



ECOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE ET D'ANALYSE DES SYSTÈMES

Regroupement des articles par thèmes

Réalisé par :
Mohamed Reda CHENNA
Anass GRINI
Ayoub EZZIDANI

Enseignant :
Mohamed LAZAAR

6 juin 2022



Remerciements :

Nous voudrions tout d'abord adresser toute notre gratitude à notre professeur Mohammed Lazaar, pour sa confiance, sa disponibilité et surtout cette opportunité pour bien maîtriser le processus de traitement des articles texte écrit en arabe afin d'appliquer des modèles de clustering et initier notre carrière par un projet intéressant et relevant du monde réel.

Nous désirons aussi remercier tous ceux qui contribuaient à la réussite de ce projet pour leurs conseils et leurs connaissances qu'ils nous ont partagé.



Résumé

Les données massives (Big data) possèdent un important potentiel scientifique, spécifiquement dans les domaines du Data Mining, Machine Learning et traitement des langues naturelles NLP.

Ce projet consiste en première lieu de collecter les données et le contenu des articles en arabe à partir des sites de presse notamment Aljazeera et Hespress (Data Scraping).

Après la collection des données vient l'étape de regroupement de ces articles en se basant sur le thème de chacun d'eux.

Le travail est de classifier les articles de presse en entrée sous format de document texte, puis donner une probabilité de confiance sur l'appartenance un parmi les themes suivants : *Politique, Culture, Sport, Tamazight, Science - Technologie*

Mots-clé : *NLP, Clustering, Bag of Words*

Table des matières

1	Introduction	1
1.1	Objectif du projet	1
1.2	Problématique soulevée	1
1.3	Hypothèse de solution	2
2	Description de dataset	3
2.1	Scraping	3
2.2	Exploration des données et Visualisation	3
2.2.1	Création d'un dataframe	3
2.2.2	Visualisation	4
2.3	Extraction de caractéristiques	4
3	Traitement du text arabe [NLP]	6
3.0.1	Supprimer les caractères spéciaux	6
3.0.2	Suppression de tous les mots d'arrêt	7
3.0.3	Lemmatisation	8
3.0.4	Stemming	8
3.0.5	Tokenization	8
3.0.6	Words cloud	9
3.0.7	Vectorization	9
4	les méthode de clustering	10
4.1	K-means	10
4.1.1	Algorithme	10
4.2	Clustering spectrale	10
4.2.1	Algorithme de base	10
5	Résultats	11
5.1	Résultats de clustering	11
5.1.1	K-means	11
5.1.2	Clustering spectrale	12
5.1.3	Comparaison	12
5.2	Méthodes de validation	13
5.2.1	K-means	13
5.2.2	Clustering spectrale	14

Chapitre 1

Introduction

1.1 Objectif du projet

Le projet consiste à collecter le contenu des articles en arabes à partir de deux différents sites (*AL Jazeera* et *Hespress*) et faire un regroupement de ces articles de presse en se basant sur leurs thèmes.

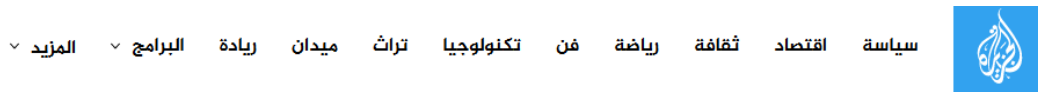


FIGURE 1.1 – Les Thèmes de 'Al Jazeera'

Les étapes de notre travail se présente chronologiquement comme suit :

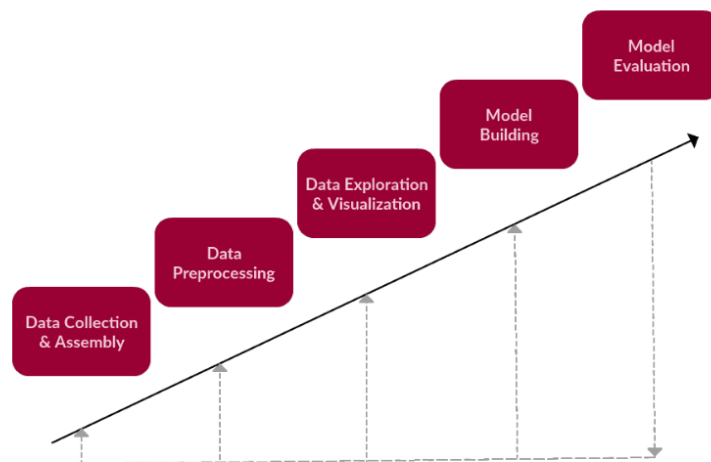


FIGURE 1.2 – Les étapes de travail

Pour ce qui concerne le première point on a besoin des données textes représentant chaque article, malheureusement les sites déjà mentionné n'offrent pas la possibilité de télécharger les articles. Donc on passe au choix d'extraction des données à l'aide des techniques de '*Scrapping*'.

1.2 Problématique soulevée

La problématique soulevée pour ce projet est d'automatiser la tâche de diriger les nouveaux articles écrits en langue arabe vers leur rubrique correspondante, une tâche qui peut être faite manuellement risque de prendre du temps et de faire des erreurs, mais l'automatisation de cette tâche va remédier ce problème.

En deuxième lieu le traitement automatique des fax, le filtrage des courriers électroniques, et bien d'autres applications de fouille de texte arabe sont aussi des problématiques soulevées.

1.3 Hypothèse de solution

La solution consiste en générale a analyser chaque article en le décomposant en mots et en appliquant à ces mots plusieurs traitemtn de langage(NLP :Natural Language Preprocessing) afin de pouvoir prédire sa classe et l'associer à la catégorie convenable.

Chapitre 2

Description de dataset

2.1 Scraping

Afin de récupérer les articles de presse arabes, on a conçu un *Crawler* qui parcourt les articles de <https://www.aljazeera.net/news/> et <https://www.hespress.com> respectivement et extrait le contenu text de ces articles.

On s'est intéressé par Cinq (5) catégories d'articles : *Culture, Economie, Politique, Science et Technologie* et *Tamazight*.

En fin de tâche du *Scraper*, on obtient Cinq (5) dossiers au Cinq catégories respectivement, contenant des fichiers texte (.txt)

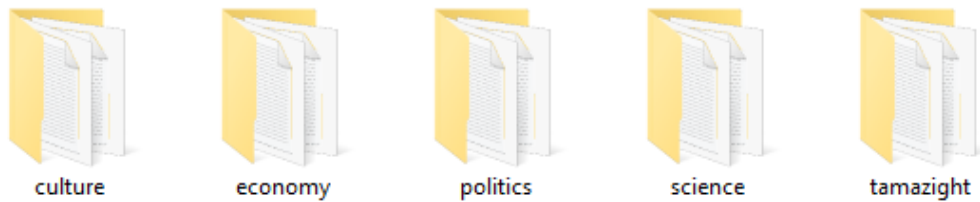


FIGURE 2.1 – Datasets

Chaque dossier contient donc plusieurs fichiers textes contenant chacun un article extrait à partir de la catégorie correspondante des sites.

Vous trouverez l'implémentation du scraper dans le référentiel suivant : <https://github.com/Taylor-X01/News-Categories-Clustering/tree/master/scraper>

2.2 Exploration des données et Visualisation

2.2.1 Création d'un dataframe

Le premier traitement effectué est de regrouper les données collectées sous forme de dataframe qu'on peut par la suite exploiter dans nos étapes suivantes. Le résultat se présente comme suite :

(3295, 2)

	content	category
0	... مهرجان "فيكام" بمكناس يسعى إلى خلق سوق مغربية	culture
1	... من \r\n\r\n ملقحي تشكيلي يروج للثقافة الصحراوية	culture
2	... رحيل الشاعر العراقي حسب الشيخ جعفر عن 80 عاما	culture
3	... الصديق المحتاج" .. واتصاب بتكشف طريقة جديدة لخد	science
4	... لحماية القصر.. عوغل تسمح للأهل بإزالة صور أينا	science
5	... حولت أحلام العديد من المبتكرين لواقع.. تأثير أ	science
6	... "الوسيد تتغلب على صلا بشكل رسمي في مجال "المدى	science
7	... نشطاء يحولون على الانتخابات المقبلة لتسريع ال	tamazight
8	... مقتل 31 شخصا في تدافع قرب إحدى الكنائس جنوبي ن	politics
9 بتكيران بضع أخطاء الحركة الإسلامية تحت المجهر	politics

FIGURE 2.2 – Dataframe

2.2.2 Visualisation

Les fréquences des articles de chaque thème peut se visualisé comme suit :

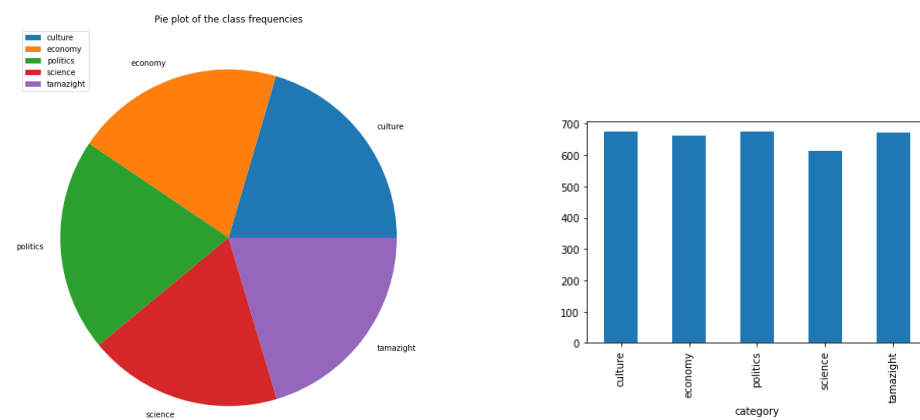


FIGURE 2.3 – Les fréquences des classes

2.3 Extraction de caractéristiques

Le jeu de données ne contient aucune valeur manquante :

```
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  -
0    text    430 non-null    object
1    label    430 non-null    object
dtypes: object(2)
```

FIGURE 2.4 – Les valeurs manquantes

Après la visualisation des classes et de nombre d'articles dans chacune de ces classes on peut les labelisé comme suit :

	label	labelld
0	culture	0
1	politics	1
6	tamazight	2
10	science	3
11	economy	4

FIGURE 2.5 – Labellisation des classes

Chapitre 3

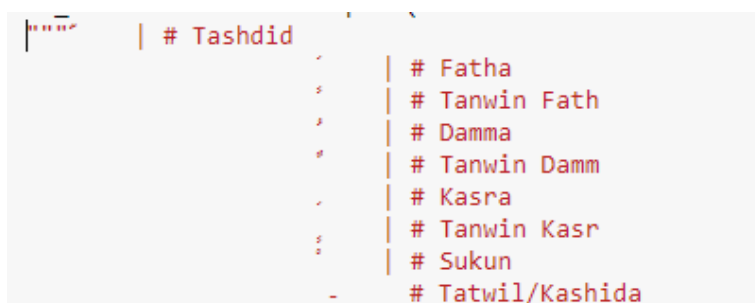
Traitement du text arabe [NLP]

Le traitement du langage naturel est l'un des domaines qui connaît la croissance la plus rapide au monde. Le NLP fait son chemin dans un certain nombre de produits et de services que nous utilisons dans notre vie quotidienne. Les étapes les plus importantes d'un pipeline NLP sont le traitement et le nettoyage du texte, y compris le Stemming (déradinement) et la Lemmatisation.

Certaines caractéristiques du langage, comme la ponctuation et les mots courants tels que "مع", "في", "ال", contribuent souvent à structurer le document, mais n'apportent pas grand-chose au sens. Il est donc préférable de les supprimer avant d'analyser les données textuelles et de les introduire dans notre pipeline de traitement du langage naturel.

3.0.1 Supprimer les caractères spéciaux

La première étape effectuée dans ce traitement a été de supprimer les caractères spéciaux qui n'influencent pas le sens de texte. ci-dessous des exemples des caractères spéciaux :



ّ	# Tashdid
َ	# Fatha
ً	# Tanwin Fath
ُ	# Damma
ٌ	# Tanwin Damm
ِ	# Kasra
ٍ	# Tanwin Kasr
ْ	# Sukun
-	# Tatwil/Kashida

FIGURE 3.1 – Exemple des caractères spéciaux

Après la suppression le résultats de chaque document peut être visualiser comme suit :

	content	category	cleaned_content
0	...مارسيل خليفة يمتنى الأصل المغربي ويتسك بالما	culture	...مارسيل خليفة يمتنى الأصل المغربي ويتسك بالما
1	...بركة يبسط خطة الحكومة لحماية المغاربة من "الوض	politics	...بركة يبسط خطة الحكومة لحماية المغاربة الوضعية
2	...جامعة الحسن الثاني ثمن ثرات الشاوية \r\n\r\n	culture	...جامعة الحسن الثاني ثمن ثرات الشاوية نظم مختصر
3	...المعارضة ترفض احتقار البرلمان ومسؤول حكومي نحت	politics	...المعارضة ترفض احتقار البرلمان ومسؤول حكومي نحت
4	...الحكومة تدارس مشاريع مراسيم يتعقد يوم الخميس	politics	...الحكومة تدارس مشاريع مراسيم يتعقد يوم الخميس
5	...مناورات عسكرية تعزز التعاون الأمني بين الجيشين	politics	...مناورات عسكرية تعزز التعاون الأمني الجيشين الم
6	...دائشون يطالبون رفع التمييز اللغة الأمازيغية الم	tamazight	...دائشون يطالبون رفع التمييز اللغة الأمازيغية الم
7	...نظم \r\n\r\n تطوان تحضن المهرجان الدولي للعود	culture	...نظم \r\n\r\n تطوان تحضن المهرجان الدولي للعود
8	...المعارضة تكفي بالتسويق داخل "النواب".. وخلافا	politics	...المعارضة تكفي بالتسويق داخل النواب وخلافات ال
9	...وزاره الفلاحة والمعهد الملكي بذلان عراجل ا	tamazight	...وزاره الفلاحة والمعهد الملكي بذلان عراجل الت

FIGURE 3.2 – Suppression les caractères spéciaux

3.0.2 Suppression de tous les mots d'arrêt

Un mot d'arrêt est un mot couramment utilisé (tel que "ال", "في", "مع") qu'un moteur de recherche a été programmé pour ignorer, à la fois lors de l'indexation des entrées pour la recherche et lors de leur récupération comme résultat d'une requête de recherche. Nous ne voudrions pas que ces mots occupent de l'espace dans notre base de données, ni qu'ils prennent le précieux temps de traitement. Pour cela, nous pouvons les supprimer facilement, en stockant une liste de mots que vous considérez comme des mots d'arrêt. NLTK (Natural Language Toolkit) en python a une liste de mots d'arrêt stockée dans 16 langues différentes.

En appliquant ce traitement on aura :

	content	category	cleaned_content	cleaned_content_list
0	...مارسيل خليفة يمتنى الأصل المغربي ويتسك بالما	culture	...مارسيل خليفة يمتنى الأصل المغربي ويتسك بالما	..., 'مارسيل', 'خليفة', 'يمتنى', 'الأصل', ']
1	...بركة يبسط خطة الحكومة لحماية المغاربة من "الوض	politics	...بركة يبسط خطة الحكومة لحماية المغاربة الوضعية	...بركة', 'يبسط', 'خطة', 'الحكومة', 'ال', ']
2	...جامعة الحسن الثاني ثمن ثرات الشاوية \r\n\r\n	culture	...جامعة الحسن الثاني ثمن ثرات الشاوية نظم مختصر	..., 'جامعة', 'الحسن', 'الثاني', 'ثمن', ']
3	...المعارضة ترفض احتقار البرلمان ومسؤول حكومي نحت	politics	...المعارضة ترفض احتقار البرلمان ومسؤول حكومي نحت	...المعارضة', 'يرفض', 'احتقار', 'البرلمان', ']
4	...الحكومة تدارس مشاريع مراسيم يتعقد يوم الخميس	politics	...الحكومة تدارس مشاريع مراسيم يتعقد يوم الخميس	...الحكومة', 'تدارس', 'مشاريع', 'مراسيم', ']
5	...مناورات عسكرية تعزز التعاون الأمني بين الجيشين	politics	...مناورات عسكرية تعزز التعاون الأمني الجيشين الم	...مناورات', 'عسكرية', 'تعزز', 'التعاون', ']
6	...دائشون يطالبون رفع التمييز اللغة الأمازيغية الم	tamazight	...دائشون يطالبون رفع التمييز اللغة الأمازيغية الم	..., 'دائشون', 'يطالبون', 'رفع', 'التمييز', ']
7	...نظم \r\n\r\n تطوان تحضن المهرجان الدولي للعود	culture	...نظم \r\n\r\n تطوان تحضن المهرجان الدولي للعود	...تطوان', 'تحضن', 'المهرجان', 'الدولي', ']
8	...المعارضة تكفي بالتسويق داخل "النواب".. وخلافا	politics	...المعارضة تكفي بالتسويق داخل النواب وخلافات ال	...المعارضة', 'تكفي', 'بالتسويق', 'داخل', ']
9	...وزاره الفلاحة والمعهد الملكي بذلان عراجل ا	tamazight	...وزاره الفلاحة والمعهد الملكي بذلان عراجل الت	...وزاره', 'الفلاحة', 'والمعهد', 'الملكي', ']

FIGURE 3.3 – cleaned data

3.0.3 Lemmatisation

La lemmatisation désigne un traitement lexical apporté La lemmatisation est une autre technique utilisée pour réduire les mots à une forme normalisée. Dans la lemmatisation, la transformation utilise un dictionnaire pour ramener les différentes variantes d'un mot à sa racine. Ainsi, avec cette approche, nous sommes en mesure de réduire les inflexions non triviales telles que "تكتبون", "اكتبوا" à la racine "كتب".

3.0.4 Stemming

Stemming est le processus qui consiste à réduire un mot à sa racine. Prenons un exemple. Considérons trois mots, "branché", "branchement" et "branches". Ils peuvent tous être réduits au même mot "branche". Après tout, tous les trois véhiculent la même idée de quelque chose qui se sépare en plusieurs chemins ou branches. Là encore, cela permet de réduire la complexité tout en conservant l'essence du sens porté par ces trois mots.

3.0.5 Tokenization

La tokenisation est une tâche courante dans le traitement du langage naturel (NLP). Il s'agit d'une étape fondamentale à la fois dans les méthodes traditionnelles de NLP, comme le vecteur de comptage, et dans les architectures avancées basées sur le Deep Learning, comme les transformateurs.

Les tokens sont les éléments constitutifs du langage naturel.

La tokenisation est une façon de séparer un morceau de texte en unités plus petites appelées tokens. Les tokens peuvent être des mots, des caractères ou des sous-mots. Par conséquent, la tokénisation peut être classée en trois types : la tokénisation des mots, des caractères et des sous-mots (caractères n-gram).

Les occurrences des mots de chaque classe peut être exprimé avec u la figure ci-dessous. Plus l'écriture est grande plus que le nombre d'occurrences de mot est grand :

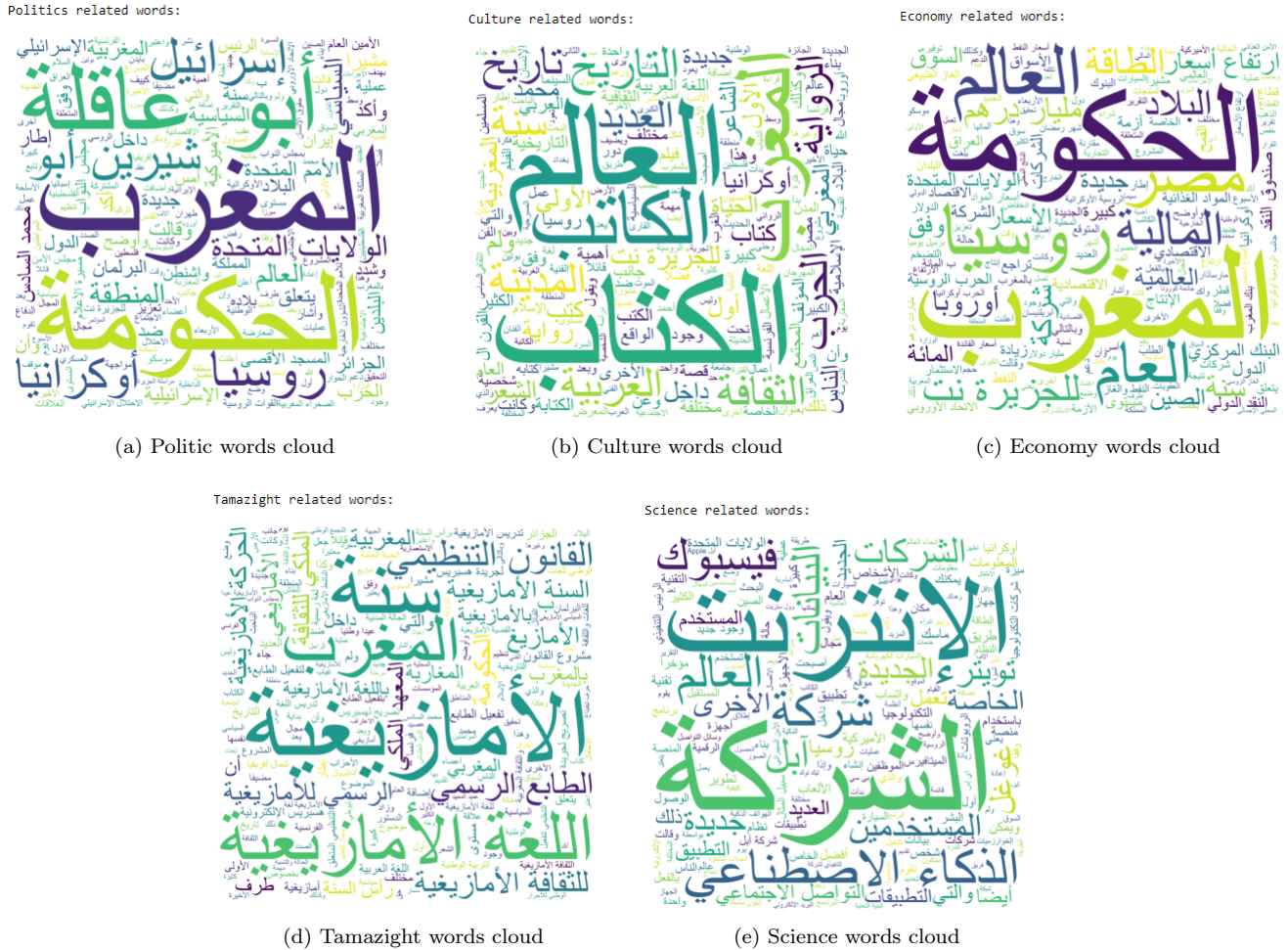


FIGURE 3.4 – Words cloud de chaque catégorie des documents

3.0.7 Vectorization

La vectorization est une technique utilisée pour convertir les données textes en caractéristiques numériques exploitable dans les modèles de machine learning. Les trois techniques les plus utilisées sont le Bag of Words, tf-idf vectorization et word embedding. Dans notre projet nous avons utilisé tf-idf et Bag of words et nous avons comparé les résultats obtenus pour chacune de ces méthodes

$$\text{tf}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad idf(t) = \ln\left(\frac{1+n}{1+df(t)}\right) + 1$$

FIGURE 3.5 – Algorithm Vectorisation $TF-IDF = TF \times IDF$

Chapitre 4

les méthode de clustering

4.1 K-means

Le partitionnement en k-means est une méthode de partitionnement de données et un problème d'optimisation combinatoire. Étant donnés des points et un entier k , le problème est de diviser les points en k groupes, souvent appelés clusters, de façon à minimiser une certaine fonction. On considère la distance d'un point à la moyenne des points de son cluster ; la fonction à minimiser est la somme des carrés de ces distances.

Il existe une heuristique classique pour ce problème, souvent appelée méthodes des k-moyennes, utilisée pour la plupart des applications. Le problème est aussi étudié comme un problème d'optimisation classique, avec par exemple des algorithmes d'approximation.

4.1.1 Algorithme

- Choisir k points qui représentent la position moyenne des partitions $m_1^{(1)}, \dots, m_k^{(1)}$ initiales (au hasard par exemple) ;
- Répéter jusqu'à ce qu'il y ait convergence :
 - affecter chaque observation à la partition la plus proche (c.-à-d. effectuer une partition de Voronoï selon les moyennes) :
$$S_i^{(t)} = \left\{ \mathbf{x}_j : \|\mathbf{x}_j - \mathbf{m}_i^{(t)}\| \leq \|\mathbf{x}_j - \mathbf{m}_{i^*}^{(t)}\| \forall i^* = 1, \dots, k \right\}$$
 - mettre à jour la moyenne de chaque cluster :
$$\mathbf{m}_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{\mathbf{x}_j \in S_i^{(t)}} \mathbf{x}_j \mathbf{m}_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{\mathbf{x}_j \in S_i^{(t)}} \mathbf{x}_j.$$

4.2 Clustering spectrale

4.2.1 Algorithme de base

- 1 Calculer le Laplacien \mathbf{L} (ou le Laplacien normalisé).
- 2 Calculer les k premiers vecteurs propres (les vecteurs propres correspondant aux plus petites valeurs propres de \mathbf{L})
- 3 Considérez la matrice formée par les premiers vecteurs propres ; la **l -ème** ligne définit les caractéristiques du noeud du graphe \mathbf{l} .
- 4 Regroupez les nœuds du graphe en fonction de ces caractéristiques (par exemple, à l'aide d'un regroupement de type k-means).

Le clustering spectral est étroitement lié à la réduction de la dimensionnalité non linéaire, et les techniques de réduction de la dimension telles que l'intégration localement linéaire peuvent être utilisées pour réduire les erreurs dues au bruit ou aux valeurs aberrantes.

Chapitre 5

Résultats

5.1 Résultats de clustering

5.1.1 K-means

Le clustering avec k-means a donné les résultats suivant :

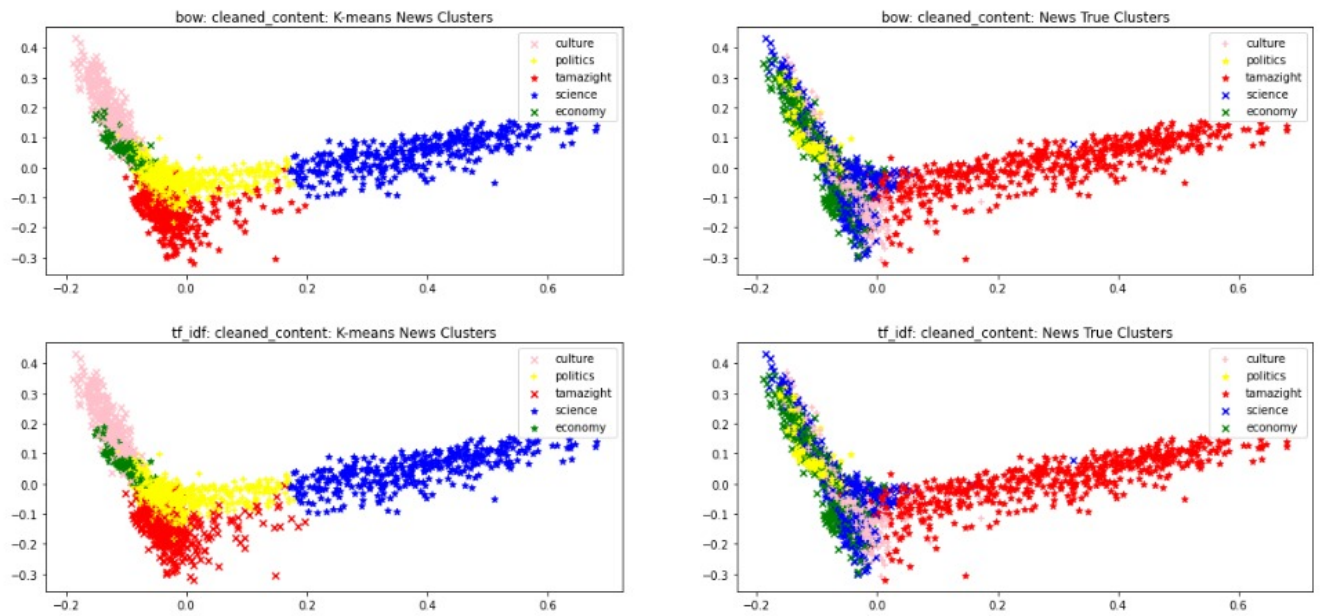


FIGURE 5.1 – Les classes de k-means

5.1.2 Clustering spectrale

Le clustering avec clustering spectrale a donnée les résultats suivant :

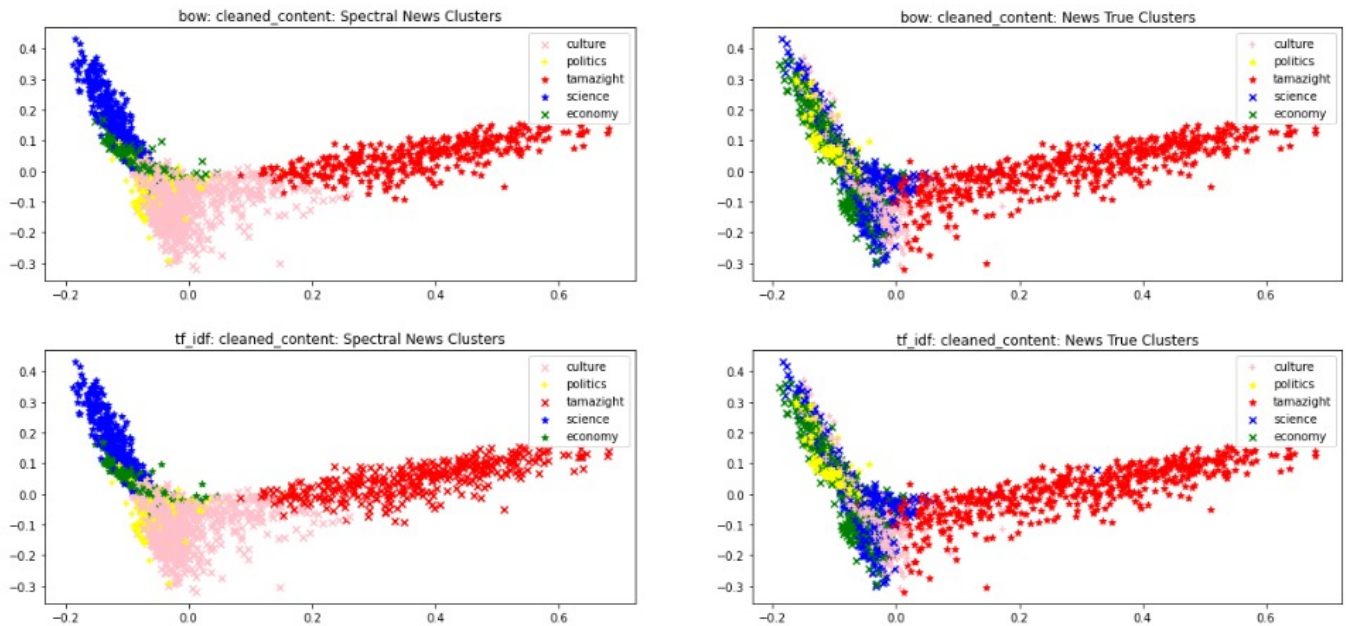


FIGURE 5.2 – Les classes de Clustering spectrale

5.1.3 Comparaison

Avec l'inconvénient majeur de clustering spectrale à savoir une grande sensibilité au bruit et plus généralement une difficulté à détecter des clusters en contact, Il a donnée des résultats qui contient des confusion entre les articles de tamazight/culture et économie/politique et ça revient à la similarité de contenu de ces articles.

5.2 Méthodes de validation

5.2.1 K-means

L'indice de davies bouldin pour le K-means a donné les résultats suivant :

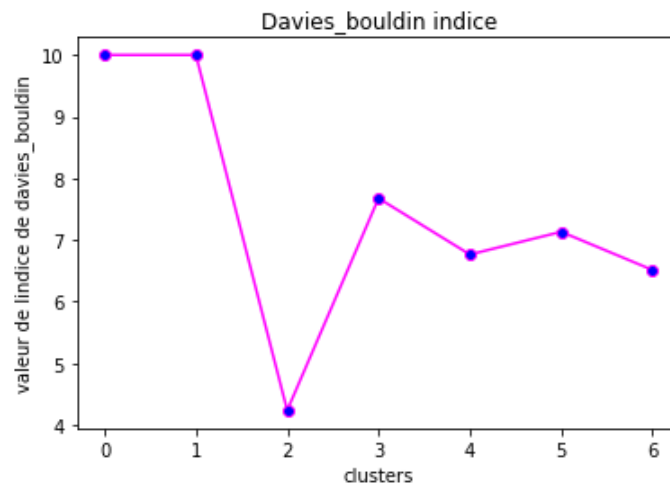


FIGURE 5.3 – davies bouldin scores pour k-means

La plus petite valeur est donnée par 2 clusters. Donc on peut fixer le nombre de clusteres à **2**.

L'indice de silhouette pour le K-means a donné les résultats suivant :

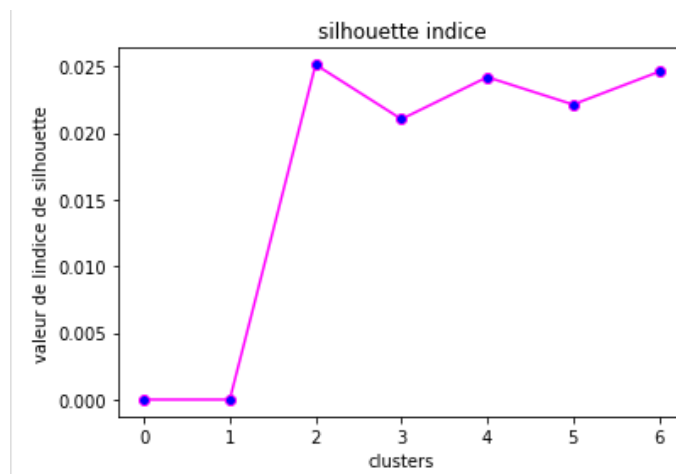


FIGURE 5.4 – silhouette scores pour k-means

La plus grande valeur est donnée par 2. Donc on peut fixer le nombre de clusteres à **2**.

5.2.2 Clustering spectrale

L'indice de davies bouldin pour le Clustering spectrale a donné les résultats suivant :

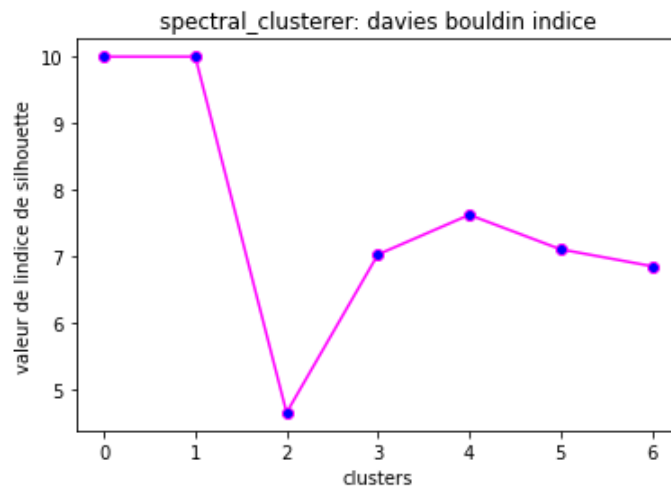


FIGURE 5.5 – davies bouldin scores pour Clustering spectrale

La plus petite valeur est donnée par 2. Donc on peut fixer le nombre de clusteres à **2**.

L'indice de silhouette pour le Clustering spectrale a donné les résultats suivant :

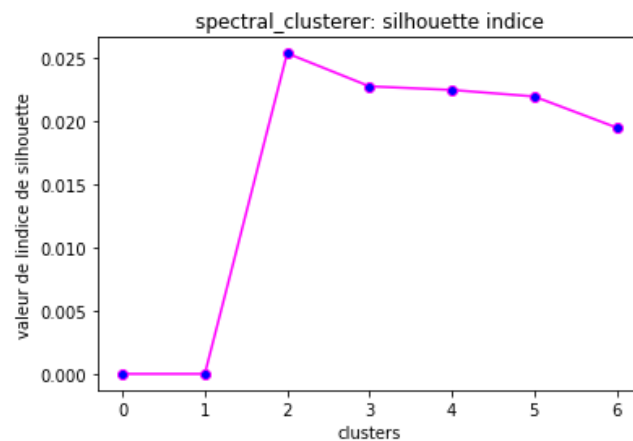


FIGURE 5.6 – silhouette scores pour Clustering spectrale

La plus grande valeur est donnée par 2. Donc on peut fixer le nombre de clusteres à **2**.

Conclusion

Problèmes rencontrés

- Manque de bibliothèque python pour la manipulation et traitement des données de la langue arabe
- Les articles de Tamazight et culture sont similaire, de même pour les articles de politiques et économie ce qui a diminué le score de clustering

Bilan

Une grande collection de documents peut fournir des informations utiles aux gens, mais c'est aussi un défi de trouver les informations utiles à partir d'une grande collection de documents.

Les techniques de fouille de texte mises en œuvre avec succès aident à identifier la catégorie de chaque document textuel dans laquelle il s'inscrit.

Ce projet est a comme objectif de catégoriser les documents écrits dans la langue Arabe, Puisque la plupart des systèmes de catégorisation de texte existants se concentrent sur la catégorisation de texte dans la langue anglaise, ils sont incapables de catégoriser avec précision les documents écrits dans d'autres langues avec des résultats significatifs. La langue Arabe a été utilisée comme étude de cas pour construire une approche de catégorisation de texte.

Cette recherche a réalisé plusieurs expériences intéressantes basées sur la catégorisation de textes. L'étape de pré traitement des données a préparé les données pour le processus de clustering, en supprimant les mots qui ont conduit à l'interférence du regroupement des documents, plusieurs technique ont été appliqué sur les données traitées pour obtenir la matrice de similarité et les techniques de clustering k-means ont été appliquées sur la matrice de similarité pour valider les résultats obtenus et pour garantir la complétude. Enfin, les résultats obtenus ont été correctement analysés et discutés avec les recommandations possibles. Cette recherche a considéré plusieurs tâches liées à la catégorisation de texte comme mentionné dans le champ de recherche.

Les résultats de cette recherche sont généralement applicables à n'importe quel domaine contenant des données textuelles et aident à traiter de grandes quantités de données. Cette approche peut être appliquée de manière pratique et étendue à des applications dans des domaines tels que l'analyse des réclamations d'assurance, le traitement automatique des fax, le filtrage des courriers électroniques, et bien d'autres applications de fouille de texte.

Bibliographie

- [1] <https://www.kdnuggets.com/2018/03/text-data-preprocessing-walkthrough-python.html>.
- [2] https://www.researchgate.net/publication/323433518_Arabic_text_clustering_using_improved_clustering_algorithms_with_dimensionality_reduction/figures?lo=1.
- [3] <https://towardsdatascience.com/all-the-news-17fa34b52b9d>.
- [4] <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html>.