Rapport

Préparé par:

Fatma Gharouel
Amal Maatoug
Farah Riahi

Encadré par:

Dr. Mariem Gzara

2023-2024

Traitement de données

KALIMAT

Le code source est accessible via ce lien:

https://github.com/amal2535/Traitement-de-donnees/tree/main

I. Analyse de données :

1. Introduction :

- **Contexte :**

Dans un monde de plus en plus axé sur les données, le besoin d'outils efficaces pour analyser et comprendre l'information dans des langues diverses est essentiel. L'arabe, en tant que langue riche et complexe, présente des défis et des opportunités uniques pour l'analyse de données. L'introduction de **KALIMAT,** un Corpus Arabe Polyvalent, répond à ce besoin en fournissant une ressource complète pour le traitement naturel du langage en arabe.

- **Objectifs :**

Les objectifs principaux de notre projet visent à habiliter les chercheurs, les développeurs et les amateurs de langues avec un outil robuste pour l'analyse de données en langue arabe. Nous aspirons à accomplir les points suivants :

- **Classification des Catégories :** Mettre en œuvre un système de classification de textes capable d'assigner automatiquement chaque article à l'une des six catégories définies : culture, économie, actualités locales, actualités internationales, religion, et sports.
- **Compréhension Améliorée :** Faciliter une compréhension approfondie de la langue arabe à travers des articles diversifiés couvrant les domaines culturels, économiques, locaux, internationaux, religieux et sportifs.

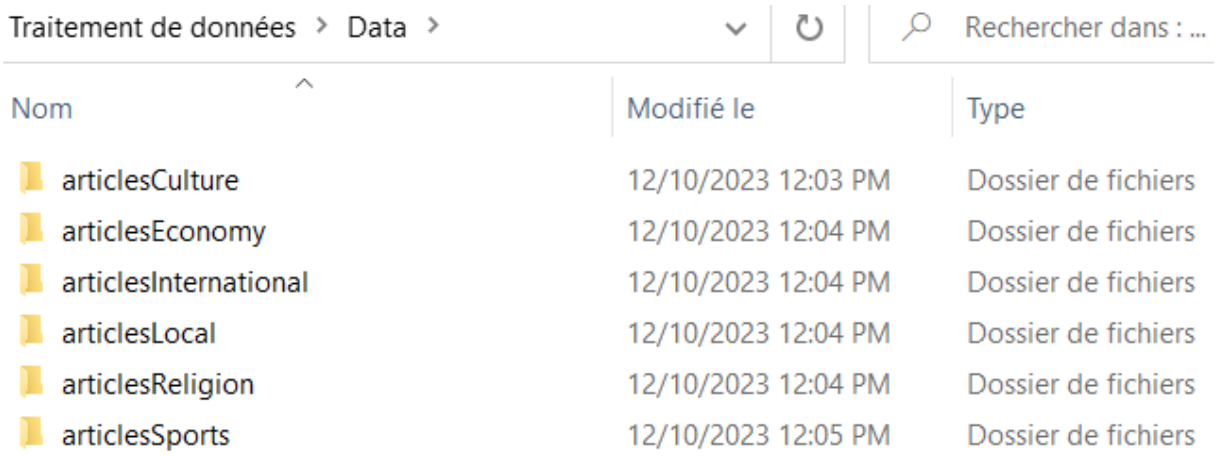
2. Description de projet :

- **Architecture :**

L'architecture est conçue de manière à assurer la modularité, la flexibilité et la facilité d'utilisation.

Voici une vue d'ensemble de l'architecture générale du projet :

Structure des Dossiers :



➔ Les articles collectés sont répartis en six catégories distinctes : culture, économie, actualités locales, actualités internationales, religion et sports.

- **Fonctionnalités principales :**

1. Nettoyage du Texte :

La fonctionnalité de nettoyage du texte est conçue pour garantir des données de qualité. Les étapes de nettoyage incluent la suppression des caractères spéciaux, la normalisation des caractères, et la gestion des espaces pour garantir une cohérence dans le texte brut.

2. Nettoyage Arabe Spécifique :

Une fonctionnalité spécifique à l'arabe consiste en un nettoyage textuel ciblé pour garantir la précision et la pertinence des analyses. Cela inclut la suppression des caractères non-arabes et le traitement spécifique des espaces arabes.

3. Stemming et Stopwords :

Pour une analyse linguistique approfondie, on intègre le stemming arabe et la suppression des mots vides (stopwords) pour réduire les mots à leur racine et éliminer les termes courants sans pertinence.

4. Analyse de Fréquence des Mots :

L'analyse de fréquence des mots permet d'identifier les termes les plus fréquemment utilisés dans chaque catégorie d'articles. Cette fonctionnalité offre une vision globale des tendances lexicales dans les différents domaines, contribuant ainsi à une compréhension approfondie du contenu.

5. Visualisation Graphique :

La visualisation graphique vise à fournir des représentations visuelles intuitives des données. Cela inclut des graphiques de barres pour illustrer la distribution des catégories d'articles, des nuages de mots pour mettre en évidence les termes clés, et des graphiques de réduction de dimensionnalité pour visualiser la structure latente des données.

6. Réduction de Dimensionnalité :

La fonction de réduction de dimensionnalité utilise des techniques telles que l'Analyse en Composantes Principales (ACP) pour compresser l'espace des caractéristiques. Cela facilite la visualisation des données dans un espace tridimensionnel ou bidimensionnel, permettant ainsi une compréhension plus concise des relations entre les articles.

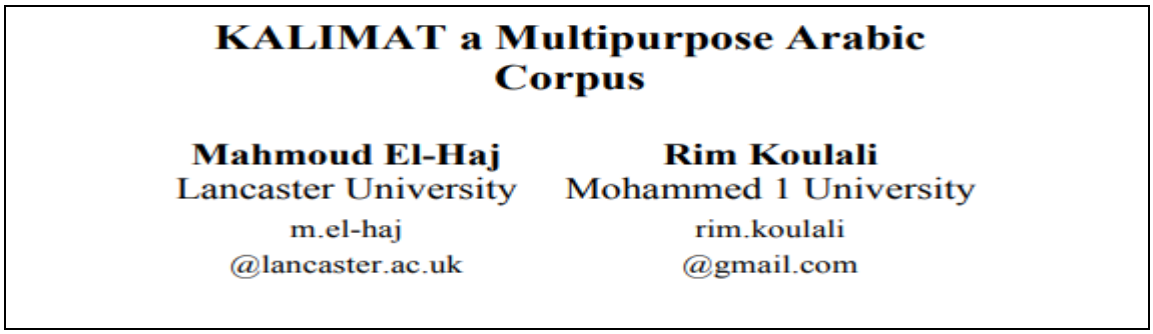
➔ Ces fonctionnalités principales visent à fournir une vue complète pour l'exploration, l'analyse et la compréhension des données en langue arabe.

3. Collecte de données :

- **Source de données :**

Origine des Données : Les données utilisées proviennent du corpus KALIMAT 1.0, une ressource linguistique arabe polyvalente.
Méthode de Collecte : 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.

Référence : 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.


KALIMAT a Multipurpose Arabic Corpus

Mahmoud El-Haj
Lancaster University
m.el-haj
@lancaster.ac.uk

Rim Koulali
Mohammed 1 University
rim.koulali
@gmail.com

- **Raisons de la Pertinence :**

1. **Diversité des Catégories :** Les articles couvrent un large éventail de sujets, allant de la culture à l'économie en passant par les actualités locales et internationales, la

religion et les sports. Cette diversité permet d'explorer les spécificités linguistiques dans différents contextes.

2. **Contexte Riche :** En provenance d'un journal omanais, les articles reflètent le contexte régional et fournissent des perspectives riches sur des événements locaux et internationaux, ainsi que sur des aspects culturels, économiques et religieux.

- **Catégories :**

Description :

1. **Culture :**

   - **Contenu :** Cette catégorie englobe des articles liés à la culture, aux arts, à la littérature, à la musique, au cinéma, et à d'autres aspects de l'héritage culturel. Elle peut inclure des critiques, des événements culturels, des interviews d'artistes, et des analyses culturelles.

2. **Économie :**

   - **Contenu :** 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.

3. **Actualités Locales :**

   - **Contenu :** Cette catégorie se concentre sur les événements et les actualités qui se déroulent à l'échelle locale. Cela peut inclure des nouvelles communautaires, des événements locaux, des annonces gouvernementales, des faits divers, et des reportages sur la vie quotidienne au niveau local.

4. **Actualités Internationales :**

   - **Contenu :** Les actualités internationales englobent les événements et les développements qui se produisent à l'échelle mondiale. Cela peut inclure des rapports sur des affaires internationales, des conflits, des accords diplomatiques, des nouvelles mondiales, et des analyses des relations internationales.

5. **Religion :**

- **Contenu :** Les articles de cette catégorie abordent des sujets liés à la religion, aux pratiques spirituelles, aux événements religieux, aux enseignements, et aux discussions sur la foi. Ils peuvent également inclure des analyses théologiques et des informations sur les communautés religieuses.

6. **Sports :**

- **Contenu :** Les articles sportifs couvrent une variété de disciplines sportives, des résultats des matchs aux performances des athlètes, aux analyses tactiques, et aux grands événements sportifs. Les interviews d'athlètes, les rapports de matchs, et les nouvelles sur les équipes font partie de cette catégorie.

4. Analyse des résultats :

- **Lecture de notre Data :**

Data Brut :

Commentaire :

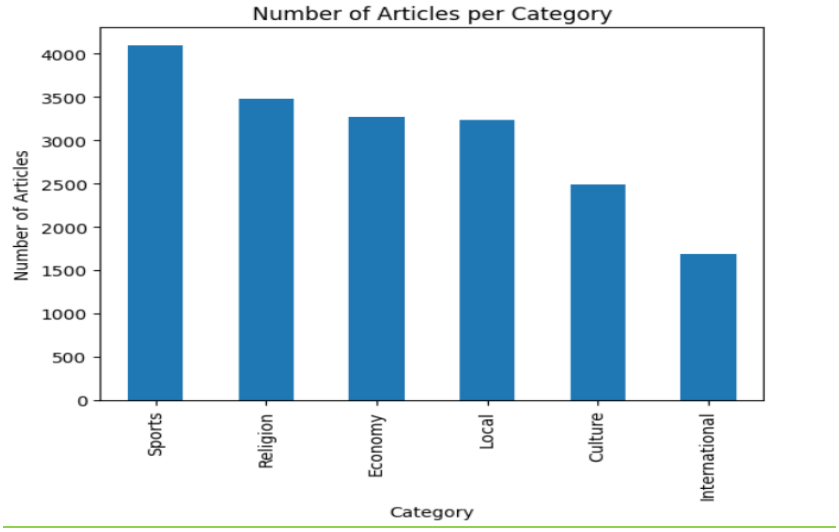
- L'affichage se fait de la même manière pour chaque catégorie : on affiche quelques mots de chaque article appartenant à cette catégorie
- Et pour mieux comprendre, on met notre data dans un data frame pour une meilleure visibilité

	Category	Article Text
0	Sports	...اتحاد الكرة يتلقى دعوة من اللجنة التنظيمية 2 ا
1	Sports	... يكثف اتحاد العاب القوى حاليا اتصالاته المكثفة
2	Sports	... تدخل اليوم أندية الدرجة الثانية لكرة القدم في
3	Sports	...قمة صلالة إذا كانت هناك قمة مثيرة في المنطقة ا
4	Sports	... فنجاء صحار مباراة فنجاء وصحار لا تقل أهمية عن
...	...	...
18251	Local	...احتفل مستشفى النهضة امس تحت رعاية احمد بن عبدا
18252	Local	... تشارك السلطنة ممثلة في وزارة النقل والاتصالات
18253	Local	...كتب صلاح بن سعيد العبري : جاء تنظيم المؤتمر ال
18254	Local	...صور من عبدالله باعلوي : تقوم صباح اليوم لجنة م
18255	Local	...مسقط العمانية : يفتتح فى الثلاثين من شهر سبتمب

- **Exploration & analyse des résultats :**

  Compter le nombre d'articles dans chaque catégorie :

```
Number of articles per category:
Sports: 4095
Religion: 3475
Economy: 3265
Local: 3237
Culture: 2495
International: 1689
```



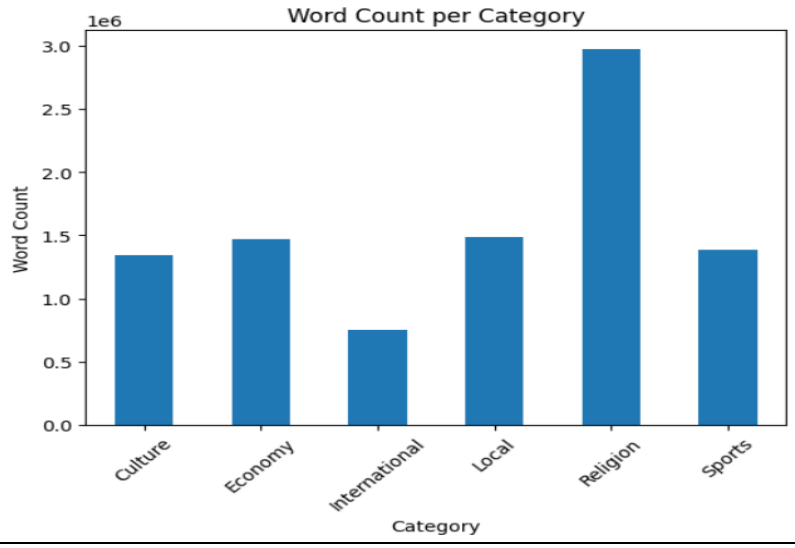
Commentaire :

- Cette distribution nous offre une vision quantitative de la répartition des articles dans notre ensemble de données.

- **Analyse comparative :** En examinant la distribution des articles par catégorie, nous pouvons identifier des disparités significatives dans la représentation des thèmes.
  • La catégorie Sports présente le plus grand nombre d'articles, avec un total de 4095 articles. Cela indique une forte représentation de l'actualité sportive dans notre ensemble de données.
  • La catégorie International a le plus petit nombre d'articles, totalisant seulement 1689. Cela suggère une représentation relativement moindre des actualités internationales dans notre ensemble de données.

Compter le nombre de mots dans chaque catégorie :

```
Word count in each category:
Category
Culture          1344078
Economy          1472513
International     751815
Local            1484586
Religion         2972644
Sports           1388148
```



Compter le nombre de mots dans chaque article de chaque catégorie :

```
Category: Culture
Article 7361: Word Count: 2732
Article 7362: Word Count: 835
Article 7363: Word Count: 691
Article 7364: Word Count: 1972
Article 7365: Word Count: 329
Article 7366: Word Count: 552
Article 7367: Word Count: 128
Article 7368: Word Count: 187
Article 7369: Word Count: 568
Article 7370: Word Count: 545
Article 7371: Word Count: 86
Article 7372: Word Count: 707
Article 7373: Word Count: 1290
Article 7374: Word Count: 1108
Article 7375: Word Count: 406
Article 7376: Word Count: 932
Article 7377: Word Count: 138
```

```
Category: Sports
Article 1: Word Count: 101
Article 2: Word Count: 119
Article 3: Word Count: 379
Article 4: Word Count: 201
Article 5: Word Count: 189
Article 6: Word Count: 165
Article 7: Word Count: 136
Article 8: Word Count: 214
Article 9: Word Count: 190
Article 10: Word Count: 219
Article 11: Word Count: 183
Article 12: Word Count: 139
Article 13: Word Count: 236
Article 14: Word Count: 319
Article 15: Word Count: 389
Article 16: Word Count: 414
Article 17: Word Count: 272
```

Commentaire :

- En analysant le nombre de mots dans chaque catégorie, nous pouvons tirer des insights sur la longueur et la complexité des articles.

- Cette analyse souligne des variations significatives dans la longueur des articles entre les catégories. Par exemple, la catégorie "Religion" se distingue avec le plus grand nombre de mots, ce qui pourrait indiquer une couverture approfondie des sujets religieux. À l'inverse, la catégorie "International" semble contenir des articles plus concis.

Les mots les plus fréquents de chaque catégorie :

```
Category: Culture
Word: في | Count: 40167
Word: . | Count: 39824
Word: من | Count: 32424
Word: على | Count: 15906
Word: ( | Count: 13950
Word: ) | Count: 13832
Word: التي | Count: 9320
Word: أن | Count: 8922
Word: عن | Count: 8083
Word: ان | Count: 7585
```

```
Category: Economy
Word: في | Count: 39295
Word: من | Count: 37212
Word: . | Count: 31655
Word: على | Count: 19694
Word: ان | Count: 13779
Word: الى | Count: 13415
Word: التي | Count: 8474
Word: في | Count: 7875
Word: خلال | Count: 6367
Word: هذه | Count: 6130
```

```
Category: International
Word: . | Count: 26198
Word: في | Count: 24911
Word: من | Count: 18119
Word: على | Count: 11033
Word: ان | Count: 9823
Word: الى | Count: 5524
Word: ( | Count: 5511
Word: ) | Count: 5443
Word: : | Count: 4821
Word: التي | Count: 4630
```

```
Category: Local
Word: من | Count: 39535
Word: في | Count: 39122
Word: . | Count: 24636
Word: على | Count: 20983
Word: التي | Count: 10249
Word: بن | Count: 10127
Word: الى | Count: 9977
Word: ان | Count: 9813
Word: : | Count: 7909
Word: هذه | Count: 7573
```

```
Category: Religion
Word: . | Count: 83200
Word: من | Count: 73785
Word: في | Count: 72737
Word: الله | Count: 48049
Word: : | Count: 42091
Word: ( | Count: 36712
Word: ) | Count: 35532
Word: على | Count: 35428
Word: أن | Count: 29520
Word: لا | Count: 20369
```

```
Category: Sports
Word: في | Count: 47679
Word: . | Count: 32845
Word: من | Count: 31172
Word: على | Count: 21337
Word: ان | Count: 10348
Word: الى | Count: 10304
Word: ( | Count: 9325
Word: ) | Count: 9304
Word: التي | Count: 7925
Word: بن | Count: 7719
```

Commentaire :

- En examinant les mots les plus fréquents dans chaque catégorie, nous pouvons identifier les termes qui caractérisent le mieux le contenu de chaque domaine.
- Cette analyse des mots fréquents offre un aperçu des thèmes dominants dans chaque catégorie. Par exemple, dans la catégorie "Religion", on observe une fréquence élevée de mots tels que "الله" (Dieu) et "أن" (que), soulignant probablement des discussions théologiques. L'utilisation de tels résultats peut orienter davantage votre compréhension du contenu spécifique de chaque catégorie.

- **Nettoyage & prétraitements des données :**

Le processus de nettoyage et de prétraitement des données en arabe a été réalisé de manière approfondie pour garantir la qualité et la cohérence des données utilisées. Voici les étapes clés du processus :

1. Suppression des Symboles & des Caractères Non-Arabes :

- On a commencé par l'élimination de tout texte qui n'est pas en arabe et qui n'est pas un nombre (Principalement les symboles et les ponctuations)

Input : Article Text
Output : Cleaned Article Text

	Category	Article Text	Word Count	Cleaned Article Text
0	Sports	...اتحاد الكرة يتلقى دعوة من اللجنة التنظيمية ال 2	101	...اتحاد الكرة يتلقى دعوة من اللجنة التنظيمية ال 2
1	Sports	... يكثف اتحاد العاب القوى حاليا اتصالاته المكثفة	119	... يكثف اتحاد العاب القوى حاليا اتصالاته المكثفة
2	Sports	... تدخل اليوم أندية الدرجة الثانية لكرة القدم في	379	... تدخل اليوم أندية الدرجة الثانية لكرة القدم في
3	Sports	...قمة صلالة إذا كانت هناك قمة مثيرة في المنطقة ا	201	...قمة صلالة إذا كانت هناك قمة مثيرة في المنطقة ا
4	Sports	... فنجاء صحار مباراة فنجاء وصحار لا تقل أهمية عن	189	... فنجاء صحار مباراة فنجاء وصحار لا تقل أهمية عن
...	...	...	...	...
18251	Local	...احتفل مستشفى النهضة امس تحت رعاية احمد بن عبدا	502	...احتفل مستشفى النهضة امس تحت رعاية احمد بن عبدا
18252	Local	... تشارك السلطنة ممثلة في وزارة النقل والاتصالات	198	... تشارك السلطنة ممثلة في وزارة النقل والاتصالات
18253	Local	...كتب صلاح بن سعيد العبري : جاء تنظيم المؤتمر ال	371	...كتب صلاح بن سعيد العبري جاء تنظيم المؤتمر الدو
18254	Local	...صور من عبدالله باعلوي : تقوم صباح اليوم لجنة م	201	...صور من عبدالله باعلوي تقوم صباح اليوم لجنة متا
18255	Local	...مسقط العمانية : يفتتح فى الثلاثين من شهر سبتمب	554	... مسقط العمانية يفتتح فى الثلاثين من شهر سبتمبر

Compter le nombre de mots supprimés dans chaque catégorie :

```
Total Deleted Words: 73974
Deleted Words by Category:
Category: Culture
Total Deleted Words: 10200
Deleted Words:
Word: '.' | Count: 2398
Word: '(' | Count: 2038
Word: ')' | Count: 2033
Word: ':' | Count: 1979
Word: '-' | Count: 582
Word: '!' | Count: 380
Word: '...' | Count: 263
Word: ',' | Count: 217
Word: '_' | Count: 28
Word: '11.30' | Count: 7
Word: '7.30' | Count: 6
Word: '8.30' | Count: 5
Word: '1.30' | Count: 5
```

```
Category: Culture
Total Deleted Words: 10200
----------------------
Category: Economy
Total Deleted Words: 16172
----------------------
Category: International
Total Deleted Words: 6625
----------------------
Category: Local
Total Deleted Words: 10444
----------------------
Category: Religion
Total Deleted Words: 16723
----------------------
Category: Sports
Total Deleted Words: 13810
----------------------
```

Compter le nombre de mots supprimés dans chaque article de chaque catégorie :

```
Article 1:
------------------
Article 2:
Word: '.' | Count: 1
------------------
Article 3:
Word: '.' | Count: 1
Word: '4.45' | Count: 1
------------------
Article 4:
Word: '.' | Count: 1
------------------
Article 5:
Word: '.' | Count: 1
------------------
Article 6:
Word: '.' | Count: 1
------------------
```

```
Word: ':' | Count: 1
------------------
Article 51:
Word: '.' | Count: 1
Word: ':' | Count: 1
------------------
Article 52:
Word: '28-4' | Count: 1
Word: '22-4' | Count: 1
Word: '23-4' | Count: 1
Word: '.' | Count: 1
Word: '24-4' | Count: 1
Word: '25-4' | Count: 1
Word: '27-4' | Count: 1
Word: ':' | Count: 1
------------------
Article 53:
Word: '.' | Count: 1
```

Commentaire :

- Ces chiffres reflètent le niveau de pré-traitement nécessaire pour chaque catégorie. Des différences significatives entre les catégories peuvent indiquer des variations dans la qualité initiale des données ou des caractéristiques linguistiques spécifiques à chaque domaine.

2. Nettoyage des Mots Vides (Stopwords) :

Si on explore le data résultant :

- **Les mots les plus fréquents pour chaque catégorie (après 1er nettoyage):**

```
Category: Culture
Word: في | Count: 40167
Word: من | Count: 32424
Word: على | Count: 15906
Word: التي | Count: 9320
Word: أن | Count: 8922
Word: عن | Count: 8083
Word: ان | Count: 7585
Word: الذي | Count: 6500
Word: هذا | Count: 6166
Word: الى | Count: 6074
```

```
Category: Economy
Word: في | Count: 39295
Word: من | Count: 37212
Word: على | Count: 19694
Word: ان | Count: 13779
Word: الى | Count: 13415
Word: التي | Count: 8474
Word: فى | Count: 7875
Word: خلال | Count: 6367
Word: هذه | Count: 6130
Word: عن | Count: 5948
```

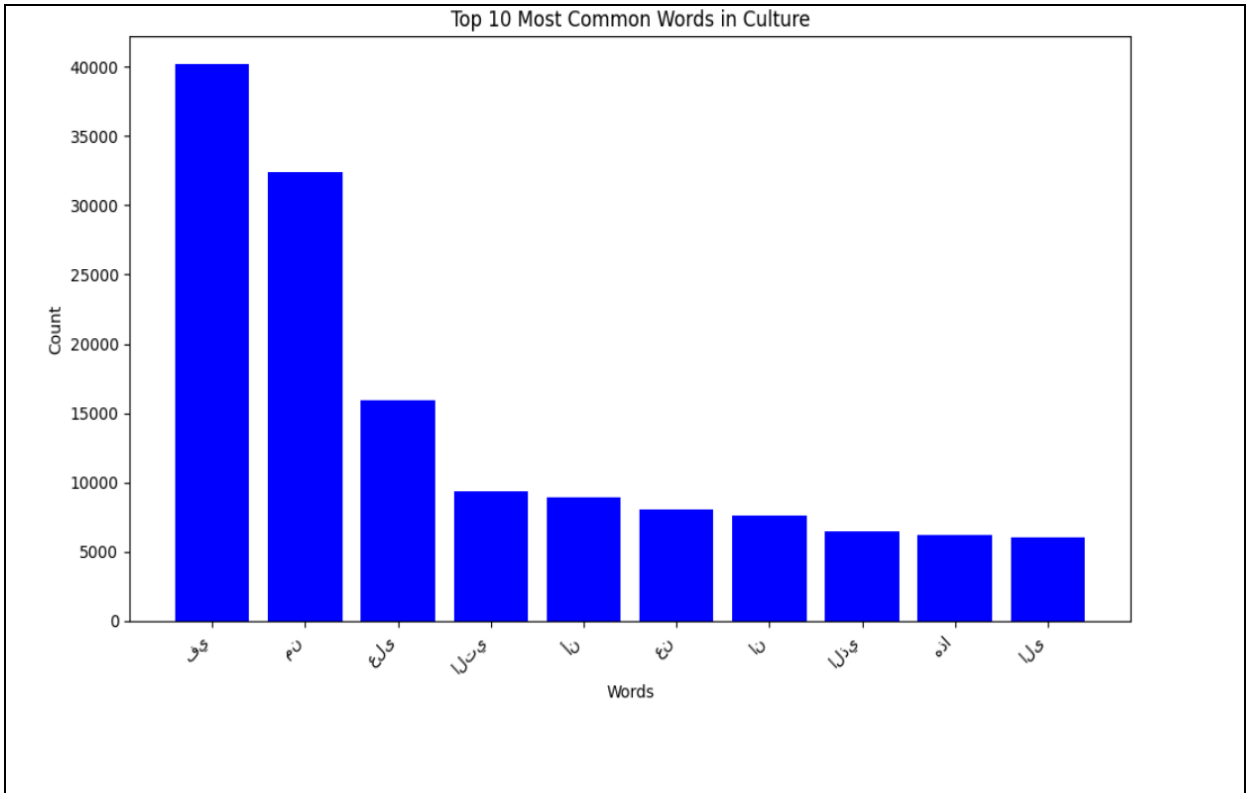
```
Category: International
Word: في | Count: 24911
Word: من | Count: 18119
Word: على | Count: 11033
Word: ان | Count: 9823
Word: الى | Count: 5524
Word: التي | Count: 4630
Word: عن | Count: 4381
Word: أن | Count: 4113
Word: فى | Count: 3223
Word: الذي | Count: 3049
```

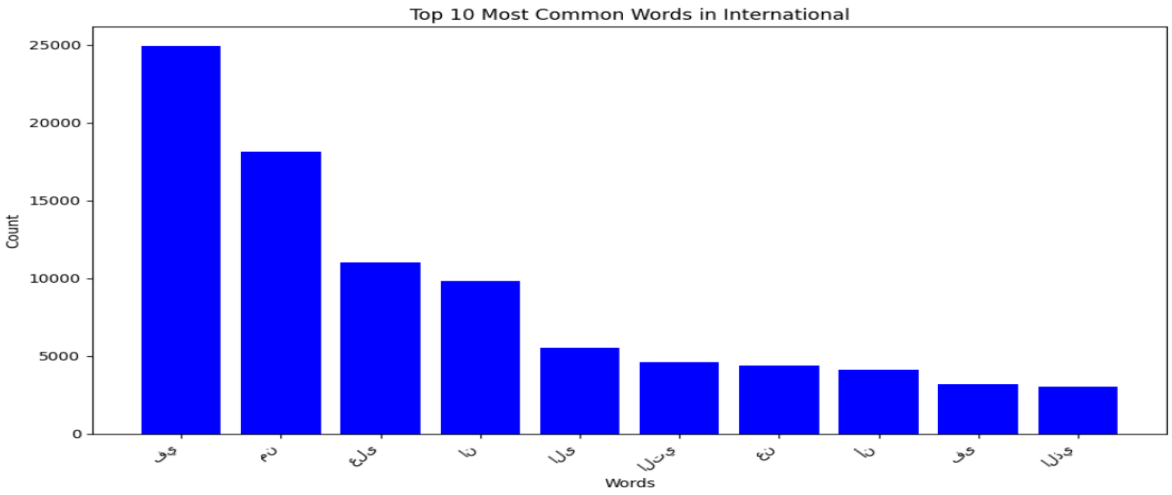
```
Category: Local
Word: من | Count: 39535
Word: في | Count: 39122
Word: على | Count: 20983
Word: التي | Count: 10249
Word: بن | Count: 10127
Word: الى | Count: 9977
Word: ان | Count: 9813
Word: هذه | Count: 7573
Word: هذا | Count: 6800
Word: حيث | Count: 6146
```

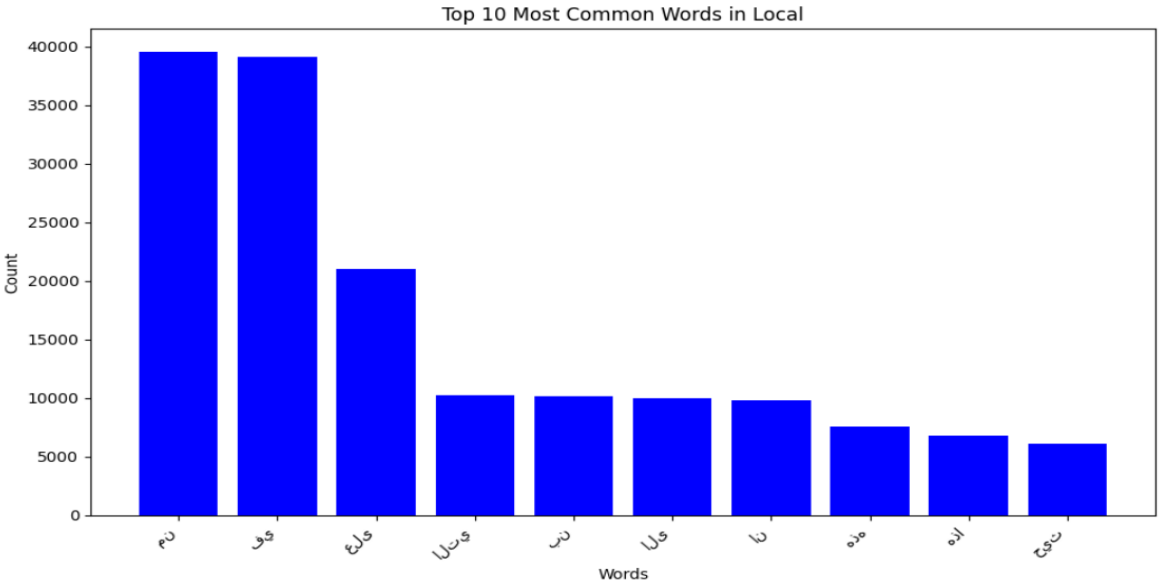
```
Category: Religion
Word: من | Count: 73785
Word: في | Count: 72737
Word: الله | Count: 48049
Word: على | Count: 35428
Word: أن | Count: 29520
Word: لا | Count: 20369
Word: ما | Count: 19563
Word: عن | Count: 18907
Word: عليه | Count: 18746
Word: إلى | Count: 18639
```

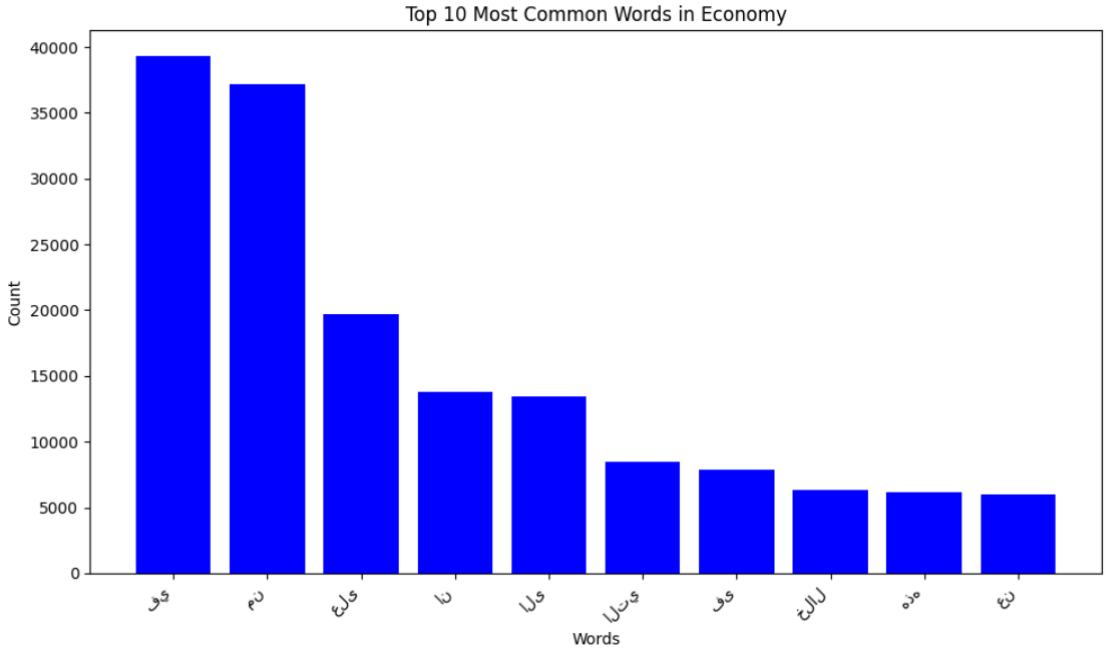
```
Category: Sports
Word: في | Count: 47679
Word: من | Count: 31172
Word: على | Count: 21337
Word: ان | Count: 10348
Word: الى | Count: 10304
Word: التي | Count: 7925
Word: بن | Count: 7719
Word: الذي | Count: 7357
Word: مع | Count: 6198
Word: بعد | Count: 5339
```

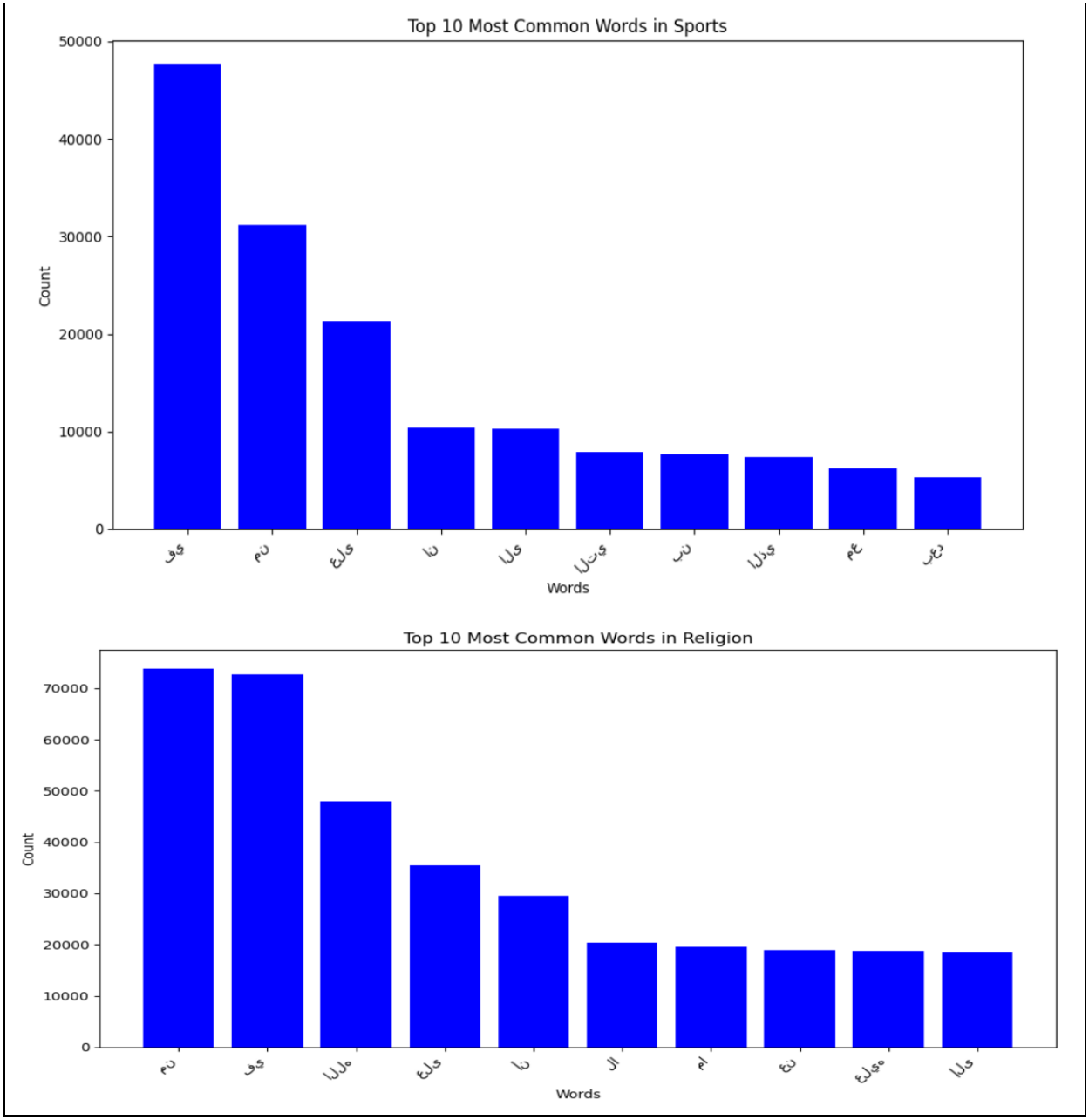
➔ **On remarque que la majorité des mots fréquents sont des stopwords :**




Top 10 Most Common Words in International
Count
Words
25000
20000
15000
10000
5000
0


Top 10 Most Common Words in Local
Count
Words
40000
35000
30000
25000
20000
15000
10000
5000
0


Top 10 Most Common Words in Economy
Count
Words
40000
35000
30000
25000
20000
15000
10000
5000
0



➔ **Les stopwords sont des mots très courants qui apparaissent fréquemment dans un texte mais qui ne portent souvent pas de signification importante.**

➔ **Elimination des stopwords !!**

Input : Cleaned Article Text

Output : Cleaned Article Text (2eme iteration)

	Category	Article Text	Word Count	Cleaned Article Text	Cleaned Arabic Text 2
0	Sports	...اتحاد الكرة يتلقى دعوة من اللجنة التنظيمية 2 ا	101	...اتحاد الكرة يتلقى دعوة من اللجنة التنظيمية 2 ا	...اتحاد الكرة يتلقى دعوة اللجنة التنظيمية 2 اتحا
1	Sports	... يكثف اتحاد العاب القوى حاليا اتصالاته المكثفة	119	... يكثف اتحاد العاب القوى حاليا اتصالاته المكثفة	... يكثف اتحاد العاب القوى حاليا اتصالاته المكثفة
2	Sports	... تدخل اليوم أندية الدرجة الثانية لكرة القدم في	379	... تدخل اليوم أندية الدرجة الثانية لكرة القدم في	...تدخل اليوم أندية الدرجة الثانية لكرة القدم جول
3	Sports	...قمة صلالة إذا كانت هناك قمة مثيرة في المنطقة ا	201	...قمة صلالة إذا كانت هناك قمة مثيرة في المنطقة ا	...قمة صلالة كانت قمة مثيرة المنطقة الداخلية صلال
4	Sports	... فنجاء صحار مباراة فنجاء وصحار لا تقل أهمية عن	189	... فنجاء صحار مباراة فنجاء وصحار لا تقل أهمية عن	...فنجاء صحار مباراة فنجاء وصحار تقل أهمية باقي ا
...	...	...	...	...	...
18251	Local	...احتفل مستشفى النهضة امس تحت رعاية احمد بن عبدا	502	...احتفل مستشفى النهضة امس تحت رعاية احمد بن عبدا	... احتفل مستشفى النهضة امس رعاية احمد بن عبدالله
18252	Local	... تشارك السلطنة ممثلة في وزارة النقل والاتصالات	198	... تشارك السلطنة ممثلة في وزارة النقل والاتصالات	...تشارك السلطنة ممثلة وزارة النقل والاتصالات قطا
18253	Local	...كتب صلاح بن سعيد العبري : جاء تنظيم المؤتمر ال	371	...كتب صلاح بن سعيد العبري جاء تنظيم المؤتمر الدو	...كتب صلاح بن سعيد العبري جاء تنظيم المؤتمر الدو
18254	Local	...صور من عبدالله باعلوي : تقوم صباح اليوم لجنة م	201	...صور من عبدالله باعلوي تقوم صباح اليوم لجنة متا	...صور عبدالله باعلوي تقوم اليوم لجنة متابعة متطل
18255	Local	...مسقط العمانية : يفتتح فى الثلاثين من شهر سبتمب	554	... مسقط العمانية يفتتح فى الثلاثين من شهر سبتمبر	...مسقط العمانية يفتتح فى الثلاثين شهر الحالى بمر

18256 rows × 5 columns

3. Stemming avec ISRIStemmer :

- L'algorithme de stemming ISRIStemmer a été appliqué pour réduire les mots à leur racine, favorisant ainsi la cohérence et réduisant la dimensionnalité des données.

Input : Cleaned Article Text

Output : Stemmed text

	Category	Article Text	Word Count	Cleaned Article Text	Cleaned Arabic Text 2	Stemmed Text
0	Sports	اتحاد الكرة يتلقى دعوة من اللجنة التنظيمية ا 2...	101	اتحاد الكرة يتلقى دعوة من اللجنة التنظيمية ا 2...	اتحاد الكرة يتلقى دعوة اللجنة التنظيمية 2 ...اتحا	تحد كرة لقى دعة لجن نظم 2 تحد قوى كثف ...اتصالاته
1	Sports	يكثف اتحاد العاب القوى حاليا اتصالاته ... المكثفة	119	يكثف اتحاد العاب القوى حاليا اتصالاته ... المكثفة	يكثف اتحاد العاب القوى حاليا اتصالاته ... المكثفة	كثف تحد عاب قوى حال اتصالاته كثف عدد دول ...هدف ت
2	Sports	تدخل اليوم أندية الدرجة الثانية لكرة القدم في ...	379	تدخل اليوم أندية الدرجة الثانية لكرة القدم في ...	تدخل اليوم أندية الدرجة الثانية لكرة القدم ...جول	دخل اليوم ندي درج ثني لكر قدم جول همة تضح ... خلل
3	Sports	قمة صلالة إذا كانت هناك قمة مثيرة في ...المنطقة ا	201	قمة صلالة إذا كانت هناك قمة مثيرة في ...المنطقة ا	قمة صلالة كانت قمة مثيرة المنطقة ...الداخلية صلال	قمة صلل كانت قمة ثير نطق دخل صلل خرى ...حضن قمة ق
4	Sports	فنجاء صحار مباراة فنجاء وصحار لا تقل أهمية ... عن	189	فنجاء صحار مباراة فنجاء وصحار لا تقل أهمية ... عن	فنجاء صحار مباراة فنجاء وصحار تقل أهمية ...باقي ا	نجء صحر برا نجء صحر تقل همي بقي بري ستى ...وفي طم
...	...	...	...	...	...	...
18251	Local	احتفل مستشفى النهضة امس تحت رعاية ...احمد بن عبدا	502	احتفل مستشفى النهضة امس تحت رعاية ...احمد بن عبدا	احتفل مستشفى النهضة امس رعاية احمد ... بن عبدالله	حفل شفى نهض امس رعي حمد بن عبدالل خنج ... دير عام
18252	Local	تشارك السلطنة ممثلة في وزارة النقل ... والاتصالات	198	تشارك السلطنة ممثلة في وزارة النقل ... والاتصالات	تشارك السلطنة ممثلة وزارة النقل ...والاتصالات قطا	شرك سلط مثل وزر نقل تصل قطع تصل برد ادر ...برد خل
18253	Local	كتب صلاح بن سعيد العبري : جاء تنظيم ...المؤتمر ال	371	كتب صلاح بن سعيد العبري جاء تنظيم ...المؤتمر الدو	كتب صلاح بن سعيد العبري جاء تنظيم ...المؤتمر الدو	كتب صلح بن سعد عبر جاء نظم ؤمر دول نطق ...شرق وسط
18254	Local	صور من عبدالله باعلوي : تقوم صباح اليوم ...لجنة م	201	صور من عبدالله باعلوي تقوم صباح اليوم ...لجنة متا	صور عبدالله باعلوي تقوم اليوم لجنة متابعة ...متطل	صور عبدالل علو تقم اليوم لجن تبع تطلب خطة ... درس
18255	Local	مسقط العمانية : يفتتح فى الثلاثين من شهر ...سبتمب	554	مسقط العمانية يفتتح فى الثلاثين من شهر ... سبتمبر	مسقط العمانية يفتتح فى الثلاثين شهر ...الحالى بمر	سقط عمن فتح فى ثلث شهر حلى ركز عمن ...دولى عرض عر

18256 rows × 6 columns

Data préparée :

➔ Ces étapes de nettoyage et de prétraitement ont été essentielles pour garantir la qualité des données, réduire le bruit, et créer une base solide pour les analyses linguistiques avancées réalisées.

- **Exploration du Data préparée :**

Les mots les plus fréquents dans chaque catégorie :

Category: Culture
Word: علم, Count: 7866
Word: عمل, Count: 7804
Word: ان, Count: 7671
Word: كتب, Count: 7450
Word: عرب, Count: 7134
Word: جمع, Count: 7026
Word: عرض, Count: 6605
Word: الى, Count: 6085
Word: شعر, Count: 5927
Word: قدم, Count: 5809

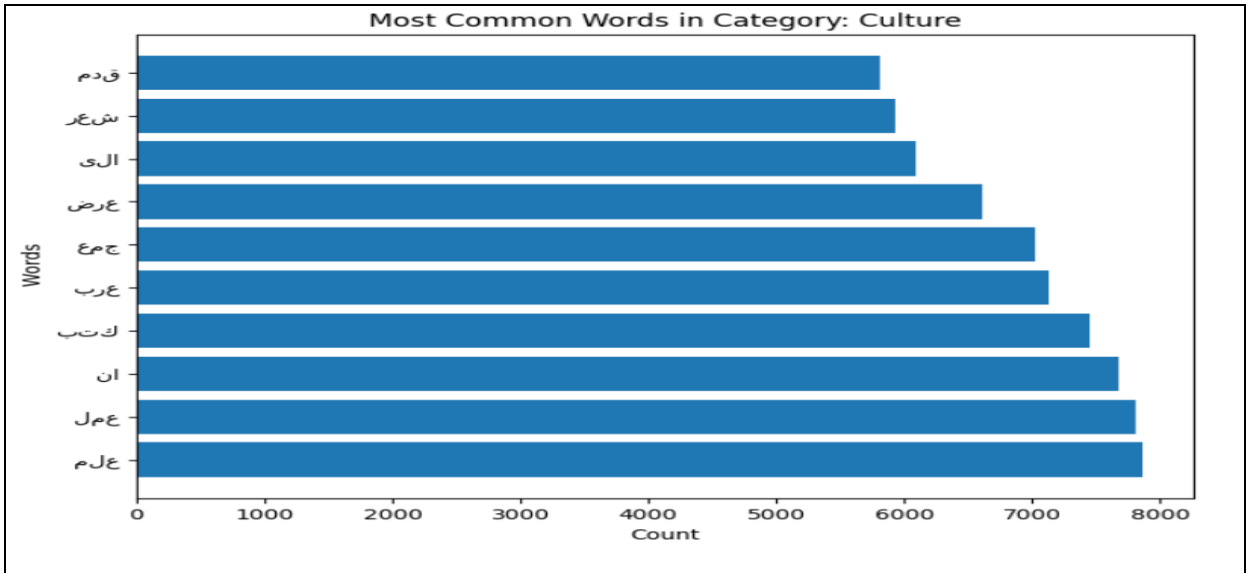
Category: Economy
Word: عمل, Count: 18477
Word: ان, Count: 13802
Word: الى, Count: 13522
Word: شرك, Count: 10787
Word: دول, Count: 10605
Word: علم, Count: 8638
Word: في, Count: 7875
Word: عدد, Count: 7730
Word: جمع, Count: 7704
Word: صنع, Count: 7330

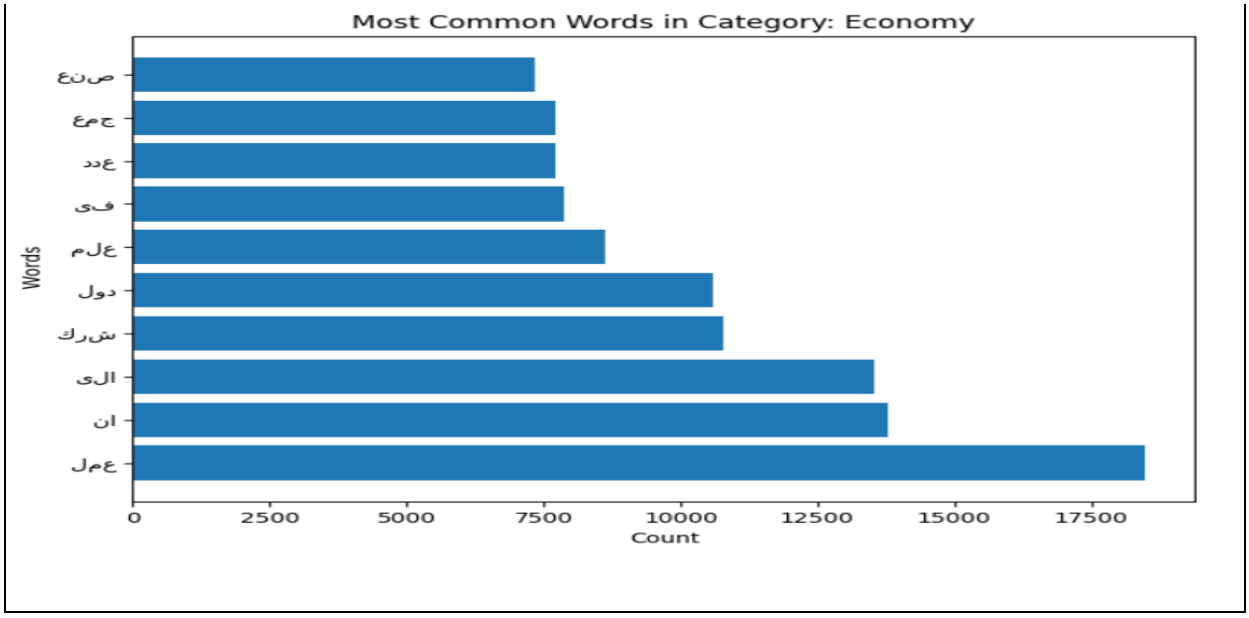
Category: International
Word: ان, Count: 9839
Word: امر, Count: 5837
Word: الى, Count: 5527
Word: عرق, Count: 5100
Word: عمل, Count: 4652
Word: رئس, Count: 4314
Word: جمع, Count: 4098
Word: حكم, Count: 4038
Word: وقل, Count: 3286
Word: في, Count: 3223

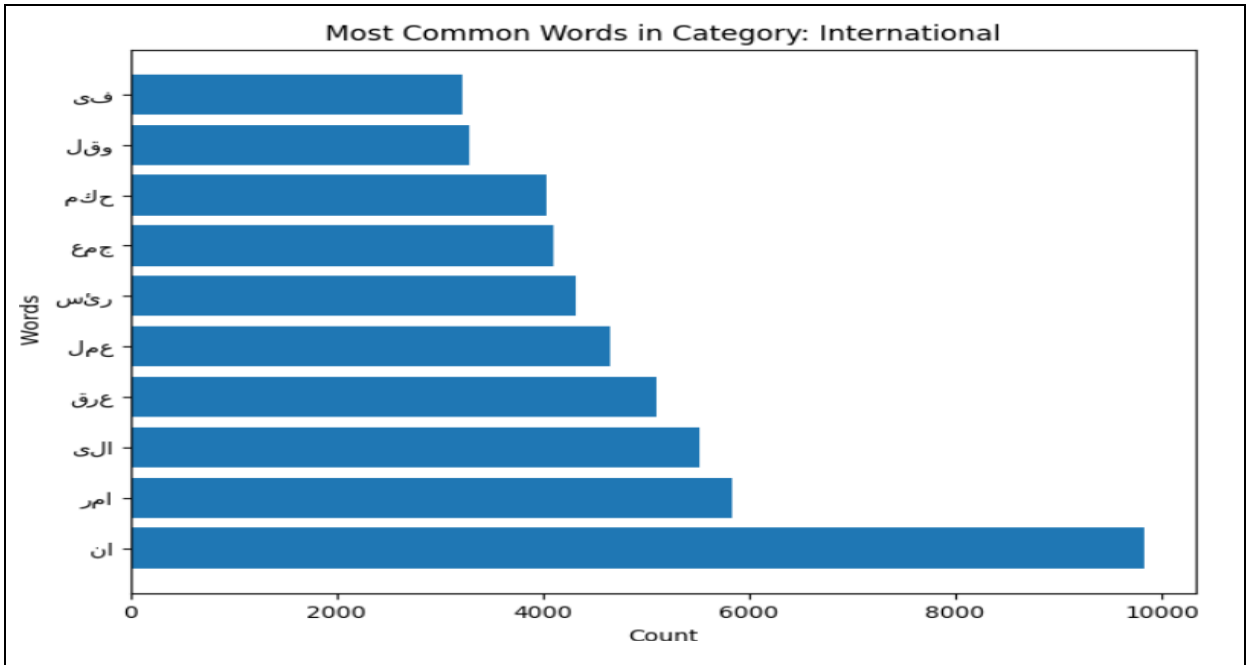
Category: Local
Word: علم, Count: 17239
Word: جمع, Count: 13426
Word: عمل, Count: 12906
Word: درس, Count: 10789
Word: بن, Count: 10127
Word: الى, Count: 9996
Word: ان, Count: 9828
Word: سلط, Count: 9058
Word: سعد, Count: 8949
Word: طلب, Count: 8704

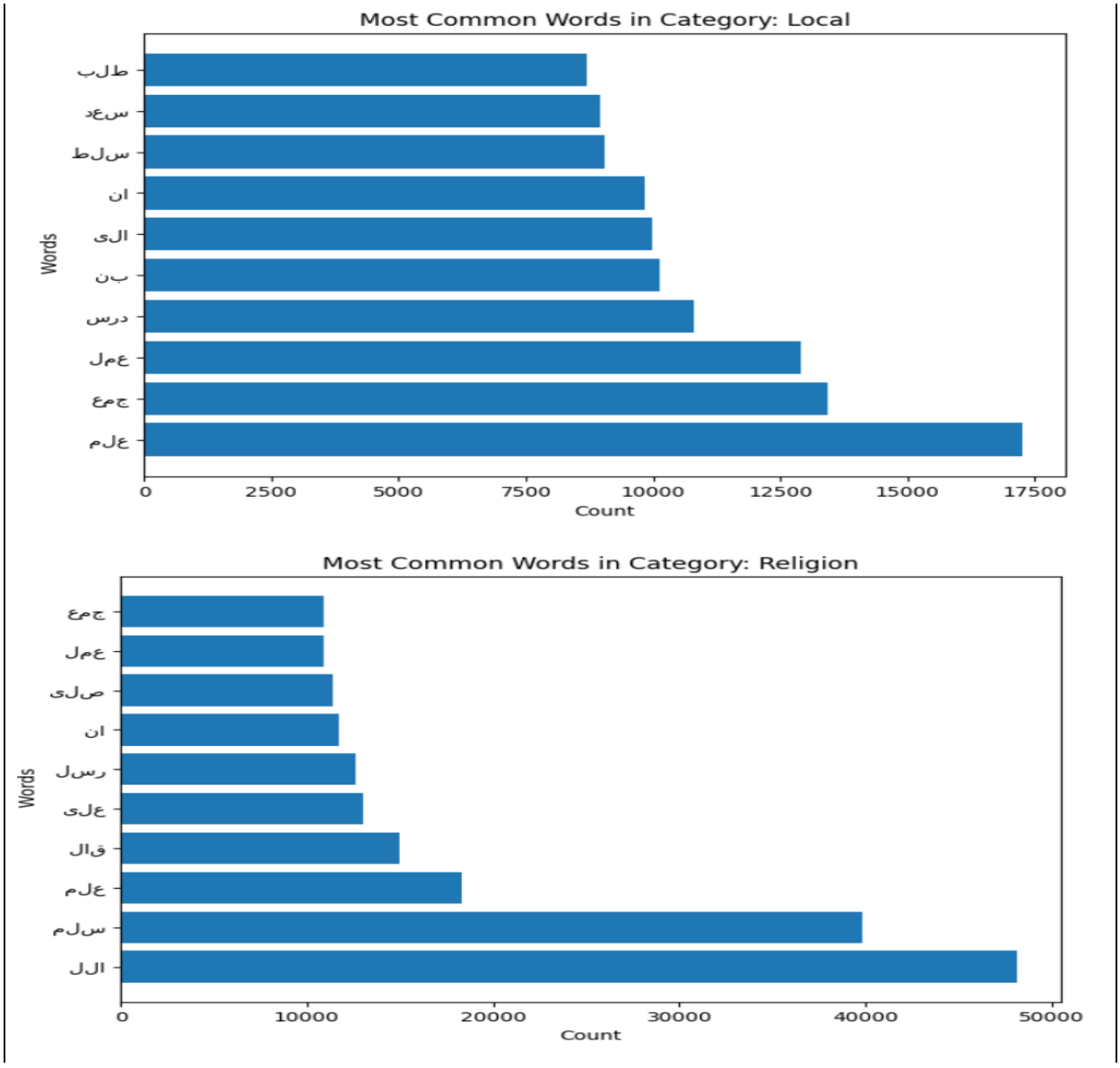
Category: Religion
Word: الل, Count: 48089
Word: سلم, Count: 39860
Word: علم, Count: 18349
Word: قال, Count: 15003
Word: على, Count: 13050
Word: رسل, Count: 12596
Word: ان, Count: 11710
Word: صلى, Count: 11431
Word: عمل, Count: 10924
Word: جمع, Count: 10909

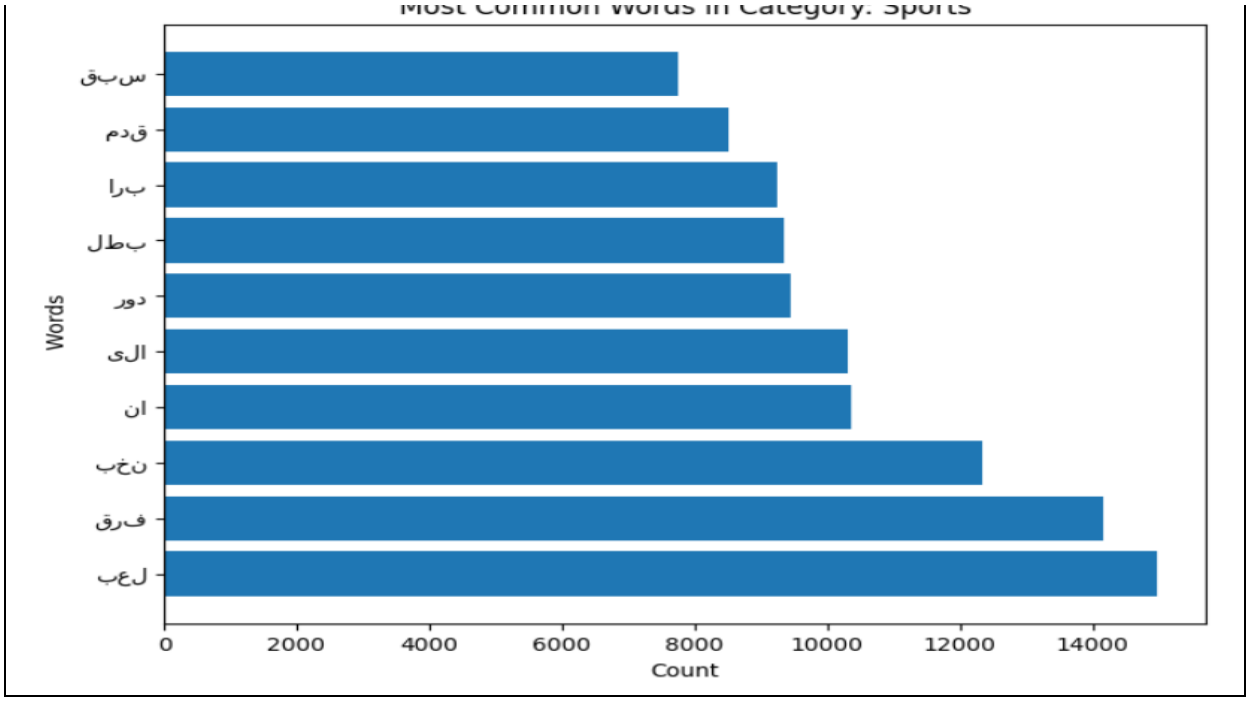
Category: Sports
Word: لعب, Count: 14959
Word: فرق, Count: 14149
Word: نخب, Count: 12335
Word: ان, Count: 10355
Word: الى, Count: 10309
Word: دور, Count: 9450
Word: بطل, Count: 9346
Word: برا, Count: 9239
Word: قدم, Count: 8518
Word: سبق, Count: 7741




Most Common Words in Category: Economy
Words
عنص
عمج
ددع
يف
ملع
لود
كرش
يلا
ان
لمع
0
2500
5000
7500
10000
12500
15000
17500
Count


Most Common Words in Category: International
Words
يف
لقو
مكح
عمج
سئر
لمع
قرع
يلا
رما
ان
0
2000
4000
6000
8000
10000
Count


Most Common Words in Category: Local
Words
Count
0
2500
5000
7500
10000
12500
15000
17500
Most Common Words in Category: Religion
Words
Count
0
10000
20000
30000
40000
50000



➔ **Nos données sont maintenant préparées, la structure de data reste non structurée, d'où le besoin de la vectorisation de Data.**

4. Application de TFIDF :

- La matrice TF-IDF (Term Frequency-Inverse Document Frequency) est une représentation numérique de documents textuels qui tient compte de l'importance relative des termes dans ces documents au sein d'une collection plus large de documents.

Input : Stemmed text.
Output : tfidf_df.

TF-IDF Matrix:

	00	000	0000	000000	0004	0041	0051	01	0100000	01876	...	يوليوالجاري	يوميليدي	يونغ	يونيفيرسدادي	يوهيكاتشو	يوو	يبا	يبف	بيفدوكيموفا
0	0.174836	0.01212	0.0	0.0	0.0	0.0	0.0	0.015474	0.0	0.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
1	0.000000	0.00000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
2	0.000000	0.00000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
3	0.000000	0.00000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
4	0.000000	0.00000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4090	0.000000	0.00000	0.0	-1.0	-1.0	-1.0	-1.0	0.000000	-1.0	-1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4091	0.000000	0.00000	0.0	-1.0	-1.0	-1.0	-1.0	0.000000	-1.0	-1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4092	0.000000	0.00000	0.0	-1.0	-1.0	-1.0	-1.0	0.000000	-1.0	-1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4093	0.000000	0.00000	0.0	-1.0	-1.0	-1.0	-1.0	0.000000	-1.0	-1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4094	0.000000	0.00000	0.0	-1.0	-1.0	-1.0	-1.0	0.000000	-1.0	-1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5. Reduction de données :

- **Technique utilisée est le schéma TF-IDF (Term Frequency-Inverse Document Frequency) :**

Term Frequency (TF - Fréquence du terme) : Cela mesure la fréquence d'un terme dans un document particulier. Elle est calculée comme le nombre de fois qu'un terme apparaît dans un document, divisé par le nombre total de termes dans ce document. Cela donne une idée de l'importance relative d'un terme dans un document.

Inverse Document Frequency (IDF - Fréquence Inverse du Document) : Cela mesure l'importance d'un terme dans l'ensemble du corpus (ensemble de tous les documents). Les termes fréquents dans l'ensemble du corpus ont un IDF plus faible, tandis que les termes moins fréquents ont un IDF plus élevé. Cela permet de mettre en évidence les termes qui sont distinctifs pour un document particulier par rapport à l'ensemble du corpus.

TF-IDF : La valeur TF-IDF d'un terme dans un document est le produit de sa fréquence dans ce document (TF) et de l'inverse de sa fréquence dans l'ensemble du corpus (IDF). Cela donne une mesure qui tient compte de l'importance du terme dans le document spécifique et dans l'ensemble du corpus.

Le résultat de la vectorisation TF-IDF est une représentation numérique des documents texte, où chaque document est représenté par un vecteur de valeurs TF-IDF pour chaque terme du vocabulaire. Cette représentation est souvent utilisée comme entrée pour des modèles d'apprentissage automatique, tels que des classificateurs.

Input : tfidf_df.
Output : tfidf_reduced_df

Dimensions de data avant la réduction :

```
Dimension avant ACP: (18256, 39961)
```

Réduction :

```
        00  000  0000        10       100  1000  103        11  110        12  \
0      0.0  0.0   0.0  0.000000  0.000000   0.0  0.0  0.000000  0.0  0.000000
1      0.0  0.0   0.0  0.000000  0.000000   0.0  0.0  0.000000  0.0  0.000000
2      0.0  0.0   0.0  0.000000  0.015336   0.0  0.0  0.012475  0.0  0.000000
3      0.0  0.0   0.0  0.000000  0.000000   0.0  0.0  0.000000  0.0  0.000000
4      0.0  0.0   0.0  0.022588  0.000000   0.0  0.0  0.000000  0.0  0.000000
...    ...  ...   ...       ...       ...   ...  ...       ...  ...       ...
14599  0.0  0.0   0.0  0.000000  0.000000   0.0  0.0  0.000000  0.0  0.000000
14600  0.0  0.0   0.0  0.000000  0.000000   0.0  0.0  0.000000  0.0  0.000000
14601  0.0  0.0   0.0  0.000000  0.000000   0.0  0.0  0.000000  0.0  0.000000
14602  0.0  0.0   0.0  0.047910  0.000000   0.0  0.0  0.000000  0.0  0.099647
14603  0.0  0.0   0.0  0.000000  0.000000   0.0  0.0  0.000000  0.0  0.000000

       ...  يوم  يول      يوص      يورانيوم  يور  يوح  يوج  يهم      يهل  \
0      ...  0.0  0.000000  0.0  0.0  0.0       0.0  0.084229  0.000000  0.0
1      ...  0.0  0.000000  0.0  0.0  0.0       0.0  0.000000  0.000000  0.0
2      ...  0.0  0.013955  0.0  0.0  0.0       0.0  0.000000  0.000000  0.0
3      ...  0.0  0.000000  0.0  0.0  0.0       0.0  0.000000  0.038384  0.0
4      ...  0.0  0.000000  0.0  0.0  0.0       0.0  0.000000  0.000000  0.0
...    ...  ...       ...  ...  ...  ...       ...       ...       ...  ...
14599  ...  0.0  0.000000  0.0  0.0  0.0       0.0  0.000000  0.000000  0.0
14600  ...  0.0  0.000000  0.0  0.0  0.0       0.0  0.000000  0.000000  0.0
14601  ...  0.0  0.000000  0.0  0.0  0.0       0.0  0.000000  0.000000  0.0
14602  ...  0.0  0.000000  0.0  0.0  0.0       0.0  0.000000  0.000000  0.0
14603  ...  0.0  0.000000  0.0  0.0  0.0       0.0  0.000000  0.000000  0.0
...
14602  0.0
14603  0.0

[14604 rows x 5000 columns]
```

Dimensions de data après la réduction :

```
[14604 rows x 5000 columns]
```

Finalement on a obtenu : Data prête.

6. Conclusion :

➔ Le projet a démontré l'efficacité de diverses techniques d'analyse de données appliquées à des textes en langue arabe. Il a également souligné l'importance du nettoyage préalable des données pour obtenir des résultats précis.

➔ **Futures tâches :**

Classification Automatique : Implémenter des modèles de classification automatique basés sur l'apprentissage automatique pour prédire la catégorie d'un texte.

➔ En implémentant ces tâches, le projet peut devenir une ressource encore plus puissante pour l'analyse de données en langue arabe.

II. Apprentissage Automatique :

1. Algorithmes d'apprentissage non-supervisé :

- **Clustering avec K-means :**

L'algorithme K-means est un algorithme de clustering qui vise à regrouper des points de données similaires dans des clusters.

Son utilité principale réside dans la segmentation des données en groupes homogènes, facilitant ainsi l'analyse exploratoire des données.

- **Choix du nb de cluster :** au nombre de catégories :

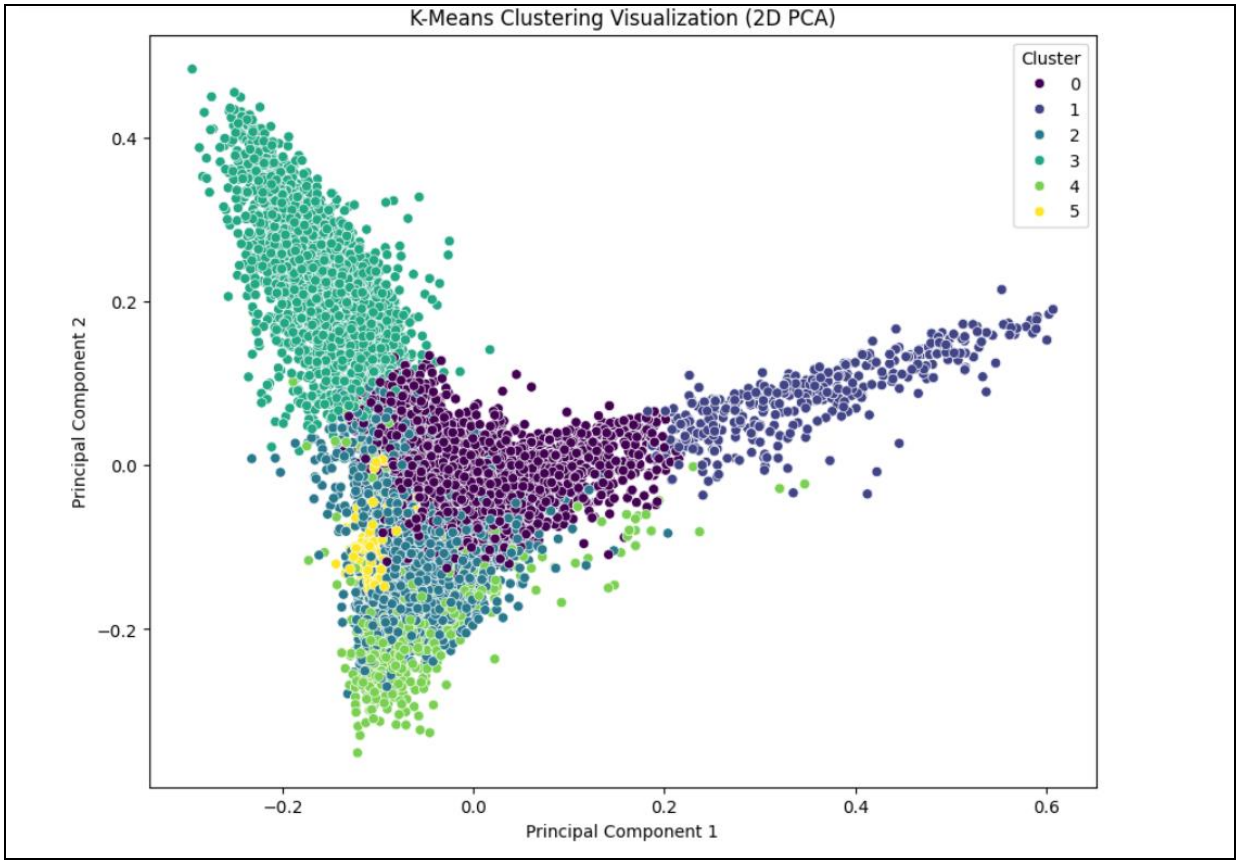
Six catégories ('Sports', 'Economy', 'Culture', 'Religion', 'International', 'Local')

➔ L'application de l'algorithme de clustering K-means se fait aux données réduites par l'ACP.

➔ Chaque ligne correspond à un document, et la colonne 'Cluster' indique le cluster auquel le document est assigné :

	Cluster
0	0
1	3
2	1
3	2
4	0
...	...
14599	0
14600	0
14601	0
14602	3
14603	2

➔ Visualisation des clusters dans un graphique en nuage de points en deux dimensions (PC1 et PC2) en utilisant Seaborn. Chaque point représente un document, et la couleur indique le cluster auquel il appartient.



➔ La visualisation des clusters dans un graphique en nuage de points en deux dimensions (PC1 et PC2) en utilisant Seaborn.

➔ Chaque point représente un document et la couleur indique le cluster auquel il appartient.

➔ La visualisation montre que les catégories Sport et Economie sont les mieux séparés

- ## Clustering avec DBSCAN :

DBSCAN est un algorithme de clustering non supervisé en apprentissage automatique puissant pour la détection de clusters, en particulier dans des ensembles de données où la densité des clusters varie. Cependant, il nécessite un ajustement prudent des paramètres et peut ne pas être optimal dans toutes les situations, en particulier lorsque les clusters ont des densités très différentes

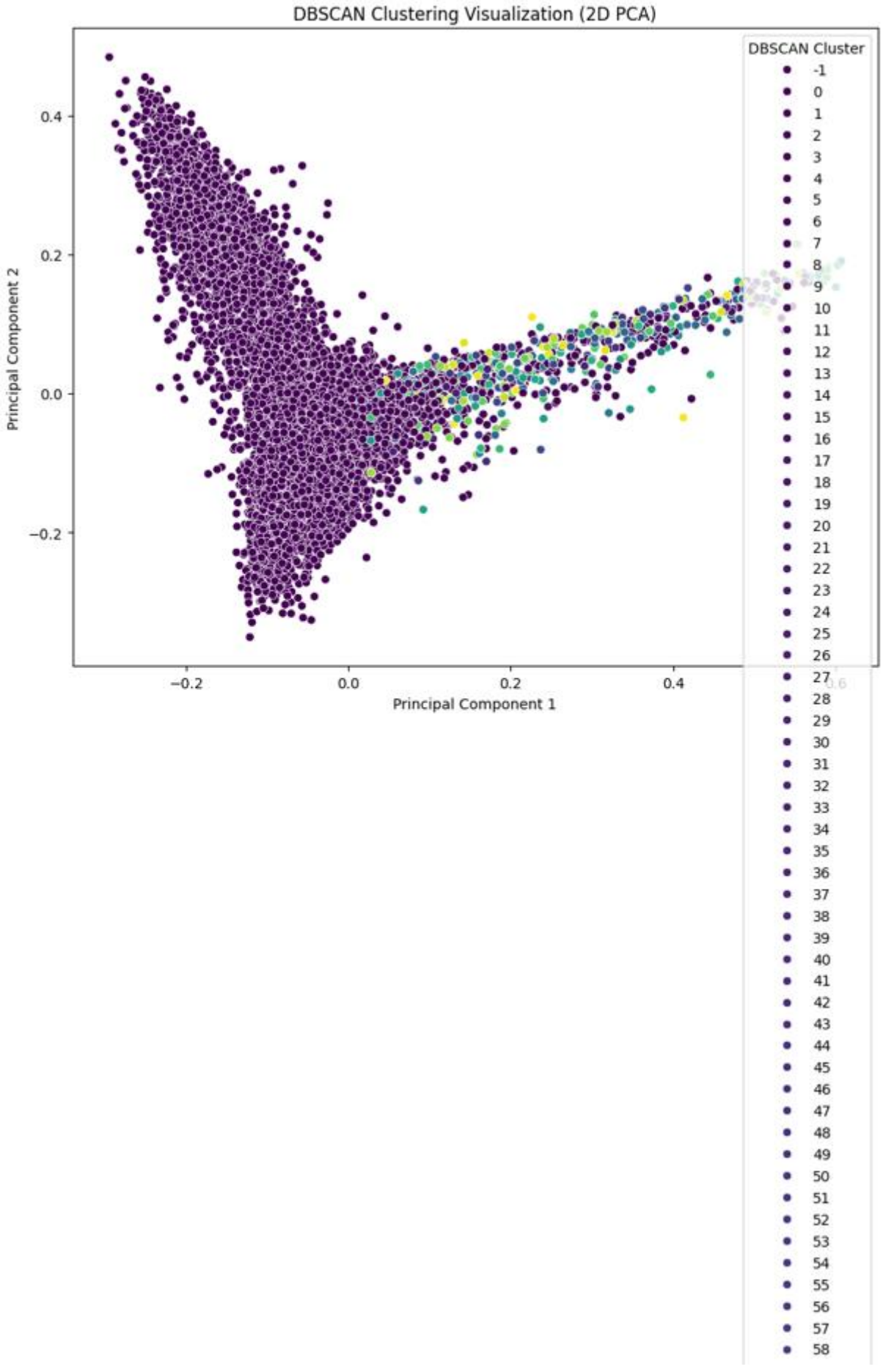
➔ L'algorithme DBSCAN est appliquée sur La matrice réduite TF-IDF (X_train_tfidf_reduced) normalisée en utilisant StandardScaler.

La normalisation est souvent importante pour les algorithmes basés sur la densité comme DBSCAN.

➔ **Ajustement de paramètres :** Les paramètres importants sont eps (rayon maximal pour considérer deux points comme voisins) et min_samples (nombre minimal de points dans un voisinage pour former un cluster). Ces valeurs doivent être ajustées en fonction des caractéristiques spécifiques des données.

```
# Apply DBSCAN clustering
eps_value = 0.5
min_samples_value = 5
```

➔ **Visualisation des données :** on utilise la bibliothèque seaborn pour créer une visualisation en nuage de points (scatter plot) en 2D des clusters DBSCAN. Chaque point représente un document, et chaque couleur un cluster


DBSCAN Clustering Visualization (2D PCA)
DBSCAN Cluster
-1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
0.4
0.2
0.0
−0.2
Principal Component 2
−0.2
0.0
0.2
0.4
Principal Component 1

➔ Les points sont colorés en fonction de leur cluster, avec les axes représentant les deux premières composantes principales (PC1 et PC2) de la réduction de dimensionnalité (PCA) des données.

➔ Le nombre de clusters est égale à 383

2. Algorithmes d'apprentissage supervisé :

- **Classification avec KNN :**

➔ K-Nearest Neighbors (KNN) est un algorithme d'apprentissage supervisé utilisé pour la classification et la régression.

➔ KNN est un algorithme simple et polyvalent, mais il a ses limites, en particulier en termes de coût computationnel, sensibilité à la dimensionnalité et nécessité d'une normalisation soigneuse des données.

➔ KNN est appliqué sur le texte stemmed et suivant les catégories des articles qui sont au nombre de 6 : ('Sports', 'Economy', 'Culture', 'Religion', 'International', 'Local')

➔ Les données sont divisées en ensembles d'entraînement (X_train, y_train) et de test (X_test, y_test). 80% des données sont utilisées pour l'entraînement et 20% pour les tests.

➔ Le nombre de neigbhors est fixé à : **n_neighbors=3**

```
Accuracy: 0.9233296823658269
Classification Report:
              precision    recall  f1-score   support

      Culture       0.87      0.91      0.89       527
      Economy       0.88      0.87      0.87       653
International       0.92      0.94      0.93       341
        Local       0.87      0.83      0.85       635
     Religion       0.98      1.00      0.99       698
       Sports       0.98      0.98      0.98       798

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

Interprétation du résultat :

- L'accuracy est la mesure globale de la performance du modèle.
  Dans ce cas, le modèle atteint une accuracy d'environ 92,33%, ce qui signifie que 92,33% des classifications sont correctes.
- La précision mesure la proportion d'instances correctement classées parmi celles qui ont été prédites comme appartenant à une classe spécifique.

Classe	Précision
Sport	**0.98**
Culture	**0.87**
Economy	**0.88**
Religion	**0.98**
International	**0.92**
Local	**0.87**

- Le rappel mesure la proportion d'instances correctement classées parmi toutes les instances réelles qui appartiennent à une classe spécifique.

Classe	Rappel
Sport	**0.98**
Culture	**0.91**
Economy	**0.87**
Religion	**1**
International	**0.94**
Local	**0.83**

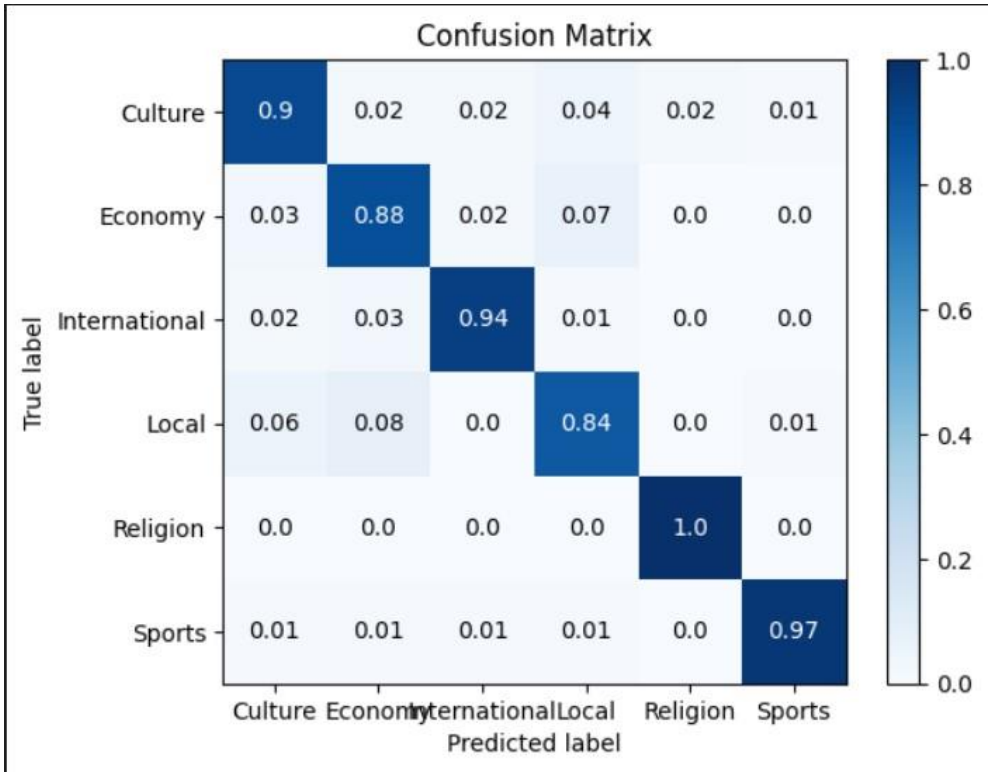
- Le F1-score est une moyenne pondérée de la précision et du rappel. Il est particulièrement utile lorsque les classes sont déséquilibrées.

Classe	F1-score
Sport	**0.98**
Culture	**0.89**
Economy	**0.87**
Religion	**0.99**
International	**0.93**
Local	**0.85**

- La ligne "macro avg" représente la moyenne non pondérée des métriques (précision, rappel, F1-score) pour chaque classe.

- La ligne "weighted avg" représente la moyenne pondérée des métriques en fonction du support de chaque classe.

Matrice de confusion :



La matrice de confusion est une représentation normalisée qui montre la performance d'un modèle de classification sur plusieurs classes. Chaque ligne de la matrice correspond à une classe réelle, tandis que chaque colonne correspond à la classe prédite. Les valeurs dans la matrice représentent la proportion d'échantillons réels de chaque classe correctement (diagonale principale) ou incorrectement (hors diagonale principale) classifiés.

En analysant la matrice de confusion :

Culture :

90% des échantillons réels de Culture ont été correctement classés comme Culture. Il y a des confusions avec Economy (2%), Local (4%), et Sports (1%).

Economy :

88% des échantillons réels d'Économie ont été correctement classés comme Économie.
Il y a des confusions avec Culture (3%) et Local (7%).

International :

94% des échantillons réels d'International ont été correctement classés comme International.
Il y a des confusions très faibles avec les autres catégories.

Local :

84% des échantillons réels de Local ont été correctement classés comme Local.
Il y a des confusions avec Culture (6%), Economy (8%), et Sports (1%).

Religion :

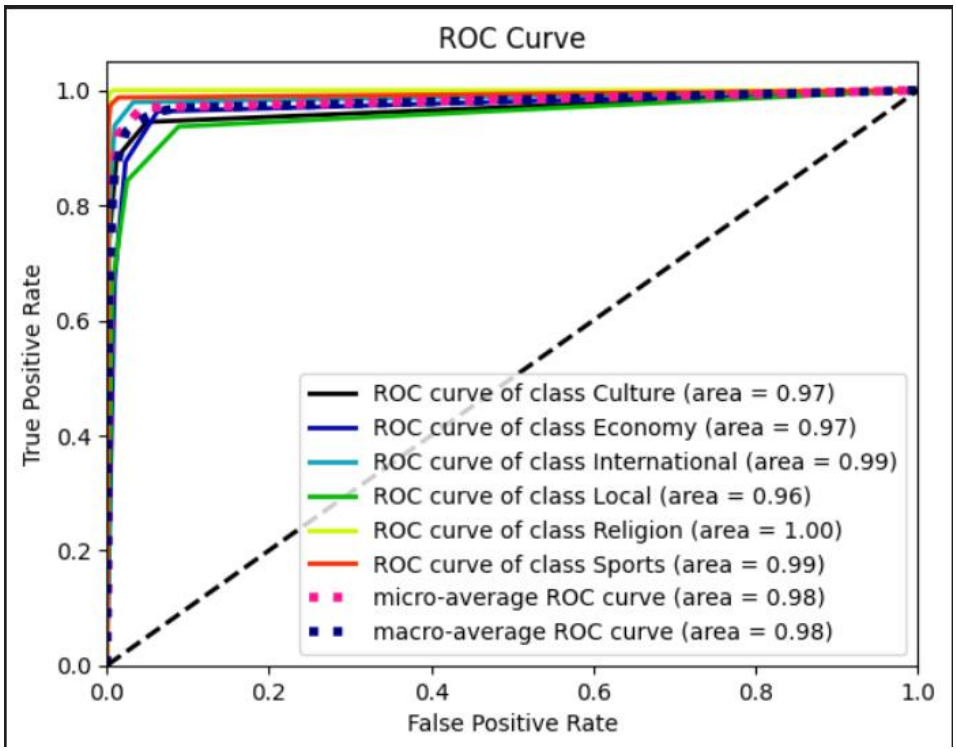
100% des échantillons réels de Religion ont été correctement classés comme Religion.
Aucune confusion avec d'autres catégories.

Sports :

97% des échantillons réels de Sports ont été correctement classés comme Sports.
Il y a des confusions avec Culture (1%) et Local (1%).

➡ La diagonale principale de la matrice montre les taux de classification corrects pour chaque classe, tandis que les valeurs en dehors de la diagonale indiquent les erreurs de classification. Globalement, la matrice de confusion semble montrer une performance solide du modèle, avec des taux élevés de classification correcte pour la plupart des classes.

La courbe ROC (Receiver Operating Characteristic):



La courbe ROC (Receiver Operating Characteristic) est un graphique qui illustre la performance d'un modèle de classification binaire à différents seuils de probabilité. Elle est souvent utilisée pour évaluer la capacité d'un modèle à discriminer entre les classes positives et négatives.

- **True Positive Rate** (Taux de vrais positifs) : C'est le taux de vrais positifs par rapport à tous les exemples réels positifs. En d'autres termes, il mesure la capacité du modèle à classer correctement les exemples positifs.

- **False Positive Rate** (Taux de faux positifs) : C'est le taux de faux positifs par rapport à tous les exemples réels négatifs. Il mesure la capacité du modèle à éviter de classer incorrectement les exemples négatifs.

Chaque courbe ROC représentée dans le graphique correspond à une classe spécifique du modèle de classification multiclasse, et elle est associée à une "aire sous la courbe" (AUC). L'AUC est un indicateur de la performance globale du modèle, où une valeur plus proche de 1 indique une meilleure performance.

- ROC curve of class **Culture (area = 0.97)** : Cette courbe montre la performance du modèle pour la classe Culture, avec une AUC de 0.97.

- ROC curve of class **Economy (area = 0.97)** : Cette courbe montre la performance du modèle pour la classe Economy, avec une AUC de 0.97.

- ROC curve of class **International (area = 0.99)** : Cette courbe montre la performance du modèle pour la classe International, avec une AUC de 0.99.

- ROC curve of class **Local (area = 0.96)** : Cette courbe montre la performance du modèle pour la classe Local, avec une AUC de 0.96.

- ROC curve of class **Religion (area = 1)** : Cette courbe montre la performance du modèle pour la classe Religion, avec une AUC de 1, ce qui indique une classification parfaite.

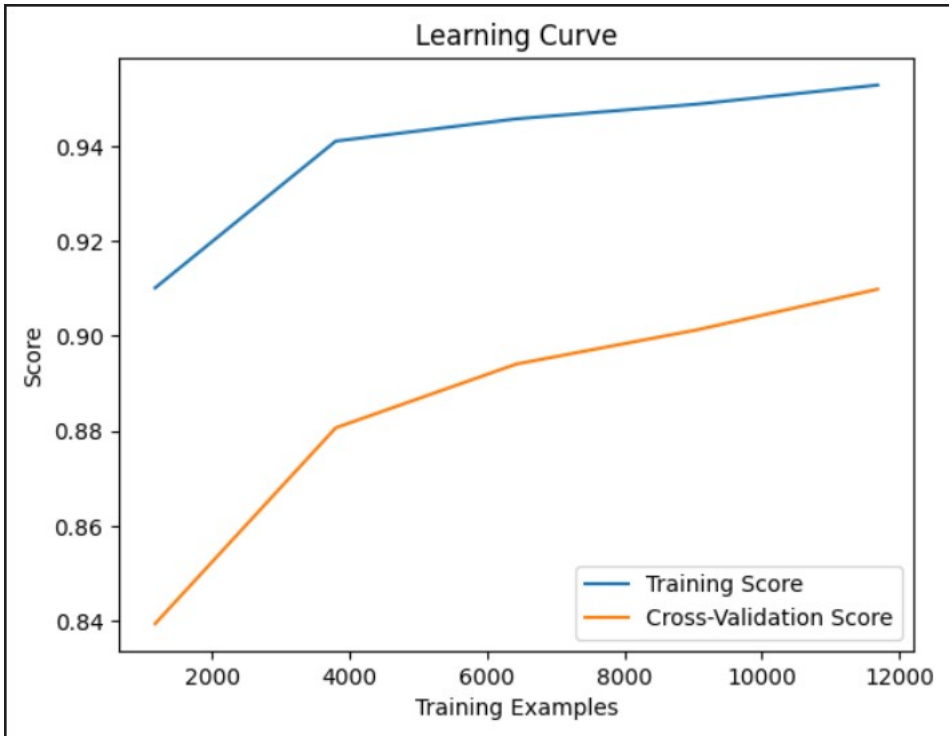
- ROC curve of class **Sports** (**area = 0.99**) : Cette courbe montre la performance du modèle pour la classe Sports, avec une AUC de 0.99.

Micro-average ROC curve (area = 0.98) : Cette courbe représente la performance globale agrégée sur toutes les classes à l'aide d'une approche micro-moyenne.

Macro-average ROC curve (area = 0.98) : Cette courbe représente la performance globale agrégée sur toutes les classes à l'aide d'une approche macro-moyenne.

➡ En résumé, la courbe ROC et les AUC spécifiques à chaque classe donnent une indication visuelle de la capacité du modèle à discriminer entre les différentes classes. Des AUC élevées (plus proches de 1) suggèrent une meilleure performance du modèle.

La courbe d'apprentissage (learning curve) :



La courbe d'apprentissage (learning curve) est un graphique qui montre comment la performance d'un modèle évolue en fonction du nombre d'exemples d'entraînement. Elle est généralement utilisée pour évaluer si un modèle bénéficierait de l'ajout de plus de données d'entraînement et pour détecter des problèmes tels que le surajustement ou le sous-ajustement.

Interprétation du graphique :

Training Score : La courbe de la performance du modèle sur l'ensemble d'entraînement. Elle indique comment le modèle "apprend" à partir des données d'entraînement à mesure que la taille de l'ensemble d'entraînement augmente.

Cross-Validation Score : La courbe de la performance du modèle sur l'ensemble de validation croisée. Elle mesure la capacité du modèle à généraliser sur des données qu'il n'a pas vues pendant l'entraînement.

Training Examples : Le nombre d'exemples d'entraînement. L'axe horizontal montre la taille de l'ensemble d'entraînement.

En analysant la courbe d'apprentissage :

On observe que les deux courbes, celle de l'entraînement et celle de la validation croisée, **convergent vers un score élevé**, ce qui est une bonne chose. Cela suggère que le modèle bénéficie de l'ajout de plus de données d'entraînement et qu'il n'y a pas de surajustement significatif.

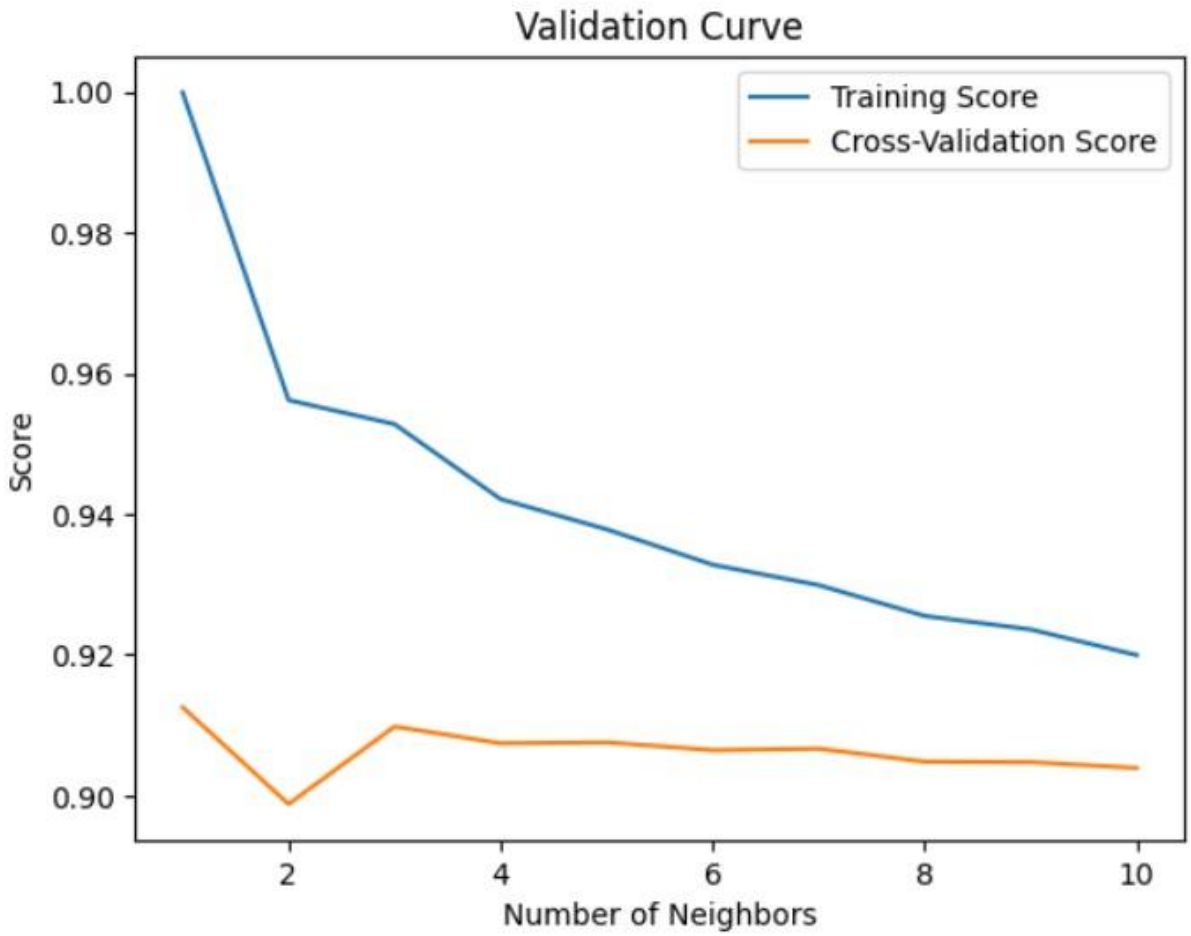
Cependant, il y a quelques points à noter :

- Tout d'abord, la courbe d'entraînement atteint un score de 0,94 après seulement 2 000 exemples d'entraînement. Cela suggère que le modèle pourrait être en train de surajuster les données d'entraînement. Pour confirmer ou infirmer cette hypothèse, il serait utile de voir comment la courbe d'entraînement évolue au-delà de 2 000 exemples d'entraînement.

- Deuxièmement, la courbe de validation croisée atteint un score de 0,92 après 10 000 exemples d'entraînement. Cela suggère que le modèle pourrait encore bénéficier d'un peu plus de données d'entraînement. Si la courbe de validation croisée était capable d'atteindre un score de 0,94 ou plus, cela suggérerait que le modèle est bien généralisable à de nouvelles données.

En conclusion, la courbe d'apprentissage suggère que le modèle est performant.

La courbe de validation (Validation Curve) :



La courbe de validation (Validation Curve) est un graphique qui permet d'analyser comment la performance d'un modèle varie en fonction des différents paramètres du modèle. Elle est souvent utilisée pour déterminer la meilleure valeur à utiliser pour un hyperparamètre particulier.

Analysons le graphique :

Training Score : La courbe montre la performance du modèle sur l'ensemble d'entraînement en fonction du nombre de voisins. Elle indique comment le modèle s'ajuste aux données d'entraînement lorsque le nombre de voisins change.

Cross-Validation Score : La courbe montre la performance du modèle sur l'ensemble de validation croisée en fonction du nombre de voisins. Elle mesure la capacité du modèle à généraliser sur des données qu'il n'a pas vues pendant l'entraînement.

Number of Neighbors : L'axe horizontal montre la valeur du paramètre que vous avez fait varier, dans ce cas, le nombre de voisins.

Analysons les résultats :

Pour le Training Score : On observe que le score d'entraînement est presque parfait (proche de 1) quel que soit le nombre de voisins. Cela suggère que le modèle est capable de s'ajuster très bien aux données d'entraînement, mais cela pourrait également indiquer un surajustement si le score de validation n'est pas aussi élevé.

Pour le Cross-Validation Score : On observe un point de score de validation maximisé autour de 2 à 4 voisins, après quoi le score diminue légèrement. Cela suggère que le modèle fonctionne mieux avec un nombre limité de voisins.

➡ En interprétant ce graphique, il semble que le nombre optimal de voisins se situe autour de 2 à 4, où le score de validation croisée est maximisé. Au-delà de ce nombre, le modèle commence à surajuster les données d'entraînement, ce qui se reflète dans la diminution du score de validation. Vous pourriez choisir un nombre de voisins dans cette plage pour optimiser les performances de votre modèle k-plus proches voisins.

- **Décision Tree :**

➔ **Précision globale :** La précision globale du modèle d'arbre de décision est d'environ 85,4%, ce qui est un bon résultat. Cela signifie que le modèle classe correctement la catégorie correcte dans environ 85,4% des cas.

```
Accuracy: 0.8543263964950711
Classification Report:
                precision    recall  f1-score   support

       Culture       0.81      0.75      0.78       527
       Economy       0.79      0.78      0.78       653
 International       0.84      0.86      0.85       341
         Local       0.73      0.74      0.73       635
      Religion       0.96      1.00      0.98       698
        Sports       0.94      0.96      0.95       798

      accuracy                           0.85      3652
     macro avg       0.84      0.85      0.84      3652
  weighted avg       0.85      0.85      0.85      3652
```

➔ **Précision par catégorie :** On a également fourni une évaluation détaillée par catégorie, montrant la précision, le rappel et le score F1 pour chaque classe.

➔ Les F1-Scores pour chaque classe sont relativement élevés, variant de 0,73 à 0,98, suggérant un bon équilibre entre la précision et le rappel.

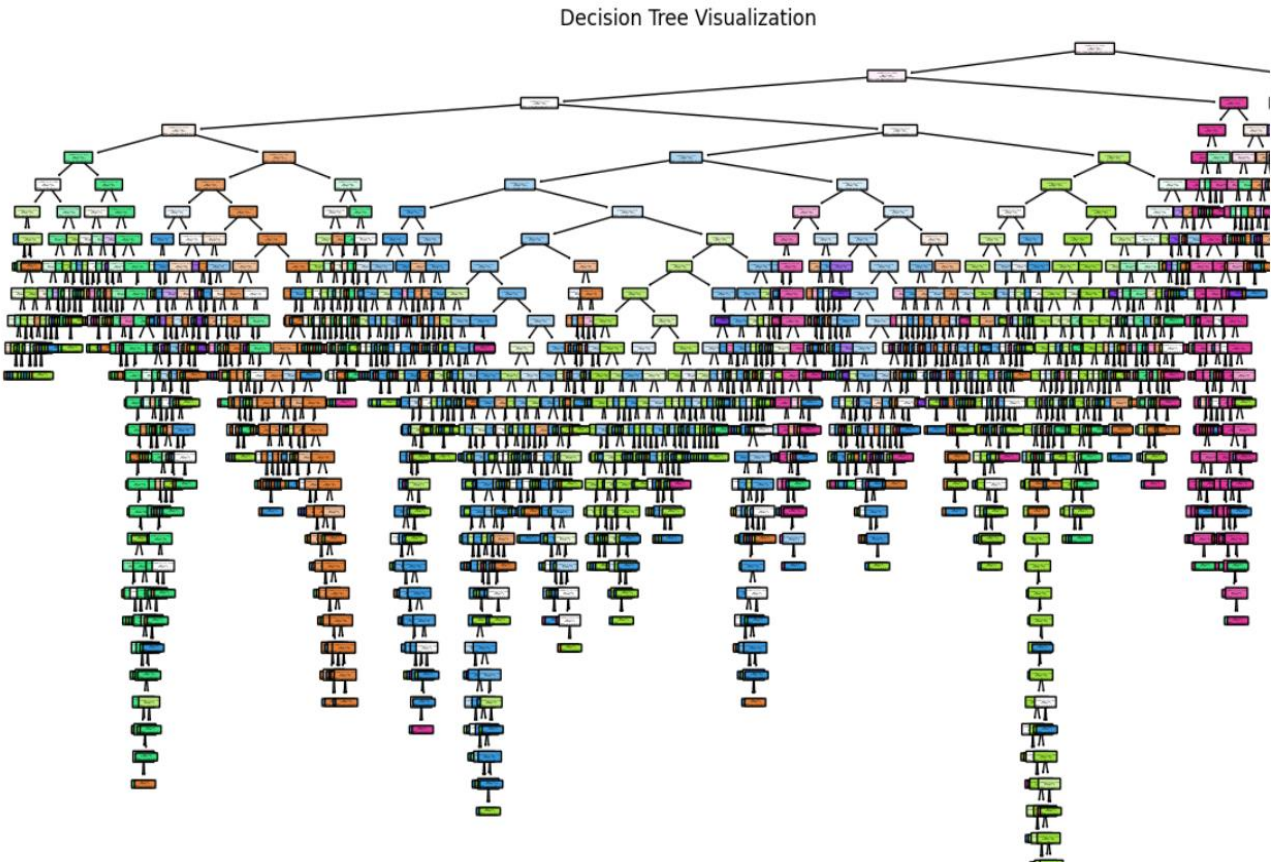
➔ La plus haute précision est observée pour "Religion" (96 %), indiquant que lorsque le modèle prédit la classe "Religion", il a raison 96 % du temps.

➔ Les rappels les plus élevés sont observés pour "Religion" et "International" (tous deux à 100 % et 86 %, respectivement), indiquant que le modèle capture très bien toutes les instances de ces classes.

➔ Les moyennes macro et pondérées de la précision, du rappel et du F1-Score sont d'environ 0,84 à 0,85, indiquant une bonne performance globale sur l'ensemble des classes.

➔ La visualisation de l'arbre de décision est utile pour comprendre comment le modèle prend des décisions.

➔ Voici une vue globale sur notre decision Tree :



- **Réseaux de neurones :**

Construction du modèle de réseau neuronal :

Un modèle de réseau neuronal séquentiel est construit avec trois couches : une couche d'entrée (Dense), une couche cachée et une couche de sortie avec une fonction d'activation softmax.

Le modèle est compilé avec l'optimiseur 'adam' et la perte 'sparse_categorical_crossentropy'.

Entraînement du modèle :

Le modèle est entraîné sur l'ensemble d'entraînement avec 10 époques et une taille de lot (batch size) de 32.

Les performances du modèle sont évaluées sur un ensemble de validation.

Évaluation du modèle sur l'ensemble de test :

Le modèle est évalué sur l'ensemble de test, et l'exactitude du modèle est affichée.

Prédictions et rapport de classification :

Les prédictions sont effectuées sur l'ensemble de test.

Le rapport de classification est affiché, montrant la précision, le rappel et le score F1 pour chaque classe.

```
Epoch 1/10
366/366 [==============================] - 1s 1ms/step - loss: 0.6480 - accuracy: 0.8089 - val_loss: 0.2801 - val_accuracy: 0.9083
Epoch 2/10
366/366 [==============================] - 0s 1ms/step - loss: 0.2560 - accuracy: 0.9099 - val_loss: 0.2629 - val_accuracy: 0.9117
Epoch 3/10
366/366 [==============================] - 0s 1ms/step - loss: 0.2373 - accuracy: 0.9169 - val_loss: 0.2474 - val_accuracy: 0.9165
Epoch 4/10
366/366 [==============================] - 0s 1ms/step - loss: 0.2264 - accuracy: 0.9165 - val_loss: 0.2427 - val_accuracy: 0.9182
Epoch 5/10
366/366 [==============================] - 0s 1ms/step - loss: 0.2144 - accuracy: 0.9225 - val_loss: 0.2500 - val_accuracy: 0.9158
Epoch 6/10
366/366 [==============================] - 0s 1ms/step - loss: 0.2068 - accuracy: 0.9262 - val_loss: 0.2360 - val_accuracy: 0.9226
Epoch 7/10
366/366 [==============================] - 0s 1ms/step - loss: 0.1991 - accuracy: 0.9288 - val_loss: 0.2340 - val_accuracy: 0.9243
Epoch 8/10
366/366 [==============================] - 0s 1ms/step - loss: 0.1912 - accuracy: 0.9312 - val_loss: 0.2243 - val_accuracy: 0.9267
Epoch 9/10
366/366 [==============================] - 0s 1ms/step - loss: 0.1844 - accuracy: 0.9333 - val_loss: 0.2217 - val_accuracy: 0.9240
Epoch 10/10
366/366 [==============================] - 0s 1ms/step - loss: 0.1757 - accuracy: 0.9367 - val_loss: 0.2157 - val_accuracy: 0.9278
115/115 [==============================] - 0s 772us/step - loss: 0.2129 - accuracy: 0.9302
Accuracy: 0.930175244808197
115/115 [==============================] - 0s 705us/step
Classification Report:
              precision    recall  f1-score   support
...
     accuracy                           0.93      3652
    macro avg       0.93      0.93      0.93      3652
 weighted avg       0.93      0.93      0.93      3652
```

Interprétation des résultats :

Le modèle de réseau neuronal atteint une exactitude (accuracy) d'environ 93% sur l'ensemble de test.

Le rapport de classification fournit des mesures détaillées pour chaque classe, montrant de bonnes performances en termes de précision, rappel et score F1 pour chaque catégorie.

L'évolution des performances du modèle (loss et accuracy) est affichée pour chaque époque d'entraînement.

➡ **Ces résultats suggèrent que le modèle de réseau neuronal est capable de bien généraliser à partir des données d'entraînement et de faire des prédictions précises sur de nouvelles données.**