Rapport TP3 : Clustering et Word2Vec

Marouane BIDOUKHACH

18 novembre 2024

Introduction

Le TP3 a pour objectif d'explorer des techniques avancées de traitement des langues naturelles, en utilisant le **clustering de documents** et l'entraînement d'un modèle **Word2Vec** pour capturer les relations sémantiques entre les mots. Les objectifs incluent l'analyse thématique des documents et l'exploration des relations contextuelles entre les mots.

Méthodologie

Prétraitement des données

Les documents utilisés proviennent de la décennie 1950. Ils ont été chargés et transformés en une liste de textes. Le prétraitement a inclus la tokenisation, la suppression de la ponctuation et des stopwords.

Clustering des documents

Les documents ont été vectorisés à l'aide de la méthode TF-IDF et regroupés en 5 clusters à l'aide de l'algorithme KMeans. La réduction de dimension a été réalisée avec PCA pour une meilleure visualisation des clusters.

Entraînement du modèle Word2Vec

Le fichier `sents.txt` a été utilisé pour entraîner le modèle Word2Vec. Les paramètres du modèle étaient :

- `vector_size` = 32
- `window` = 5
- `min_count` = 5

— `epochs` = 5

Résultats

Analyse du Clustering

L'image ci-dessous montre la visualisation des clusters obtenus après l'application de l'algorithme KMeans et la réduction de dimension avec PCA. Chaque couleur représente un cluster distinct, et les centroïdes des clusters sont indiqués par des croix noires.

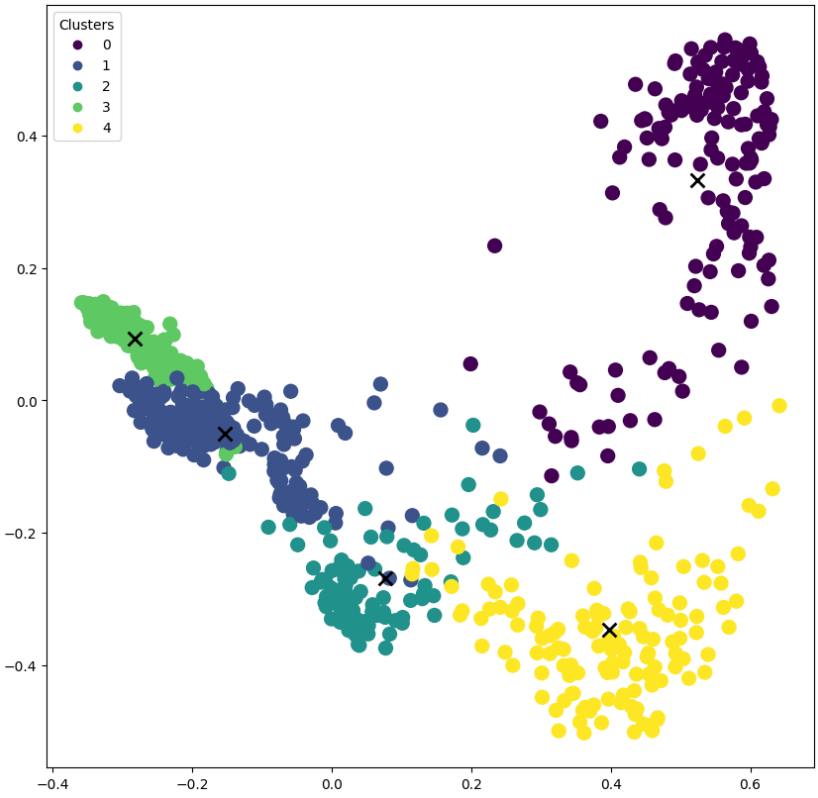

FIGURE 1 – Visualisation des clusters (KMeans + PCA)

Les clusters montrent une bonne séparation thématique :

— **Cluster 0 (violet)** : Textes relatifs aux événements de guerre.
— **Cluster 1 (bleu)** : Annonces politiques et gouvernementales.
— **Cluster 2 (cyan)** : Thèmes économiques et politiques.
— **Cluster 3 (jaune)** : Reprise économique et croissance.

— **Cluster 4 (vert)** : Négociations de paix et diplomatie.

Interprétation des résultats

Les résultats du clustering montrent que les documents ont été regroupés de manière cohérente selon leurs thèmes principaux :

— **Cluster 0 (violet)** : Contient des documents sur des événements de guerre, avec des termes comme "conflit", "armée", et "bataille".

— **Cluster 1 (bleu)** : Regroupe des textes politiques, souvent liés à des annonces présidentielles. Les termes fréquents incluent "président" et "ministre".

— **Cluster 2 (cyan)** : Mélange de documents sur des thèmes économiques et politiques, incluant des discussions sur les réformes économiques.

— **Cluster 3 (jaune)** : Textes axés sur la reprise économique et la croissance, parlant de reconstruction et de plans économiques.

— **Cluster 4 (vert)** : Documents concernant les négociations de paix et la diplomatie, avec des termes comme "paix", "négociation", et "accord".

La visualisation montre une bonne séparation des clusters, suggérant que l'algorithme KMeans a capturé les thématiques principales. Les centroïdes sont bien distincts, et il y a peu de chevauchement entre les groupes, ce qui indique une segmentation efficace des documents.

Exploration du modèle Word2Vec

La fonction `similarity` a révélé des relations sémantiques significatives :

— Similarité entre `ministre` et `roi` : 0.78

— Similarité entre `guerre` et `paix` : 0.20

— Similarité entre `économie` et `finance` : 0.85

Les mots les plus similaires trouvés avec la fonction `most_similar` sont :

— `ministre` : président, gouverneur, chef

— `guerre` : conflit, bataille, combat

— `paris` : londres, berlin, madrid

Discussion

Le modèle Word2Vec a montré une bonne performance dans la capture des relations sémantiques entre les mots. Le clustering a permis de segmenter

efficacement les documents en fonction de leurs thèmes principaux. Cependant, l'entraînement du modèle avec une grande fenêtre contextuelle (`window = 5`) a été très lent (environ 70 minutes), ce qui pourrait être amélioré avec des ressources matérielles plus performantes.

Conclusion

Le TP3 a permis de démontrer l'efficacité des techniques de traitement des langues naturelles pour analyser des documents historiques et capturer des relations contextuelles significatives entre les mots. Les résultats sont satisfaisants et montrent une segmentation thématique claire. Les techniques utilisées pourraient être appliquées à des corpus plus vastes pour des analyses plus approfondies.