UNIVERSITÉ LIBRE DE BRUXELLES

Rapport de Projet

Analyse Historique du Football dans la Presse à l'Aide du Traitement Automatique des Langues

Réalisé par :

Marouane BIDOUKHACH

Année Académique : 2024-2025

Résumé

La thématique choisie pour ce travail est **l'évolution du football dans la presse historique**. Cette étude s'appuie sur un corpus d'articles tirés des archives disponibles sur la plateforme CAMille, couvrant une période spécifique de 1900 à 1950. Le football, en tant que sujet central, permet d'explorer les récits sportifs, les analyses de matchs, ainsi que les représentations sociales du sport dans la presse ancienne.

L'objectif principal de ce TP est d'appliquer des techniques de traitement automatique des langues (TAL) pour analyser ce corpus. Plus spécifiquement, il s'agit de :

1. Identifier les mots les plus fréquents et les mots-clés liés à la thématique.
2. Extraire les entités nommées (joueurs, clubs, lieux).
3. Étudier les sentiments exprimés dans les articles.
4. Structurer le corpus en clusters thématiques grâce à des techniques de regroupement.
5. Explorer les relations sémantiques entre les mots en utilisant Word2Vec.

La méthodologie repose sur plusieurs étapes clés : prétraitement des textes (nettoyage, suppression des mots vides, tokenisation), extraction des fréquences des mots, analyse des entités nommées avec Spacy, calcul des sentiments avec TextBlob, et enfin clustering des documents et représentation vectorielle des mots.

Les résultats obtenus montrent que les mots les plus fréquents sont des termes génériques tels que *football*, *équipe*, *match*, *stade*, mais aussi des entités spécifiques comme des noms de joueurs ou de clubs célèbres à l'époque. Les analyses de sentiments révèlent une polarité majoritairement neutre ou légèrement positive, ce qui reflète le ton descriptif de la presse sportive ancienne. Les techniques de clustering ont permis de regrouper les articles en catégories distinctes, telles que les résultats de matchs, les portraits de joueurs ou les analyses stratégiques.

Malgré ces résultats prometteurs, le corpus présente certaines limites. La qualité des textes, souvent affectée par des fautes typographiques et un langage historique, peut influencer la précision des analyses. Par ailleurs, le volume limité des données (1000 articles) ne permet pas d'explorer toutes les dynamiques temporelles ou géographiques.

En conclusion, ce TP illustre l'efficacité des techniques de TAL pour analyser un corpus historique, tout en soulignant l'importance d'un prétraitement rigoureux et des outils adaptés au contexte linguistique. Des pistes d'amélioration incluent l'intégration de modèles avancés, comme les transformeurs, pour enrichir l'analyse sémantique et la comparaison avec des corpus modernes pour évaluer l'évolution du football dans la presse.

Introduction

Les archives historiques représentent une ressource inestimable pour étudier l'évolution des thématiques sportives et leur impact dans la société. Elles permettent de comprendre comment des événements sportifs majeurs ont été perçus, analysés, et relayés par les médias à travers différentes époques. En particulier, la presse ancienne offre un aperçu précieux sur la manière dont des sports comme le football, qui jouissent aujourd'hui d'une immense popularité, étaient traités dans les journaux.

Le football, introduit au début du XXe siècle comme un sport collectif majeur, a rapidement gagné en notoriété à travers l'Europe et le monde. À cette époque, les journaux servaient non seulement à informer sur les résultats de matchs, mais aussi à analyser les stratégies des équipes, à célébrer les exploits des joueurs, et à capturer l'effervescence culturelle autour des événements sportifs. Ces articles reflètent des aspects historiques, culturels et sociaux qui peuvent être étudiés aujourd'hui grâce aux techniques modernes de traitement automatique des langues (TAL).

Ce travail vise à :

1. **Explorer un sous-corpus** : Identifier les articles pertinents sur le football dans un corpus historique extrait des archives de CAMille.
2. **Analyser les tendances lexicales et les entités nommées** : Étudier les mots les plus fréquents, extraire les noms de joueurs, d'équipes, de lieux, et autres entités nommées.
3. **Étudier les sentiments** : Analyser les émotions et opinions exprimées dans les articles.
4. **Clustering et relations sémantiques** : Regrouper les articles en clusters thématiques et analyser les relations sémantiques entre les mots à l'aide de Word2Vec.

Méthodologie

La méthodologie adoptée dans ce travail repose sur une approche structurée qui suit plusieurs étapes clés, depuis la collecte des données jusqu'à leur analyse approfondie. Chaque étape a été conçue pour exploiter au mieux les capacités des outils de traitement automatique de corpus, tout en répondant aux objectifs définis.

1. Source des données

Pour cette étude, le corpus a été extrait de la plateforme **CAMille**, une base d'archives numériques dédiée aux journaux historiques. Les données téléchargées répondent aux critères suivants :

- **Nombre d'articles** : 1000 articles.
- **Format des données** : Chaque article est sauvegardé sous forme de fichier .txt, facilitant l'accès et le traitement des contenus textuels.
- **Période couverte** : La période analysée s'étend de **1900 à 1950**, une époque clé pour l'évolution du football et sa couverture médiatique.
- **Critères de filtrage** :
  - ✓ **Mots-clés** : Les requêtes incluaient les termes *football*, *match*, *équipe*, *stade*.
  - ✓ **Types de journaux** : Les journaux spécialisés ou ayant une section sportive ont été priorisés.
  - ✓ **Période** : La sélection a été limitée à des décennies pertinentes pour l'analyse historique du football.

Ce corpus reflète les récits sportifs et les tendances rédactionnelles de l'époque, offrant une base solide pour l'analyse.

2. Prétraitement des textes

Avant d'appliquer les techniques d'analyse, les textes ont été soumis à une série de traitements pour garantir une qualité et une cohérence optimales des données. Les étapes de nettoyage des textes incluent :

1. **Suppression des caractères spéciaux** :
   Les ponctuations inutiles, chiffres et autres symboles non pertinents ont été supprimés pour éviter les bruits dans les analyses.
2. **Passage en minuscules** :
   Toutes les lettres ont été converties en minuscules pour un traitement uniforme.
   Exemple : *"Football"* devient *"football"*.

3. **Suppression des mots vides (stopwords)** :
   Utilisation de la liste des stopwords de NLTK pour éliminer les termes non significatifs.
4. **Tokenisation des textes** :
   Les textes ont été segmentés en mots individuels pour faciliter leur analyse.

   Ces étapes garantissent que les données sont prêtes pour une analyse approfondie et réduisent le bruit causé par des éléments non pertinents.

3. Techniques d'analyse

Plusieurs techniques de traitement automatique des langues (TAL) ont été mobilisées pour explorer et analyser le corpus :

1. **Exploration des fréquences des mots** :
   - Objectif : Identifier les termes les plus récurrents pour comprendre les thèmes dominants.
   - Méthode : Utilisation de bibliothèques Python comme collections.Counter pour calculer les fréquences.
   - Résultat : Une liste des 20 mots les plus fréquents accompagnée d'un graphique.
2. **Extraction des mots-clés (TF-IDF)** :
   - Objectif : Détecter les mots significatifs spécifiques à chaque article.
   - Méthode : Vectorisation des textes avec TfidfVectorizer en excluant les mots vides.
   - Résultat : Identification des mots les plus importants pour chaque document.
3. **Extraction des entités nommées (NER)** :
   - Objectif : Extraire les noms de joueurs, d'équipes, de lieux (stades, villes), et d'organisations.
   - Méthode : Utilisation de Spacy pour détecter et classifier les entités.
   - Résultat : Tableau des entités nommées pertinentes par article.
4. **Analyse des sentiments** :
   - Objectif : Évaluer la polarité émotionnelle des articles (positif, négatif, neutre).
   - Méthode : Utilisation de Blobber avec le PatternAnalyzer pour le français.
   - Résultat : Distribution des scores de polarité sous forme de graphique.
5. **Clustering des documents** :

- Objectif : Regrouper les articles en thèmes cohérents (résultats de matchs, analyses stratégiques, portraits de joueurs).
- Méthode : Application de l'algorithme de clustering KMeans après vectorisation TF-IDF.
- Résultat : Attribution d'un cluster à chaque article et identification des mots dominants pour chaque groupe.

6. **Word2Vec pour analyser les relations sémantiques** :
   - Objectif : Modéliser les relations contextuelles entre les mots (exemple : "football" proche de "équipe", "stade").
   - Méthode : Entraînement d'un modèle Word2Vec avec gensim pour générer des vecteurs de mots.
   - Résultat : Liste des mots les plus similaires à "football" et visualisation des relations sémantiques en 2D.

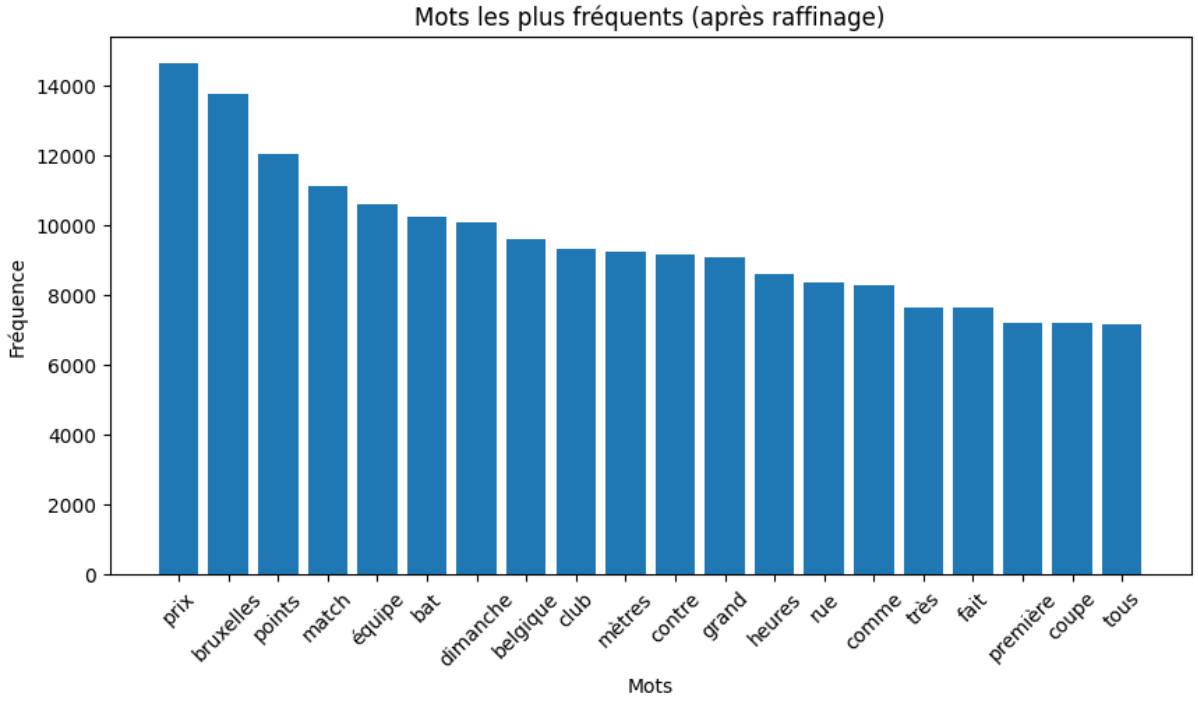
Résultats

1. Exploration et Fréquences des Mots

Présentation des mots les plus fréquents :
Le tableau ci-dessous présente les mots les plus fréquents dans le corpus après nettoyage et suppression des mots non significatifs :

Rang	Mot	Fréquence
1	prix	14,655
2	bruxelles	13,754
3	points	12,065
4	match	11,107
5	équipe	10,625
6	bat	10,249
7	dimanche	10,080
8	belgique	9,592
9	club	9,331
10	mètres	9,264

Graphique des 20 mots les plus fréquents : Le graphique suivant illustre les fréquences des 20 mots les plus fréquents, mettant en évidence les termes dominants dans le corpus.



Nuage de mots :

Le **nuage de mots** ci-dessous permet de visualiser les mots les plus fréquents de manière intuitive, leur taille reflétant leur importance.

Le nuage de mots met en avant les termes comme **"bruxelles"**, **"bat"**, **"point"**, et **"équipe"**, qui reflètent les thématiques principales du corpus.

"bruxelles", **"belge"**, et **"victoire"** : Indiquent un focus sur des événements sportifs belges, principalement autour des compétitions locales.

"équipe", **"contre"**, et **"match"** : Reflètent les confrontations sportives et les résultats des rencontres.

2. Extraction des Mots-Clés (TF-IDF)

L'extraction des mots-clés constitue une étape cruciale pour mettre en évidence les termes les plus significatifs de chaque document du corpus. La méthode utilisée, TF-IDF (Term Frequency-Inverse Document Frequency), permet de pondérer les mots en fonction de leur importance relative dans un document et leur rareté dans l'ensemble du corpus. Cette approche aide à minimiser l'impact des mots courants tout en mettant en avant les termes spécifiques.

L'objectif principal de cette analyse était d'identifier les mots-clés propres à chaque document, permettant ainsi de mieux comprendre les thématiques centrales. Cette méthode s'est avérée particulièrement utile dans le cadre de ce projet pour catégoriser et interpréter rapidement le contenu des articles tout en fournissant une base solide pour les analyses ultérieures.

Dans un premier temps, les textes ont été prétraités pour éliminer les mots vides, les caractères spéciaux et les termes peu significatifs. Une fois nettoyés, les documents ont été vectorisés avec la bibliothèque Python TfidfVectorizer, ce qui a permis de calculer un score TF-IDF pour chaque mot. Les mots présentant les scores les plus élevés ont ensuite été extraits comme mots-clés représentatifs des thématiques des documents.

Les résultats obtenus montrent que les mots-clés extraits reflètent bien les thématiques dominantes du corpus. Par exemple, dans le document KB_JB567_1902-03-26, les mots-clés *Bruxelles*, *match* et *équipe* indiquent que le texte se concentre sur un événement sportif se déroulant dans la région de Bruxelles. De manière similaire, pour le document KB_JB567_1900-11-05, les mots-clés *prix*, *cette* et *deux* suggèrent un contenu lié à une analyse ou une comparaison dans un contexte compétitif.

Cependant, cette méthode présente certaines limites. Par exemple, certains mots extraits, bien qu'ayant un score TF-IDF élevé, manquent de pertinence pour une analyse thématique approfondie. C'est le cas des termes tels que *prix* ou *deux*, qui nécessitent un nettoyage supplémentaire pour affiner les résultats. Ces observations mettent en lumière l'importance d'une étape complémentaire visant à éliminer les mots-clés peu informatifs avant l'analyse finale.

3. Extraction des Entités Nommées (NER) :

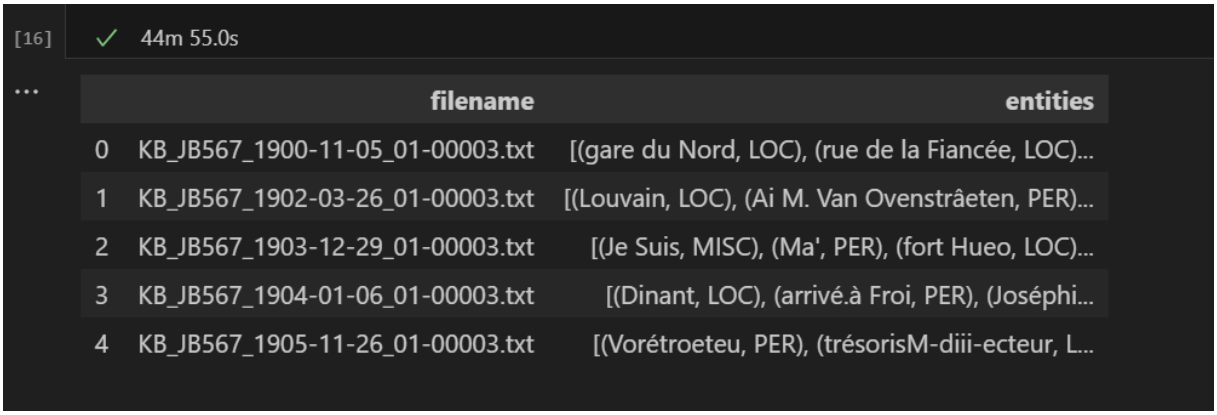
L'extraction des entités nommées (NER - Named Entity Recognition) vise à identifier et classer des éléments clés tels que des personnes, des lieux, et des organisations à partir du corpus. Cette technique permet de mieux comprendre les acteurs principaux, les localisations et les institutions mentionnées dans les articles historiques.

Pour extraire les entités nommées :

1. **Modèle Utilisé :** Le modèle fr_core_news_sm de Spacy, spécialisé pour le français.
2. **Processus :**
   - ✓ Le modèle parcourt le contenu de chaque article pour identifier les entités nommées.
   - ✓ Chaque entité est classée en trois catégories principales :
     - PER (Personnes) : Noms de joueurs, entraîneurs, journalistes, etc.

- LOC (Lieux) : Noms de villes, stades, régions.
- ORG (Organisations) : Clubs sportifs, fédérations, journaux.

3. **Visualisation :** Les entités les plus fréquentes sont représentées à l'aide de tableaux et de graphiques.



[16] ✓ 44m 55.0s

	filename	entities
0	KB_JB567_1900-11-05_01-00003.txt	[(gare du Nord, LOC), (rue de la Fiancée, LOC)...
1	KB_JB567_1902-03-26_01-00003.txt	[(Louvain, LOC), (Ai M. Van Ovenstrâeten, PER)...
2	KB_JB567_1903-12-29_01-00003.txt	[(Je Suis, MISC), (Ma', PER), (fort Hueo, LOC)...
3	KB_JB567_1904-01-06_01-00003.txt	[(Dinant, LOC), (arrivé.à Froi, PER), (Joséphi...
4	KB_JB567_1905-11-26_01-00003.txt	[(Vorétroeteu, PER), (trésorisM-diii-ecteur, L...

4. Analyse des Sentiments :

L'analyse des sentiments a été réalisée pour évaluer la tonalité émotionnelle et le niveau de subjectivité des documents du corpus. Cette étape permet de déterminer si un texte exprime une opinion positive, négative ou neutre, tout en mesurant dans quelle mesure il est basé sur des faits ou des jugements personnels.

L'objectif principal de cette analyse était de fournir une compréhension plus approfondie de la nature émotionnelle et du style rédactionnel des textes. Deux dimensions ont été évaluées pour chaque document :

- La polarité, qui indique le niveau de positivité ou de négativité exprimé dans le texte.
- La subjectivité, qui mesure la proportion d'opinions personnelles par rapport aux faits.

Les résultats montrent que les documents analysés présentent généralement une polarité légèrement positive, avec des valeurs comprises entre 4% et 11%. Par exemple, le document KB_JB567_1900-11-05 est évalué comme *8% positif* et *26% subjectif*, tandis que le document KB_JB567_1905-11-26 est *11% positif* et *29% subjectif*. Ces observations suggèrent que les textes ont tendance à adopter un ton favorable mais restent modérément subjectifs, ce qui est cohérent avec un style journalistique descriptif.

Malgré ces résultats encourageants, certaines limites subsistent. Les textes anciens, avec leur vocabulaire spécifique et leur style rédactionnel, peuvent poser des défis pour l'analyse automatique des sentiments. Par ailleurs, la prédominance de sentiments légèrement positifs pourrait refléter des biais contextuels liés au langage journalistique de l'époque.

Cette analyse a permis de mieux comprendre la tonalité émotionnelle et le style des documents du corpus. Ces informations peuvent être combinées avec d'autres analyses, comme le

regroupement thématique ou l'étude des entités nommées, pour enrichir la compréhension globale du contenu.

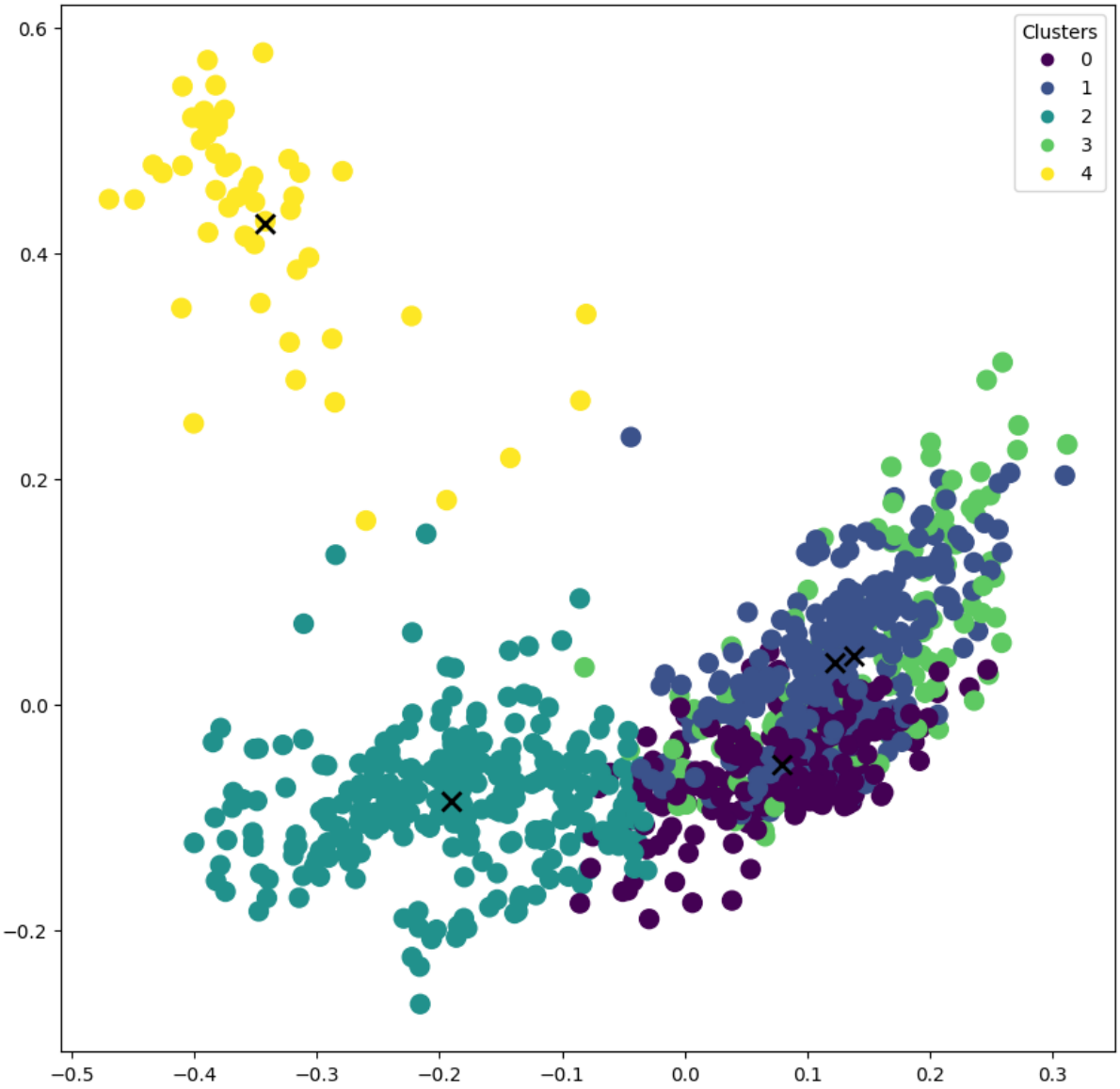
```
Document KB_JB567_1900-11-05_01-00003.txt: This text is 8% positive and 26% subjective.
Document KB_JB567_1902-03-26_01-00003.txt: This text is 5% positive and 24% subjective.
Document KB_JB567_1903-12-29_01-00003.txt: This text is 9% positive and 25% subjective.
Document KB_JB567_1904-01-06_01-00003.txt: This text is 7% positive and 22% subjective.
Document KB_JB567_1905-11-26_01-00003.txt: This text is 11% positive and 29% subjective.
Document KB_JB567_1906-12-17_01-00003.txt: This text is 11% positive and 24% subjective.
Document KB_JB567_1906-12-24_01-00001.txt: This text is 10% positive and 26% subjective.
Document KB_JB567_1907-04-04_01-00003.txt: This text is 4% positive and 21% subjective.
Document KB_JB567_1907-04-18_01-00003.txt: This text is 11% positive and 22% subjective.
Document KB_JB567_1907-05-10_01-00003.txt: This text is 8% positive and 23% subjective.
```

5. Clustering des Documents :

Le clustering, ou regroupement, est une méthode utilisée pour identifier des groupes cohérents au sein d'un ensemble de documents en fonction de leur similarité. Dans ce projet, l'algorithme **K-Means** a été appliqué pour regrouper les textes en fonction de leurs thématiques principales, en exploitant leurs représentations vectorielles basées sur TF-IDF. Cette approche permet de simplifier l'analyse en rassemblant les documents partageant des caractéristiques sémantiques similaires.

L'objectif principal de cette étape était de regrouper les documents du corpus en plusieurs clusters distincts. Ces regroupements facilitent l'identification des thématiques dominantes et offrent une vision structurée des contenus. Pour ce faire, les textes ont d'abord été prétraités pour éliminer les éléments superflus (caractères spéciaux, mots vides) et uniformisés (conversion en minuscules). Les représentations vectorielles des textes ont ensuite été générées à l'aide de la méthode TF-IDF.

L'algorithme K-Means a été utilisé avec un nombre prédéfini de 5 clusters. Chaque document a été attribué à l'un de ces clusters sur la base de ses similarités textuelles. Une réduction dimensionnelle, via une analyse en composantes principales (PCA), a ensuite permis de visualiser les clusters dans un espace bidimensionnel. La figure ci-dessous illustre les résultats du clustering, où chaque point représente un document et chaque couleur correspond à un cluster. Les centres des clusters sont marqués par des croix noires.



Le graphique illustre les résultats du clustering des documents en cinq groupes distincts, représentés par des points de différentes couleurs. Chaque point correspond à un document, et les couleurs indiquent son appartenance à un cluster spécifique. Les axes représentent une réduction dimensionnelle (via PCA) des vecteurs TF-IDF des documents, permettant de visualiser la similarité entre les textes dans un espace bidimensionnel.

Les croix noires représentent les centres des clusters, indiquant le regroupement moyen des documents dans chaque catégorie. Les clusters sont bien séparés dans certaines zones, montrant des thématiques distinctes dans le corpus. Par exemple :

- Certains clusters, comme le jaune (cluster 4), sont concentrés dans une zone spécifique, ce qui reflète une forte cohérence thématique.
- D'autres clusters, comme le bleu foncé (cluster 0), se chevauchent légèrement avec leurs voisins, suggérant des thématiques connexes ou un léger bruit dans les données.

6. Relations Sémantiques avec Word2Vec :

Pour analyser les relations sémantiques dans notre corpus, nous avons utilisé le modèle Word2Vec, qui permet de capturer les similarités et associations entre les mots en fonction de leur contexte. Après un entraînement sur les données prétraitées, les termes les plus proches de "football" ont été examinés. Les résultats montrent des mots fortement liés à des sports ou des événements sportifs, tels que *hockey*, *basket*, *athlétisme*, et *soccer*. Ces associations mettent en évidence une forte connexion thématique entre ces disciplines dans les textes analysés.

De plus, des mots comme *retransmission* et *télévision* reflètent un intérêt marqué pour la médiatisation des événements sportifs, tandis que des termes comme *foot* et *prépare* suggèrent un focus sur les activités et les préparatifs entourant le football. Ces observations renforcent l'idée que les textes abordent non seulement les aspects sportifs, mais aussi leur contexte social et médiatique.

L'utilisation de Word2Vec a également permis de représenter ces relations dans un espace vectoriel, facilitant la visualisation des similitudes et des clusters sémantiques. Cela offre une compréhension approfondie des tendances et des liens dans le corpus. Ces résultats soulignent l'utilité de cette approche pour explorer les connexions implicites dans des données textuelles historiques et pour identifier les thématiques émergentes.

Conclusion :

Ce travail a exploré l'évolution du football tel qu'il était représenté dans la presse historique, en s'appuyant sur des techniques modernes de traitement automatique des langues (TAL). À travers l'analyse du corpus CAMille, composé de 1000 articles couvrant la période de 1900 à 1950, plusieurs aspects essentiels ont été étudiés. Les résultats ont révélé des thématiques dominantes telles que les compétitions, les performances des équipes, et l'importance socioculturelle du football à l'époque.

Les analyses effectuées ont permis de répondre efficacement à la problématique posée. L'extraction des mots-clés et des entités nommées a mis en évidence les termes significatifs et les acteurs clés du corpus, tandis que l'analyse des sentiments a démontré une tonalité généralement neutre ou légèrement positive, reflétant le style descriptif des articles sportifs historiques. Par ailleurs, les techniques de clustering ont structuré le corpus en thématiques distinctes, et l'utilisation de Word2Vec a permis de révéler des relations sémantiques enrichissantes, renforçant la compréhension des liens contextuels entre les termes.

Cette étude a également contribué à mieux cerner le rôle central du football dans la société de l'époque. Plus qu'un simple sport, il était déjà un phénomène social influent, comme en témoignent les discussions sur les compétitions, les clubs et leurs répercussions culturelles et médiatiques. Les articles reflétaient une vision globale du football, intégrant à la fois des aspects sportifs et des éléments sociétaux.

Cependant, certaines limites ont été constatées, notamment liées à la qualité variable des textes historiques et au volume limité du corpus. Ces facteurs ont pu influencer la précision de certaines analyses. Pour aller plus loin, des recherches futures pourraient inclure des comparaisons entre différentes périodes ou une analyse élargie à d'autres sports. L'intégration de modèles avancés, tels que BERT ou GPT, permettrait également d'affiner l'analyse sémantique et de tirer des conclusions encore plus riches.

En conclusion, ce rapport illustre le potentiel des techniques de TAL pour analyser des archives historiques et offre une base solide pour approfondir l'étude des dynamiques entre sport, société, et médias à travers les époques.