Rapport: Réflexion et Planification du Pipeline sur la partie entre le prétraitement textuel et la géolocalisation des articles:

Ivan Can Arisoy

3 novembre 2023

1 Prétraitement des Données Textuelles

1.1 Document Embeddings

— Tokenisation : Utilisation du tokenizer CamemBERT pour segmenter chaque document du corpus.
— Embeddings : Passage de chaque document tokenisé à travers CamemBERT pour obtenir les embeddings.
— Moyennisation des embeddings : Moyennisation des embeddings de chaque token dans un document pour obtenir un vecteur d'embedding unique pour le document entier.
— Lemmatisation : Conversion de chaque mot à sa forme de base.
— Suppression des mots vides : Élimination des mots non pertinents.
— Vectorisation (TF-IDF) : Transformation du texte en vecteurs numériques avec TF-IDF.

2 Filtrage de Pertinence

Utilisation du modèle FlauBERT pour évaluer la pertinence du texte par rapport au lexique de la sécurité alimentaire et conservation des documents pertinents pour la question à l'étude.

3 Analyse de Sentiment*

Avant le clustering, utilisation d'un modèle pré-entraîné tel que BERT pour réaliser une analyse de sentiment sur les documents.

4 Modélisation des Sujets

4.1 Clustering

1. Conversion des embeddings documentaires en une matrice pour le clustering.
2. Utilisation d'un algorithme de clustering (par exemple, KMeans) pour regrouper les embeddings.
3. Attribution de chaque document à son cluster respectif, représentant un sujet ou thème particulier.

Enrichissement du Lexique

Il existe peu de vocabulaire du lexique de la sécurité alimentaire en Français disponible sur Internet. Une solution consisterait à enrichir le lexique manuellement en ajoutant des mots au dictionnaire. On pourrait également envisager le « web scraping » pour compléter notre lexique.

Planification des Tâches

Étape	Description	Dates
1. Prétraitement des données textuelles	Enrichissement du lexique : 1 jour Tokenization : 1 jour Embeddings et Moyennisation des embeddings : 1 jour Lemmatisation, Suppression des mots vides, et Vectorisation (TF-IDF) : 1 jour	03/11 - 06/11
2. Filtrage de pertinence	Configuration de FlauBERT : 0,5 jour Évaluation et filtrage des documents avec le lexique : 1,5 jour	07/11 - 08/11
3. Analyse de sentiment	Configuration de BERT pour l'analyse de sentiment : 0,5 jour Analyse de sentiment des documents : 0,5 jour	09/11
4. Modélisation des sujets	Clustering des embeddings documentaires : 2 jour Attribution des documents aux clusters : 1 jour	10/11 - 13/11
5. Analyse et Interprétation des Clusters	Etudes de chaque cluster : 2 jour Visualisation des clusters : 1 jour	13/11 - 15/11
6. Géolocalisation et visualisation des Articles		15/11 - 17/11