distributions et des processus de Dirichlet.

La LDA repose sur deux hypothèses clés :

- ✓ Les documents sont un mélange de sujets.
- ✓ Les sujets sont un mélange de mots.

Les sujets génèrent les mots en utilisant la distribution de probabilité. En langage statistique, les documents sont la densité (ou distribution) de probabilité des sujets et les sujets sont la densité (ou distribution) de probabilité des mots [2]. La figure ci-dessous illustre le fonctionnement de la LDA.

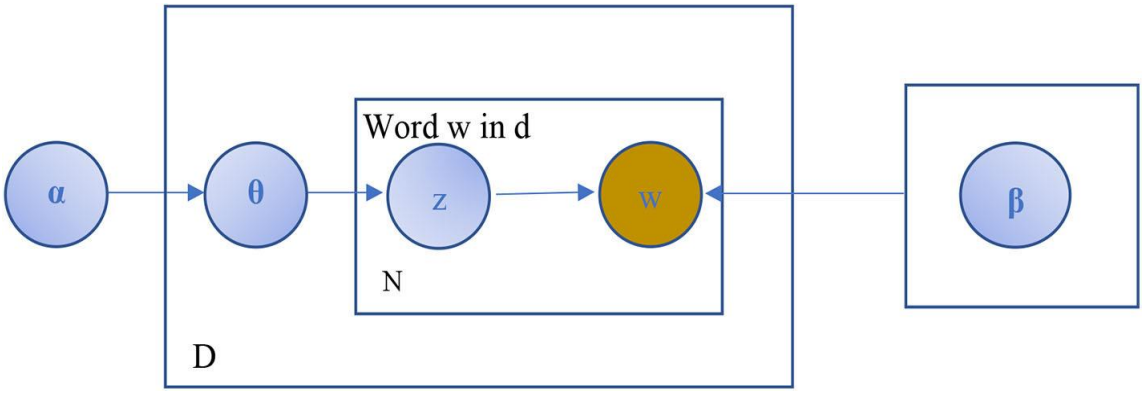

Figure 4. L'allocation latente de Dirichlet

Où :

- ✧ β: Le Dirichlet pour la distribution de mots.
- ✧ θ: Matrice T×D (Termes/Documents) contenant les distributions de thématiques spécifiques aux documents.
- ✧ Z: Un sujet pour un mot choisi dans un document.

- ✧ W: Fait référence à un mot spécifique dans N.
- ✧ D: La longueur des documents.
- ✧ N: Le nombre de mots dans le document.

Les avantages de la LDA

- Elle ne nécessite aucune donnée d'entraînement préalable.
- Elle fournit plus de données interprétables sémantiquement et fonctionne bien en présence de contrainte de temps.
- Elle peut travailler sur de longs documents et les documents de longueur mixte.
- Elle est capable d'améliorer les relations transitives entre les sujets et d'obtenir un ordre élevé de co-occurrence dans de petits documents comme dans des paragraphes et des phrases.

Les inconvénients de la LDA

- Elle nécessite l'agrégation de messages courts pour éviter la rareté des données dans les documents de petite longueur.
- Elle est incapable de modéliser les relations entre les sujets qui aident à comprendre la structure en profondeur de documents.
- Elle nécessite un nombre prédéfini de sujets (T). Si T est trop petit, les sujets sont plus généraux. Si T est trop grand, les sujets se chevaucheront l'un avec l'autre.

2.2.4. NNMF : Non-Negative Matrix Factorization

NNMF est un modèle qui vise à obtenir des sujets pour des données des textes courts en utilisant la matrice de corrélation de termes asymétrique factorisant, la matrice Termes/Documents et la représentation matricielle de sac-de-mots d'un corpus de texte.

Dans la figure 5 :

$$D_{mxn} \approx U_{mxk} V_{kxn}$$

Où, une matrice D peut être factorisée en deux matrices U et V, correspondant respectivement à K axes de coordonnées et N points dans un nouvel espace sémantique (chaque point représente un document). Avec la propriété que les trois matrices ont des éléments non négatifs [2].

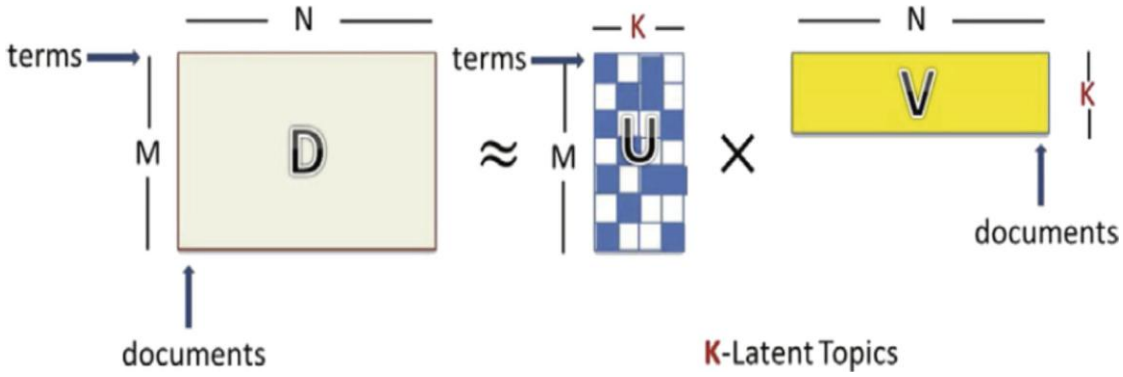

Figure 5. La factorisation non-négative de la matrice

Les avantage de la NNMF

- Le processus NNMF est rapide pour les grandes quantités de données en temps réel.
- NNMF est capable d'extraire des sujets significatifs sans informations ni connaissances préalables de la signification sous-jacente dans les données originales.
- NNMF convient aux tâches de reconnaissance de mots et de vocabulaire.

Les inconvénients de la NNMF

- La NNMF fournit parfois des résultats qui sont sémantiquement incorrects.

2.3. Conclusion

Au cours de ce chapitre, nous avons étudié les aspects les plus importants des méthodes populaires qui sont utilisées dans la modélisation thématique.

Dans le chapitre suivant, nous décrirons les étapes de conception et de mise en oeuvre de notre application de modélisation thématique.

Chapitre 03: Architecture et modélisation

3.1. Introduction

Dans ce chapitre, nous illustrons le processus de la modélisation thématique en l'appliquant à un cas concret qui inclut un corpus des textes en langue arabe. Nous expliquons les étapes du projet, les méthodes et les techniques que nous avons utilisées pour obtenir les résultats souhaités et les évaluer à la fin.

Alors, quelle est la méthodologie utilisée et comment est-elle appliquée?

3.2. Description du projet

Notre projet est une application de la modélisation thématique pour le texte arabe. Son but est de découvrir les thèmes qui apparaissent dans un corpus en analysant les mots des textes originaux. Dans ce qui suit, nous décrivons techniquement les différentes approches de l'analyse thématique, à savoir LSA, LDA et NMF. Nous avons appliqué ces modèles sur un ensemble des données textuelles et publiques qui sont disponibles en ligne pour les recherches qui sont liées au traitement automatique de la langue Arabe.

3.3. Architecture de l'application de modélisation thématique

Dans notre projet, nous avons suivi une série d'étapes pour parvenir à l'application du processus de la modélisation thématique. Ci-dessous, une figure montrant les étapes principales et importantes de notre travail :

a. Collection de données.
b. Nettoyage et prétraitement des données collectées.
c. Traitement des données en appliquant des méthodes de modélisation thématique.
d. Évaluation des résultats obtenus.

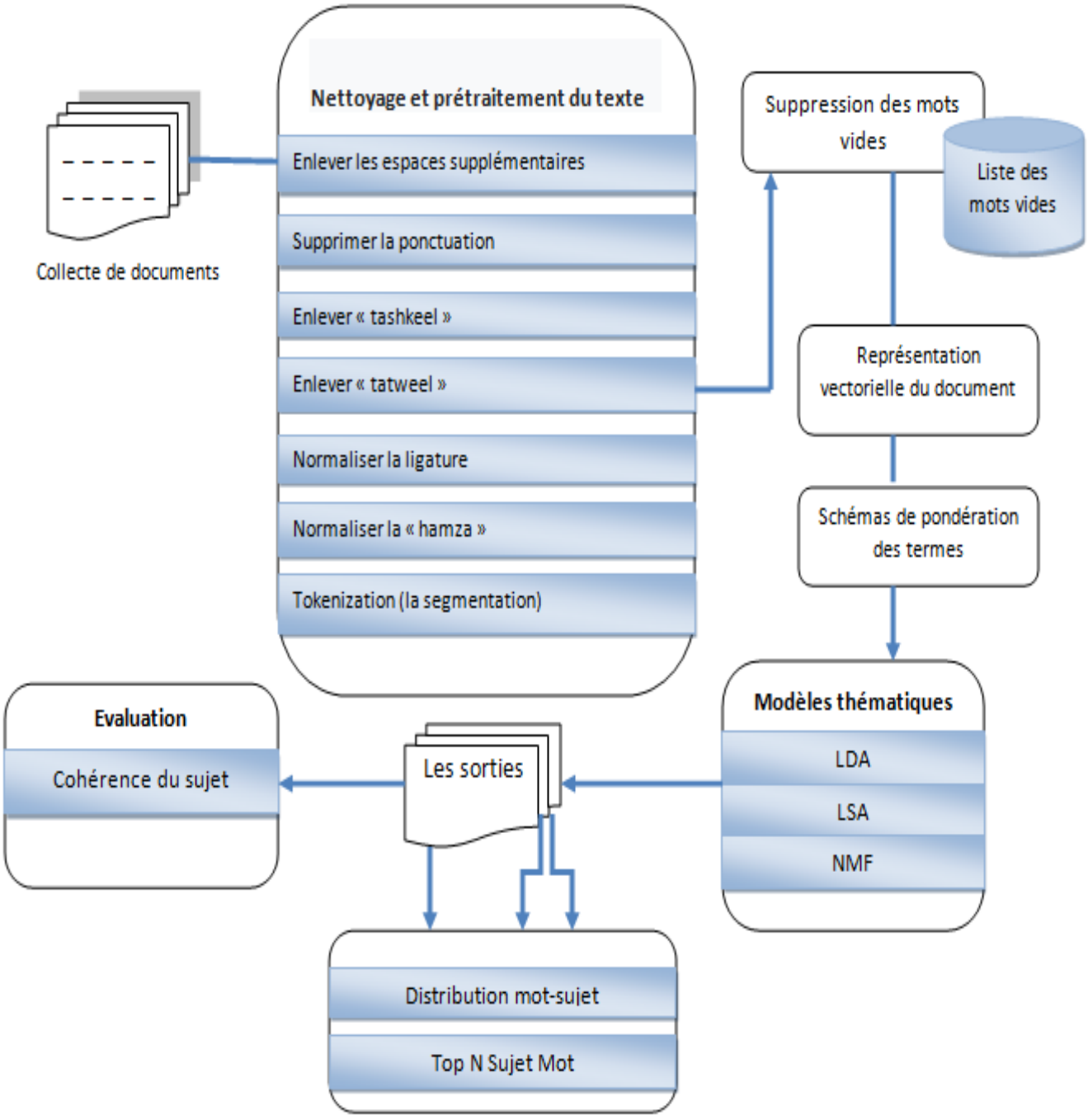

Figure 6. Architecture du système de modélisation thématique

3.4. Algorithme de la modélisation thématique du texte

L'algorithme ci-dessous décrit le processus de prétraitement et de préparation des données pour la modélisation thématique avec une brève explication des étapes.

Algorithme1 : Algorithme de préparation des données

Input : DataSet for Arabic Classification

Output : Texte préparé

```
tf-idf = []
bow = []
Pour texte ϵ DataSet for Arabic Classification faire
    Prétraitement (texte)
    bow += BoW (texte)
    tf-idf += TF-IDF (texte)
Fin
Texte préparé = Algorithme (bow, tf-idf)
Retourner Texte préparé
```

Figure 7. Algorithme de préparation des données

L’algorithme suivant décrit le processus de la modélisation thématique avec ses entrées et ses sorties finales :

Algorithme 2: Algorithme de la modélisation thématique des données

Input : Texte préparé

Output : Thèmes de texte

Thèmes de texte = modèle LDA (texte préparé)

Thèmes de texte =modèle LSA (texte préparé)

Thèmes de texte =modèle NMF (texte préparé)

Évaluer la performance (LDA, LSA, NMF)

Visualiser (Thèmes de texte)

Retourner Thèmes de texte

Figure 8. Algorithme de la modélisation thématique des données

3.5. Conception

3.5.1. Description du corpus de données

La base de données que nous avons utilisée pour notre projet est une collection de textes arabes écrites en langue arabe moderne utilisée dans les articles de journaux. Le texte contient des mots alphabétiques, numériques et symboliques. L’existence de mots numériques et symboliques dans cet ensemble de données pourrait indiquer l’efficacité et la robustesse de nombreux documents de classification et d’indexation de textes arabes.

La base de données se compose de 111,728 documents et 319, 254,124 mots structurés en fichiers texte, et collectés à partir de 3 journaux marocains arabes : Assabah, Hespress et Akhbarona en utilisant un processus d'exploration Web semi-automatique.

Les documents dans la base de données sont classés en 5 classes : sport, politique, culture, économie et diversité. Le nombre de documents et de mots pour chaque classe varie d'une classe à l'autre [17].

3.5.2. Nettoyage et prétraitement de données

Avant de commencer n'importe quelle application de modélisation thématique, les données doivent être pré-traitées pour s'assurer que tous les textes sont passés par les mêmes étapes et qu'ils sont prêts à être exploités par les différentes méthodes de modélisation thématique.

Dans ce qui suit, nous mentionnons les fonctions de prétraitement dont nos textes ont subi:

- Supprimer les espaces supplémentaires.
- Supprimer la ponctuation.
- Supprimer les mots vides : supprimer les mots inutiles d'un texte.
- Supprimer le Tatweel : supprimer l'extension de ligne entre les lettres arabes
- Supprimer les diacritiques : supprimer les voyelles d'un texte.
- La tokenization (ségmentation) : dans lequel les mots sont séparés les uns des autres par des virgules.
- Normaliser la « hamza » : le remplacement de la lettre finale " ة" par "ه " et la lettre finale lettre "ى " avec "ي ".
- Normaliser la ligature : le remplacement de ‘’ ء ’’ par ‘’ ئ ‘’.
- Supprimer les doublons : Supprimer les lignes en double.

3.5.3. Extraction des caractéristiques

Après l'étape de prétraitement, nous avons appliqué deux techniques pour la vectorisation du texte: la TF-IDF et le sac-de-mots (Bag-of-Words).

3.5.3.1. Bag-of-Words(BoW)

Le modèle BoW, en français sac-de-mots, convertit le texte en vecteurs de longueur fixe en comptant le nombre de fois que chaque mot apparaît. L'idée principale consiste à représenter une entité textuelle notée *d* (pour document) sous forme d'un vecteur indexé par les mots (termes) qu'elle contient [18].

$tf\ (t_i, d)$est la fréquence (le nombre d'occurrences) du terme t_i dans le document *d*. La nature du terme *ti* est en général, le résultat des opérations linguistiques lors de la phase de sélection des attributs dans le processus du prétraitement (tokenisation, lematisation, filtrage des mots vides, . . .).

Formellement, cette représentation peut être modélisée comme étant une projection d'un document d dans un espace de haute dimension :

$$\mathbf{d \leftrightarrow ø(d) = (tf\ (t_1, d), tf\ (t_2, d), . . ., tf\ (t_n, d)) \epsilon\ R^n}$$

Dans cette représentation, l'ordre des mots et les informations grammaticales sont ignorés. Ainsi, il est impossible de reconstruire le document original à partir de sa représentation en «sac-de-mots».

3.5.3.2. Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF est une mesure statistique numérique qui est utilisée pour évaluer l'importance d'un mot (terme) dans une collection de documents en se basant sur les occurrences de chaque mot. La TF examine la fréquence d'un mot particulier par rapport au document. La IDF examine à quel point un mot est commun (ou peu commun) dans le corpus. Il est obtenu en divisant le

nombre total de documents par le nombre de documents contenant le terme, puis en prenant le logarithme de ce quotient. Le modèle TF-IDF est défini comme suit [2] [19] :

$\mathbf{TF_{i,j}}$ = nombre d'occurrences de mot dans les documents/nombre de mots dans tous les documents.

$\mathbf{IDF_{i,j}}$= log (nombre de documents / nombre de documents contenant le mot).

$$(\mathbf{TF\_IDF})_{\mathbf{i,j}} = \mathbf{TF_{i,j}} \times \mathbf{IDF_{i,j}}$$

3.5.4. Modélisation thématique

Pour notre projet, nous avons implémenté trois méthodes de modélisation thématique : LDA, LSA et NMF. Les pseudo-codes de ces modèles sont illustrés ci-dessous :

Algorithme 3: Algorithme LDA

Entrée : données d'entraînement X, le nombre de sujets K, hyper paramètres a, B

Sortie : matrice d'affectation de sujet Z, matrice de sujet-document M, matrice de mot-sujet N

Initialiser M, N à zéros

Pour document j ϵ [1, D] faire

  Pour position de jeton i dans document j faire

    $Z_{ij} = k \sim \text{Mult}\left(\frac{1}{K}\right)$

    M_{kj} += 1 ; Nuk_{wk} += 1

  Fin Pour

Fin Pour

Répéter

Pour document j ϵ [1, D] faire

  Pour position de jeton i dans document j faire

    M_{kj} −= 1; N_{wk} −= 1

    $Z_{ij} = k' \sim p\ (Z_{ij} = k \backslash \text{rest}) =$

    $M_{k'j}$ += 1; $N_{wk'}$ += 1

  Fin Pour

Fin Pour

Jusqu'à convergence

Figure 9. Algorithme de LDA [20]

Algorithme 4: Algorithme NMF

Entrée : Matrice non négative V de taille K × N, Paramètre de rang R ϵ N, Seuil ε utilisé comme critère d'arrêt

Sortie : Matrice modèle non négative W de taille K × R, Matrice d'activation non négative H de taille R x N

Procédure : Définir les matrices non négatives $W^{(0)}$ et $H^{(0)}$ par une initialisation aléatoire ou informée. De plus set l = 0. Appliquez les règles de mise à jour suivantes (écrites en notation matricielle) :

(1) $H^{(l+1)} = H^{(l)} \bullet (((w^{(l)})^{T} v) / ((w^{(l)})^{T} w^{(l)} H^{(l)}))$

(2) $W^{(l+1)} = w^{(l)} \bullet ((V (H^{(l=1)})^{T}) / (W^{(l)} H^{(l+1)} (H^{(l+1)})^{T}))$

(3) Augmentez l par un.

Répétez les étapes (1) à (3) jusqu'à $||H^{(l)} - H^{(l-1)}|| \leq \varepsilon$ et $||W^{(l)} - W^{(l-1)}|| \leq \varepsilon$

(Ou jusqu'à ce qu'un autre critère d'arrêt soit rempli).

Enfin, posons $H = H^{(l)}$ et $W = W^{(l)}$.

Figure 10. Algorithme de NMF [20]

Algorithme 5: Algorithme LSA

Entrée : Matrice terme-document T

Sortie: matrice d'encodage/données encodées

```
T=F
K = 1
Tant que (K< nombre maximal d'itérations LSA).
    Nombre de caractéristique sélectionnées Rand de 2 ou 5.
    LSA sélectionne au hasard deux ou cinq caractéristique à partir de T en fonction de la
        valeur du nombre de fonctionnalités sélectionnées.
    Pour chaque caractéristique sélectionnée dans T.
        Si Valeur de caractéristique == 1 (1 signifie que la caractéristique est sélectionnée
            et 0 signifie qu'elle n'est pas sélectionnée)
                Valeur de caractéristique = 0
        Sinon
                Valeur de caractéristique =1
        Fin si
                Calculer la valeur de forme de T
        Si f(T) < f(F)
                F=T
        Fin si
          K = K +1
    Fin pour
Fin tant que
retourner F
```

Figure 11. Algorithme de LSA [20]

3.5.5. Métriques d'évaluation

Dans cette section, nous présentons les principales méthodes d'évaluation appliquées en modélisation thématique qui représentent des mesures appropriées pour juger la qualité d'un modèle sujet donné :

3.5.5.1. La cohérence

La cohérence est une mesure généralement utilisée pour évaluer les modèles d'analyse thématique en mesurant le degré de similarité sémantique des mots dans une thématique. Dans ce cas, les sujets sont représentés par les N premiers mots ayant la probabilité la plus élevée d'appartenir à ce sujet particulier. En bref, le score de cohérence mesure à quel point ces mots sont similaires les uns aux autres [18].

Il existe deux types très utilisés de cohérence :

- **Le score de cohérence UMass**

Il calcule la fréquence à laquelle deux mots, w_i et w_j apparaissent ensemble dans le corpus et il est défini comme [21] :

$$\boldsymbol{C_{UMass}(w_i, w_j) = log \frac{D(w_i, w_j) + 1}{D(w_i)}}$$

Où :

- ✓ D (w_i,w_j) : indique le nombre de fois où les mots w_i et w_j apparaissent ensemble dans les documents.
- ✓ D (w_i) : indique nombre de fois où le mot w_i est apparu seule.

- **Le score de cohérence UCI**

Ce score de cohérence est basé sur des fenêtres glissantes et les informations mutuelles ponctuelles de toutes les paires de mots utilisant les N premiers mots par occurrence. Au lieu de calculer la fréquence d'apparition de deux mots dans le document, on calcule la co-occurrence des mots à l'aide d'une fenêtre glissante. Cela signifie que si une fenêtre glissante a une taille de 10, pour un mot particulier$\boldsymbol{w_i}$, on n'observe que 10 mots avant et après le mot $\boldsymbol{w_i}$.

Par conséquent, si les deux mots $\boldsymbol{w_i}$ et $\boldsymbol{w_j}$ sont apparus dans le document mais qu'ils ne sont pas ensemble dans une fenêtre coulissante, ils ne seront pas pris en considération. De même, comme pour le score UMass, la cohérence UCI entre les mots $\boldsymbol{w_i}$ et $\boldsymbol{w_j}$est définie comme suit [21] :

$$\boldsymbol{C_{UCI}(w_i, w_j) = log \frac{P(w_i, w_j) + 1}{P(w_i) \cdot P(w_j)}}$$

Où :

- ✓ P(w) : est la probabilité de voir le mot w dans la fenêtre glissante.
- ✓ P(w_i,w_j) : est la probabilité d'apparition des motsw_i etw_j ensemble dans la fenêtre glissante.

3.5.5.2. La perplexité

La perplexité est une mesure statistique utilisée pour évaluer la capacité de généralisation d'un modèle de modélisation thématique. D'un point de vue théorique, cette mesure peut être interprétée comme l'inverse de la moyenne géométrique de la vraisemblance sur l'ensemble des mots de la collection.

Après entraînement d'un modèle sur un ensemble d'apprentissage, la perplexité de ce modèle est ensuite mesurée sur un ensemble de test distinct. Une perplexité basse sur l'ensemble de test signifie que le modèle est peu « surpris » par ces nouvelles données, ce qui indique une meilleure capacité du modèle à généraliser à partir des données d'apprentissage [22].

La perplexité est calculée comme suit [23]:

$$\textbf{Perplexity } (\mathbf{D_{test}}) = \exp\left\{ \frac{-\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right\}$$

Où :

- ✓ M : est la taille du corpus de test.
- ✓ N_d : est le nombre de mots.
- ✓ P (w_d) : indique la probabilité que le mot w_d soit généré dans le document.

3.6. Conclusion

Dan ce chapitre nous avons défini les objectifs de notre projet, ainsi que la méthodologie suivie pour son réalisation qui inclut les étapes de prétraitement de l'ensemble de données pour le préparer à la phase de vectorisation des textes puis l'étape de modélisation thématique qui consiste à appliquer les modèles LSA, LDA et NMF sur les données préparées et enfin l'étape de l'évaluation de la qualité de modélisation.

Dans le chapitre suivant, nous montrerons le résultat de notre application avec plus de détails.

Chapitre 04: Implémentation & Résultats

4.1. Introduction

Dans ce chapitre, nous décrivons les différentes étapes réalisées durant l'implémentation de notre application, nous commençons par l'environnement et les outils d'implémentation, puis nous présentons les fonctionnalités de notre application. Ensuite nous présentons les différents techniques de visualisation appliquées sur les sorties de nos modèles et enfin nous discutons les résultats obtenus.

4.2. Environnement et outils d'implémentation

4.2.1. Matériel

Tableau 1**.** Caractéristiques du matériel utilisé

Caractéristiques	**Poste de travail N°01**	**Poste de travail N°02**
PC	TOSHIBA	LENOVO
Système d'exploitation	Windows 10 Professionnel	Windows 10 Professionnel
Processeur	Intel(R) Core(TM) i3 CPU M 330 @ 2.13GHz	Intel(R) Core(TM) i5-2520M CPU @ 2.50GHz 2.50 GHz
RAM	4,00 Go	4,00 Go
Type de système	SE 64 bits	SE 64 bits

4.2.2. Langage de programmation Python

Dans la partie d'implémentation, Nous avons choisi python version 3.9.7 pour implémenter notre système de modélisation thématique. Python est le langage de programmation open source le plus employé par les informaticiens. Ce langage s'est propulsé en tête de la gestion d'infrastructure, d'analyse de données ou dans le domaine du développement de logiciels. En effet, parmi ses qualités, Python permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font. Il a libéré les développeurs des contraintes de formes qui occupaient leur temps avec les langages plus anciens. Ainsi, développer du code avec Python est plus rapide qu'avec d'autres langages [24].

4.2.3. Les principaux packages Python utilisés

Nous avons utilisé plusieurs packages pour réaliser ce projet, parmi lesquels:

- **Gensim :** Gensim est une bibliothèque Python open-source gratuite permettant de représenter des documents sous forme de vecteurs sémantiques, aussi efficacement (au niveau informatique) et sans difficultés (au niveau humain) que possible. Gensim est conçu pour traiter des textes numériques bruts et non structurés à l'aide d'algorithmes d'apprentissage automatique non supervisés [25].
- **Matplotlib :** est une bibliothèque du langage de programmation Python destinée à tracer et à visualiser des données sous formes de graphiques.
- **PyArabic :** une bibliothèque de langue arabe spécifique pour Python. Elle fournit des fonctions de base pour manipuler les lettres et le texte arabes, comme détecter les lettres arabes, les groupes et les caractéristiques de lettres arabes, supprimer les signes diacritiques, etc [26].
- **PyLDAvis:** une bibliothèque Python pour la visualisation interactive des modèles thématiques. PyLDAvis est conçu pour aider les utilisateurs à interpréter les thématiques dans un modèle de sujet qui a été adapté à un corpus de données textuelles. Le package extrait des informations d'un modèle de sujet LDA adapté pour informer une visualisation Web interactive [27].

4.3. Analyse exploratoire de données

4.3.1. Caractéristiques du jeu de données

Les caractéristiques de l'ensemble de données final que nous avons utilisé sont décrites ci-après :

Tableau 2. Caractéristiques de la base de données

Nom de l'ensemble de données	Dataset for arabic classification
Nombres de lignes	111 728
Nombre de colonnes	2
Type de données	Textuelles
Noms des colonnes	(Texte, Target)
Utilisation de la mémoire	295 Mo
Nombre de valeurs nulles	0

4.3.2. Le Word Cloud des données

Word Cloud est une technique pour montrer quels mots sont les plus fréquents parmi le texte donné. Plus le mot est fréquent dans le passage analysé, plus il apparaît en grand taille de police dans l'image générée.

La figure 12 présente les mots les plus fréquents dans notre corpus de données textuelles. Comme la source des données du corpus est la presse marocaine en ligne, les mots «المغربية», « المغرب», « البرنامج » sont plus fréquents que d'autres termes.

Figure 12. Les mots les plus fréquents dans la base de données.

4.4. Nettoyage & Prétraitement

Le prétraitement et le nettoyage des données sont des étapes très importants avant d'utiliser l'ensemble de données et d'y appliquer quoi que ce soit. Notre corpus en particulier a subi les opérations de prétraitement suivantes :

- Supprimer les doublons.
- Supprimer les espaces supplémentaires.
- Supprimer les ponctuations.
- Supprimer les mots vides.
- Supprimer le *Tatweel*.
- Supprimer les diacritiques.
- La tokenization.
- La normalisation de la « hamza ».
- La normalisation de la ligature.

L’exemple suivant montre le texte avant et après le prétraitement :

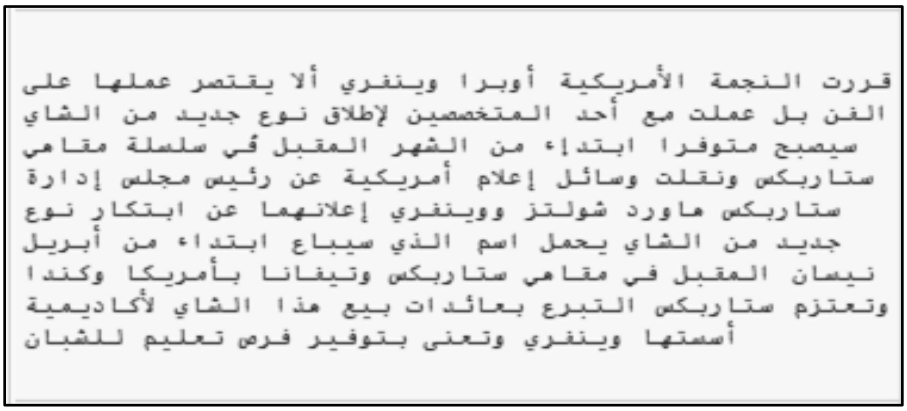
قررت النجمة الأمريكية أوبرا وينفري ألا يقتصر عملها على
الفن بل عملت مع أحد المتخصصين لإطلاق نوع جديد من الشاي
سيصبح متوفرا ابتداء من الشهر المقبل في سلسلة مقاهي
ستاربكس ونقلت وسائل إعلام أمريكية عن رئيس مجلس إدارة
ستاربكس هاورد شولتز ووينفري إعلانهما عن ابتكار نوع
جديد من الشاي يحمل اسم الذي سيباع ابتداء من أبريل
نيسان المقبل في مقاهي ستاربكس وتيفانا بأمريكا وكندا
وتعتزم ستاربكس التبرع بعائدات بيع هذا الشاي لأكاديمية
أسستها وينفري وتعنى بتوفير فرص تعليم للشبان

Figure 13. Le texte avant le prétraitement

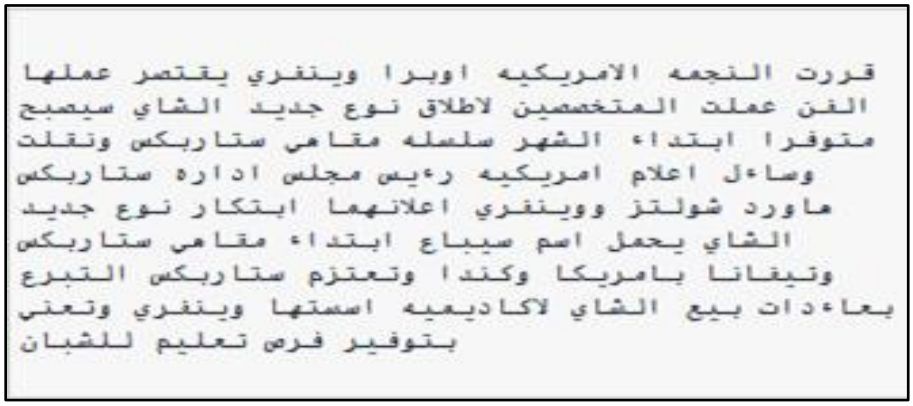
قررت النجمه الامريكيه اوبرا وينفري يقتصر عملها
الفن عملت المتخصصين لاطلاق نوع جديد الشاي سيصبح
متوفرا ابتداء الشهر سلسله مقاهي ستاربكس ونقلت
وسائل اعلام امريكيه رئيس مجلس اداره ستاربكس
هاورد شولتز ووينفري اعلانهما ابتكار نوع جديد
الشاي يحمل اسم سيباع ابتداء مقاهي ستاربكس
وتيفانا بامريكا وكندا وتعتزم ستاربكس التبرع
بعائدات بيع الشاي لاكاديميه اسستها وينفري وتعنى
بتوفير فرص تعليم للشبان

Figure 14. Le texte après le prétraitement.

4.5. La vectorisation du texte avec BoW et TF-IDF

Dans cette étape, nous avons appliqué une vectorisation textuelle basée sur deux approches, la première est TF-IDF et la seconde est BoW que nous l’avons déjà expliquédans le chapitre 3. Les deux exemples suivants de la Figure 15 montrent les vecteurs en sortie d’un texte simple après sa vectorisation à l’aide de TF-IDF et BoW :

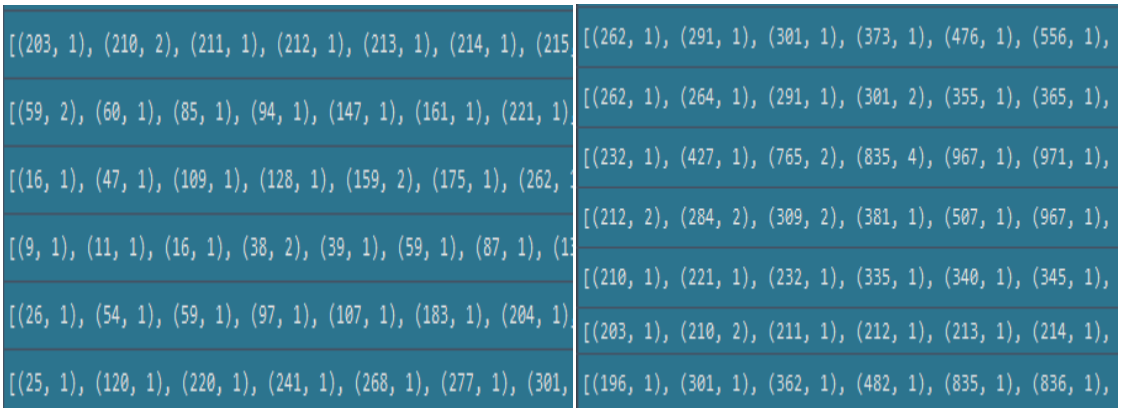

Figure 15. Le texte après la vectorisation TF-IDF (à gauche) **et** après la vectorisation BoW (à droite)

4.6. Implémentation

Les modèles de la modélisation thématique LDA, NMF et LSA de Gensim contient un ensemble de paramètres qui doivent être utilisés afin d'obtenir des résultats fiables. Les principaux paramètres que nous avons définis dans notre application sont :

- ✓ **Corpus (corpus):** le flux de vecteurs de documents ou matrice creuse de documents-termes [28].
- ✓ **Nombre de thématiques (num-topics):**le nombre de sujets latents à extraire du corpus d'apprentissage [28].
- ✓ **Id2word** : le mappage des identifiants de mots aux mots. Il est utilisé pour déterminer la taille du vocabulaire, ainsi que pour le débogage et l'impression des sujets [28].
- ✓ **Passes (passages)**: le nombre de passages est le nombre de fois qu'on doit déterminer pour parcourir ou passer sur la totalité du corpus pendant la phase d'apprentissage [28].

La section suivante montre les résultats obtenus en appliquant les méthodes de modélisation thématiques LDA, NMF et LSA avec num-topics = 04 et passes = 03. Sachant que la méthode LSA du package Gensim n'exige pas un nombre bien déterminé pour les passages. Elle a plutôt le paramètre booléen **onepass** qui doit être fixé à la valeur **False** pour forcer le multi-passage sur le corpus textuel.

4.7. Résultats

4.7.1. Les thématiques présentes dans le corpus

Les tableaux suivants représentent les résultats obtenus après l'application de LDA, LSA et NMF respectivement. Chaque tableau représente quatre thématiques avec leurs mots clés et leurs distributions. On remarque la présence d'un certain nombre de mots dans plus d'une thématique, par exemple les mots " فنان", "المغربية" et "أغنية" sont présents dans la plupart des thématiques. La proximité entre les thématiques résultats qui peuvent être combinées à une seule thématique qui est 'le divertissement', reflète les limitations de l'application des modèles thématique de base sur le texte arabe.

Modèle LDA

Tableau 3. Le résultat obtenu après l'application de LDA

Topic 00		**Topic 01**		**Topic 02**		**Topic 3**	
Mot-clé	**dist**	**Mot-clé**	**dist**	**Mot-clé**	**dist**	**Mot-clé**	**Dist**
المغربية	0,005	برنامج	0,003	المغربية	0,006	المغربية	0,004
الأغنية	0,004	البرنامج	0,003	المغربي	0,003	المغربي	0,003
عبد	0,004	ابتسام	0,002	مجموعة	0,003	المغرب	0,003
الفنان	0,003	محمد	0,002	الفنان	0,003	البرنامج	0,003
برنامج	0,003	المغربي	0,002	محمد	0,003	مجموعة	0,003
أغنية	0,003	المغرب	0,002	إنها	0,003	الجمهور	0,003
الأولى	0,003	تسكت	0,002	الفنية	0,002	القناة	0,002
محمد	0,003	الفنان	0,002	عبد	0,002	أخبارنا	0,002

Modèle LSA.

Tableau 4. Le résultat obtenu après l'application de LSA

Topic 00		Topic 01		Topic 02		Topic 3	
Mot-clé	**dist**	**Mot-clé**	**dist**	**Mot-clé**	**dist**	**Mot-clé**	**Dist**
البرنامج	0,112	أغنية	0,118	البرنامج	0,229	ابتسام	0,289
الفنان	0,095	الأغنية	0,114	الفيلم	0,159	تسكت	0,244
برنامج	0,092	البرنامج	0,107	السينمائي	0,156	المسلسل	0,141
القناة	0,089	الفنان	0,096	المهرجان	0,150	الأغنية	0,127
المغربية	0,088	كليب	0,090	السينما	0,143	أكاديمي	0,105
الأغنية	0,082	القناة	0,080	الأفلام	0,136	الفنانة	0,092
عبد	0,082	الفنانة	0,072	برنامج	0,134	مسلسل	0,091
اغنية	0,080	ابتسام	0,072	القناة	0,131	ستار	0,080

Modèle NMF

Tableau 5. Le résultat obtenu après l'application de NMF.

Topic 00		Topic 01		Topic 02		Topic 3	
Mot-clé	**dist**	**Mot-clé**	**dist**	**Mot-clé**	**dist**	**Mot-clé**	**Dist**
الاعلام	0,002	المسلسل	0,002	الله	0,002	الفنانة	0,002
الأغنية	0,001	عساف	0,001	الفنية	0,002	البرنامج	0,002
القانون	0,001	الفيلم	0,001	المهرجان	0,002	الفيلم	0,002

العمومي	0,001	الساعة	0,001	محمد	0,002	السينما	0,002
عبر	0,001	خالد	0,001	المغربية	0,001	الفنان	0,001
الفني	0,001	دائما	0,001	السينمائي	0,001	المغرب	0,001
مصر	0,001	العربية	0,001	العام	0,001	مهرجان	0,001
دفاتر	0,001	أم	0,001	الغناء	0,001	موازين	0,001

4.7.2. La performance des modèles de modélisation thématique

Le tableau suivant montre les scores de cohérence des modèles thématiques LDA, NMF et LSA :

Tableau 6. Les scores obtenus de cohérence et de perplexité

Modèles	**Cohérence**	**Perplexité**
Modèle LDA	0,2959	-11,3582
Modèle LSA	**0,5499**	/
Modèle NMF	0,2782	/

4.7.3. La performance des modèles par rapport au nombre de thématiques

Les figures ci-dessous illustrent la performance des modèles LDA, LSA et NMF en fonction du nombre de thématiques choisi. Le nombre optimal de thématiques est celui qui donne la valeur de cohérence la plus élevée.

On peut voir que le nombre optimal de thématiques pour LDA est égal à 10 thématiques, et 2 thématiques pour LSA. Concernant NMF, plus le nombre de thématiques est grand, plus le score de cohérence est élevé.

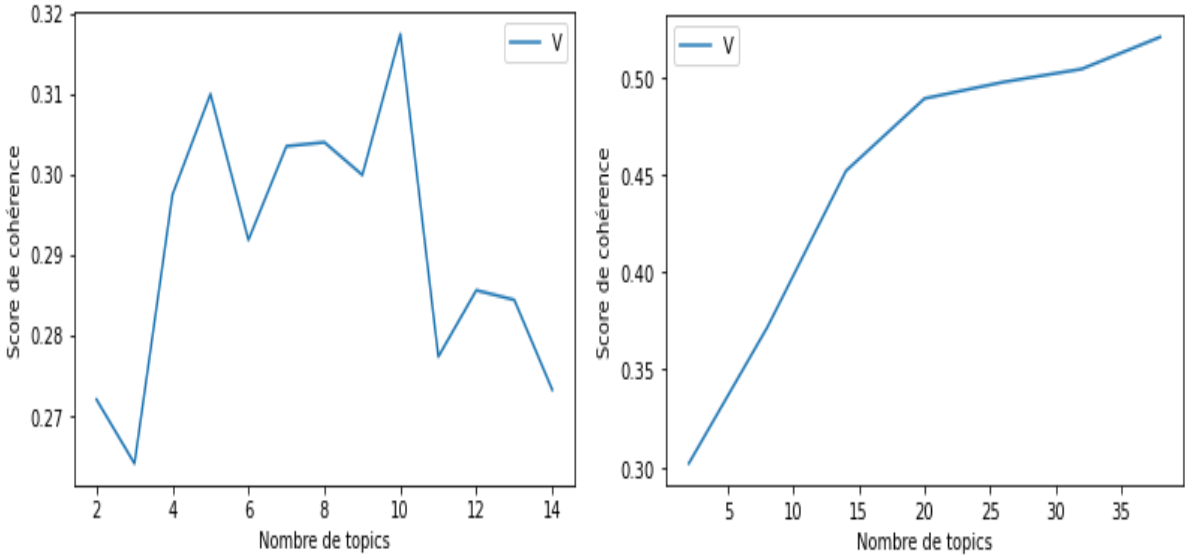

Figure 16. Le nombre optimal de thématique pour LDA (gauche) et pour NMF (droite)

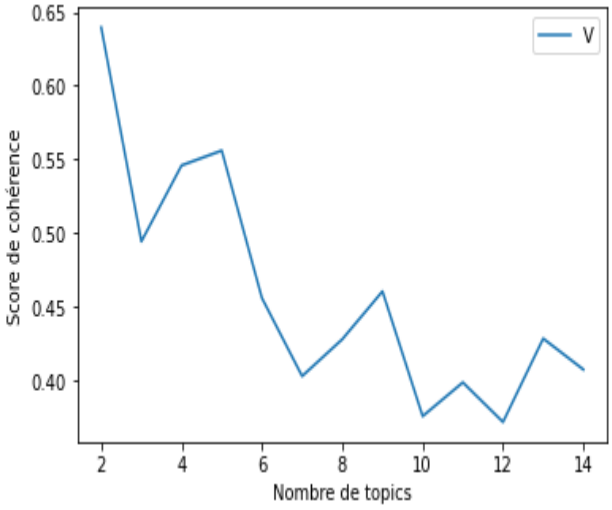

Figure 17. Le nombre optimal de thématique pour LSA

4.7.4. La performance des modèles par rapport au nombre de passages sur le corpus

Nous avons testé les modèles LDA, LSA et NMF avec des valeurs variées du paramètre «passes» pour chercher le nombre optimal de passages sur corpus qui donne un score de cohérence plus élevé.

Les résultats illustrés dans les figures suivantes montrent que, pour LDA et LSA, incrémenter le nombre de passage n'améliore pas vraiment la valeur de la cohérence. Ainsi, tandis que le LSA a généralement besoin d'un seul passage pour une bonne cohérence entre les thématiques extraites (sur la figure 19, la valeur «True» pour le paramètre «**Onepass**» du LSA produit un score de cohérence plus élevé que la valeur «False»), le LDA ne requière qu'environ 03 passages sur corpus pour obtenir son meilleur score de cohérence. Néanmoins, le NMF fait l'exception en aboutissant son meilleur score de performance avec un nombre de passages égal à 20.

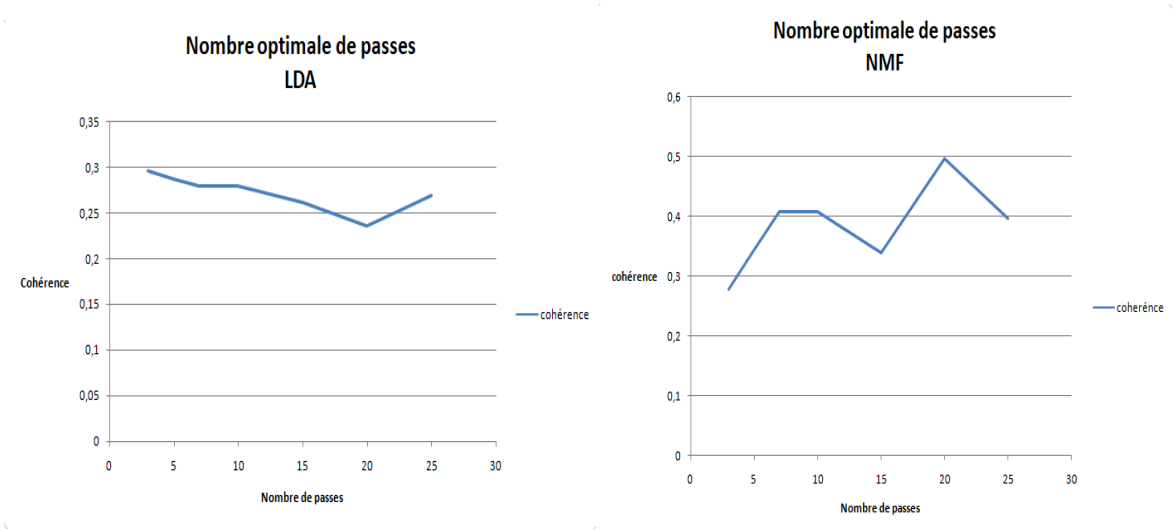

Figure 18. Le nombre optimal de passages pour LDA (gauche) et pour NMF (droite)

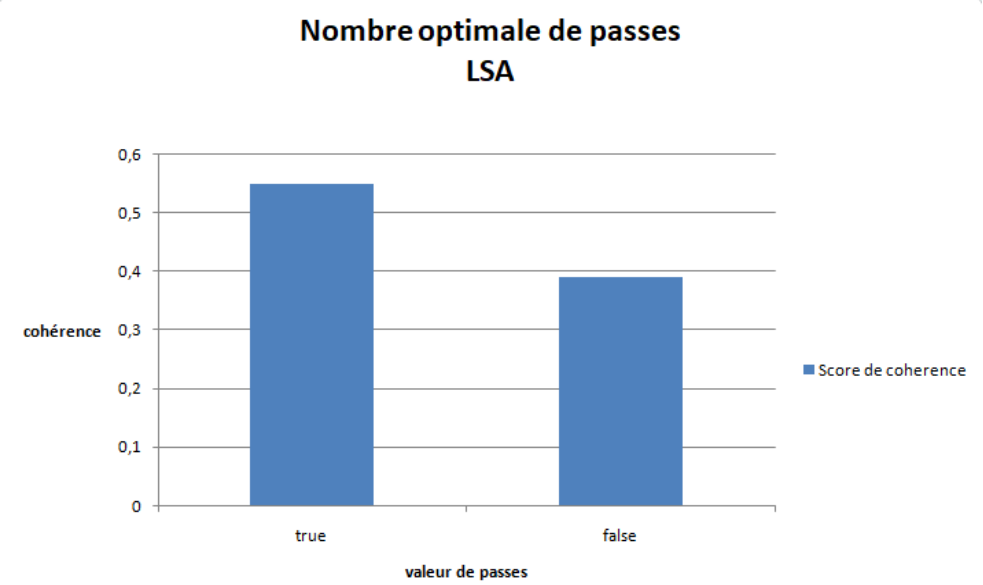

Figure 19. Le nombre optimal de passages pour LSA

4.7.5. La thématique dominante avec sa contribution dans chaque échantillon

Chaque document est composé de plusieurs thématiques. Mais, généralement, un seul sujet parmi ces thématiques est dominant. Le tableau ci-dessous présente la thématique dominante pour chaque document et montre son poids et ses mots-clés.

Tableau 7. La thématique dominante avec sa contribution dans chaque échantillon

Document	Thématique	Poids	Mot-clé	Texte
0	2	0.9969	الفيلم, عبد, مجموعه , البرنامج , المغربي ...	'استوديوهات', 'ورزازات', 'وصحراء', 'مرزوكه' , 'واثار', 'وليلي', 'الرباط', 'والبيضاء', ' انتهى' ...
1	0	0.9873	المغربيه, الفنان, القناه, الاولي, الوطنيه ...	'قررت', 'النجمه', 'الامريكيه', 'اوبرا' , 'وينفري', 'يقتصر', 'عملها', 'الفن', 'عملت', ' لاطلاق ' ...
2	0	0.994	المغربيه, الفنان, القناه, الاولي, الوطنيه ...	'اخبارنا', 'المغربيه','الوزاني', 'تصوير', 'الشملالي', 'الهب', 'النجم', 'المغربي', 'حماس', ' ازيد ' ...
3	2	0.996	الفيلم, عبد, مجموعه, البرنامج, المغربي ...	'تزال', 'صناعه', 'الجلود', 'المغرب', 'تتبع', 'طريقه' , 'التقليديه', 'واليدويه', 'وتستخدم', ' مواد ' ...

4.7.6. Les phrases les plus représentatives pour chaque thématique

Parfois dans l'analyse des textes, on souhaite obtenir des échantillons de phrases qui représentent bien un sujet donné. Le tableau suivant présente la phrase la plus exemplaire pour chaque sujet dans notre corpus de textes arabes.

Tableau 8. Les phrases les plus représentatives pour chaque thématique

Thématique	Poids	Mot-clé	Texte
0	0.9987	المغربيه, الفنان, القناه, الاولي, الوطنيه ...	... ' ناحيه ', 'الابداع', 'والخلق', 'اتوقف', 'يتوقف , ' قلبي', ' اجرت', ' الحوار', ' نورا', ' الفواري'.
1	0.9986	لجمهور, مجموعه, محمد, انه, عبد, الاغنيه ...	... 'العمومي', 'ورجال', 'السياسه', 'بهذه', 'الشروط ' , ' نفسها', ' اجرت', ' الحوار', ' امينه', ' كندي'.
2	0.9992	الفيلم, عبد, مجموعه, البرنامج, المغربي ...	... 'الخلفيه', 'والارضيه', ' الصلبه ' , 'المواجهه' , ' مزاعم', 'وادعاءات', 'خصوم', 'الوحده', ' الترابيه'.
3	0.9991	انه, برنامج, انها, المغرب, عبد, المغاربه ...	... 'موسيقيه', ' تعبيريه', ' لكبار', ' فناني', 'البلد' , 'لتعريف', 'رواد', 'الدوره', 'التوغل', 'اعماقه'.

4.7.7. La distribution de la longueur de chaque échantillon

Lorsqu'on travaille avec un grand nombre de documents, on souhaite connaître la longueur des documents dans leur ensemble et par sujet. La figure 20 trace la distribution des longueurs des documents.

D'après cette figure, le nombre moyen des mots dans les documents est 205, de sorte qu'on a plus de 150 documents qui ont une longueur entre 125 et 250. Nous remarquons que plus le nombre de mots augmente, plus le nombre de documents diminue, car la majorité des documents contiennent moins de 500 mots.

La figure 21 représente la distribution fréquentielle des mots dans chaque échantillon selon le thème dominant. Nous observons que les quatre thématiques (topics) partagent le même nombre optimal de mots dans les documents, qui est environ de 125 mots/documents.

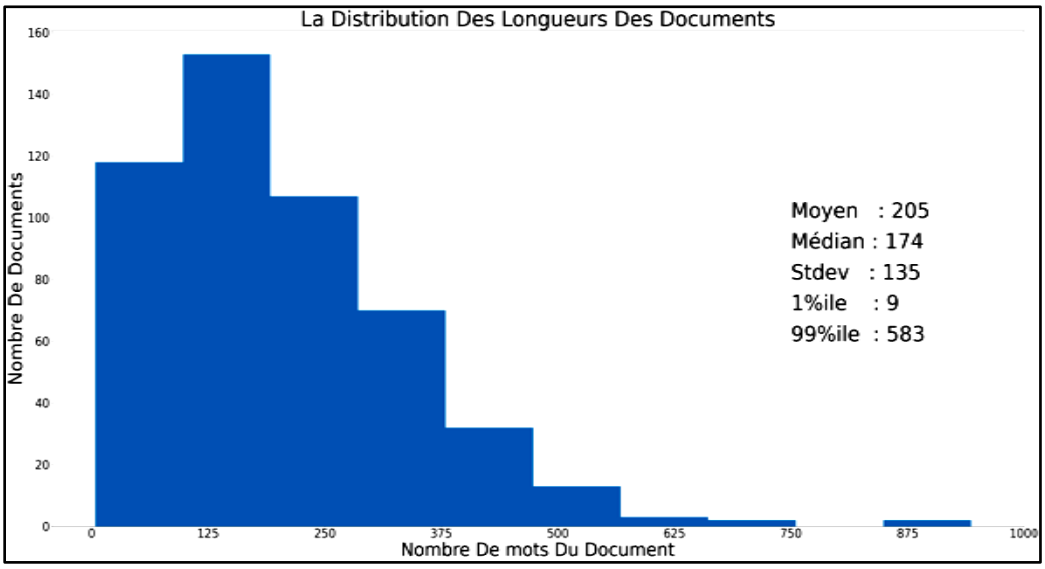

Figure 20. La distribution de fréquence de mots dans chaque document

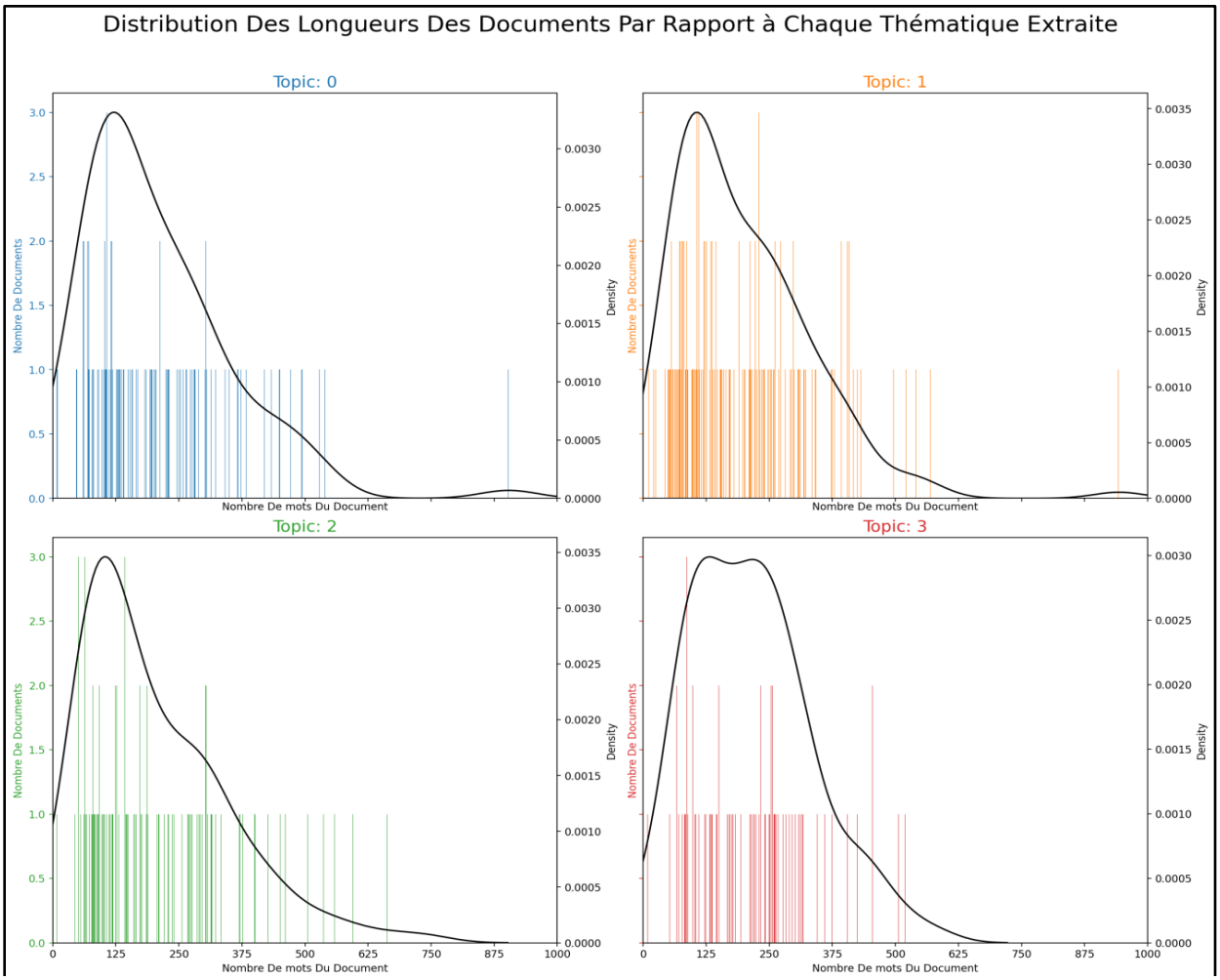

Figure 21. La distribution de fréquence de mots dans chaque document par la thématique dominante

4.7.8. Word Cloud pour les Top N mots dans chaque thématique

La figure 22 montre les mots les plus fréquents dans chaque topic. Les principaux mots sont « «المغربية», «البرنامج», «الأغنية» et «الفنان» dans topic 0, topic 1, topic 2 et topic 3 respectivement.

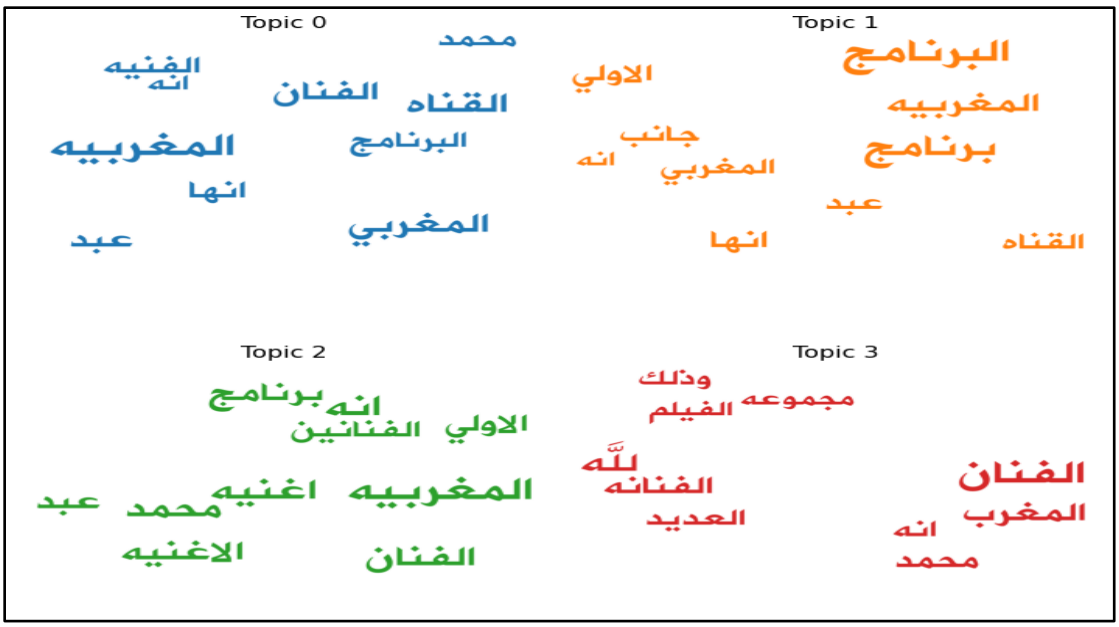

Figure 22. Word Cloud pour les Top N mots dans chaque thématique

4.7.9. La longueur et l'importance des mots-clés dans chaque thématique

La figure ci-dessous représente la longueur de chaque mot-clé et son importance (poids) pour chaque document.

Nous remarquons que pour les quatre thématiques topic 0, topic 1, topic 2 et topic 3, le mot «المغربية» représente le terme le plus important et le plus fréquent dans le corpus.

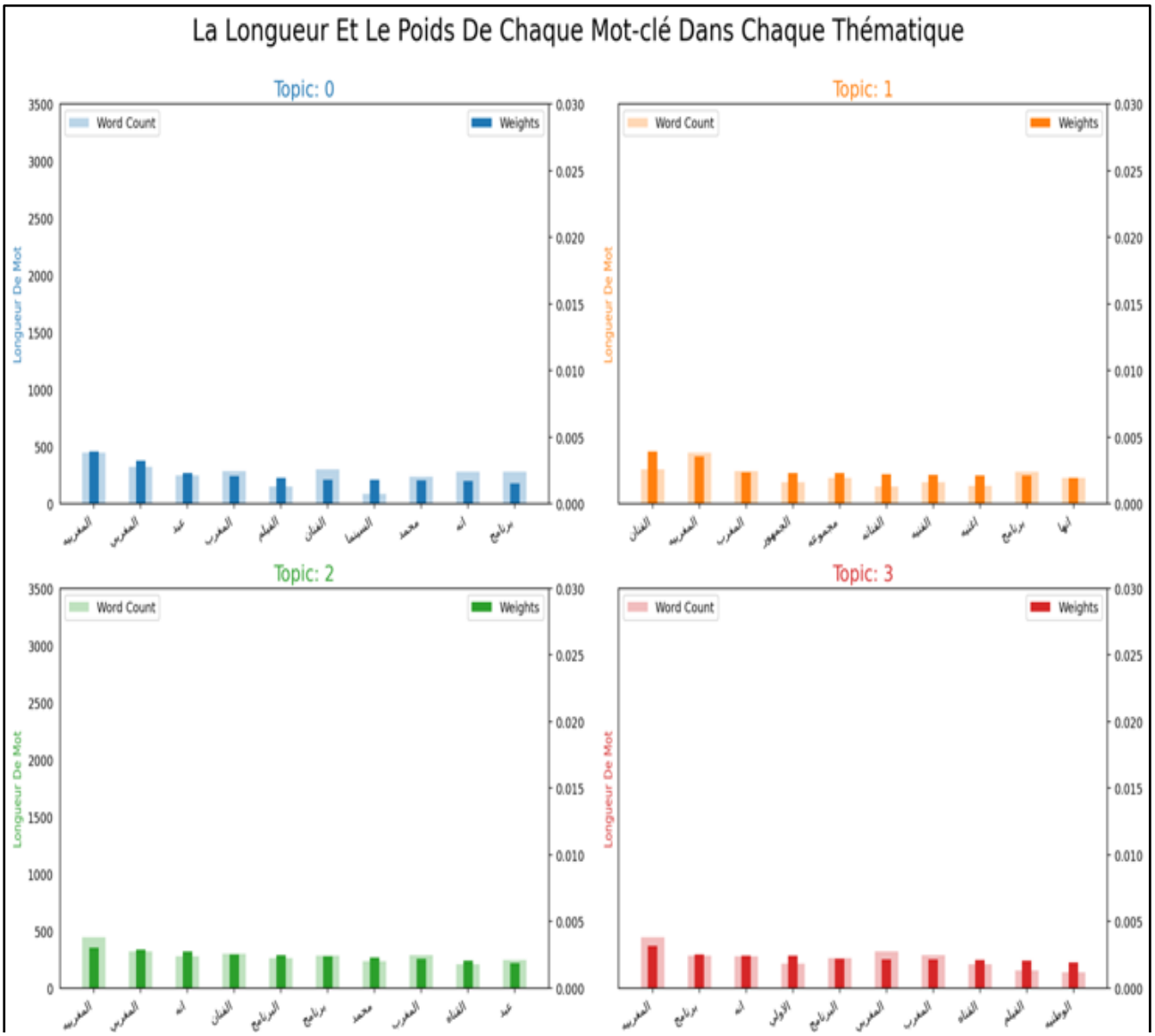

Figure 23. La longueur et le poids de chaque mot-clé dans chaque thématique

4.7.10. La visualisation par PyLDAvis

En utilisant PyLDAvis, on peut visualiser tous les termes que correspondent à un sujet (thématique). La figure suivante (figure 25) par exemple affiche les 30 termes les plus pertinents pour le sujet N°02.Chaque bulle sur le graphique (figure 24) représente une thématique. Si nous

déplaçons le curseur sur l'une des bulles, les mots et les barres seront mis à jour (figure 24). Ces mots sont les mots-clés qui forment la thématique sélectionnée [29].

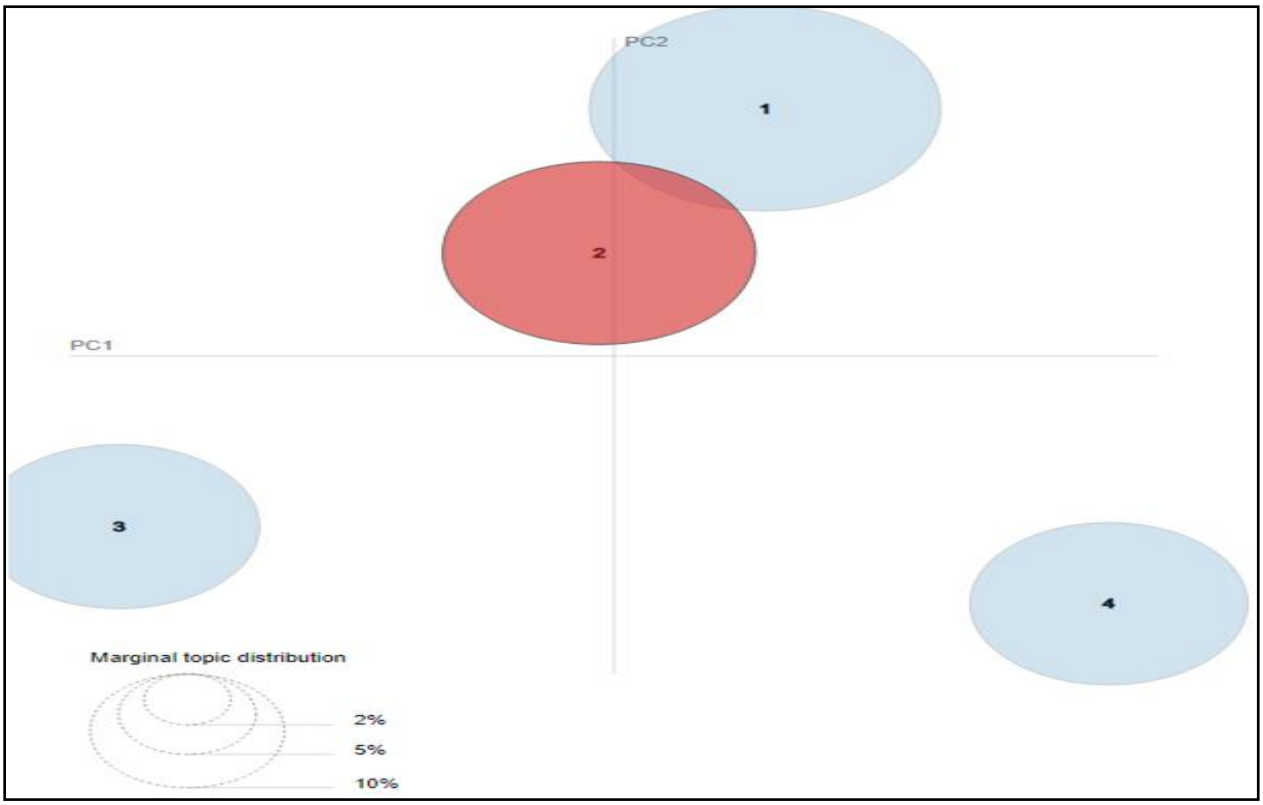

Figure 24. La thématique sélectionnée pour la visualisation

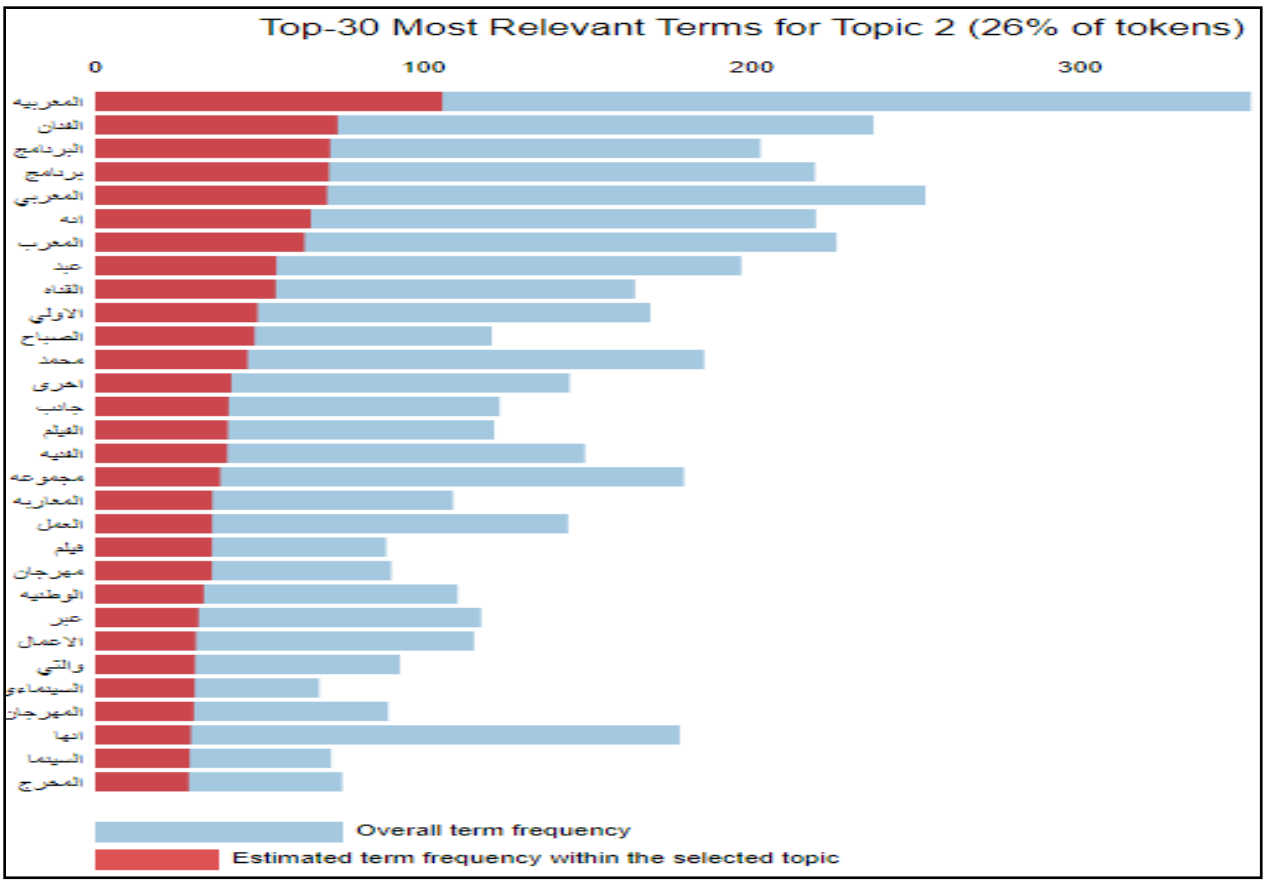

Figure 25. Les 30 termes les plus pertinents pour la thématique sélectionnée

4.8. Discussion des résultats

Nous avons déjà mentionné que la cohérence est la mesure de performance la plus fiable pour évaluer les méthodes de modélisation thématique. Donc, après avoir obtenu des résultats avec cette mesure, nous pouvons décider qu'elle est la meilleure méthode parmi les trois modèles implémentés sachant que plus la valeur de cohérence est élevée, plus la méthode est performante.

D'après les résultats illustrés dans le tableau 6, nous remarquons que la performance du modèle LSA est élevée par rapport aux autres modèles LDA et NMF.

Nous pouvons justifier cela en disant que LDA et NMF fonctionnent mieux avec des scripts courts tels que les titres contrairement à LSA, qui est capable de gérer et de traiter les scripts volumineux (comme les scripts que nous avons utilisés) avec une grande performance.

Enfin, nous concluons que la méthode LSA est le meilleur choix pour appliquer la modélisation thématique des textes arabe.

4.9. Conclusion

Nous avons terminé notre travail dans ce chapitre, dans lequel nous avons précisé les étapes de mise en œuvre et de réalisation du projet de fin d'études en appliquant différents modèles thématiques sur des textes en langue arabe tout en montrant les résultats finaux et leur interprétation afin d'en comprendre le but de cette étude et les performances des modèles appliqués.

Conclusion générale

La modélisation thématique est très importante pour découvrir les sujets qui sont produits dans une collection de données textuelles ainsi pour surpasser les difficultés de tri et de catégorisation de données, de savoir à quel domaine elles appartiennent et à quel sujet elles se rapportent.

Dans ce contexte, nous avons mis en place un système qui permet d'appliquer les méthodes populaires de modélisation thématique sur un jeu de données textuelles. Contrairement à l'utilisation habituelle qui porte sur les langues étrangères comme le français et l'anglais, nous avons appliqué ces techniques aux données textuelles en langue arabe. Ce qui était pour nous un challenge du fait des problèmes techniques et de la difficulté relative qui sont liés au traitement automatique du langage arabe.

Les limitations de ce projet incluent les techniques de visualisation et les métriques d'évaluation qui ne fonctionnent pas avec toutes les modèles. Aussi, nous n'avons pas étudié l'influence des différents paramètres des modèles utilisés, ce qui peut conduire à des résultats contestables de performance.

D'autre part, les documents arabes ont une grande taille du vocabulaire, en particulier pour les documents longs, ce qui entraîne une dimensionnalité élevée qui affecte la vitesse des modèles thématiques.

Pour la suite de ce travail, nous espérons d'améliorer les méthodes de modélisation thématique basiques pour qu'elles puissent réduire la dimensionnalité du vocabulaire arabe, et prennent en charge toutes les limitations qui en résultent.

Enfin, nous espérons qu'une grande partie de nos objectifs ont été atteints et que cette étude sera une simple référence pour tous les intéressés par la modélisation thématique en langue arabe.

Les références

[1]. VAYANSKY, Ike et KUMAR, Sathish AP. A review of topic modeling methods. Information Systems, 2020, vol. 94, p. 101582

[2]. ALBALAWI, Rania, YEAP, Tet Hin, et BENYOUCEF, Morad. Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 2020, vol. 3, p. 42.

[3]. ALGHAMDI, Rubayyi et ALFALQI, Khalid. A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl.(IJACSA), 2015, vol. 6, no 1.

[4]. LEE, Sang Su, CHUNG, Tagyoung, et MCLEOD, Dennis. Dynamic item recommendation by topic modeling for social networks. In : 2011 Eighth International Conference on Information Technology: New Generations. IEEE, 2011. p. 884-889.

[5]. MOHAMMED, Shaymaa H. et AL-AUGBY, Salam. Lsa & lda topic modeling classification: Comparison study on e-books. Indonesian Journal of Electrical Engineering and Computer Science, 2020, vol. 19, no 1, p. 353-362.

[6]. AZIZ, Saqib, DOWLING, Michael, HAMMAMI, Helmi, et al. Machine learning in finance: A topic modeling approach. European Financial Management, 2019.

[7]. LIU, Lin, TANG, Lin, DONG, Wen, et al. An overview of topic modeling and its current applications in bioinformatics. SpringerPlus, 2016, vol. 5, no 1, p. 1-22.

[8]. RAMAGE, Daniel, ROSEN, Evan, CHUANG, Jason, et al. Topic modeling for the social sciences. In : NIPS 2009 workshop on applications for topic models: text and beyond. 2009. p. 1-4.

[9]. BARDE, Bhagyashree Vyankatrao et BAINWAD, Anant Madhavrao. An overview of topic modeling methods and tools. In : 2017 International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2017. p. 745-750.

[10]. SARIOGLU, Efsun, CHOI, Hyeong-Ah, et YADAV, Kabir. Clinical report classification using natural language processing and topic modeling. In : 2012 11th International Conference on Machine Learning and Applications. IEEE, 2012. p. 204-209.

[11]. YANG, Yi, YAO, Quanming, et QU, Huamin. VISTopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. Visual Informatics, 2017, vol. 1, no 1, p. 40-47.

[12]. WITTEN, Ian H., PAYNTER, Gordon W., FRANK, Eibe, et al. Kea: Practical automated keyphrase extraction. In : Design and Usability of Digital Libraries: Case Studies in the Asia Pacific. IGI global, 2005. p. 129-152.

[13]. GRAHAM, Shawn, WEINGART, Scott, et MILLIGAN, Ian. Getting started with topic modeling and MALLET. The Editorial Board of the Programming Historian, 2012.

[14]. AKEF, Islam, ARANGO, Juan S. Munoz, et XU, Xiaowei. Mallet vs GenSim: Topic modeling for 20 news groups report. *Univ. Ark. Little Rock Law J., http://dx. doi. org/10.13140/RG*, 2016, vol. 2, no 19179, p. 39205.

[15]. HORNIK, Kurt et GRÜN, Bettina. topicmodels: An R package for fitting topic models. Journal of statistical software, 2011, vol. 40, no 13, p. 1-30.

[16]. KHERWA, Pooja et BANSAL, Poonam. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, 2020, vol. 7, no 24

[17]. BINIZ, mohamed, "DataSet for Arabic Classification", Mendeley Data, V2, 2018. doi:10.17632/v524p5dhpj.2

[18]. BELLAOUAR, M. M., & GHADA, I. E. Modélisation thématique: cas des publications scientifiques (Doctoral dissertation, جامعة غرداية), 2020.

[19]. GEORGE, Shini. Comparison of LDA and NMF Topic Modeling Techniques for Restaurant Reviews.

[20]. https://www.researchgate.net/, consulté le 18/06/2022.

[21]. Topic modeling coherence score, https://www.baeldung.com/cs/topic-modeling-coherence-score, consulté le 18/06/2022.

[22]. Thibaut Thonet. Modèles thématiques pour la découverte non supervisée de points de vue sur le Web. Web. Université Paul Sabatier - Toulouse III, 2017. Français. ffNNT : 2017TOU30167ff. fftel-01655278v2ff

[23]. Gan J, Qi Y. Selection of the Optimal Number of Topics for LDA Topic Model-Taking Patent Policy Analysis as an Example. Entropy (Basel). 2021 Oct 3;23(10):1301. doi: 10.3390/e23101301. PMID: 34682025; PMCID: PMC8534395.

[24]. Python https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1445304-python-definition-et-utilisation-de-ce-langage-informatique/, consulté le 18/06/2022.

[25]. Gensim https://radimrehurek.com/gensim/intro.html, consulté le 18/06/2022.

[26]. Pyarabic https://pyarabic.sourceforge.io/, consulté le 18/06/2022.

[27]. PyldaVis https://pyldavis.readthedocs.io/en/latest/readme.html, consulté le 18/06/2022.

[28]. LdaModel https://radimrehurek.com/gensim/models/ldamodel.html, consulté le 18/06/2022.

[29]. La visualisation par PyLDAvis https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/#6.-What-is-the-Dominant-topic-and-its-percentage-contribution-in-each-document, consulté le 18/06/2022.