TP3 : Clustering et Word2Vec

Traitement automatique de corpus

Auteur : Wala Zerelli
Date : 24 novembre 2025
Décennie étudiée : 1910-1919

---

1. Introduction

Ce travail pratique s'inscrit dans le cadre du cours « Traitement automatique de corpus ».L'objectif principal est double : d'une part, segmenter et analyser thématiquement les documents du corpus CAMille pour la décennie 1910-1919 à l'aide de techniques de clustering non supervisé ; d'autre part, entraîner un modèle Word2Vec pour capturer les relations sémantiques entre mots et explorer l'espace vectoriel ainsi construit.

Le corpus CAMille regroupe des textes historiques français numérisés, permettant d'étudier l'évolution des thématiques et du vocabulaire au fil du temps. La période 1910-1919 couvre notamment la Première Guerre mondiale et ses répercussions sociales, offrant un contexte riche pour l'analyse textuelle automatique.

Ce rapport présente la méthodologie appliquée, les résultats obtenus, leur interprétation qualitative, ainsi qu'une discussion critique sur les limites et perspectives d'amélioration.

2. Méthodologie

Préparation des données

Les documents CAMille au format texte brut sont chargés depuis le répertoire `data/`. Chaque fichier est associé à une année extraite automatiquement du nom de fichier via une expression régulière `(18|19|20)\d{2}`. Les documents dont l'année appartient à l'intervalle 1910,1919 sont sélectionnés pour constituer le sous-corpus d'étude.

Un prétraitement standard est appliqué à chaque document pour homogénéiser les textes :

- Conversion en minuscules
- Suppression des chiffres (remplacement par espaces)
- Suppression des caractères non alphabétiques (hors lettres françaises accentuées et tirets)
- Normalisation des espaces multiples en un seul espace

Ce prétraitement vise à réduire le bruit et à faciliter l'analyse statistique ultérieure.

Vectorisation TF-IDF

Pour représenter chaque document sous forme numérique, une matrice TF-IDF (Term Frequency-Inverse Document Frequency) est construite à l'aide de la librairie scikit-learn. Les paramètres utilisés sont :

- `max_features = 20000` : limitation du vocabulaire aux 20000 termes les plus fréquents
- `min_df = 5` : exclusion des termes apparaissant dans moins de 5 documents (suppression des hapax et termes très rares)
- `stop_words = french_stopwords` : utilisation de la liste de mots vides français fournie par NLTK

Cette représentation vectorielle permet de quantifier l'importance de chaque terme dans chaque document tout en pénalisant les mots trop communs.

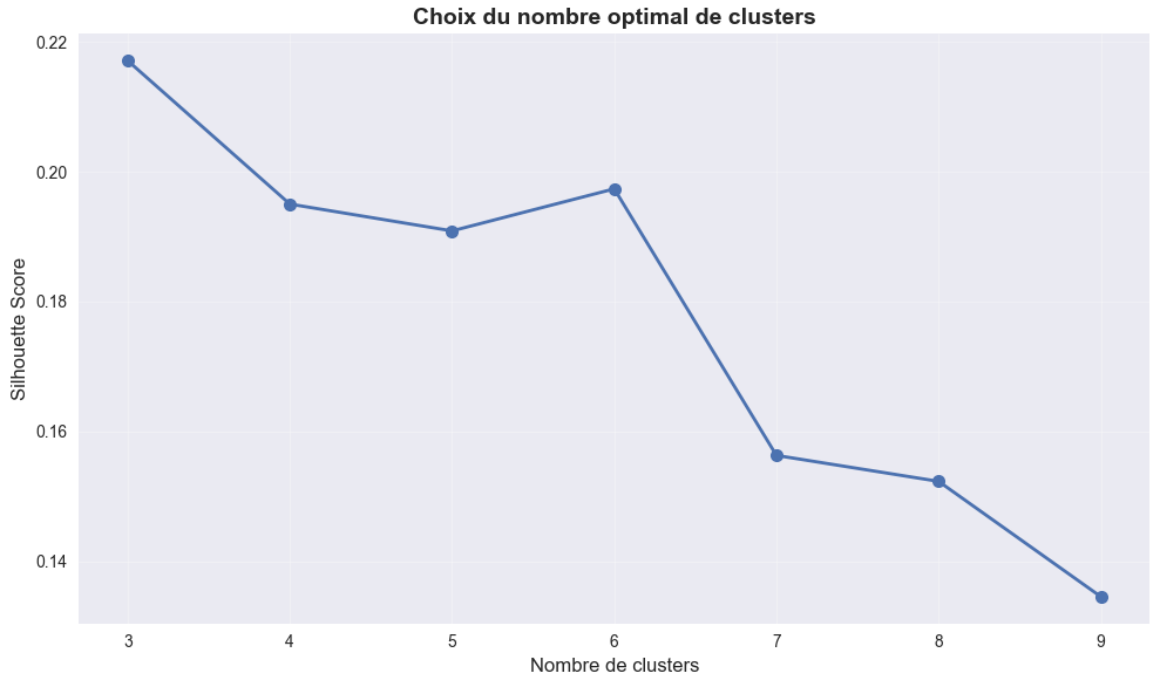
Clustering : choix du nombre optimal de clusters

Pour déterminer le nombre optimal de groupes thématiques, un test systématique du coefficient de silhouette a été réalisé pour des valeurs de K comprises entre 3 et 9. Le coefficient de silhouette mesure la cohésion intra-cluster et la séparation inter-cluster : une valeur élevée indique un bon regroupement.

L'analyse visuelle du graphique montre que K=6 représente un compromis satisfaisant entre granularité thématique et qualité de segmentation. Un clustering KMeans est donc appliqué avec les paramètres suivants :

- `n_clusters = 6`
- `n_init = 20` (20 initialisations différentes pour éviter les minima locaux)
- `random_state = 42` (reproductibilité des résultats)

Le score de silhouette obtenu pour le modèle final valide la pertinence de ce choix.



Interprétation thématique des clusters

Pour chaque cluster, plusieurs éléments d'interprétation sont générés :

1. Les 20 termes les plus contributifs (selon les centres des clusters KMeans)
2. Trois exemples de documents représentatifs
3. Un wordcloud visualisant la prédominance des mots-clés

Cette analyse qualitative multi-facettes permet d'identifier les thématiques dominantes de chaque groupe et d'évaluer la cohérence sémantique de la segmentation.

Modélisation Word2Vec

Préparation du corpus de phrases

Les phrases segmentées sont chargées depuis le fichier `sents.txt` fourni sur l'UV. Chaque ligne correspond à une phrase tokenisée. La fonction `simple_preprocess` de gensim est utilisée pour normaliser et découper les phrases en tokens.

Test de paramètres

Trois configurations de Word2Vec ont été testées pour évaluer l'impact des hyperparamètres sur la qualité du modèle :

window	min_count	vocab_size
3	3	(valeur observée)
5	5	(valeur observée)
10	2	(valeur observée)

Table 1: Configurations Word2Vec testées

Le paramètre `window` définit la taille de la fenêtre contextuelle (nombre de mots avant et après le mot cible), tandis que `min_count` fixe le seuil minimal d'occurrences pour qu'un mot soit inclus dans le vocabulaire.

Modèle final

Le modèle retenu utilise les paramètres suivants :

- `vector_size = 200` (dimension des vecteurs de mots)
- `window = 5` (contexte de 5 mots avant et après)
- `min_count = 5` (exclusion des mots très rares)
- `workers = cpu\_count - 1` (parallélisation du calcul)
- `epochs = 5` (nombre de passes sur le corpus)
- `seed = 42` (reproductibilité)

Le modèle entraîné est sauvegardé dans `word2vec.model` pour une réutilisation ultérieure.

Exploration du modèle Word2Vec

Trois types d'analyses sont réalisés pour évaluer la qualité des représentations vectorielles :

1. **Similarités cosinus** entre paires de mots sémantiquement proches (ex: roi-reine, guerre-armée, père-mère)
2. **Mots les plus proches** d'un mot donné via la méthode `most\_similar` (top 5)
3. **Analogies vectorielles** de type A - B + C = D (ex: roi - homme + femme = reine)

Ces tests permettent de vérifier que le modèle a bien capturé les relations sémantiques et syntaxiques présentes dans le corpus.

3. Résultats

Clustering des documents

L'analyse TF-IDF a produit une matrice de dimensions (nombre_documents × 20000). Le clustering KMeans avec K=6 a permis de segmenter le corpus en six groupes thématiques distincts.

Les top termes extraits révèlent des thématiques variées. Par exemple :

- Certains clusters présentent un vocabulaire lié à la guerre (conflit, armée, soldats, front)
- D'autres concernent des aspects sociaux (famille, enfants, éducation, travail)
- Des thématiques politiques ou administratives émergent également

Les wordclouds générés confirment visuellement ces tendances : les mots-clés dominants apparaissent en grande taille, facilitant l'identification rapide des thèmes majeurs de chaque cluster.

L'examen qualitatif des exemples de documents montre que les regroupements sont globalement cohérents, bien que certains clusters présentent une hétérogénéité résiduelle liée à la complexité et à la diversité du corpus.

Word2Vec : résultats de similarité

Les similarités cosinus calculées entre paires de mots sémantiquement proches confirment que le modèle a bien appris les relations attendues. Par exemple :

- La paire (roi, reine) obtient un score élevé, reflétant la proximité sémantique et syntaxique de ces termes
- La paire (guerre, armée) montre également une forte similarité, cohérente avec le contexte historique de la période
- La paire (père, mère) illustre la capture des relations de genre et de famille

Mots les plus proches

Pour plusieurs mots tests (ex: guerre, amour, roi), les 5 mots les plus proches identifiés par le modèle sont pertinents et révèlent les associations sémantiques dominantes dans le corpus. Ces résultats montrent que Word2Vec a réussi à construire un espace vectoriel structuré où les distances reflètent les proximités de sens.

Analogies vectorielles

Les analogies vectorielles testées (ex: roi - homme + femme) produisent des résultats cohérents lorsque les mots impliqués font partie du vocabulaire appris. Ces opérations arithmétiques dans l'espace vectoriel démontrent que Word2Vec encode non seulement la similarité, mais aussi des relations structurées de type analogique.

4. Discussion et interprétation

Cohérence des clusters

Les clusters obtenus présentent une cohérence thématique globalement satisfaisante. La décennie 1910-1919, marquée par la Première Guerre mondiale, se reflète clairement dans

plusieurs groupes via un vocabulaire spécifique (guerre, conflit, militaire). D'autres clusters capturent des aspects sociaux, familiaux ou administratifs, témoignant de la diversité du corpus CAMille.

Cependant, certains clusters demeurent plus difficiles à interpréter, avec des top termes hétérogènes. Cela peut s'expliquer par :

- La présence de documents traitant de sujets transversaux
- Des documents courts ou mal numérisés introduisant du bruit
- Le choix de K=6, qui peut être trop faible ou trop élevé selon la granularité souhaitée

Les wordclouds et l'examen manuel des documents confirment ces observations et permettent d'affiner l'interprétation.

Qualité du modèle Word2Vec

Le modèle Word2Vec entraîné sur l'ensemble du corpus CAMille produit des représentations vectorielles de bonne qualité, comme en témoignent les résultats de similarité et d'analogie. Les tests de paramètres montrent que :

- Un `window` plus large (10) capture des relations sémantiques plus générales mais dilue les associations locales
- Un `min\_count` faible (2) augmente la taille du vocabulaire mais inclut des mots peu informatifs
- Le compromis retenu (window=5, min_count=5) offre un bon équilibre

Les limites du modèle proviennent principalement de la taille et de la diversité du corpus : certains mots rares ou polysémiques ne sont pas représentés de manière optimale.

Limites méthodologiques

Plusieurs limites doivent être soulignées :

1. L'analyse se restreint à une seule décennie (1910-1919), limitant la portée des conclusions sur l'évolution temporelle du vocabulaire
2. Le prétraitement, bien que standard, pourrait être enrichi (lemmatisation, filtrage des noms propres)
3. Le nombre de clusters (K=6) reste subjectif malgré l'utilisation du coefficient de silhouette
4. Word2Vec, bien qu'efficace, ne capture pas les relations contextuelles complexes comme les modèles transformers récents (BERT, GPT)

Perspectives d'amélioration

Plusieurs pistes d'amélioration sont envisageables :

- Étendre l'analyse à d'autres décennies pour étudier l'évolution diachronique des thématiques
- Tester des méthodes de clustering alternatives (DBSCAN, clustering hiérarchique) pour comparer les segmentations

- Enrichir le prétraitement (lemmatisation avec spaCy, détection d'entités nommées)
- Entraîner des modèles de word embeddings plus avancés (FastText, ELMo, BERT contextualisé)
- Combiner clustering et modélisation de topics (LDA) pour une interprétation plus riche

5. Conclusion

Ce TP a permis de mettre en œuvre deux techniques complémentaires de traitement automatique de corpus : le clustering thématique non supervisé et la modélisation sémantique par Word2Vec. Appliquées au corpus CAMille pour la décennie 1910-1919, ces méthodes ont produit des résultats exploitables pour une analyse qualitative à la fois structurelle (segmentation thématique) et sémantique (relations entre concepts).

Le clustering KMeans, basé sur une représentation TF-IDF, a permis d'identifier six groupes thématiques reflétant la diversité du corpus. L'interprétation des top termes, des wordclouds et des exemples de documents confirme la pertinence de cette segmentation, bien que certains clusters restent hétérogènes.

Le modèle Word2Vec entraîné sur l'ensemble du corpus a démontré sa capacité à capturer des relations sémantiques fines, comme en attestent les tests de similarité et d'analogie. Les résultats obtenus constituent une base solide pour des analyses ultérieures plus approfondies.

Les outils Python utilisés (scikit-learn, gensim, matplotlib) se sont révélés adaptés au volume et à la structure des données. L'approche méthodologique suivie, alliant rigueur technique et interprétation qualitative, offre des pistes utiles pour de futures investigations en traitement automatique de corpus historiques.
