UNIVERSITE LIBRE DE BRUXELLES

Faculté de Lettres, Traduction et Communication

STIC-B545 – Traitement Automatique de Corpus - 202526

RAPPORT DE TRAVAIL TP 3

ALBAN Kamga Dze

Enseignant : DE WILDE Max

Assistant : Lebailly Denis

Année académique 2025-2026

1. Resumé

Ce travail pratique (TP3) vise à analyser un corpus de documents historiques à l'aide de techniques de clustering et de word embeddings. L'objectif principal est de mettre en évidence des structures thématiques dans les documents d'une décennie choisie et d'explorer les relations sémantiques entre les mots grâce à un modèle word2vec.

Pour la première partie, nous avons effectué un clustering des documents de la décennie 1940–1949 en utilisant l'algorithme KMeans sur des vecteurs TF-IDF. Cette analyse a permis d'identifier plusieurs groupes de documents présentant des similarités lexicales et thématiques, notamment des documents administratifs, des textes narratifs et des informations géographiques.

La seconde partie consistera à entraîner un modèle word2vec sur un corpus pré-segmenté en phrases, afin d'étudier les relations sémantiques entre mots et de vérifier si des patterns cohérents apparaissent entre termes fréquemment associés.

2. Clustering des documents

Pour cette étape j'ai appliqué un **k-means** avec **4 clusters** (N_CLUSTERS = 4) sur le corpus de 1006 documents de la décennie 1940–1949. Chaque document a été associé à un cluster en fonction de la similarité de son contenu.

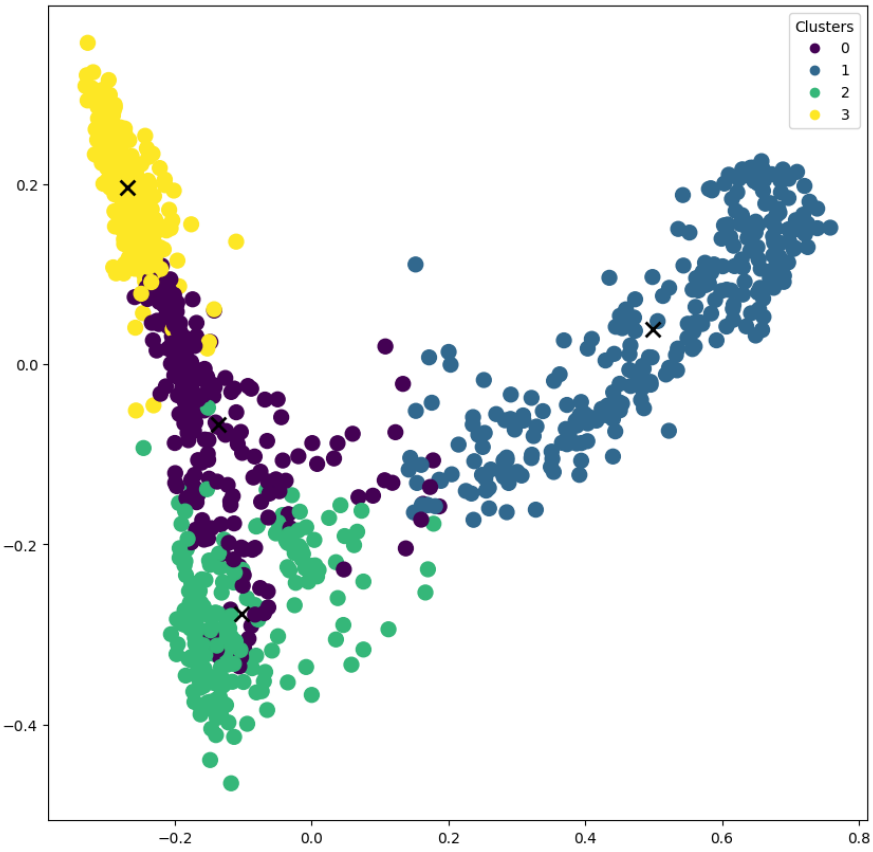


Figure 1 : clustering

❖ Analyse des clusters

Pour interpréter les résultats, nous avons extrait les 10 mots les plus fréquents par cluster après nettoyage et suppression des stopwords :

Cluster	Mots fréquents
0	van, 10, 15, 11, plus, 12, 30, 14, do, 20
1	rue, 10, fr, bruxelles, 30, 15, 000, do, 14, 50
2	ag, rossel, rue, dem, fr, ch, av, 000, tél, ec
3	plus, cette, do, être, tout, fait, comme, guerre, deux, après

❖ Observations

Cluster 0 : semble regrouper des textes contenant beaucoup de chiffres et mentions diverses (par ex., dates, numéros d'adresse ou d'annonces).

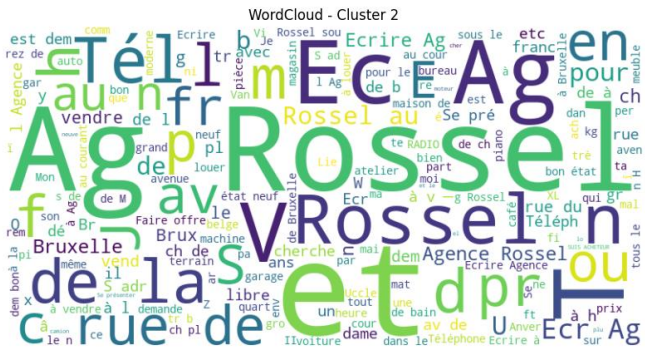
Cluster 1 : contient des textes géolocalisés ou administratifs, avec des mentions fréquentes de "rue", "Bruxelles", "fr" et d'autres codes.

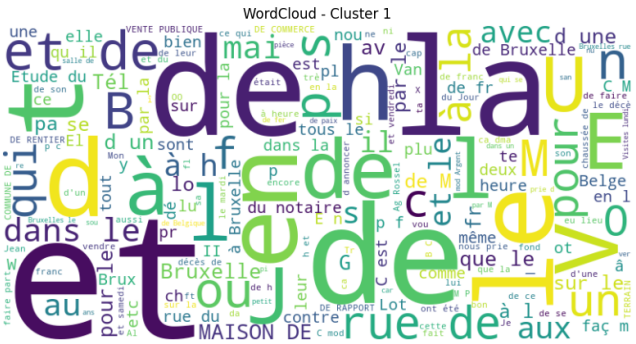
Cluster 2 : est fortement centré sur l'entreprise « AG Rossel », ce qui correspond à des textes économiques ou financiers.

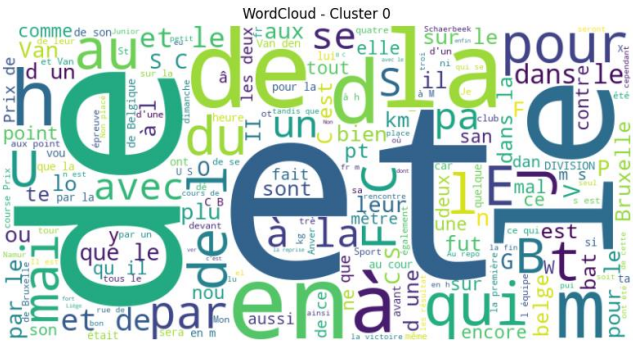
Cluster 3 : regroupe des textes plus narratifs ou généraux, avec des mots fréquents comme "plus", "cette", "être", "fait", "guerre", ce qui pourrait refléter des articles de presse plus descriptifs ou des chroniques.

Note : les clusters semblent cohérents avec les types de documents présents dans la décennie, bien que certains clusters contiennent des mots numériques ou répétitifs (*10, 15, 30, 000*), ce qui pourrait nuire à la lisibilité et nécessiter un nettoyage supplémentaire

❖ Nuages de mots (WordClouds)







Les nuages de mots ont été générés pour chaque cluster afin de visualiser la fréquence des termes de manière graphique.

- Les clusters 0, 1 et 2 sont représentatifs de leurs thèmes respectifs : chiffres et adresses, administration et localisation, entreprise AG Rossel.
- Le nuage de mots du cluster 3 l'on observe déjà que les termes généraux et narratifs dominent.

3. Exploration du modèle Word2Vec et discussion des résultats

3.1.Méthodologie d'entraînement

Pour l'entraînement du modèle Word2Vec, j'ai procédé par plusieurs essaies tout en ajustant au fur et à mesure les différents paramètres du model (vector_size, windows, min_count, …) ; Ceci permet ainsi de sauvegarder et de retenir le ou les meilleurs model entrainé.

Paramètres du modèle Word2Vec retenu :

- vector_size=32 : réduction des vecteurs à 32 dimensions.
- window=5 : contexte de 5 mots avant et après le mot cible.
- min_count=5 : les mots apparaissant moins de 5 fois sont ignorés.
- epochs=5 : nombre d'itérations sur le corpus.
- workers=4 : parallélisation sur 4 threads pour accélérer l'entraînement.

3.2.Exploration du modèle

Pour explorer les relations sémantiques entre mots, l'on utilise deux fonctions principales : **similarity ()** et **most_similar ()**

❖ **most_similar ()**

Mot cible	Voisins principaux	Observation générale
ministre	secretaire, membre, la_demission, professeur, vice_-_president, le_gouverneur, president, depute, adjoint, ancien_ministre	Les voisins sont principalement des titres, fonctions ou rôles liés au monde politique et administratif, ce qui montre que le modèle capture correctement le contexte ministériel.
roi	prince, capitaine, comte, cure, sergent, marechal, chevalier, saint, collaborateur, colonel	Les voisins sont liés à la royauté, la noblesse et les fonctions militaires ou religieuses associées, reflétant le contexte aristocratique et hiérarchique du mot "roi".
pay	monde, groupe, cas, sejour, sens, danger, milieu, voyage, triomphe	Les voisins reflètent des concepts abstraits, sociaux ou géopolitiques, capturant l'idée générale de territoire, population et événements associés au pays.

De manière globale, les trois requêtes montrent que le modèle Word2Vec parvient à regrouper des mots présentant des liens sémantiques cohérents. Pour « Roi », « Prince » ou « chevalier », les termes retournés appartiennent systématiquement au même champ lexical ou à des contextes proches. On observe également que le modèle ne renvoie pas nécessairement des synonymes stricts, mais plutôt des termes apparaissant dans des environnements similaires. Cela indique que l'entraînement s'appuie surtout sur la proximité contextuelle plutôt que sur la simple ressemblance lexicale. Ainsi, ces résultats illustrent la capacité du modèle à structurer des associations pertinentes en fonction des usages réels présents dans le corpus.

❖ **similarity ()**

Paires de mots	Similarité	Observation générale
roi – reine	0.5149	Similarité modérée ; le modèle reconnaît que les deux mots appartiennent au même champ royal, mais ils ne sont pas interchangeables.
ministre – gouvernement	0.5152	Similarité modérée ; le modèle identifie le lien institutionnel entre une personne (ministre) et l'institution (gouvernement).
guerre – bataille	0.7308	Similarité élevée ; le modèle saisit bien que ces deux concepts sont très proches dans le contexte militaire et historique.

Ces résultats suggèrent que le modèle Word2Vec capture efficacement les liens conceptuels structurants du corpus, tout en distinguant les relations hiérarchiques ou contextuelles des relations quasi synonymiques ou fortement corrélées.