RAPPORT

La Guerre de Corée
Traitement Automatique de Corpus

ETUDIANT : KARIFALA KABA

Table des matières

INTRODUCTION

Dans le contexte du cours de "Traitement Automatique de Corpus", notre démarche s'est inscrite dans une initiative intellectuelle visant à mettre en pratique les connaissances acquises. L'objectif sous-jacent à cette entreprise était de susciter une réflexion critique sur le traitement automatique de corpus dans son ensemble.

Le mandat qui nous a été confié consistait à sélectionner une thématique spécifique, présente dans les journaux entre 1831 et 1970, pour ensuite l'analyser en recourant aux techniques enseignées au fil du cours. Ces techniques englobent l'exploration, l'analyse des fréquences, l'identification des mots-clés, des entités nommées, l'analyse de sentiment, le regroupement (clustering), ainsi que des approches plus avancées telles que Word2Vec.

Notre choix s'est porté sur une analyse approfondie de la Guerre de Corée en raison de l'abondance d'articles qui lui ont été consacrés dans divers journaux belges entre 1950 et 1953. Avant d'entamer notre analyse, nous avons procédé à une revue conceptuelle succincte du traitement automatique de corpus.

Le traitement automatique de corpus, également désigné sous l'appellation de Traitement du Langage Naturel (NLP), constitue un ensemble de techniques permettant aux machines de lire, de comprendre et d'inférer la représentation des textes naturels. Il vise à traiter les langues pour différentes tâches et applications. Cette discipline résulte de la convergence des domaines de la linguistique et de l'informatique, s'efforçant de décrypter la structure du langage et de développer des modèles capables de comprendre, de décomposer et de séparer les détails significatifs des textes et des discours.

Les chercheurs en Traitement Automatique de Corpus s'emploient à reproduire la compréhension et l'utilisation du langage par les humains. Ils ont ainsi mis au point des technologies pour l'analyse lexicale et morphologique, la segmentation des mots, l'analyse sémantique, la signification des mots, la représentation des connaissances, ainsi que des approches et outils basés sur la connaissance pour le traitement automatique des langues.

Notre projet s'articule autour de l'application de ces connaissances méthodologiques à un sous-corpus spécifique portant sur la Guerre de Corée entre 1950 et 1953. Nous avons méticuleusement collecté des données, affinant les résultats en fonction du journal et de la date de publication. L'exploration du corpus a été réalisée à l'aide du notebook "explore" du module 2, offrant une analyse approfondie de la distribution du corpus et l'extraction des mots les plus fréquents.

Par la suite, nous avons mis en œuvre le traitement automatique de corpus en utilisant divers notebooks du cours. Le notebook "keywords" et "wordcloud" du module 3 a été déployé pour extraire les mots-clés, en excluant les stopwords, et analyser la distribution du vocabulaire. Le notebook "ner" a été mobilisé afin d'extraire les entités nommées, incluant les personnes, les lieux ou les organisations. L'analyse des similarités de mots a été entreprise avec le module 4, et enfin, le notebook "sentiment" du module 3 a permis de déterminer le sentiment des phrases.

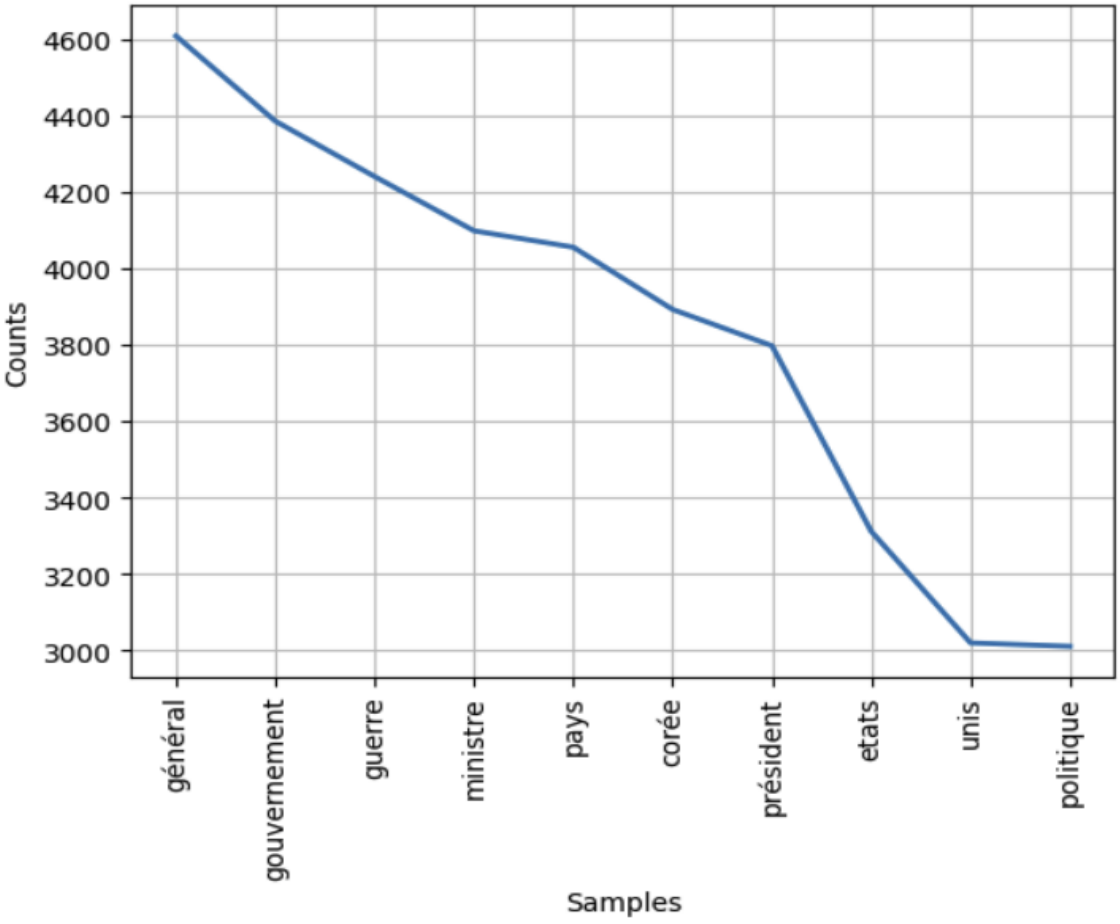
Cette approche holistique nous a offert une analyse précise et pertinente, nous éclairant sur les avantages et les limites du traitement automatique de corpus dans un contexte historique. Elle a également permis une plongée approfondie dans la thématique complexe de la Guerre de Corée.

Analyse

L'analyse du sous-corpus extrait de CAMille, se focalisant sur la Guerre de Corée, s'impose comme une démarche judicieuse. Dans un souci préliminaire, notre objectif consistait à mener une étude exhaustive de la répartition lexicale caractéristique de ce corpus. Nous sommes d'avis que la mise en lumière des termes les plus fréquents au sein du lexique offre une première perception significative de la thématique sous-jacente. À cette fin, notre choix s'est orienté de manière spécifique vers l'année 1951. Le recours au notebook "wordcloud" du module 3 a été notre choix stratégique, offrant une vue d'ensemble des composants lexicaux. Une démarche itérative a été entreprise pour intégrer des stopwords dans le code, visant ainsi à éliminer les termes jugés superflus. Nous procéderons désormais à l'exposition de l'essence du nuage de mots résultant de cette démarche :

(La suite du contenu serait développée en fournissant des explications plus détaillées, des interprétations approfondies, ainsi que des analyses contextuelles du nuage de mots obtenu.)

Il est possible de noter que le nuage de mots met en évidence des thèmes fréquemment abordés au sein du corpus, notamment des termes tels que "guerre", "gouvernement", "Corée", "général", "politique", "président", "pays", "ministre", "forces", "communiste", "soviétique", etc. Cette visualisation permet de mettre en avant les mots les plus pertinents dans notre corpus, tout en offrant une perspective sur le contexte probable de leur utilisation. Par ailleurs, l'analyse de la distribution du vocabulaire à l'aide du notebook freq (module 2) a permis d'identifier les mots les plus fréquents, lesquels ont été représentés graphiquement. Le schéma ci-dessous illustre cette analyse :



L'axe horizontal (x) du graphique représente les mots les plus fréquents au sein de notre corpus, tandis que l'axe vertical (y) indique

le nombre de fois qu'ils apparaissent. Bien que cette information soit cruciale pour la compréhension de notre corpus, elle ne constitue pas en elle-même un élément essentiel à notre analyse.

Dans une seconde phase, nous procédons à l'examen des entités nommées présentes dans le corpus. Cette démarche nous offre une perspective globale sur les organisations, les individus et les lieux ayant joué un rôle dans la Guerre de Corée. Le tableau suivant présente les entités telles que les organisations, les personnes et les lieux qui reviennent le plus fréquemment dans le corpus en lien avec la Guerre de Corée.

Organisations	**Personnes**	**Lieux**
Conseil	Mac Arthur	Corée
Reuter	Staline	Etats-Unis
Conseil de sécurité	Nehru	Belgique
Nations-Unies	M. Duvieusart	Bruxelles
Parlement	Président Truman	Américains
Sénat	M. Churchill	Etat
Chambres	M. Malik	Grande-Bretagne
Assemblée	M. Gromyko	Londres
Tokyo	Prince Royal	Moscou
Benelux	M. Spaak	Anvers
O.N.U.	Léopold III	Paris

Le tableau présenté ci-dessus met en évidence diverses organisations ayant joué un rôle dans la Guerre de Corée. Cependant, au sein de la catégorie "organisations", certains résultats apparaissent assez imprécis, tels que "Conseil", "Reuter", "Parlement", "Sénat", "Chambres", "Assemblée". En l'absence de contexte, il est difficile de déterminer de quelles organisations spécifiques il s'agit et à quel pays elles se réfèrent. Nous pouvons seulement identifier le type d'organisation mentionné. Il est également notable que "Tokyo" est considéré comme une organisation, ce que nous pensons être intentionnel dans le corpus, où Tokyo est mentionné en tant qu'organisation plutôt qu'en tant que pays.

En ce qui concerne les personnes, le tableau révèle des personnalités importantes ayant participé à la Guerre de Corée, telles que Mac Arthur, Staline, Nehru, Truman, etc. Cela contribue à une meilleure compréhension des acteurs clés dans le conflit.

Divers lieux sont également évoqués, avec la Corée en tête, comme on peut s'y attendre dans le contexte de la Guerre de Corée. Ensuite, les pays impliqués dans la guerre, tels que les États-Unis, sont mentionnés. Cependant, une observation pertinente est que l'algorithme a inclus "Américains" et "État" parmi les lieux, ce qui constitue une erreur apparente.

La troisième étape de notre analyse se concentre sur la mise en évidence des similarités lexicales. Préalablement, des étapes de regroupement (module 4) et de segmentation des phrases (module 4) ont été effectuées. Pour cette phase, une segmentation en phrases de notre corpus a été réalisée, générant un nouveau fichier appelé sents.txt. Cette approche a permis d'obtenir des résultats plus cohérents pour l'analyse des mots similaires. Nous avons ensuite examiné les termes du lexique présents dans notre nuage de mots afin d'analyser leurs mots similaires.

Mots du lexique	Mot 1	Mot 2	Mot 3	Mot 4	Mot 5
Guerre	Evénements 80%	Lutte 70%	Russie 75%	Corée 75%	Paix 65%
Gouvernement	G. Britannique 85%	G. Américain 85%	Chancelier 81%	Conseil 81%	Parlement 77%
Corée	Chine 81%	Formose 79%	Guerre 75%	Offensive 70%	Ligne 70%
Général	Chef 85%	Maréchal 76%	Secrétaire 80%	Président 77%	Cabinet 76%
Politique	Question 86%	Volonté 86%	Position 84%	Solution 84%	Nécessité 84%
Président	Prés. du Conseil 91%	Délégué 87%	Ministre 87%	Sénateur 87%	Conseiller 82%
Pays	Mouvement 70%	Problème 69%	Régime 66%	Conflit 66%	Territoire 64%
Forces	Troupes 91%	Alliés 87%	Armées 86%	Coréens 85%	Bases 84%
Communiste	Soviétique 88%	Britannique 87%	Russe 82%	Militaire 77%	Formose 73%

Ce tableau présente les mots les plus similaires à ceux figurant dans notre nuage de mots, accompagnés de leur taux de similarité en pourcentage. Il est observable que, dans la plupart des cas, l'algorithme a extrait des termes similaires en tenant compte du contexte d'utilisation, plutôt que de se baser uniquement sur l'essence intrinsèque du mot. Les mots similaires identifiés offrent ainsi une meilleure compréhension de leur utilisation dans le corpus, permettant d'appréhender le contexte et de découvrir d'autres termes connexes. À titre d'exemple, pour le mot "Pays", les mots similaires ne sont pas nécessairement des synonymes, mais plutôt des termes en relation avec le contexte de guerre dans lequel certains pays, notamment la Corée, étaient impliqués, tels que "Problème", "Régime", "Conflit", "Territoire", etc.

Enfin, nous estimons qu'effectuer une analyse des phrases du corpus à l'aide du notebook sentiment (module 3) revêt une pertinence particulière dans notre cas. L'objectif est d'analyser la perspective des journaux belges ayant rapporté les événements de la Guerre de Corée de 1950 à 1953. Cette démarche implique l'extraction d'une série de

Phrases du corpus, suivie de l'analyse de leur polarité (positive ou négative) ainsi que de leur subjectivité (subjective ou objective). Le tableau suivant présente ces phrases, accompagnées de leurs taux respectifs de polarité et de subjectivité. Ces phrases ont été sélectionnées de manière aléatoire, couvrant l'étendue du corpus du début à la fin.

Phrases	**Polarité**	**Subjectivité**
"On rapporte que, conformément à l'ordre du général Mac Arthur, des bases aériennes de Corée du Nord ont été bombardées."	Neutre	Objectif
"Le dynamitage du pont eut même lieu alors que les troupes de la Corée du Sud étaient entassées, les pare-chocs au contact sur le tablier du pont, avec le personnel de la mission militaire américaine."	Neutre	19% subjectif
"Dans la capitale provisoire, Suwon, qui débordait de camions, de matériel japonais, américain, ou pris aux rouges, les Américains et les Coréens essayaient frénétiquement de réorganiser les troupes fraîches et les régiments démembrés en unités aptes au combat."	11% positive	21% subjective
"Le pape Pie XII a pris aujourd'hui une décision importante qu'on croit viser principalement les gouvernements communistes d'Europe orientale, en décrétant l'excommunication « de ceux qui conspirent contre les autorités ecclésiastiques légitimes »."	12% positive	13% subjective
"Le porte-parole a signalé que l'U.R.S.S. n'a pas encore répondu à la demande faite par la Grande- Bretagne au gouvernement soviétique pour qu'il use de son influence sur la Corée du nord en vue de mettre un terme aux combats."	5% négative	13% subjective
"La campagne de meetings de protestation contre l'intervention américaine en Corée du nord se développe dans toute l'U.R.S.S., où de nombreuses réunions se sont tenues, hier notamment, dans les régions de Stalingrad et de Leningrad."	6% positive	8% subjective
"Lundi et mardi, des appareils du porte-avions britannique et ceux du porte-avions américain « Valey Force », attaquèrent Pyongyang."	8% positive	5% subjective
"Le commandant en chef de la marine américaine en Extrême-Orient annonce que « la flotte combinée anglo-américaine a attaqué les installations de l'aérodrome de Pyongyang, capitale de la Corée du nord, les 3 et 4 juillet."	7% positive	7,5% subjective
"C'est la première fois, depuis deux ans, que des diplomates soviétiques et orientaux acceptent une invitation officielle à la légation des Etats-Unis."	1% positive	8% subjective

À la lumière des résultats obtenus, il apparaît que les phrases rédigées dans les journaux belges au sujet de la Guerre de Corée sont généralement positives et subjectives. Cette observation peut sembler surprenante, étant donné qu'il s'agit de phrases décrivant des événements passés. On aurait pu anticiper des expressions plus objectives et des énoncés plutôt neutres. Cependant, en ce qui concerne la polarité, compte tenu du contexte de guerre, on aurait également pu envisager des phrases plus négatives que positives.

Notamment, il est intéressant de noter que la majorité des phrases positives tendent à refléter un point de vue sudiste. Cependant, une phrase affichant une polarité positive semble pencher vers une perspective nordiste. Il est donc possible d'estimer que, dans ce cas, l'algorithme n'a pas pris en compte le contexte global des phrases, mais s'est contenté d'analyser les phrases individuellement.

Cette constatation souligne l'importance de considérer le contexte historique et culturel lors de l'analyse de la polarité des phrases, surtout dans le cadre d'événements historiques tels que la Guerre de Corée. Ces nuances soulignent la complexité de l'interprétation automatique du sentiment, mettant en évidence la nécessité de développer des approches plus sophistiquées pour contextualiser correctement les informations historiques.

CONCLUSION

Dans le cadre de notre projet, nous avons entrepris l'analyse d'un corpus traitant de la Guerre de Corée sur la période de 1950 à 1953 en utilisant des techniques de Traitement Automatique de Corpus. Ces méthodes ont été appliquées pour traiter les textes et fichiers composant notre corpus, facilitant ainsi l'utilisation du traitement automatique des langues. La thématique a été extraite de "CaMille", un centre d'archives sur les médias et l'information visant à faciliter des recherches ciblées et avancées, accessible aux chercheurs et au public.

Au cours de nos analyses, nous avons identifié l'un des principaux avantages du traitement automatique de corpus, à savoir la possibilité d'analyser une grande quantité d'informations textuelles. La numérisation des sources historiques, telle que réalisée dans CaMille, permet de préserver ces sources pour un accès futur, tout en offrant la possibilité de les consulter et de les exploiter, comme démontré dans ce projet.

L'exploitation de ces sources a été rendue possible grâce aux techniques de Traitement Automatique de Corpus, permettant d'explorer les corpus, analyser la distribution du vocabulaire, extraire les mots-clés, reconnaître les entités nommées, évaluer la polarité et la subjectivité des phrases, extraire les similarités de mots, et bien plus encore. Cette analyse s'est avérée particulièrement efficace, offrant une rapidité d'analyse inégalée, ce qui constitue un avantage significatif pour les historiens.

Cependant, malgré les nombreux avantages du Traitement Automatique de Corpus, certaines limites ont été identifiées au cours de notre analyse. Des incohérences et imprécisions dans les résultats peuvent être attribuées à la qualité du corpus, notamment la présence de nombreux mots vides. De plus, des lacunes ont été constatées dans l'analyse du contexte, soulignant l'importance cruciale de celui-ci pour une interprétation correcte des résultats.

En conclusion, bien que le Traitement Automatique de Corpus présente certaines limites, celles-ci sont négligeables par rapport aux avantages considérables qu'il offre aux historiens et à toute personne travaillant avec des corpus. La plateforme numérique des archives de la presse belge, CAMille, a joué un rôle central en nous permettant d'extraire une thématique pertinente, à savoir la Guerre de Corée. L'accès direct à ces sources constitue un atout inestimable, représentant une véritable révolution dans le domaine de la recherche historique.

Bibliographie

- Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing : an introduction. Journal of the American Medical Informatics Association, 18(5), 544-551.

- CAMille Centre d'archive sur les Médias et l'information, https ://www.camille-ulb-kbr.be.

- Chowdhary, K. (2020). Natural language processing. Fundamentals of artificial
  Intelligence, 603-649.