



# PROJET TRAITEMENT DE DONNEES

Préparé par :

**Guiza Asma**

**Yosra Omran**

**Fatma Bouzgarrou**

**Meriam Maâtallah**

**2023-2024**



### 1. Présentation Générale du projet :

#### 1.1. Description de projet :

Le processus de catégorisation de texte en groupes organisés est connu sous le nom de classification de texte « data mining », également appelée étiquetage de texte ou catégorisation de texte. La classification de texte en fouille de données peut automatiquement analyser le texte et attribuer un ensemble de tags ou catégories prédéfinis en fonction de son contenu en utilisant le traitement du langage naturel (NLP).

#### 1.2. Objectifs :

Les objectifs primordiaux de notre projet s'articulent autour de la volonté de doter les chercheurs, développeurs et passionnés de langues d'un outil robuste dédié à l'analyse de données en langue arabe. Nous ambitionnons de concrétiser les points suivants :

- **Classification des Catégories** : Implémenter un système de classification de textes capable d'attribuer automatiquement chaque article à l'une des six catégories prédéfinies : culture, économie, actualités locales, actualités internationales, religion et sports.
- **Compréhension Améliorée** : Faciliter une appréhension approfondie de la langue arabe à travers des articles diversifiés couvrant les domaines culturels, économiques, locaux, internationaux, religieux et sportifs.
- **Le nettoyage des données** : est une étape cruciale visant à préparer les données pour l'analyse et la modélisation. Cette phase implique le traitement des données brutes afin de les rendre plus cohérentes, complètes et adaptées à l'utilisation dans les modèles d'apprentissage automatique.

#### 1.3. Contexte :

À l'ère actuelle, où le paysage informationnel est de plus en plus dominé par les données, l'impératif d'outils efficaces pour l'analyse et la compréhension de l'information dans des langues diverses est incontournable. L'arabe, en tant que langue à la fois riche et complexe, offre des défis stimulants et des opportunités uniques dans le domaine de l'analyse de données. L'avènement de KALIMAT, un Corpus Arabe Polyvalent, répond de manière proactive à cette nécessité en mettant à disposition une ressource complète dédiée au traitement naturel du langage en arabe

### 1.4. Méthodologie :

Méthode CRISP (Cross-Industry Standard Process for Data Mining) : La méthode CRISP est une approche standardisée en data mining qui se compose de plusieurs étapes bien définies pour résoudre des problèmes de traitement et d'analyse de données. CRISP-DM est un processus composé de six phases différentes :

**-Compréhension du domaine :** Acquisition d'une connaissance approfondie du domaine de problème et de ses besoins.

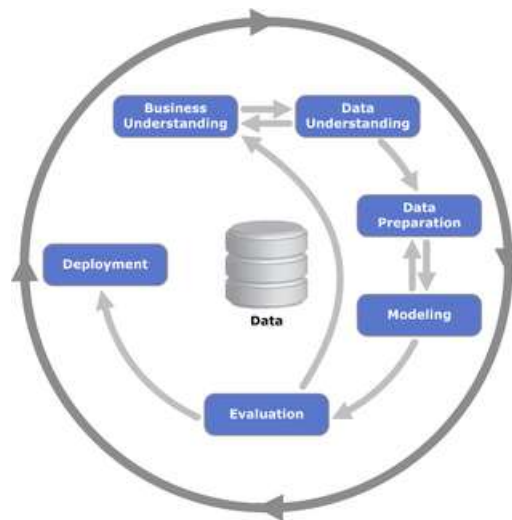
**-Collecte de données :** Rassemblement des données pertinentes pour l'analyse.

**-Préparation des données :** Nettoyage, transformation et structuration des données.

**-Modélisation :** Sélection et application de techniques de modélisation pour découvrir des modèles.

**-Évaluation :** Évaluation des modèles pour leur performance et leur adéquation.

**-Déploiement :** Mise en œuvre des résultats du modèle dans l'environnement opérationnel.



## 2. Les ensembles des données :

<b>Origine des Données</b>	Les données utilisées proviennent du corpus KALIMAT 1.0, une ressource linguistique arabe polyvalente.
<b>Méthode de Collecte</b>	Les données ont été extraites à partir de 20 291 articles du journal omanais Alwatan (Abbas et al., 2011). Chaque article a été annoté et étiqueté manuellement pour six catégories principales : culture, économie, actualités locales, actualités internationales, religion et sports.
<b>Référence</b>	Abbas, M., Guellil, I., Belalem, G., & Rosso, P. (2011). KALIMAT at CLEF 2011: Arabic Information Retrieval Experiments using UMA-PHRASEBOOK. CLEF (Notebook Papers/LABs/Workshops), 138-143.

### 3. Description des données :

<u>Culture :</u>	Cette catégorie englobe des articles embrassant un large éventail de sujets liés à la culture, aux arts, à la littérature, à la musique, au cinéma, ainsi qu'à d'autres aspects de notre patrimoine culturel. Elle peut comprendre des critiques, des comptes rendus d'événements culturels, des entretiens avec des artistes, et des analyses approfondies de divers aspects culturels.
<u>Sport :</u>	La catégorie dédiée aux articles sportifs explore une diversité de disciplines sportives, allant des résultats des matchs aux performances individuelles des athlètes, en passant par les analyses tactiques et les temps forts des grands événements sportifs. Elle inclut également des entretiens avec des athlètes, des rapports détaillés sur les matchs, ainsi que des actualités concernant les équipes, offrant ainsi une couverture

	complète de l'univers sportif.
<u>International :</u>	Les actualités internationales englobent les événements et les évolutions qui se déroulent à l'échelle mondiale. Cette catégorie s'étend à des rapports sur les affaires internationales, les conflits, les accords diplomatiques, les nouvelles mondiales, ainsi que des analyses approfondies des relations internationales. Elle offre ainsi une couverture complète des enjeux et des développements qui façonnent la scène mondiale.
<u>Economie</u>	Les articles liés à l'économie traitent des sujets financiers, des marchés, des entreprises, de l'emploi, des politiques économiques, et d'autres aspects liés aux activités économiques. L'analyse des tendances économiques, des rapports financiers, et des entrevues avec des experts peuvent être inclus.
<u>Local</u>	Cette catégorie met l'accent sur les événements et les actualités se déployant au niveau local. Elle englobe un large spectre d'informations, comprenant des nouvelles communautaires, des faits divers, des annonces gouvernementales, des reportages sur la vie quotidienne locale, ainsi que la couverture d'événements et d'initiatives spécifiques à la région.
<u>Religion :</u>	Cette catégorie explore des sujets touchant à la religion, aux pratiques spirituelles, aux événements religieux, aux enseignements, ainsi qu'aux discussions entourant la foi. Elle englobe également des analyses théologiques et des informations relatives aux communautés religieuses, offrant ainsi une perspective approfondie sur les dimensions spirituelles et philosophiques.

### 4. Réalisation :

Dans cette partie, nous allons détailler la mise en œuvre du projet, depuis la collecte des données jusqu'à l'application d'un modèle de classification et l'analyse des résultats.

#### 4.1. Lecture et Prétraitement des données :

Le processus de réalisation commence par la lecture des données brutes à partir de fichiers texte stockés dans des dossiers organisés par catégories. Les principales étapes de cette phase sont :

- **Chargement des Ressources NLP** : Les ressources nécessaires de la bibliothèque NLTK sont téléchargées, comprenant les modules de tokenization, les stopwords, ainsi que les stemmer et lemmatiser spécifiques à l'arabe.
- **Parcours des Fichiers** : Les fichiers texte sont parcourus dans une structure de dossiers organisée par catégories. Seuls les fichiers avec l'extension ".txt" sont pris en compte.
- **Extraction de Contenu** : Pour chaque fichier, seul un échantillon (10%) de son contenu est extrait pour des raisons de traitement plus rapide.
- **Nettoyage du Contenu** : Les caractères de nouvelle ligne (\n), les chiffres, et la ponctuation sont supprimés du contenu.
- **Tokenization** : Le contenu est tokenisé en mots individuels.
- **Suppression des Stopwords** : Les stopwords (mots courants sans grande signification) dans la langue arabe sont retirés.
- **Stemming et Lemmatization** : Les mots sont soumis à un processus de stemming pour réduire à leur racine, puis à la lemmatization pour les ramener à leur forme canonique.
- **Construction de la DataFrame** : Les données prétraitées, y compris le contenu, la catégorie, et les tokens, sont stockées dans une liste pour être ultérieurement converties en une DataFrame.



	Catégorie	Nom du Fichier \
0	articlesCulture	culturecapr1.txt
1	articlesCulture	culturecapr1005.txt
2	articlesCulture	culturecapr1006.txt
3	articlesCulture	culturecapr1007.txt
4	articlesCulture	culturecapr1008.txt

0	لم	الرحبي	تنطلق	اليوم	الدورة	البرامجية
1	صل	العلوي	شاركت	السلطنة	صباح	امس دول ا
2	عروض	على	مسرح	الشباب	وعرض	في الرستاق ثم
3	الد	عبداللطيف	حين	يناقش	الموضوع	الثقاف
4	صباح	أمس	بقاعة	الموسيقى	في	جامعة السلطان

0	الم	الرحبي	تنطلق	اليوم	الدورة	البر
1	يصل	العلوي	شاركت	السلطنة	صباح	امس
2	عروض	على	مسرح	الشباب	وعرض	في الر
3	ه	خالد	عبداللطيف	حين	يناقش	الموضوع
4	و صباح	أمس	بقاعة	الموسيقى	في	جامعة

tokens\_lemmat

### 4.2. Analyse des Données Textuelles :

Après la construction de la DataFrame, différentes analyses sont effectuées pour comprendre la nature des données textuelles :

## Projet de traitement de données

- **Statistiques Globales** : Le nombre total de mots dans l'ensemble du corpus est calculé, ainsi que le nombre total de mots dans chaque document.

```
Nombre total de mots : 1385009
Nombre total de mots dans chaque document 0
1      130
2      105
3      295
4       55
...
18251   32
18252   25
18253   36
```

- **Analyse par Catégorie** : Des statistiques sont générées, montrant le nombre total de mots par catégorie.

```
Nombre total de mots par catégorie Ca
articlesCulture      195551
articlesEconomy      229760
articlesInternational 111437
articlesLocal        238476
articlesReligion     401954
```

- **Mots Fréquents** : Les mots les plus fréquents dans l'ensemble du corpus sont identifiés.

```
Nombre total de mots distincts 17561
Mots les plus fréquents 12825
11833      علم
11403      جمع
...
1      بنه‌ایه شهر
1      کت
```



## Projet de traitement de données

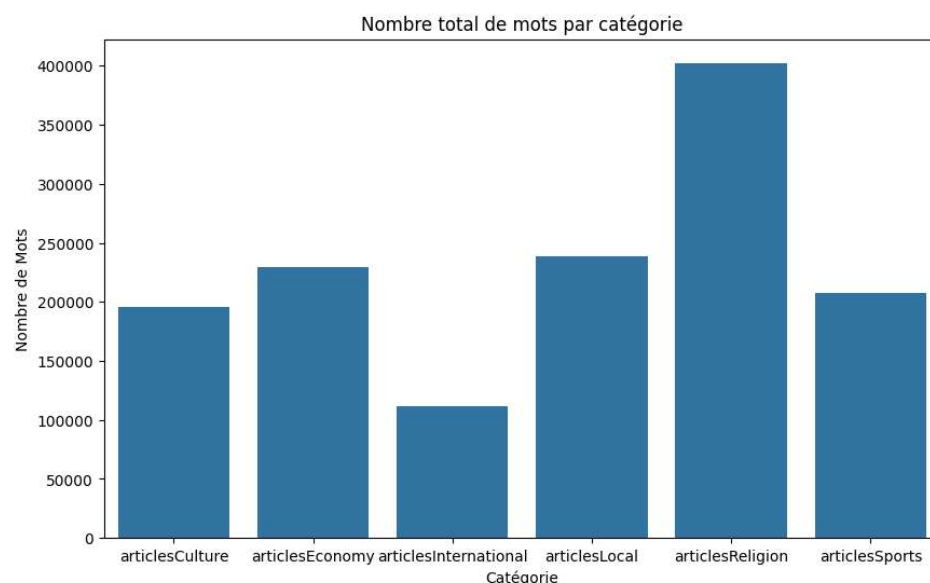
- **Mots Fréquents par Catégorie :** Les mots les plus fréquents et les moins fréquents sont identifiés pour chaque catégorie.

Mots les plus fréquents par catégorie :		
	Catégorie	Mot le plus fréquent
0	articlesCulture	كتب
1	articlesEconomy	عمل
2	articlesInternational	ان
3	articlesLocal	بن
4	articlesReligion	ال
5	articlesSports	نخب

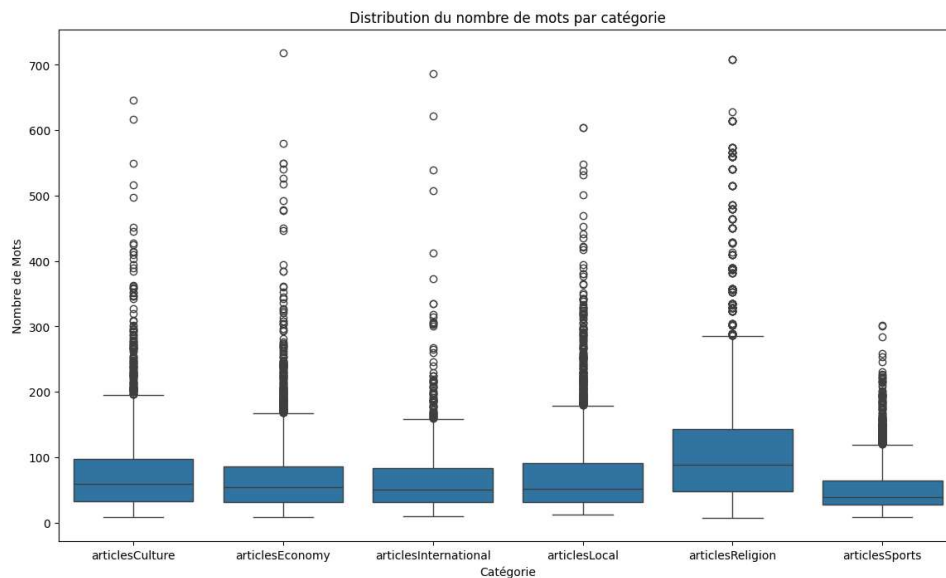
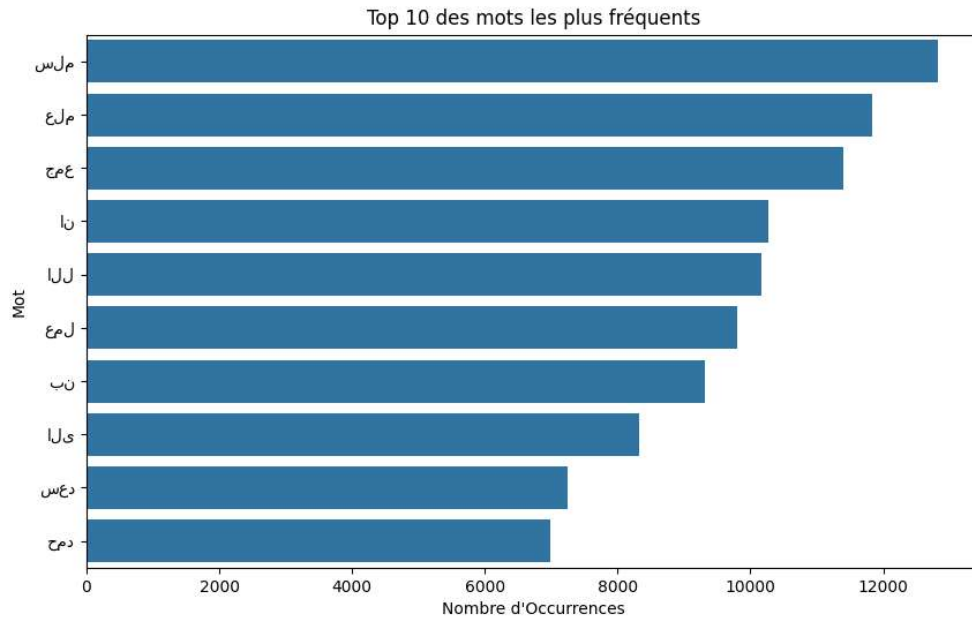
  

Mots les moins fréquents par catégorie :		
	Catégorie	Mot le moins fréquent
0	articlesCulture	بذ
1	articlesEconomy	ووض
2	articlesInternational	رنور
3	articlesLocal	بتغ
4	articlesReligion	تيق
5	articlesSports	دور الاول

- **Visualisations :** Des graphiques à barres et des boîtes à moustaches sont utilisés pour visualiser le nombre total de mots par catégorie et explorer la distribution des mots.



## Projet de traitement de données



### 4.3. Traitement des mots fréquents :

- **Fréquences et Seuil :** Les fréquences de chaque mot sont calculées, et un seuil de fréquence est spécifié (1000 fois). Les mots qui dépassent ce seuil sont identifiés comme mots fréquents.

	Catégorie	Nom du Fichier \
0	articlesCulture	culturecapr1.txt
1	articlesCulture	culturecapr1005.txt
2	articlesCulture	culturecapr1006.txt
3	articlesCulture	culturecapr1007.txt
4	articlesCulture	culturecapr1008.txt

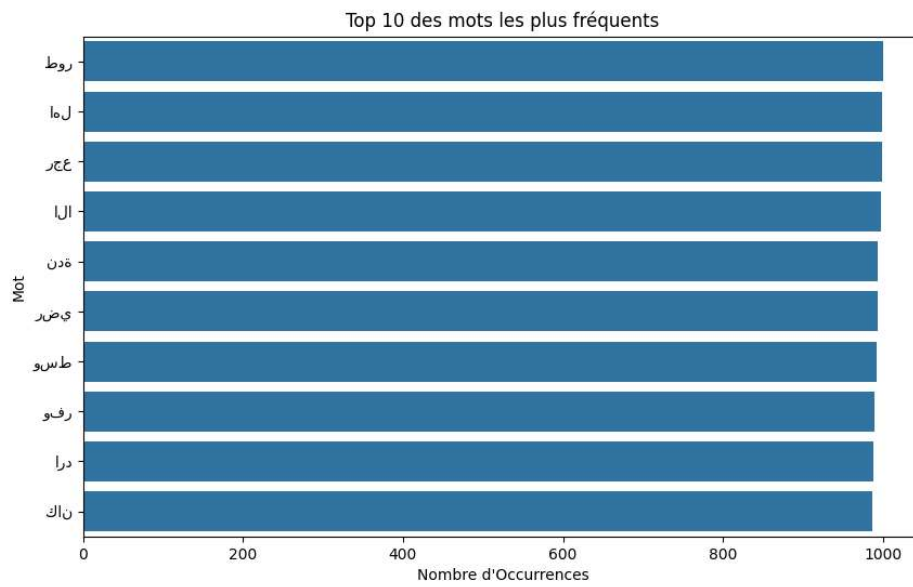
  

	tokens_sans_mots_fr
0	مچ و یقزو ذعة و رنمچ و سمر و طول و برل و ایو [
1	م و ترث و اقم و قرم و رعی و عنتر و ترث و احتف [
2	نزی حقلت و حواله و ذعة و سمر و طول و برل و ایو [

- **Suppression des Mots Fréquents :** Les mots fréquents sont retirés de chaque liste de tokens\_lemmatization.

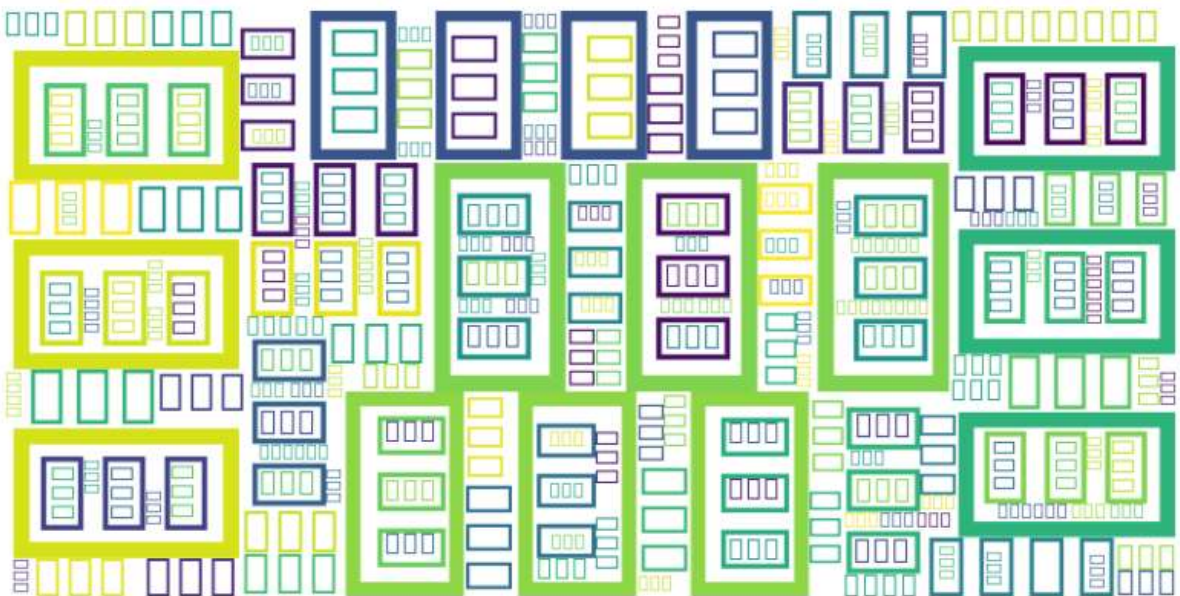
Nombre total de mots dans la colonne 'tokens_sans_mots_freq'	
1000	طور
999	اهل
999	رجع
997	الا
994	ندة
...	
1	كشمول
1	وأعتبروه
1	مستشاريه
1	لاستخدامه

## Projet de traitement de données



### 4.4. Visualisation avec WordCloud :

Un nuage de mots est généré à partir des tokens débarrassés des mots fréquents. Cette visualisation donne une idée des termes les plus importants dans l'ensemble du corpus.



### 4.5. Vectorisation avec TF-IDF:

**Construction de la Matrice TF-IDF :** Les données sont vectorisées en utilisant la méthode TF-IDF pour représenter numériquement l'importance des mots dans chaque document par rapport à l'ensemble du corpus.

```
Nombre d'attributs TF-IDF : 17267
Pourcentage de zéros dans la matrice TF-IDF : 99.82
```

### 4.6. Réduction de dimensionnalité avec PCA:

**Réduction en 50 Composantes Principales :** Une analyse en composantes principales (PCA) est appliquée pour réduire la dimensionnalité à 50 composantes principales, conservant l'essentiel de l'information tout en réduisant la complexité.

	PC1	PC2	PC3	PC4	PC5	PC6
0	-0.020996	0.024307	-0.100189	0.026325	-0.024400	-0.046536
1	-0.016432	0.030055	-0.049197	0.009919	0.037462	0.003502
2	-0.024972	0.013533	-0.021140	-0.000118	-0.032753	0.050949
3	-0.014678	-0.048745	-0.087587	0.013977	0.004911	-0.004931
4	-0.013548	0.011841	-0.026140	0.000737	-0.002506	-0.004563

	PC8	PC9	PC10	...	PC41	PC42	PC43
0	-0.072773	-0.023236	0.036045	...	-0.064382	0.031409	0.032831
1	-0.007571	-0.006185	-0.010107	...	-0.005005	0.002906	-0.031730
2	-0.044540	-0.065351	-0.004315	...	-0.047631	0.006639	-0.012370
3	0.010695	-0.032444	-0.036569	...	-0.018744	0.023973	-0.007998
4	-0.015188	0.001749	-0.014129	...	0.010683	0.032243	-0.014072

	PC45	PC46	PC47	PC48	PC49	PC50
0	-0.018854	0.071499	-0.067046	-0.044187	0.021331	-0.075296
1	0.001607	-0.010492	-0.031858	-0.024790	-0.023857	-0.065172
2	0.010762	0.021090	-0.010441	0.022407	0.014342	0.001234

```
Indice de silhouette : 0.024805748917039
```

	PC1	PC2	PC3	PC4	PC5	PC6
0	-0.020996	0.024307	-0.100189	0.026325	-0.024400	-0.046536
1	-0.016432	0.030055	-0.049197	0.009919	0.037462	0.003502
2	-0.024972	0.013533	-0.021140	-0.000118	-0.032753	0.050949
3	-0.014678	-0.048745	-0.087587	0.013977	0.004911	-0.004931
4	-0.013548	0.011841	-0.026140	0.000737	-0.002506	-0.004563

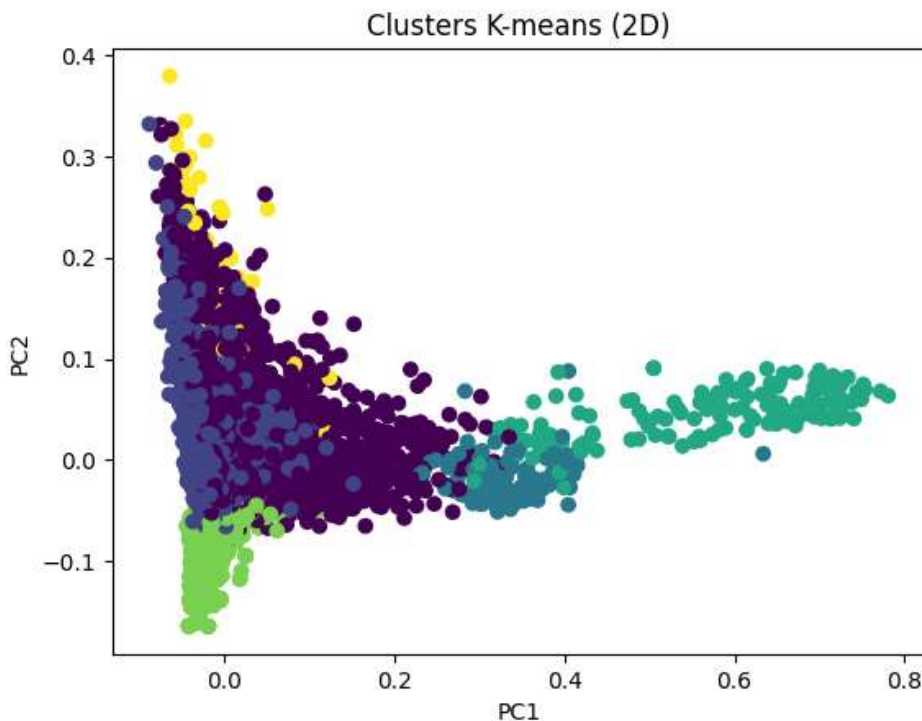
	PC8	PC9	PC10	...	PC42	PC43	PC44
0	-0.072773	-0.023236	0.036045	...	0.031409	0.032831	0.007600
1	-0.007571	-0.006185	-0.010107	...	0.002906	-0.031730	-0.029100
2	-0.044540	-0.065351	-0.004315	...	0.006639	-0.012370	-0.018900
3	0.010695	-0.032444	-0.036569	...	0.023973	-0.007998	0.047400
4	-0.015188	0.001749	-0.014129	...	0.032243	-0.014072	0.000900

	PC46	PC47	PC48	PC49	PC50	Cluster
0	0.071499	-0.067046	-0.044187	0.021331	-0.075296	0
1	-0.010492	-0.031858	-0.024790	-0.023857	-0.065172	0
2	0.021090	-0.010441	0.022407	0.014342	0.001234	0

### 4.7. Clustering avec K-means :

- **Choix du Nombre de Clusters :** Le nombre de clusters est spécifié à 6.
- **Application du K-means :** Le modèle K-means est entraîné sur les données réduites par PCA.
- **Indice de Silhouette :** L'indice de silhouette est calculé pour évaluer la cohérence des clusters formés.
- **Visualisation des Clusters :** Une visualisation en deux dimensions des clusters est réalisée en utilisant les deux premières composantes principales de la PCA.



Cluster	Dominant_Category
0	articlesReligion
1	articlesLocal
2	articlesEconomy
3	articlesSports
4	articlesLocal
5	articlesEconomy

#### **Interprétation :**

D'après cette visualisation et cette exploration, on remarque que le cluster le mieux séparé est celui regroupant les articles sur l'économie et les articles sur le sport, tandis que le cluster des articles sportifs et celui des articles économiques ne sont pas bien distingués.



### 4.8. NMF :

L'algorithme NMF (Non-Negative Matrix Factorization) utilisé pour extraire des thèmes à partir de documents (exploration des termes les plus importants par thème)

```
Theme 0:
قرء رحم لأن خطب سيج امة زوج علماء

Theme 1:
الة قرن ائة داول ورق رجب سند كمة

Theme 2:
عرف رعى عضء رتمج هيئ دئر ربو وفد

Theme 3:
ل جول ونس ائي صفر بري قطر صين رصد

Theme 4:
وؤشر بيس سعد خقض بنك ائة رجع اما
```

Calculer la divergence KL entre les distributions de probabilité des termes :

La divergence KL (Kullback-Leibler) est une mesure de la différence entre deux distributions de probabilités

```
Theme 0: KL Divergence = -990.8250563568463
Theme 1: KL Divergence = -109.67365766306963
Theme 2: KL Divergence = -449.26673543019973
Theme 3: KL Divergence = -424.80662327014835
Theme 4: KL Divergence = -103.57830809948625
Theme 5: KL Divergence = -146.89649356487155
Theme 0: Top Terms = و رحم و لأن و خطب و سيج و امة و زوج و علماء
Theme 1: Top Terms = و قرن و ائة و داول و ورق و رجب و سند و كمة
Theme 2: Top Terms = و رعى و عضء و رتمج و هيئ و دئر و ربو و وفد
Theme 3: Top Terms = ل و ونس و ائي و صفر و بري و قطر و صين و رصد
```



### Interprétation :

**Theme 0** : KL Divergence : -990.8250563568456

La divergence KL très négative indique une bonne correspondance entre la distribution des termes dans ce thème et la distribution réelle dans les documents. Cela suggère que le **thème est bien défini**.

**Theme 1** : KL Divergence : -109.67365766306943

Une divergence KL négative, bien que moins importante que pour le thème 0, indique toujours une correspondance raisonnable entre la distribution des termes dans le thème et la distribution réelle.

**Theme 2** : KL Divergence : -449.26673543019933

La divergence KL négative suggère une correspondance, mais la valeur moins élevée que pour le thème 0 peut indiquer une certaine variabilité ou ambiguïté dans ce thème.

**Theme 3** : KL Divergence : -424.8066232701478

Une divergence KL négative indique une correspondance, mais la valeur moins élevée que pour le thème 0 suggère une variabilité ou ambiguïté dans ce thème.

**Theme 4** : KL Divergence : -103.57830809948592

La divergence KL négative indique une correspondance, bien que la valeur soit moins élevée que pour le thème 0.

**Theme 5** : KL Divergence : -146.89649356487237

La divergence KL négative indique une correspondance, mais la valeur est moins élevée que pour le thème 0.

### 4.9. Apprentissage Supervisé : KNN :

Précision du modèle KNN : 0.7719058050383352

Rapport de Classification :

	precision	recall	f1-score	s
articlesCulture	0.77	0.54	0.64	
articlesEconomy	0.72	0.72	0.72	
articlesInternational	0.80	0.71	0.75	
articlesLocal	0.60	0.77	0.68	
articlesReligion	0.88	0.99	0.93	
articlesSports	0.91	0.79	0.85	
accuracy			0.77	

- La précision globale du modèle est d'environ 0.77, ce qui signifie que le modèle est correct dans 77% des prédictions sur l'ensemble de test.
- En analysant chaque classe :
  - Pour la classe "articlesCulture", la précision est de 0.77, ce qui indique que 77% des échantillons prédits comme "articlesCulture" le sont effectivement.
  - Pour la classe "articlesEconomy", la précision est de 0.72.
  - Pour la classe "articlesInternational", la précision est de 0.80.
  - Pour la classe "articlesLocal", la précision est de 0.60.
  - Pour la classe "articlesReligion", la précision est de 0.88.
  - Pour la classe "articlesSports", la précision est de 0.91.
- Le rappel (recall) est également une métrique importante. Il mesure la capacité du modèle à trouver tous les échantillons positifs. Un rappel élevé indique que le modèle parvient à identifier la plupart des occurrences réelles de la classe
- 

### 4.10. Apprentissage Supervisé : Cross-Validation :

```
Scores de validation croisée : [0.76834611 0.78033416 0.77704738 0.77704738]
Moyenne des scores de validation croisée : 0.7772244127027269
```

La validation croisée est une technique essentielle pour évaluer les performances d'un modèle de machine learning sur plusieurs sous-ensembles de données. Dans ce cas le KNN :

Interprétation des résultats :

- Les scores de validation croisée pour chaque itération sont [0.768, 0.780, 0.777, 0.777, 0.783].
- La moyenne des scores de validation croisée est d'environ 0.777.

Cela signifie que, en moyenne, le modèle KNN a une précision d'environ 77.7% sur l'ensemble de données lorsqu'il est évalué avec la validation croisée. La validation croisée permet de réduire les effets de la variabilité des données d'entraînement et fournit une évaluation plus robuste des performances du modèle.

### 4.11. Apprentissage Supervisé : GridSearch :

```
Meilleurs hyperparamètres : {'n_neighbors': 7, 'weights':  
Meilleure précision : 0.8645569635094004  
Précision sur l'ensemble de test : 0.8707557502738226
```

Interprétation des résultats :

1. **Meilleurs hyperparamètres** : Les meilleurs hyperparamètres trouvés par la recherche sur grille sont {'n\_neighbors': 7, 'weights': 'distance'}. Cela signifie que le modèle KNN optimal utilise 7 voisins (k=7) et attribue un poids inversement proportionnel à la distance lors de la prise de décision.
2. **Meilleure précision sur l'ensemble d'entraînement** : La meilleure précision obtenue lors de la validation croisée avec les meilleurs hyperparamètres est d'environ 86.5%. Cela représente la précision moyenne sur les ensembles d'entraînement lors de la recherche sur grille.
3. **Précision sur l'ensemble de test** : Après avoir trouvé les meilleurs hyperparamètres, le modèle est évalué sur un ensemble de test distinct. La précision sur cet ensemble de test est d'environ 87.1%.

**Interprétation globale** : Le modèle KNN avec les hyperparamètres optimisés semble bien généraliser aux données non vues, comme indiqué par la précision d'environ 87.1% sur l'ensemble de test. Cela suggère que le modèle est capable de faire des prédictions précises sur de nouvelles données.

### 4.12. Apprentissage Supervisé : Réseau de neurone:

```
366/366 [=====] - 58s 148ms/step - loss: 0.9463 - accuracy: 0.6347 - val_loss: 0.530  
Epoch 2/5  
366/366 [=====] - 54s 147ms/step - loss: 0.3256 - accuracy: 0.8888 - val_loss: 0.342  
Epoch 3/5  
366/366 [=====] - 51s 139ms/step - loss: 0.1380 - accuracy: 0.9589 - val_loss: 0.333  
Epoch 4/5  
366/366 [=====] - 53s 146ms/step - loss: 0.0695 - accuracy: 0.9807 - val_loss: 0.354  
Epoch 5/5  
366/366 [=====] - 64s 175ms/step - loss: 0.0361 - accuracy: 0.9908 - val loss: 0.370
```

Voici une interprétation des résultats :

1. **Entraînement du modèle** :
  - Le modèle est entraîné sur 5 époques.
  - La précision sur l'ensemble d'entraînement augmente progressivement au fil des époques, atteignant une précision finale d'environ 99.1%.

## Projet de traitement de données

- La perte (loss) sur l'ensemble d'entraînement diminue également, indiquant une bonne convergence du modèle.

### 2. Validation du modèle :

- Pendant l'entraînement, le modèle est validé sur un sous-ensemble de validation (20% de l'ensemble d'entraînement).
- La précision sur l'ensemble de validation atteint environ 89.8% à la dernière époque.

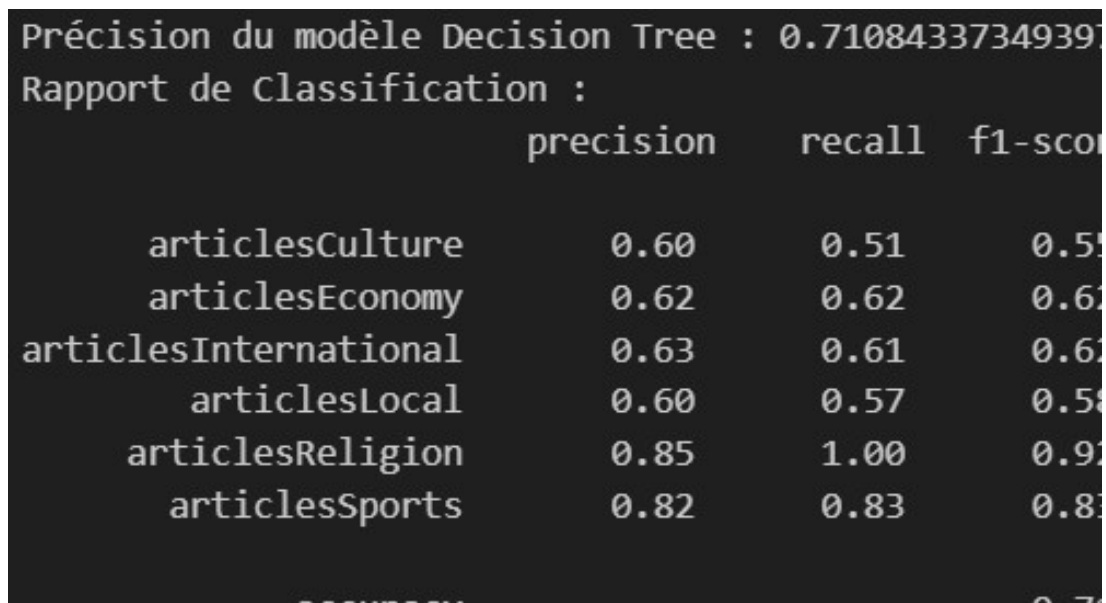
### 3. Évaluation sur l'ensemble de test :

- Une fois l'entraînement terminé, le modèle est évalué sur l'ensemble de test.
- La précision sur l'ensemble de test est d'environ 88.9%.

### Interprétation globale :

- Le modèle a une excellente performance sur l'ensemble d'entraînement, mais une précision légèrement inférieure sur l'ensemble de test, indiquant une possible surajustement (overfitting).
- La précision sur l'ensemble de test de 88.9% est néanmoins très respectable, suggérant que le modèle généralise bien à de nouvelles données.

## 4.13. Apprentissage Supervisé : DecisionTree:



```
Précision du modèle Decision Tree : 0.71084337349397
Rapport de Classification :
```

	precision	recall	f1-score
articlesCulture	0.60	0.51	0.55
articlesEconomy	0.62	0.62	0.62
articlesInternational	0.63	0.61	0.62
articlesLocal	0.60	0.57	0.58
articlesReligion	0.85	1.00	0.92
articlesSports	0.82	0.83	0.83

accuracy 0.71

Les résultats du modèle de l'arbre de décision montrent une précision globale (accuracy) de 71.08%, ce qui indique que le modèle a correctement classé environ 71% des échantillons de test. En

examinant les métriques par classe, on peut voir que la précision, le recall et le f1-score varient d'une classe à l'autre.

- La classe "articlesReligion" a une performance très élevée avec une précision de 85%, un recall de 100% et un f1-score de 92%. Cela signifie que le modèle a bien performé pour cette classe, mais il est important de noter que ces résultats peuvent être dus à un déséquilibre de classe (beaucoup plus d'échantillons pour cette classe par rapport aux autres).
- Les classes "articlesEconomy" et "articlesInternational" ont des performances similaires avec des précisions, recalls et f1-scores autour de 60-63%. Ces classes semblent être un peu plus difficiles à prédire que la classe "articlesReligion".
- La classe "articlesSports" a une bonne performance avec une précision de 82%, un recall de 83% et un f1-score de 83%.
- Les classes "articlesCulture" et "articlesLocal" ont des performances légèrement inférieures avec des précisions d'environ 60% et des recalls d'environ 50-57%.

```
Règles de l'arbre de décision :
|--- 0.06 => كرة
|   |--- 0.04 => دنا
|   |   |--- 0.06 => لقب
|   |   |   |--- 0.02 => سعر
|   |   |   |   |--- 0.03 => علماء
|   |   |   |   |   |--- 0.07 => اثني
|   |   |   |   |   |   |--- 0.07 => اقتصادية
|   |   |   |   |   |   |   |--- 0.04 => اله
|   |   |   |   |   |   |   |   |--- 0.08 => صفر
|   |   |   |   |   |   |   |   |   |--- 0.04 => ربو
|   |   |   |   |   |   |   |   |   |   |--- 0.03 => قوت
|   |   |   |   |   |   |   |   |   |   |   |--- truncated bran
|   |   |   |   |   |   |   |   |   |   |   |--- 0.03 < قوت
|   |   |   |   |   |   |   |   |   |   |   |   |--- truncated bran
|   |   |   |   |   |   |   |   |   |   |   |   |--- 0.04 < ربو
|   |   |   |   |   |   |   |   |   |   |   |   |   |--- 0.06 => فشل
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |--- truncated bran
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |--- 0.06 < فشل
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |--- class: article
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |--- 0.08 < صفر
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |--- 0.09 => تطع
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |--- 0.28 => نجح
```

Les règles de l'arbre de décision que vous avez fournies sont une représentation textuelle de la structure de l'arbre. Chaque ligne représente une règle ou une condition pour prendre une décision à un nœud spécifique de l'arbre. Voici une interprétation simplifiée des premières parties de ces règles :

1. Si la caractéristique "كرة" est inférieure ou égale à 0.06 :
  - Si la caractéristique "دنا" est inférieure ou égale à 0.04 :
  - Si la caractéristique "لقب" est inférieure ou égale à 0.06 :
  - Si la caractéristique "سعر" est inférieure ou égale à 0.02 :
  - Si la caractéristique "علماء" est inférieure ou égale à 0.03 :
  - Si la caractéristique "اٲي" est inférieure ou égale à 0.07 :
  - Si la caractéristique "اقتصادية" est inférieure ou égale à 0.07 :
  - Si la caractéristique "اله" est inférieure ou égale à 0.04 :
  - Si la caractéristique "صفر" est inférieure ou égale à 0.08 :
2. Si la caractéristique "ربو" est inférieure ou égale à 0.04 :
  - Si la caractéristique "قوت" est inférieure ou égale à 0.03 :
    - Une branche tronquée de profondeur 411 (c'est-à-dire une branche qui a été coupée à 411 niveaux de profondeur)
  - Si la caractéristique "قوت" est supérieure à 0.03 :
    - Une branche tronquée de profondeur 31
3. Si la caractéristique "ربو" est supérieure à 0.04 et la caractéristique "فشل" est inférieure ou égale à 0.06 :
  - Une branche tronquée de profondeur 15
4. Si la caractéristique "ربو" est supérieure à 0.04 et la caractéristique "فشل" est supérieure à 0.06 :
  - La classe de l'article est "articlesReligion"
    - Si la caractéristique "صفر" est supérieure à 0.08 et la caractéristique "تطع" est inférieure ou égale à 0.09 :
5. Si la caractéristique "نجح" est inférieure ou égale à 0.28 :
  - Une branche tronquée de profondeur 4
6. Si la caractéristique "نجح" est supérieure à 0.28 :

## **Projet de traitement de données**

- La classe de l'article est "articlesInternational" ... (le reste des règles est tronqué)