Sommaire

ANALYSE DES DONNEES TEXTUELLES DE LA BASE DE DONNEES "KALIMAT"

Description du projet :

Le projet "Analyse des données textuelles de la base de données Kalimat" vise à explorer et à analyser une collection de données textuelles diverses, accessibles via la base de données Kalimat. Cette collection est composée de six fichiers distincts, chacun représentant un domaine d'articles particulier. Voici un aperçu des domaines inclus :

Articles sur la culture : Cette catégorie contient des articles sur la culture, les arts, la littérature et d'autres sujets connexes. L'objectif est d'analyser les tendances culturelles et artistiques dans les articles de cette catégorie.

Articles sur l'économie : Cette section comprend des articles traitant de questions économiques, financières et commerciales. L'analyse porte sur les fluctuations économiques, les tendances du marché et les préoccupations économiques actuelles.

Articles internationaux : Les articles internationaux couvrent des événements et des sujets mondiaux, tels que les relations internationales, la politique mondiale et les affaires étrangères. L'objectif est de comprendre les relations internationales et les problèmes mondiaux.

Articles locaux : Les articles locaux se concentrent sur des événements et des sujets régionaux ou locaux. L'analyse portera sur les préoccupations locales, les événements communautaires et les tendances régionales.

Articles sur la religion : Cette catégorie couvre les sujets liés à la religion, aux croyances et aux pratiques religieuses. L'objectif est de comprendre la dynamique religieuse et les questions spirituelles abordées dans les articles.

Articles sur le sport : Les articles sur le sport traitent de l'actualité sportive, des compétitions, des performances des athlètes et des événements sportifs. L'analyse portera sur les tendances sportives, les succès des équipes et des athlètes, ainsi que sur les intérêts du public.

Objectif du projet :

L'objectif principal de ce projet est d'explorer, d'analyser et de tirer des enseignements d'une variété de données textuelles provenant de la base de données Kalimat. Nous cherchons à comprendre les tendances, les préoccupations et les principaux sujets dans les domaines de la culture, de l'économie, de l'international, du local, de la religion et du sport. Cette analyse nous permettra de mieux comprendre les dynamiques au sein de chaque domaine et de mettre en évidence les informations significatives pour une prise de décision éclairée.

Ensemble de données : Kalimat

L'ensemble de données Kalimat est une collection de textes arabes comprenant un total de 20 291 articles répartis en six catégories spécifiques : Culture, Économie, Nouvelles internationales, Nouvelles locales, Religion et Sports. Ces articles ont été extraits du journal omanais Alwatan par Abbas et al. en 2011.

La répartition des articles dans chaque catégorie est la suivante :

- Culture: 2 495 articles, totalisant 1 359 210 mots, avec une moyenne de 544 mots par article.

- Économie: 3 265 articles, totalisant 3 122 565 mots, avec une moyenne de 956 mots par article.

- Nouvelles internationales: 1 689 articles, totalisant 855 945 mots, avec une moyenne de 506 mots par article.

- Actualités locales: 3 237 articles, totalisant 1 460 462 mots, avec une moyenne de 452 mots par article.

- Religion: 3 474 articles, totalisant 1 555 635 mots, avec une moyenne de 448 mots par article.

- Sports: 4 095 articles, totalisant 9 813 366 mots, avec une moyenne de 2 397 mots par article.

Le processus de création de Kalimat a été appliqué à l'ensemble de la collecte de données. Tout d'abord, les documents ont été résumés à l'aide de deux résumeurs arabes, Gen-Summ et Cluster-based. Gen-Summ est un résumeur de document unique basé sur le modèle VSM (Salton et al. 1975) qui prend un document arabe et sa première phrase et renvoie un résumé extractif. Cluster-based est un résumeur multi-documents qui traite tous les documents à résumer comme un seul sac de phrases. Les phrases de tous les documents sont regroupées en utilisant différents nombres de grappes, puis un résumé est créé en sélectionnant uniquement les phrases de la plus grande grappe.

Au total, 2 057 résumés extractifs ont été générés pour chaque catégorie, y compris un résumé pour chaque tranche de 10, 100 et 500 articles, ainsi qu'un résumé pour tous les articles de chaque catégorie. En combinant le nombre total d'articles et le nombre total de mots dans

chaque catégorie, l'ensemble de données Kalimat offre une vue détaillée et complète du paysage journalistique couvert. Ces informations peuvent être d'une grande valeur pour les chercheurs et les analystes de données, leur permettant de se plonger dans des sujets spécifiques et d'explorer les nuances linguistiques et thématiques au sein de chaque catégorie. L'ensemble de données offre donc une riche opportunité d'analyse approfondie et de perspectives dans le domaine de la linguistique computationnelle arabe et de l'analyse de texte.

1.Lecture de données:

Cette étape joue un rôle crucial dans la gestion et la préparation des données pour l'analyse ultérieure. Le premier code extrait et structure les articles des fichiers ZIP, en les organisant dans des DataFrames distincts pour chaque catégorie. L'objectif est de faciliter l'exploration et l'analyse des articles en créant une représentation tabulaire claire avec des colonnes "Catégorie" et "Article". Le résultat final, consolidé dans le DataFrame 'combined_df', offre une vue d'ensemble organisée des articles de différentes catégories.

puis complète cette approche en générant des fichiers textes consolidés pour chaque catégorie. Il extrait le contenu textuel des articles, le combine en un seul texte et l'enregistre dans des fichiers individuels. Cette méthode simplifie la gestion des données et permet d'accéder facilement au contenu complet de chaque catégorie. L'ensemble du processus contribue à une préparation efficace des données, facilitant l'analyse future et la compréhension des tendances spécifiques à chaque domaine, qu'il s'agisse du sport, de la culture, des sujets locaux, de la religion, de l'économie ou des affaires internationales.

2.Exploration des données :

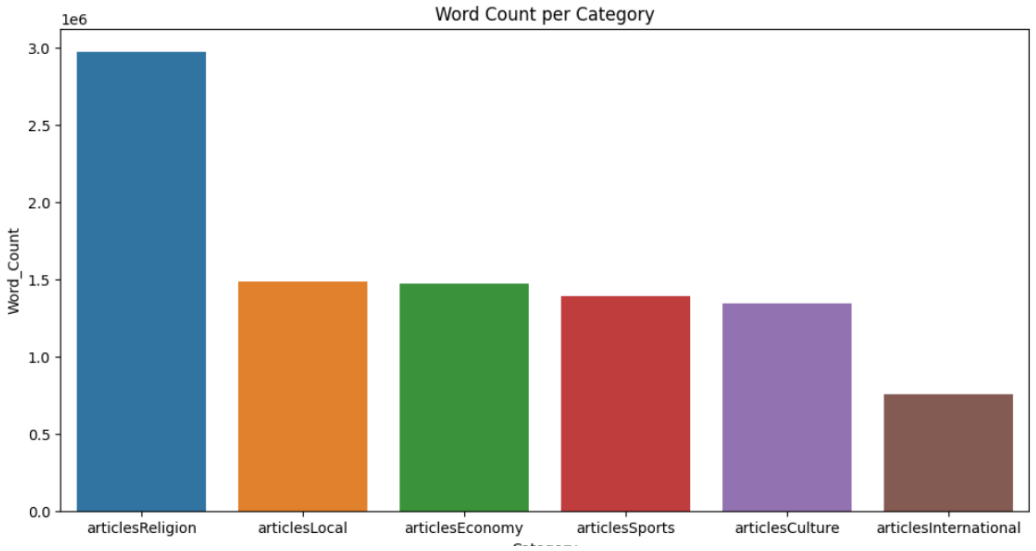
L'exploration approfondie des données a permis de générer le tableau ci-dessus, qui fournit des statistiques clés sur la structure et la longueur du contenu pour chaque catégorie d'article. Ce tableau est un instantané instructif, qui présente des paramètres essentiels tels que le nombre de mots, le nombre de caractères, le nombre moyen de caractères par mot, le nombre brut de caractères et le nombre de mots vides. Chaque ligne du tableau représente une catégorie spécifique d'articles, offrant une vue comparative de leurs caractéristiques textuelles.

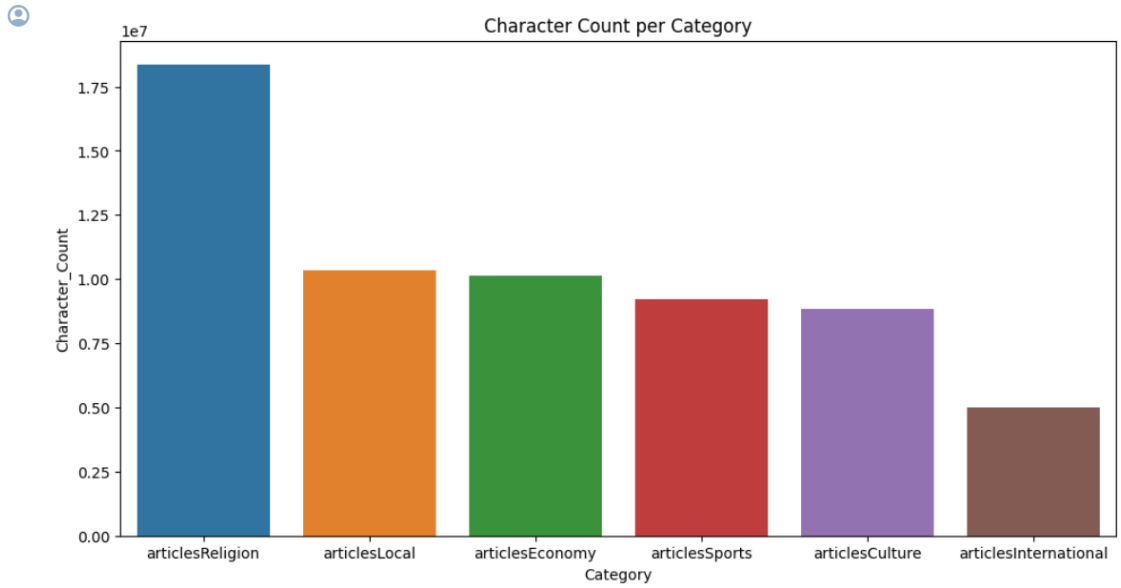
L'analyse de ces données permet de dégager plusieurs tendances. La catégorie "Religion" se distingue par un contenu particulièrement volumineux, avec un nombre de mots et de caractères significativement élevé. À l'inverse, la catégorie "Local" présente un nombre moyen de caractères par mot plus élevé, indiquant peut-être une utilisation plus complexe ou technique du langage. D'autres catégories, telles que "Économie", "Sports", "Culture" et "International", révèlent également des nuances uniques dans leurs statistiques respectives.

Total number of characters in the dataset: 61885432

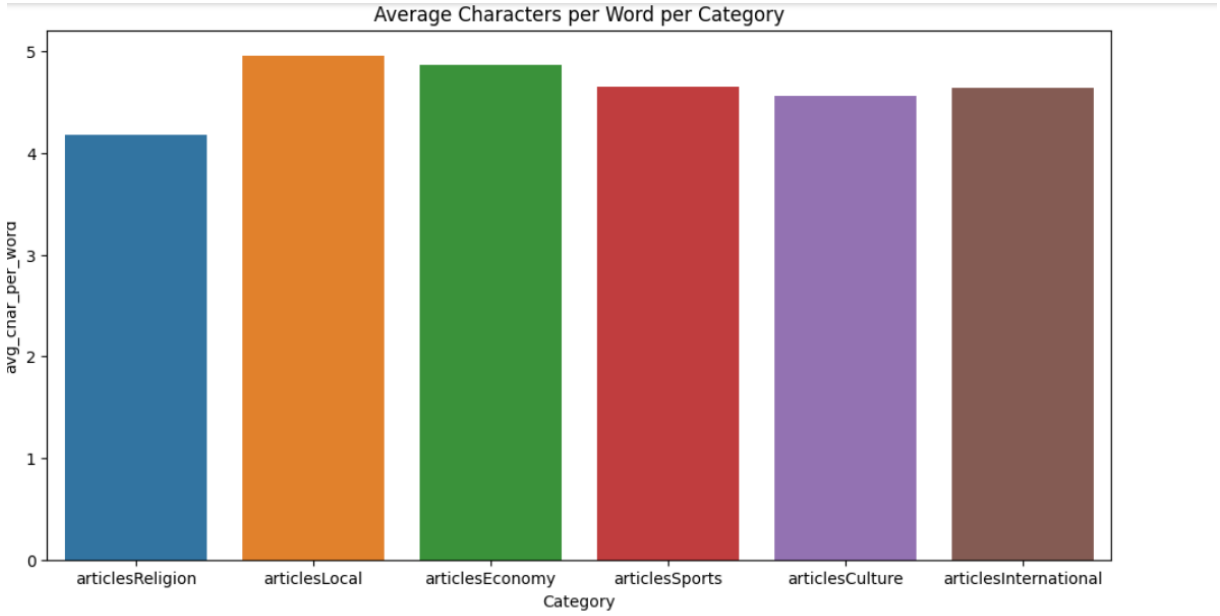
	Category	Content	Word_Count	Character_Count
0	articlesSports	اللجن\n\r\nمن\n\r\nدعوء\n\r\nيتلقى\n\r\nالكرء\n\r\nاتحاد...	1388149	9232840
1	articlesCulture	\r...اليوم\n\r\nانتطلق\n\r\n:\r\n\n\r\nالرحيي\n\r\nسالم\n\r\nكتب	1344069	8830574
2	articlesLocal	\...الجهد\n\r\nوينل\n\r\nالمشي\n\r\nعادء\n\r\nانتشجيع\n\r\nان	1484587	10336787
3	articlesReligion	\...واضرب\n\r\n)\n\r\nوتعالى\n\r\nتبارك\n\r\nالله\n\r\nيقول	2972638	18369765
4	articlesEconomy	\...اعمال\n\r\nامس\n\r\nصباح\n\r\nالشورى\n\r\nمجلس\n\r\nعقد	1472514	10119923
5	articlesInternational	...وكا\n\r\nعواصم\n\r\nابوتوار\n\r\nاسعيد\n\r\nمن\n\r\nالرباط	751815	4995543

Pour compléter l'analyse statistique présentée dans le tableau ci-dessus, une visualisation graphique permet de mieux comprendre les différences entre les catégories d'articles. La courbe ci-dessous illustre l'évolution du nombre de caractères et de mots pour chaque catégorie. Comme on peut le constater, la catégorie "Religion" se distingue nettement par une courbe ascendante, indiquant un contenu sensiblement plus long en termes de caractères et de mots. À l'inverse, la catégorie "International" présente une courbe descendante, suggérant un contenu plus succinct. Cette représentation visuelle offre une perspective instantanée sur les variations de longueur entre les catégories, complétant ainsi l'analyse quantitative précédente.



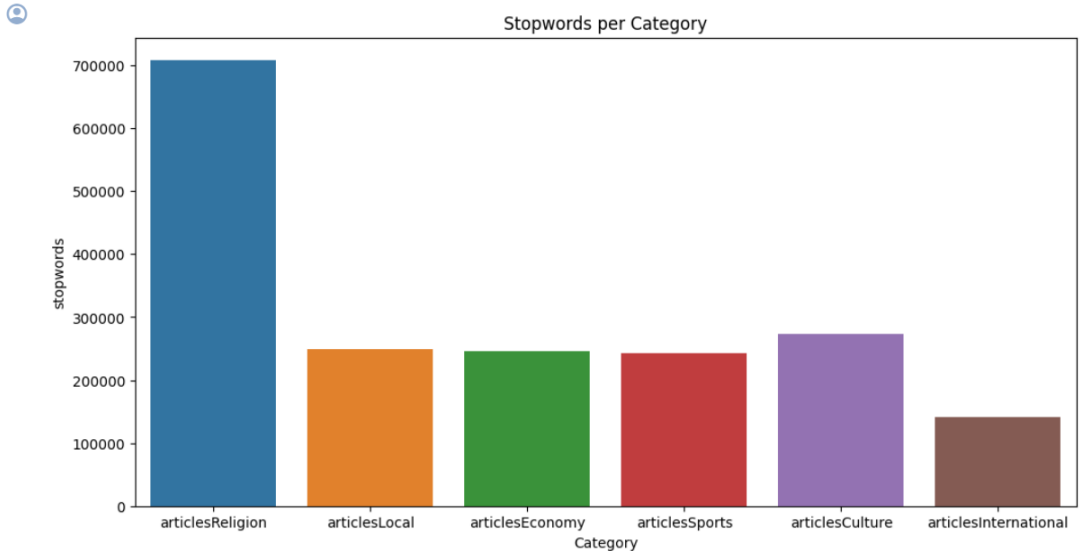


Pour compléter notre exploration de données, la figure ci-dessous montre le nombre moyen de caractères par mot dans chaque catégorie d'article. Bien que les valeurs moyennes soient relativement proches, des nuances significatives apparaissent à l'examen du graphique. Les catégories "Local" et "Économie" affichent des moyennes légèrement plus élevées, ce qui suggère l'utilisation potentielle d'un vocabulaire plus complexe ou plus technique que dans les autres catégories. En revanche, la catégorie "Religion" présente la moyenne la plus basse, ce qui indique une densité de mots relativement plus faible en termes de caractères par mot.



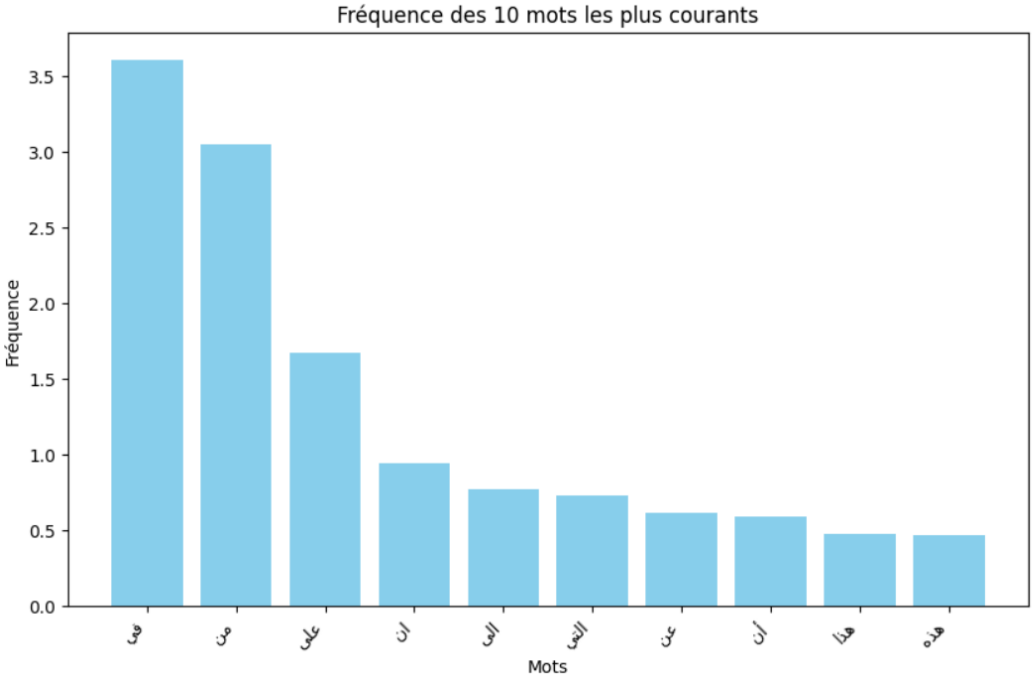
La figure ci-dessous illustre la fréquence des mots vides dans chaque catégorie d'article. L'histogramme met en évidence la disparité de l'utilisation des mots vides entre les différents thèmes. Les catégories "Économie" et "Local" présentent des fréquences relativement plus élevées de mots vides, ce qui indique peut-être un style éditorial plus formel ou technique. En revanche, la catégorie "International" présente une fréquence

plus faible de mots vides, ce qui suggère une utilisation potentiellement plus concise et directe du langage.



L'exploration approfondie de la base de données a révélé des insights essentiels sur les mots les plus fréquemment utilisés avant la suppression des mots vides. L'histogramme affiche la fréquence des dix mots les plus courants, parmi lesquels figurent des termes tels que, 'من', 'التي', 'ان', 'إلى', 'على', 'في', et 'فى'. Ces mots sont particulièrement prédominants dans le corpus d'articles, soulignant leur récurrence significative dans le langage utilisé.

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
Mots les plus fréquents dans l'intersection de toutes les catégories : {'من', 'على', 'في', 'الى', 'التي', 'ان'}
```



La matrice TF-IDF (Term Frequency-Inverse Document Frequency) fournit une représentation numérique des articles du corpus, soulignant l'importance relative des

mots dans chaque document. Chaque valeur de la matrice représente le poids TF-IDF d'un mot spécifique pour un document donné.

Valeurs faibles : Les valeurs proches de zéro indiquent que le mot a une fréquence relativement faible dans le document, ou qu'il est fréquent dans l'ensemble du corpus, n'apportant pas de distinction significative à ce document particulier.

Valeurs élevées : les valeurs élevées suggèrent que le mot est spécifique au document et qu'il peut jouer un rôle important dans sa caractérisation. Ces mots ont une fréquence significative dans le document et sont moins fréquents dans l'ensemble du corpus.

La matrice comprend un ensemble étendu de mots uniques (252 990 colonnes) extraits du corpus. Les lignes représentent différents articles et les colonnes des mots spécifiques. En examinant les valeurs de la matrice, vous pouvez identifier les mots qui contribuent le plus à la caractérisation de chaque article.

	00	000	0000	000000	0004	000450	000479	000620388	00078	000809	...	بيفتر	بيفدوكيموفا	بيفو	بيل	بيلديرين	بيمنع	بيهم	بين	بينا	بيونج
0	0.000129	0.000500	0.000436	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.001987	0.003170	0.000852	0.000000	0.000000	0.000107	0.000107	0.000107	0.000000	0.000107	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	0.003169	0.007395	0.001541	0.000000	0.000000	0.000000	0.000000	0.000000	0.000099	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	0.002127	0.000155	0.001237	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000174	0.000087	0.000523	0.000000	0.000087	0.000087	0.000000	0.000000	0.000000	0.000087
4	0.008122	0.001624	0.003101	0.000166	0.000166	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000273	0.000000	0.000000	0.000333	0.000166	0.000166	0.000000
5	0.006138	0.005288	0.001983	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000174	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

6 rows × 252479 columns

3. Prétraitement des données

-Tokinisation

La tokenisation, comme le montre l'exemple ci-dessus, est une étape essentielle du traitement du langage naturel (NLP), qui consiste à diviser un texte en unités discrètes appelées "tokens".

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
                                                Content  \
3  يقول\r\nالله\r\nتبارك\r\nوتعالى\r\n(\r\nواضرب\r...
2  ان\r\nتشجيع\r\nعادة\r\nالمشي\r\nوبذل\r\nالجهد\r...
4  عقد\r\nمجلس\r\nالشورى\r\nصباح\r\nامس\r\nاعمال\r...
0  اتحاد\r\nالكرة\r\nيتلقى\r\nدعوة\r\nمن\r\nاللجن...
1  كتب\r\nسالم\r\nالرحيبي\r\n:\r\nتنطلق\r\nاليوم\r...
5  الرباط\r\nمن\r\nسعيد\r\nبونوار\r\nعواصم\r\nوكا...

                                  Tokenized_Content
3  [يقول, الله, تبارك, وتعالى, (, واضرب, لهم, مثل...
2  [ان, تشجيع, عادة, المشي, وبذل, الجهد, البدني, ...
4  [عقد, مجلس, الشورى, صباح, امس, اعمال, الجلسة, ...
0  [اتحاد, الكرة, يتلقى, دعوة, من, اللجنة, التنظي...
1  [كتب, سالم, الرحيبي, :, تنطلق, اليوم, الدورة, ا...
5  [الرباط, من, سعيد, بونوار, عواصم, وكالات, :, ق...
```

-suppresion des stop word :

L'élimination des mots vides est une étape fondamentale dans le prétraitement des données textuelles. Dans notre analyse, nous avons appliqué cette opération à tous les articles, en éliminant les mots vides qui, bien que courants dans la langue arabe, ont peu de signification contextuelle.

La capture d'écran ci-dessous montre un exemple des résultats obtenus après l'élimination des mots vides

```
stop words after remove =  0
```

-Normalisation :

Dans le cadre du prétraitement des données, nous avons inclus une étape cruciale de normalisation du texte arabe. Cette étape vise à rendre la représentation du texte plus cohérente et simplifiée en éliminant les redondances et en normalisant des caractères spécifiques. Le code de normalisation que nous avons mis en place utilise des expressions régulières et des substitutions pour atteindre cet objectif.

Pour illustrer l'impact de cette normalisation, nous avons inclus un exemple de texte arabe avant et après l'application de la normalisation. Cette procédure garantit l'uniformité de la représentation du texte, ce qui facilite les étapes ultérieures d'analyse et de traitement des données. La normalisation du texte arabe est essentielle pour garantir la cohérence des données et améliorer la qualité des résultats obtenus au cours des différentes étapes du projet.

```
Original Text: #أنا شَدَّةٌ وَأُحِبُّ اللُّغَةَ الْعَرَبِيَّةَ
Normalized Text: #انا شده واحب اللغه العربيه
```

-Suppression des caractères spéciaux :

La suppression des caractères spéciaux dans le cadre du prétraitement des données vise à éliminer les symboles non alphanumériques qui ne contribuent pas de manière significative à la signification du texte. La capture d'écran ci-dessous illustre les résultats obtenus après cette opération, avec un exemple représentatif avant et après la suppression des caractères spéciaux dans différentes catégories d'articles.

	Category	Content	Word_Count	Character_Count	char_count	avg_char_per_word	stopwords	normalized_content	text	clean_content
3	articlesReligion	يقول الله تبارك وتعالى (واضرب مثلا ...	2972638	18369765	18369765	4.178451	708432	نقول الله تبارك وتعالى (واضرب مثلا اصحاب القر...	نقول الله تبارك وتعالى (واضرب مثلا اصحاب القر...	يقول الله تبارك وتعالى واضرب مثلا ... اصحاب
2	articlesLocal	ان تشجيع عادة المشي وبذل الجهد البدني ...	1484587	10336787	10336787	4.960561	248895	ان تشجع عاده المشى وبذل الجهد البدنى ومماارسته...	ان تشجع عاده المشى وبذل الجهد البدنى ومماارسته...	ان تشجيع عادة المشي وبذل الجهد ... البدني
5	articlesEconomy	عقد مجلس الشورى امس اعمال الجلسة ...	1472514	10119923	10119923	4.870336	246689	عقد مجلس الشورى امس اعمال الجلسه الخامسه دور ا ...	عقد مجلس الشورى امس اعمال الجلسه الخامسه دور ا ...	عقد مجلس الشورى امس اعمال الجلسة ... الخامس
4	articlesSports	اتحاد الكرة يلتقى دعوة اللجنة ...	1388149	9232840	9232840	4.648244	242204	اتحاد الكره يلتقى دعوه اللجنه التنظمسه 2 اتحا ...	اتحاد الكره يلتقى دعوه اللجنه التنظمسه 2 اتحا ...	اتحاد الكرة يلتقى دعوة اللجنة ... التنظيمية
0	articlesCulture	كتب سالم الرحبى : تتطلق اليوم الدورة ...	1344069	8830574	8830574	4.568180	273679	كتب سالم الرحبى : تتطلق اليوم الدوره البرامجيه ...	كتب سالم الرحبى : تتطلق اليوم الدوره البرامجيه ...	كتب سالم الرحبى تتطلق اليوم الدورة ... البرا
1	articlesInternational	الرئاسة سنجير بونوار عواصم وكالات : ...	751815	4995543	4995543	4.642408	140748	الرئاسة سنجد بونوار عواصم وكالات : قالت معسادر م ...	الرئاسة سنجد بونوار عواصم وكالات : قالت معسادر م ...	الرئاسة سنجير بونوار عواصم وكالات قالت معس ...

-Stemming :

L'étape de Stemming et de substitution a considérablement simplifié la représentation des mots dans notre ensemble de données. La capture des résultats montre clairement comment cette étape a contribué à la normalisation de la langue.

	Category		Processed_Article	Target
0	articlesCulture	1	كتب سلم رحب طلق اليوم دور رمح جدد يفز ذعة رنمح...	
1	articlesCulture	1	كتب يصل علي شرك سلط امس دول علم حفل بيم ترث عل...	
2	articlesCulture	1	ربع عرض سرح شبب عرض رستاق عرض نزى عرض جمع سلط ...	
3	articlesCulture	1	حور خلد عبداللطيف نقش وضع ثقف علم جمع ضلع وقت ...	
4	articlesCulture	1	فتح بقع وسق جمع سلط قبس عرض فنن يحى شيخ وذل رع...	

-TFIDF

La matrice TF-IDF ci-dessous représente une transformation cruciale de notre corpus d'articles. Chaque ligne de cette matrice correspond à un seul article, tandis que chaque colonne représente un mot spécifique de tous nos articles. Cette représentation numérique met en évidence l'importance relative des mots dans chaque document, ce qui facilite l'identification des termes spécifiques qui définissent le contenu de chaque article.

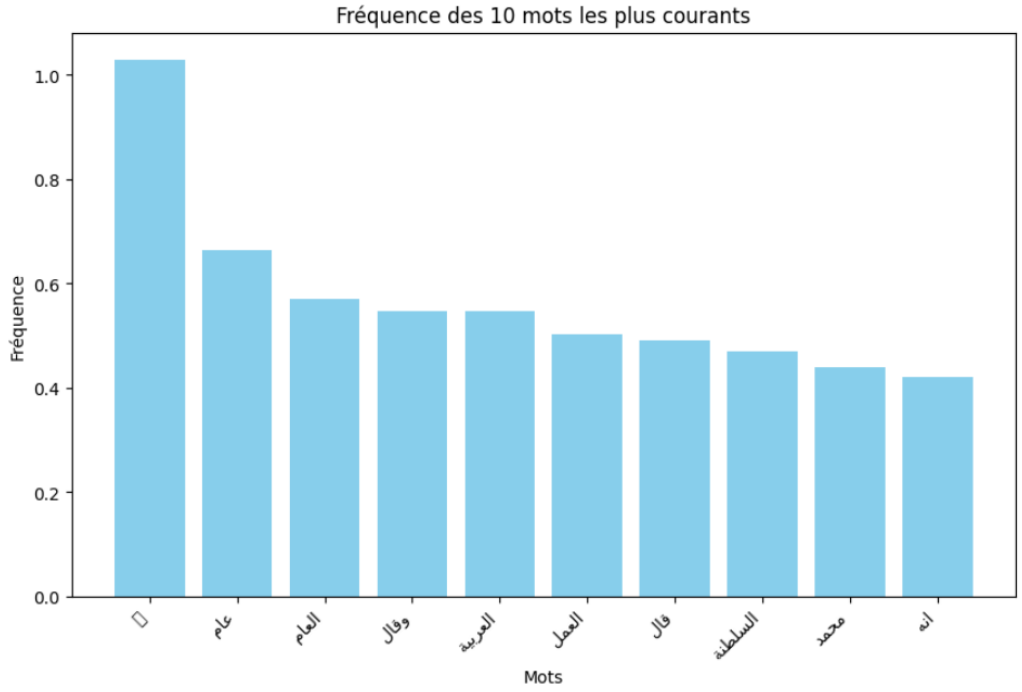
Les valeurs de cette matrice reflètent les poids TF-IDF attribués à chaque mot dans un document particulier. Des valeurs faibles, proches de zéro, suggèrent que le mot a une fréquence relativement faible dans le document ou qu'il est répandu dans tout le corpus, n'apportant aucune contribution distinctive à ce document spécifique. En revanche, des valeurs plus élevées indiquent que le mot est spécifique à ce document, jouant potentiellement un rôle clé dans sa caractérisation.

Notre matrice TF-IDF est composée d'un vaste ensemble de mots uniques (79 545 colonnes) extraits de notre corpus. En analysant attentivement les valeurs de cette matrice, vous pouvez identifier les mots qui ont le plus d'impact sur la définition du contenu de chaque article. La capture d'écran ci-jointe donne un aperçu de cette matrice, ce qui facilite la visualisation des poids TF-IDF et la compréhension des termes clés qui contribuent à l'unicité de chaque document.

	00	001	0010	002	003	004	0041	00450	00479	005	...	يتاح	يتدوتسى	يتفوتز	يته	يوهى	يوى	يند	يرا	يس	يقاء
0	0.000393	0.000000	0.000000	0.000011	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000154	0.000000
1	0.001648	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000029	0.000029	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000024	0.000000
2	0.003403	0.000056	0.000000	0.000092	0.000028	0.000046	0.000000	0.000000	0.000000	0.000084	...	0.000028	0.000028	0.000028	0.000028	0.000000	0.000000	0.000000	0.000000	0.000000	0.000028
3	0.001158	0.000000	0.000029	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	0.002521	0.000000	0.000000	0.000000	0.000000	0.000027	0.000032	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000032	0.000032	0.000032	0.000000	0.000000	0.000000
5	0.003219	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000051	0.000000	0.000000

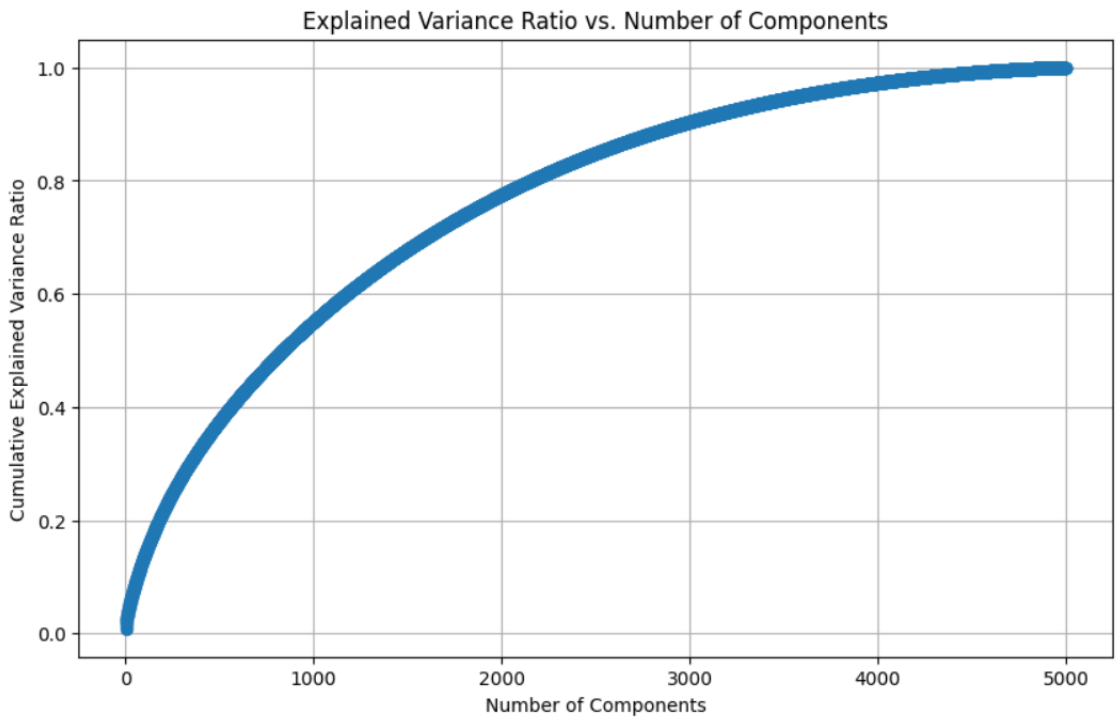
6 rows × 79545 columns

L'exploration approfondie du corpus d'articles arabes a permis de mettre en évidence les mots les plus fréquemment utilisés, avant même la suppression des mots vides. L'histogramme présenté donne un aperçu clair de la fréquence des dix mots les plus courants dans l'ensemble du corpus. Vous pouvez voir cette distribution de fréquence dans l'histogramme ci-dessus.



4-Data Reduction

-ACP :



Cette courbe, dérivée de l'analyse en composantes principales (ACP) appliquée à notre ensemble de données, offre un aperçu crucial de la manière dont les différentes composantes principales contribuent à expliquer la variabilité. Dérivée du code fourni, elle illustre graphiquement le pourcentage cumulé de variance expliquée en fonction du nombre de composantes principales.

Un examen attentif de cette courbe révèle que, dès les premières composantes, une croissance rapide indique une explication significative de la variabilité totale des données. Cependant, la courbe atteint un plateau autour du nombre de 5000 composantes, ce qui indique que l'ajout de composantes supplémentaires apporte des gains marginaux en termes d'explication de la variabilité, mais ne contribue pas de manière significative.

Le choix optimal du nombre de composantes à retenir est déterminé par l'identification du point où la courbe atteint un plateau. Cela indique que

l'ajout de composantes supplémentaires n'améliore pas sensiblement notre compréhension de la variabilité des données.

6.Classification:

-Classification SVM

-Le modèle de classification a démontré de solides performances avec une précision globale de 93,84 %. Cette mesure représente la proportion totale de prédictions correctes parmi toutes les prédictions effectuées. En d'autres termes, le modèle a correctement classé près de 94 % des instances de l'ensemble de données.

- ***Rapport de classification :***

Classification Report:

	precision	recall	f1-score	support
0	0.89	0.93	0.91	640
1	0.90	0.90	0.90	480
2	0.99	0.99	0.99	827
3	0.99	1.00	0.99	728
4	0.90	0.84	0.87	641
5	0.93	0.95	0.94	336
accuracy			0.94	3652
macro avg	0.93	0.93	0.93	3652
weighted avg	0.94	0.94	0.94	3652

cette rapport de classification fournit une analyse détaillée par classe :

Classe 0 : précision 89%, rappel 93% et score F1 91%. Le modèle a bien identifié cette classe, mais il y a eu quelques fausses alertes.
Classe 1 : précision de 90 %, rappel de 90 % et score F1 de 90 %. Une performance équilibrée, indiquant une bonne capacité à prédire cette classe.
Classe 2 : 99 % de précision, 99 % de rappel et 99 % de score F1. Excellente performance avec une identification précise de cette classe.
Classe 3 : 99 % de précision, 100 % de rappel et 99 % de score F1. Excellente performance avec l'identification précise de toutes les instances de cette classe.

Classe 4 : précision de 90 %, rappel de 84 % et score F1 de 87 %. Performance légèrement inférieure, indiquant une sensibilité réduite dans la prédiction de cette classe.
Classe 5 : précision de 93 %, rappel de 95 % et score F1 de 94 %. Solide performance avec une bonne identification de cette classe.

Matrice de confusion :

```
Confusion Matrix:
[[593   3   0   1  36   7]
 [  9 433   2   6  14  16]
 [  2   0 816   0   9   0]
 [  0   3   0 725   0   0]
 [ 55  36   6   4 540   0]
 [ 11   4   0   0   1 320]]
```

cette matrice de confusion offre une vue plus détaillée des performances par classe. Par exemple, la classe 4 présente un nombre plus élevé de faux négatifs, ce qui indique que le modèle a tendance à sous-estimer cette classe.

-Utilisation de modèle

Après l'application du modèle sur le nouveau texte, la catégorie prédite est "articlesEconomy". Cela signifie que, selon l'analyse du contenu du texte en arabe, le modèle a estimé que le sujet principal du texte est lié à l'économie.

Cette prédiction repose sur la compréhension du modèle des caractéristiques et des schémas dans les articles d'économie qu'il a appris pendant l'entraînement. La catégorie "articlesEconomy" a été choisie comme la catégorie la plus probable pour le nouveau texte, indiquant une correspondance significative avec les caractéristiques associées à cette catégorie.

-Classificateur KNN :

Après avoir appliqué le modèle KNN (K-Nearest Neighbors) à l'ensemble de données, les résultats de la classification montrent de solides performances avec une précision globale de 90,77 %. Examinons de plus près les principales mesures :

Précision globale

```
Accuracy with KNN: 90.77%
```

La précision globale mesure la proportion totale de prédictions correctes parmi toutes les prédictions effectuées. Dans ce cas, le modèle KNN a réussi à classer correctement près de 91 % des instances de l'ensemble de données de test.

Rapport de classification :

Classification Report:

	precision	recall	f1-score	support
0	0.85	0.93	0.89	640
1	0.91	0.78	0.84	480
2	0.98	0.97	0.98	827
3	0.91	0.99	0.95	728
4	0.86	0.81	0.83	641
5	0.91	0.91	0.91	336
accuracy			0.91	3652
macro avg	0.91	0.90	0.90	3652
weighted avg	0.91	0.91	0.91	3652

Les résultats par classe indiquent les performances du modèle pour chaque catégorie spécifique.

Classe 0 : Précision 85%, Rappel 93% et Score F1 89%. Le modèle a correctement identifié cette classe avec une précision convenable.

Classe 1 : précision de 91 %, rappel de 78 % et score F1 de 84 %. Précision élevée, mais sensibilité légèrement inférieure dans la prédiction de cette classe.

Classe 2 : 98 % de précision, 97 % de rappel et 98 % de score F1. Excellente performance avec une identification précise de cette classe.

Classe 3 : précision de 91 %, rappel de 99 % et score F1 de 95 %. Précision élevée et sensibilité exceptionnelle pour cette classe.

Classe 4 : précision 86 %, rappel 81 % et score F1 83 %. Performance décente, mais avec une légère baisse de la précision et du rappel.

Classe 5 : Précision 91%, Rappel 91% et Score F1 91%. Bonne performance de prédiction dans cette classe.

Matrice de confusion :

```
Confusion Matrix:
[[594   1   0   7  33   5]
 [ 13 375   5  43  25  19]
 [  2   0 802   2  17   4]
 [  0   4   0 719   5   0]
 [ 74  25   6  15 520   1]
 [ 14   5   3   4   5 305]]
```

La matrice de confusion fournit des détails sur les vrais positifs, les faux positifs, les vrais négatifs et les faux négatifs pour chaque classe. Elle montre que le modèle a réussi à classer la majorité des cas, mais qu'il a rencontré des difficultés pour prédire certaines classes.

-Utilisation de modèle

Après avoir appliqué le modèle KNN au nouveau texte arabe, la catégorie prédite est "articlesReligion". Cela signifie que, sur la base du contenu du texte, le modèle a estimé que le sujet principal était lié aux aspects religieux.

-Utilisation de Random Forest :

Après avoir appliqué le modèle Random Forest à l'ensemble de données, nous avons obtenu des résultats encourageants en termes de précision, de rappel et de score F1 pour chaque catégorie. Ces résultats démontrent l'efficacité du modèle pour la classification des textes arabes.

-Précision Globale

```
Accuracy with Random Forest: 92.06%
```

La précision globale du modèle est élevée, ce qui indique sa capacité à classer correctement la grande majorité des instances de l'ensemble de données.

-Performances par catégorie :

```
Classification Report (Random Forest):
              precision    recall  f1-score   support

           0       0.87      0.90      0.88       640
           1       0.90      0.86      0.88       480
           2       0.98      0.97      0.98       827
           3       0.98      1.00      0.99       728
           4       0.84      0.84      0.84       641
           5       0.94      0.92      0.93       336

    accuracy                           0.92      3652
   macro avg       0.92      0.91      0.92      3652
weighted avg       0.92      0.92      0.92      3652
```

La catégorie "Articles religieux" a obtenu des résultats particulièrement bons, avec une précision de 98 %, un rappel de 100 % et un score F1 de 99 %. Cela suggère que le modèle est très compétent pour détecter les aspects religieux dans les textes arabes.

Les autres catégories ont également affiché de bonnes performances, avec des scores de précision, de rappel et de F1 élevés, ce qui souligne la robustesse du modèle dans la classification multi-catégorielle.

-Analyse de confusion :

```
Confusion Matrix (Random Forest):
[[574   5   0   1  51   9]
 [ 15 411   6   8  31   9]
 [  2   1 805   0  18   1]
 [  0   3   0 725   0   0]
 [ 55  30  12   6 538   0]
 [ 14   6   1   0   6 309]]
Predicted Category (Random Forest): articlesReligion
```

La matrice de confusion nous montre que le modèle a eu quelques difficultés à distinguer certaines catégories, notamment entre "Articles locaux" et "Articles de sport". Cela peut être dû à des similitudes thématiques entre ces catégories.

-Utilisation de la régression logistique :

Les résultats obtenus après l'application de l'algorithme de régression logistique à l'ensemble de données sont très encourageants. Le modèle a démontré une précision globale de 93,54%, soulignant son efficacité dans la classification des textes arabes en différentes catégories.

Précision globale : 93,54

```
Accuracy with Logistic Regression: 93.54%
```

La précision globale du modèle est élevée, ce qui indique sa capacité à classer correctement la grande majorité des instances de l'ensemble de données.

Performance par catégorie :

```
Classification Report (Logistic Regression):
              precision    recall  f1-score   support

           0       0.89      0.92      0.91       640
           1       0.90      0.89      0.90       480
           2       0.99      0.98      0.99       827
           3       0.97      0.99      0.98       728
           4       0.89      0.85      0.87       641
           5       0.94      0.95      0.95       336

    accuracy                           0.94      3652
   macro avg       0.93      0.93      0.93      3652
weighted avg       0.94      0.94      0.94      3652
```

Les performances par catégorie sont équilibrées, avec des scores élevés pour la précision, le rappel et le score F1. En particulier, la catégorie "Articles religieux" a été identifiée avec une précision de 97 %, un rappel de 99 % et un score F1 de 98 %.

Analyse de confusion :

```
Confusion Matrix (Logistic Regression):
[[588   5   1   1  38   7]
 [  8 429   2  12  16  13]
 [  3   0 812   1  11   0]
 [  1   3   0 722   2   0]
 [ 51  35   4   5 546   0]
 [  8   5   1   1   2 319]]
```

La matrice de confusion révèle que le modèle a eu quelques difficultés à distinguer certaines catégories, en particulier entre les "articles locaux" et les "articles de sport". Ceci peut être attribué aux similitudes thématiques entre ces catégories.