Objet : Quatrième réunion avec nos encadrants

Axel CAROT

20 janvier 2024

1 Présentation des Avancées

Lors de notre récente réunion avec les commanditaires, nous avons présenté les progrès réalisés dans notre projet d'analyse textuelle sur la sécurité alimentaire. Nous avons abordé plusieurs aspects techniques et méthodologiques clés.

Nettoyage et Traitement des Données : Nous avons commencé par nettoyer les données textuelles, une étape cruciale pour assurer la qualité de notre analyse. Ensuite, nous avons procédé à la tokenisation des textes en utilisant le modèle linguistique Camembert.

Clustering et Identification des Sujets Principaux : Nous avons implémenté un clustering pour déterminer les sujets les plus fréquents dans les textes. Bien que cette étape soit encore en phase expérimentale, nous avons réussi à mettre en place le processus de base.

Analyse des Composantes Principales (ACP) : Pour réduire la dimensionnalité des données, nous avons utilisé l'ACP, en choisissant 100 composantes principales qui résument environ 95% de l'information. Cela a facilité l'identification du nombre optimal de clusters.

Exploration des Clusters : Chaque cluster a été exploré pour examiner le nombre d'articles et la fréquence des mots-clés. Nous avons remarqué une répartition inégale des articles par cluster et une prédominance de stop-words, soulignant le besoin d'un filtrage plus poussé.

Défis Techniques et Réflexions : Nous avons rencontré des difficultés techniques, notamment dans l'extraction des caractéristiques et l'évaluation des clusters. Ces défis nous ont conduit à réfléchir sur l'amélioration de nos méthodes, notamment en envisageant l'utilisation de modèles de topic modeling avancés comme BERT.

2 Travail à faire

Pour la suite du projet, nous avons identifié plusieurs pistes de travail et de réflexion :

Amélioration du Nettoyage des Données : Il est essentiel de peaufiner notre processus de nettoyage pour éliminer efficacement les stop-words et autres éléments non pertinents.

Optimisation du Clustering : Nous devons revoir notre approche de clustering pour obtenir une répartition plus équilibrée et significative des articles.

Utilisation de Modèles Avancés : L'adoption de modèles de topic modeling basés sur BERT pourrait nous aider à obtenir une représentation plus précise et visuelle des sujets.

Évaluation des Clusters : Nous devons développer une méthode robuste pour évaluer l'efficacité de nos clusters, en tenant compte de la diversité et de la pertinence des sujets.

Intégration des Résultats et Ajustements : Il sera crucial d'intégrer nos diverses analyses pour une compréhension holistique des sujets liés à la sécurité alimentaire, et de faire des ajustements en fonction des feedbacks des commanditaires.