Rapport sur le Projet NLP : Traitement de Langage Naturel pour l'Analyse de Texte Médical

Sarra Baghdadi / Nicolas Droulers

December 22, 2023

1 Introduction

Ce projet s'inscrit dans le cadre du traitement de langage naturel (NLP) appliqué à des données médicales. L'objectif principal est d'utiliser des techniques avancées pour nettoyer et analyser les textes liés aux différents types de cancer et finalement faire la classification.

2 base de données

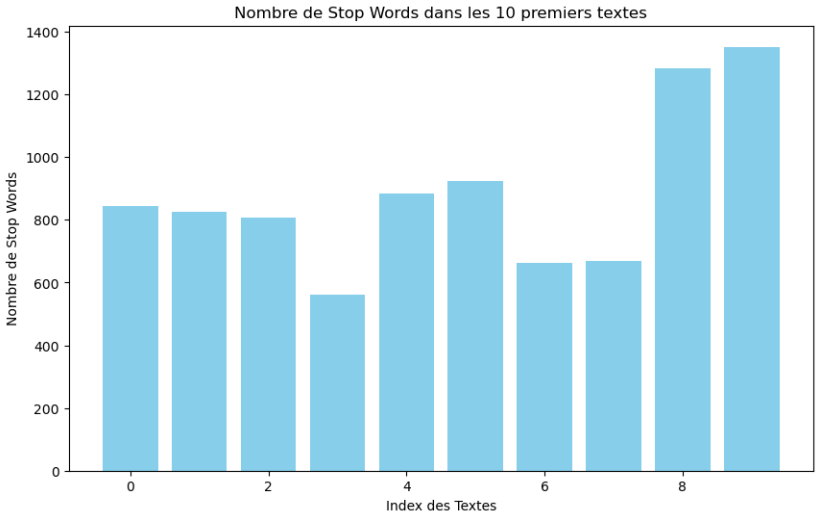
Pour ce projet de traitement de lanagage, nous avons choisi une base de données qui comporte des textes lié à trois types de cancer : Thyroid Cancer,Colon Cancer,Lung Cancer. Il y'a une colonne qui comporte le type de cancer et une colonne description qui comporte un article de maximum 6 pages qui parle de ce type de cancer avec des cas cliniques.

3 Nettoyage de Texte

Dans cette phase initiale, nous avons appliqué plusieurs étapes de nettoyage pour assurer la qualité des données :

1. **Suppression de Valeurs Dupliquées :** Élimination des doublons pour garantir l'intégrité des données. Dans la phase de nettoyage de texte, l'élimination des doublons a été une étape essentielle visant à garantir l'intégrité des données. Cette opération consiste à supprimer les entrées textuelles en double au sein de la base de données.

2. **Suppression des Stop Words :** Les mots fréquents mais non informatifs ont été éliminés. Les "stop words" sont des mots très fréquents dans une langue donnée (comme "et", "le", "de", etc.) qui ne portent généralement pas de signification particulière et ne contribuent pas significativement à l'analyse du contenu sémantique d'un texte. Dans le graphique ci dessous , nous pouvons voir le nomnre de Stop words qui apparaissent dans la dix permiers textes de notre base de données.



3. **Suppression des Caractères Spéciaux :** Élimination des caractères non alphanumériques. L'objectif principal de cette étape est de simplifier le texte en ne conservant que les éléments alphanumériques qui portent une signification linguistique ou numérique. Cette opération peut également contribuer à éviter des erreurs potentielles lors de l'application d'algorithmes d'analyse textuelle, car certains de ces caractères spéciaux pourraient perturber le processus d'extraction d'informations.

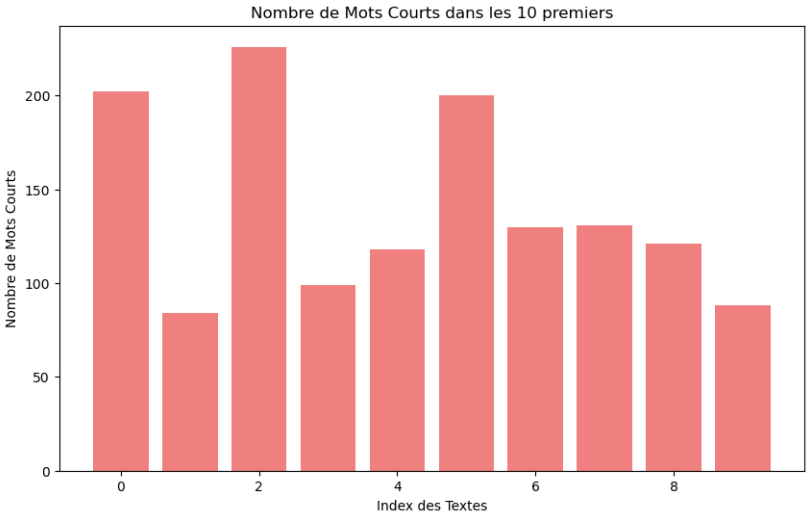
4. **Mise en Minuscule :** Uniformisation de la casse pour une analyse cohérente. Cette normalisation permet d'éviter toute ambiguïté liée à la casse des lettres. En effet, sans cette étape, les algorithmes d'analyse de texte pourraient interpréter différemment les mots écrits en majuscules et en minuscules, considérant par exemple "Médecine" et "médecine" comme deux termes distincts.

5. **Suppression des Mots Courts :** Rejet des mots courts avec un graphique représentant les 10 premiers textes. L'objectif est d'améliorer la qualité des données en éliminant le bruit introduit par des mots qui ont une contribution limitée à la compréhension du contenu sémantique d'un texte. En éliminant les mots courts, on peut souvent mettre en évidence des termes plus substantiels et pertinents pour l'analyse.

   Dans le graphique ci-dessous nous pouvons visualiser les mots courts pour les dix premiers textes.

4 Analyse Lexicale et Sémantique

Cette section se concentre sur l'analyse du texte après le nettoyage initial :



1. **Suppression des Mots Rares :** Élimination des termes peu fréquents pour réduire le bruit. La suppression des mots rares permet de se concentrer sur les termes plus fréquents et plus représentatifs du domaine ou du sujet abordé.

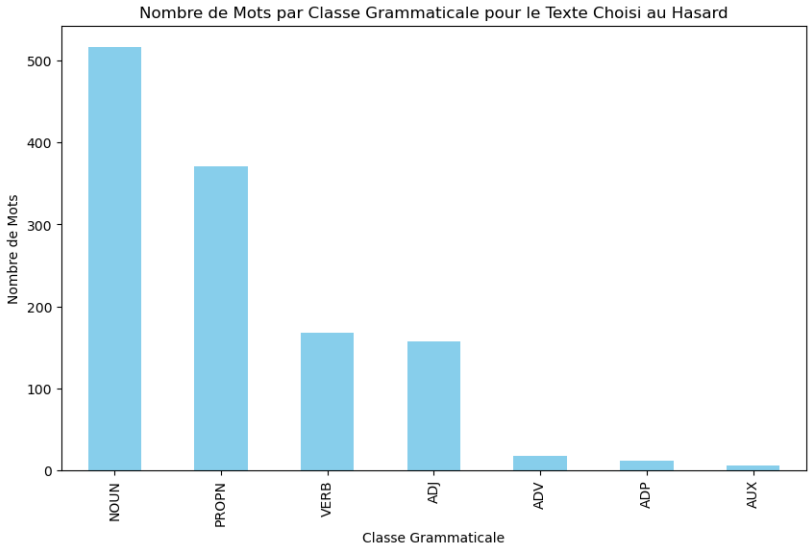
2. **Stemming et Lemmatisation :** Application de techniques pour normaliser la base.

   L'étape de stemming et lemmatisation vise à normaliser les termes dans la base de données en les ramenant à leur forme racine ou à leur forme canonique. Ces techniques sont utilisées pour réduire la diversité des termes apparentés, facilitant ainsi l'analyse textuelle en considérant différentes formes d'un mot comme une seule entité.

   Le stemming consiste à réduire les termes à leur racine ou à leur "stem". Par exemple, les mots "manger", "mangeait" et "mangeront" seraient tous ramenés à la racine "mang". Le stemming est une opération plus simpliste qui peut conduire à des résultats plus agressifs, mais il est souvent plus rapide à appliquer.

   La lemmatisation va un peu plus loin en utilisant des connaissances linguistiques pour ramener les termes à leur forme canonique ou "lemme". Par exemple, les formes verbales comme "manger", "mangeait" et "mangeront" seraient toutes ramenées à "manger", leur forme de base. La lemmatisation est généralement plus précise, mais peut nécessiter davantage de ressources computationnelles.

   L'application de ces techniques de normalisation a pour objectif d'améliorer la cohérence dans la représentation des termes, de réduire la dimensionnalité des données textuelles, et de favoriser la similarité entre termes apparentés. Cela facilite l'analyse en regroupant les différentes formes d'un



mot sous une seule catégorie, contribuant ainsi à une meilleure compréhension du contenu textuel.

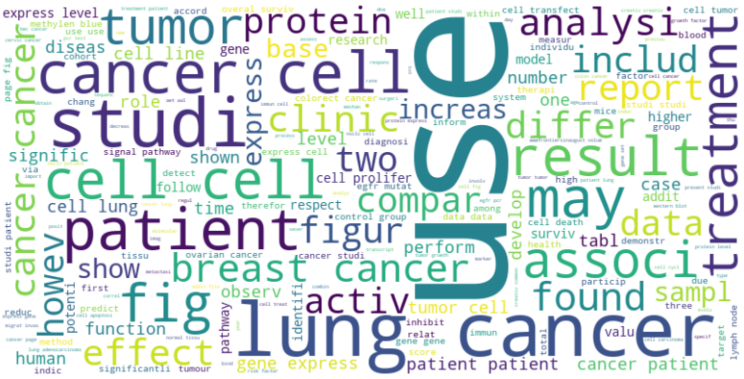
3. **Identification des classes gramaticale des mots :** L'identification des classes grammaticales des mots, également appelée étiquetage grammatical (part-of-speech tagging en anglais), est une étape du traitement de langage naturel visant à attribuer à chaque mot d'un texte une étiquette correspondant à sa classe grammaticale. Dans le graph ci-dessous nous pouvons voir les classes gramaticales des mots d'un texte tiré au hasard dans notre base de données.

4. **Named Entity Recognition (NER) :** L'objectif principal de la NER est d'extraire des informations structurées à partir du texte en identifiant et en classifiant les entités nommées présentes. Cette tâche est cruciale pour comprendre le contexte et la signification du texte, en particulier dans le domaine du traitement de texte médical.

5 Analyse Statistique et Recherche de Mot

Cette section se penche sur des analyses statistiques et des recherches spécifiques :

1. **Nuage de Mots :** Nous avons essayer de visualiser les mots de notre base de données en regardant leur fréquences d'apparition.



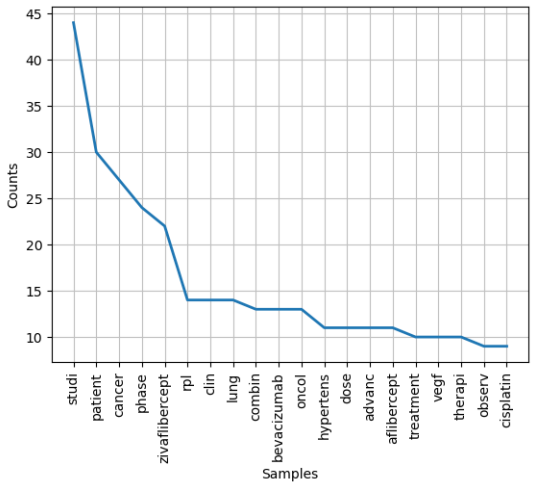
Dans le graph ci dessous , nous pouvons voir notre nuage de mot de notre base de données.

2. **Fouille de Texte :** Afin de faire la fouille de texte, nous avons fait quelques tests , tel que le nombre d'apparition du mot "thyroide" dans la base.

6 Association de Mots et Classification

Cette partie explore les relations sémantiques et les méthodes de classification :

1. **Association de Mots :** Identification des relations sémantiques et des co-occurrences fréquentes entre les mots. Cette analyse permet de découvrir comment certains mots sont liés les uns aux autres dans le contexte du corpus textuel.



Dans la figure ci-dessous nous pouvons voir cette association de mots pour un texte tiré aléatoirement dans notre base de données.

2. **Classificateur Bayesien :**

   Le classificateur bayésien est un modèle probabiliste basé sur le théorème de Bayes. Il attribue des probabilités aux différentes classes d'un ensemble en se basant sur la probabilité conditionnelle des caractéristiques observées, ce qui le rend efficace pour la classification de textes et d'autres données catégorielles. Nous avons utilisé ce modèle pour faire la classification de notre base de données. Nous avons abouti à une accuracy de 0.55.

   Avec la version BernoulliNB du modèle Bayésien nous trouvons un performance améliorée, soit un accuracy de 0.69 en général.

3. **Modèle LSTM :** Le modèle LSTM (Long Short-Term Memory) est une architecture de réseau de neurones récurrents (RNN) utilisée pour traiter des séquences de données, notamment dans le domaine du traitement de langage naturel. Contrairement aux RNN classiques, les LSTM peuvent mieux gérer les dépendances à long terme dans les séquences en utilisant des mécanismes de portes.

   Nous avons implémenté ce modèle avec deux couches LSTM de 128 neurones, et une couche dense, ainsi que 3 couches dropout entre les différentes couches, atteignant une précision de 0.68.

7 Conclusion

En conclusion, ce rapport détaille les différentes étapes du projet de traitement de langage, du nettoyage de texte à l'analyse statistique et à la classification.

Les techniques de traitement de langage naturel appliquées ont permis d'obtenir des informations précieuses à partir des données médicales, ouvrant la voie à une compréhension approfondie et à des classifications précises dans le domaine du cancer. Les modèles Bayésien et LSTM ont montré des performances encourageantes, soulignant le potentiel de ces approches pour l'analyse de texte médical.

Parmi les différents modèles que nous avons pu tester, c'est le modèle LSTM à deux couches LSTM et le modèle Bernoulli Bayésien qui nous donnent les meilleur performances sur les ensembles de test (0.68 accuracy pour le premier et 0.69 pour le deuxième).

Cependant les performance restent un peu médiocres pour ces types de modèles on pourrait s'attendre à plus. Le problème ici est la taille de notre jeu de données qui n'est pas assez grand dès que nous enlevons les doublons, nous arrivons à du overfitting très vite lorsque le nombre d'epochs augmente dans le cas des modèles LSTM.

Finalement dans notre cas spécifique, on pourrait préférer le modèle Bayésien Bernoulli au Modèle LSTM vu qu'il offre une performance similaire tout en étant beacoup moins lourd à entrainer.