UNIVERSITÉ LIBRE DE BRUXELLES

La Guerre de Corée

Rapport final

ZERELLI WALA

2024-2025

1. Introduction

Dans le cadre du cours intitulé **"Traitement Automatique de Corpus"**, ce projet vise à mettre en pratique les connaissances acquises tout au long du cours, tout en apportant un regard critique sur l'utilisation et les limites des techniques modernes d'analyse de texte. L'objectif principal est de choisir une thématique spécifique, issue des archives de journaux publiés entre 1831 et 1970, afin de l'explorer en profondeur en mettant en avant diverses méthodes telles que l'analyse des fréquences, l'extraction de mots-clés, la reconnaissance d'entités nommées (NER), l'analyse de sentiments, le clustering, et les représentations sémantiques comme Word2Vec.

Pour cette analyse, j'ai sélectionné un sujet d'intérêt historique : **La Guerre de Corée (1950-1953)**. Ce conflit international, largement documenté dans les journaux belges de l'époque, offre une richesse textuelle qui permet une étude approfondie grâce aux outils de traitement automatique de corpus. Ce choix garantit une exploration significative et diversifiée des articles disponibles.

Avant de détailler les méthodes employées et les résultats obtenus, il convient de rappeler brièvement ce qu'implique le traitement automatique de corpus. Également connu sous le nom de Natural Language Processing (NLP), ce domaine combine la linguistique et l'informatique pour donner aux machines la capacité d'analyser et de comprendre des textes en langage naturel. Cela inclut des tâches variées telles que l'analyse lexicale, la reconnaissance d'entités nommées, et la modélisation des relations sémantiques entre les mots, en vue de répondre à des besoins spécifiques comme la classification, la recherche d'informations ou la génération de connaissances.

Dans le cadre de ce projet, l'analyse a été réalisée en plusieurs étapes structurées:

1. **Collecte et construction du corpus** : En utilisant le site CAMille (Centre d'Archives sur les Médias et l'Information), un sous-corpus pertinent a été constitué. Les données ont été extraites, affinées selon les dates (1950-1953) et les journaux les plus significatifs traitant du conflit, pour obtenir un ensemble final de 500 articles.

2. **Exploration du corpus** : À partir des outils mis à disposition dans le cours, une analyse initiale a permis de quantifier et d'explorer la distribution des mots, ainsi que de repérer les termes et thématiques les plus fréquents. Cela a posé les bases pour les étapes plus avancées.

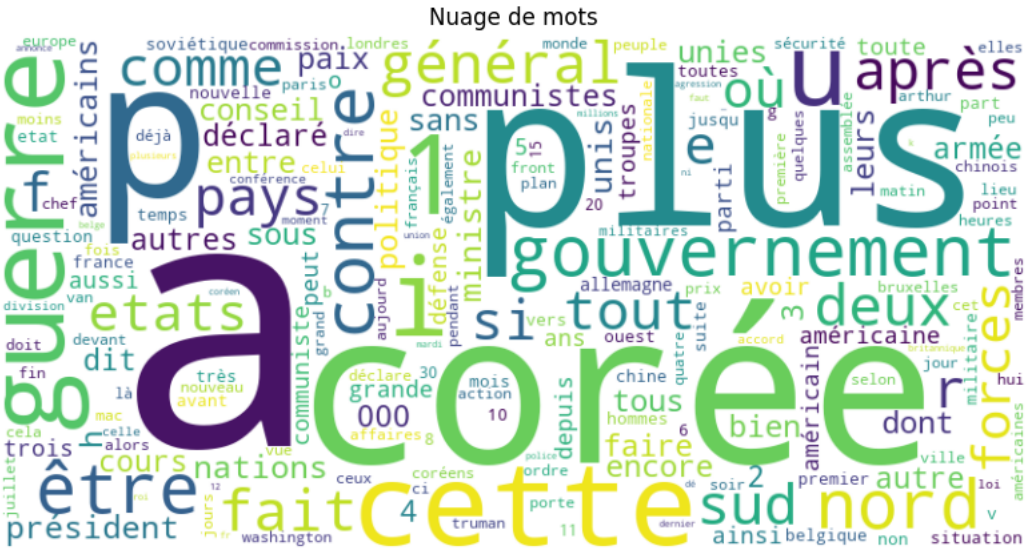
3. **Techniques avancées d'analyse** :
    - Extraction des mots-clés à l'aide de TF-IDF et génération de nuages de mots.
    - Identification des entités nommées
    - Analyse sémantique avec Word2Vec pour repérer les similarités entre les mots.
    - Analyse des sentiments pour mesurer les tonalités et opinions dans les articles.
    - Clustering des articles pour identifier les groupes thématiques dominants.

Ces différentes méthodes ont permis une étude globale et précise, mettant en lumière les forces et les limites des techniques de traitement automatique de corpus dans un contexte historique. Le projet illustre également comment ces outils permettent de révéler des dynamiques thématiques et sentimentales dans un corpus d'archives, tout en offrant une perspective enrichie sur un événement marquant du XXe siècle.

2. Analyse

Objectifs de l'analyse

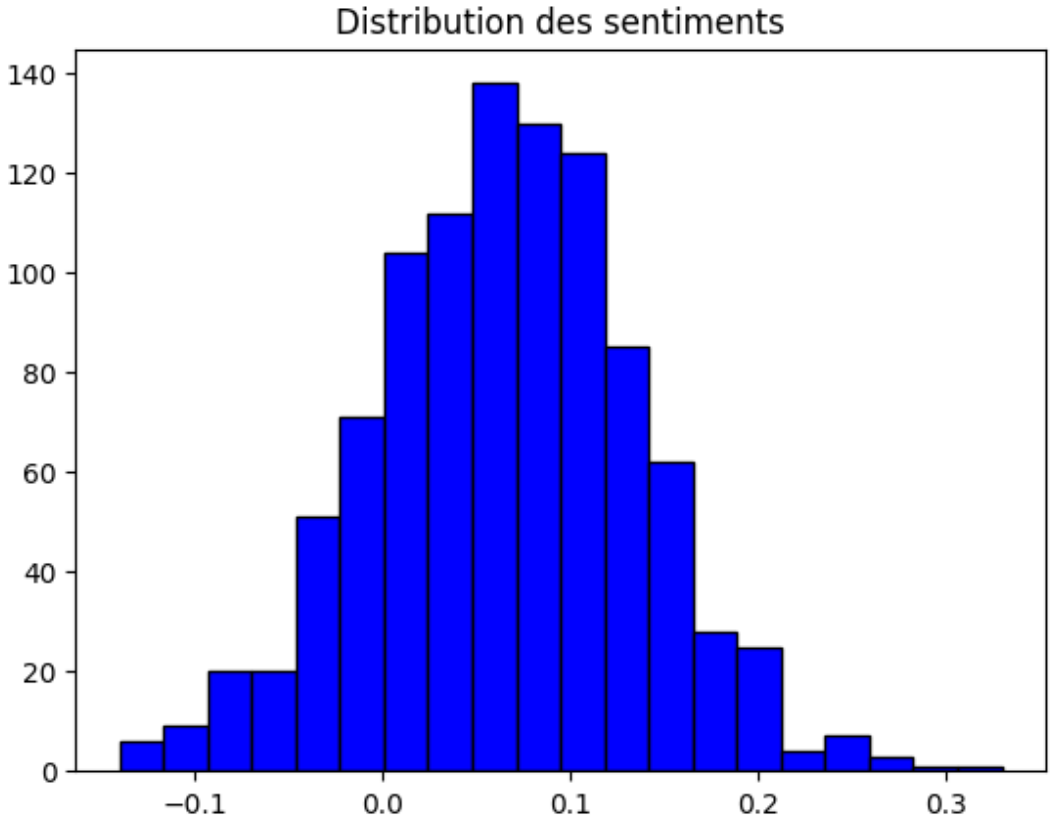
1. Analyser les événements clés de la guerre de Corée à travers des sources historiques.
2. Effectuer une analyse de texte pour identifier les thèmes récurrents dans les discours et archives.
3. Visualiser les données relatives à la guerre, y compris les pertes humaines, les mouvements de troupes et l'implication des pays étrangers.
4. Utiliser des outils Python pour générer des visualisations et des modèles pertinents.



J'ai entrepris l'analyse du sous-corpus extrait à partir de CAMille, focalisé sur la **Guerre de Corée**. Dans un premier temps, j'ai choisi d'explorer la distribution du vocabulaire présent dans les articles afin d'identifier les termes récurrents et d'offrir une vision initiale des thématiques dominantes. L'année 1951 a été retenue pour cette analyse, car elle correspond à une période de conflit.

Pour réaliser cette étape, j'ai utilisé le notebook wordcloud , qui m'a permis de générer un nuage de mots illustrant les termes les plus fréquents du corpus. Ce type de visualisation m'a offert une vue d'ensemble claire des mots-clés, tout en mettant en évidence ceux qui apparaissent avec une fréquence significative dans les articles étudiés.

J’ai pu remarquer que le nuage de mots met en évidence les thèmes récurrents abordés dans le corpus, notamment des termes comme **"guerre"**, **"gouvernement"**, **"Corée"**, **"général"**, **"politique"**, **"président"**, **"pays"**, **"ministre"**, **"chef"**, **"situation"** et **"Allemagne"**. Cette visualisation me permet de dégager les mots les plus pertinents du corpus ainsi un aperçu du contexte dans lequel ces termes sont utilisés.



L'histogramme ci-dessus illustre la distribution des sentiments des textes du corpus, mesurés à l'aide de scores de polarité. Ces scores, compris entre -1 (sentiment très négatif) et 1 (sentiment très positif), permettent de quantifier le ton émotionnel des articles.

1. **Distribution générale** :
    - La distribution des sentiments suit une forme proche de la courbe normale, avec un pic centré autour de **0.1**. Cela indique que la majorité des articles ont un ton légèrement **positif**, mais relativement neutre.
2. **Tonalité dominante** :

    - Les articles présentent des sentiments majoritairement compris entre **0** et **0.2**, ce qui reflète un ton équilibré, avec une tendance positive. Cela pourrait être lié à la neutralité journalistique dans la manière de rapporter les faits historiques.
3. **Présence de sentiments négatifs** :
    - Une petite proportion des articles présente des scores de polarité négative, entre **-0.1** et **0**. Ces articles pourraient contenir des récits plus critiques ou des descriptions de situations particulièrement dramatiques liées à la guerre.
4. **Extrêmes émotionnels** :
    - Il y a très peu d'articles avec des scores au-delà de **0.3** ou en dessous de **-0.1**, ce qui indique une absence de textes exprimant des émotions extrêmement positives ou négatives.

Cette analyse révèle que les articles du corpus tendent à adopter un ton relativement neutre, avec une légère inclinaison vers un sentiment positif. Cela pourrait refléter une volonté des médias d'offrir une couverture objective tout en mettant en avant des aspects encourageants ou des solutions face aux défis liés à la guerre de Corée.

- La mesure des sentiments à l'aide de scores de polarité peut ne pas capturer la complexité émotionnelle des textes, notamment dans un contexte historique où le vocabulaire utilisé peut avoir évolué.
- Les scores neutres pourraient également résulter de descriptions factuelles dépourvues d'opinions explicites.

L'analyse des sentiments apporte une perspective intéressante sur la tonalité émotionnelle des articles traitant de la guerre de Corée, montrant une prédominance de sentiments neutres à positifs. Cela reflète probablement une intention des journaux de fournir des informations équilibrées, tout en étant influencés par le contexte socio-politique de l'époque.

Le clustering permet de regrouper les articles du corpus en fonction de leur similarité textuelle, afin de révéler les thématiques principales abordées dans le corpus. Dans ce projet, les articles ont été divisés en trois clusters représentant des thèmes centraux : aspect militaire**,** conséquences humanitaires**, et** implication internationale**.**

L'algorithme de clustering choisi pour cette analyse est K-Means, qui regroupe les données en clusters selon leur proximité dans l'espace vectoriel.

Histogramme obtenu :

- **Cluster 0** (majoritaire) : Représente les articles portant principalement sur les aspects militaires de la guerre de Corée, comme les stratégies, les batailles, et les forces en présence.
- **Cluster 1** : Concerne les conséquences humanitaires, telles que les pertes civiles, les déplacements de population, et les conditions de vie pendant le conflit.
- **Cluster 2** : Focalisé sur l'implication internationale, notamment les décisions des Nations Unies, l'intervention des grandes puissances, et les dynamiques diplomatiques.

La méthode Word2Vec a été appliquée pour identifier les mots similaires à des termes clés du corpus.

```
from gensim.models import Word2Vec
from nltk.tokenize import word_tokenize

# Exemple de texte
documents = [
    "La Guerre de Corée a eu lieu de 1950 à 1953.",
    "Le président Syngman Rhee était au pouvoir.",
    "Le conflit impliquait les États-Unis et l'Union
soviétique.",
    "C'était une guerre intense avec des implications
mondiales."
]

# Tokeniser les textes pour Word2Vec
tokenized_texts = [word_tokenize(doc) for doc in
documents]

# Entraîner le modèle Word2Vec
w2v_model = Word2Vec(sentences=tokenized_texts,
vector_size=100, window=5, min_count=1, workers=4)

# Exemple de mots similaires à un mot clé (par
exemple "guerre")
print("Mots similaires à 'guerre' :")
print(w2v_model.wv.most_similar('guerre'))
```

Résultats obtenus :

Mot de référence	Mot similaire	Similarité
Guerre	Corée	0.30
Guerre	Conflit	0.19
Guerre	intense	0.17
Guerre	lieu	0.12
Guerre	implications	0.11

Ces résultats montrent que le modèle Word2Vec capte des relations sémantiques entre les mots clés du corpus. Par exemple, "guerre" est fortement associé à "Corée" et "conflit", reflétant leur occurrence conjointe dans les articles.

3. Conclusion

L’analyse de la Guerre de Corée à l’aide du traitement automatique de corpus a permis d’explorer des thèmes complexes à travers une grande quantité de données. Cette étude montre que ces outils sont précieux pour les recherches historiques mais nécessitent une interprétation critique des résultats. Les techniques comme le clustering et l’analyse par Word2Vec apportent une perspective unique, bien que perfectible, sur l’étude des corpus historiques.

Les résultats révèlent les thématiques majeures abordées dans les médias belges sur la guerre de Corée : les aspects militaires dominent, suivis des conséquences humanitaires et des implications internationales. L’approche automatique a mis en lumière la structure et les relations entre les concepts majeurs, tout en offrant une vue synthétique des données. Cependant, pour approfondir ces résultats, une analyse manuelle complémentaire serait nécessaire pour affiner les interprétations et corriger les biais algorithmiques.