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

Ivan Can Arisoy

10 novembre 2023

Les Étapes Principales

1 Acquisition des données

2 Traitement des Données Textuelles

Préparation des données textuelles à l'analyse. Les étapes sont les suivantes :
— Tokenization : Découpage du texte en mots ou en phrases.
— Lemmatization : Réduction des mots à leur racine.
— Stopword Removal : Élimination des mots très courants qui n'apportent pas de valeur ajoutée pour l'analyse.
— Vectorization (TF-IDF) : Transformation du texte en valeurs numériques pour faciliter le traitement par les algorithmes d'apprentissage automatique.

3 Filtrage de Pertinence

Nous utiliserons le Word Embedding avec Word2Vec pour évaluer la pertinence des articles. Une similarité sémantique sera calculée, avec un seuil défini pour ne retenir que les articles les plus pertinents.

4 Modélisation des Sujets

Nous appliquerons des méthodes de modélisation de sujets pour classifier les thèmes dans notre collection d'articles. La vérification automatisée des sujets sera assurée par un classificateur personnalisé formé sur des sujets étiquetés, avec la possibilité d'utiliser BERT pour une intégration contextuelle plus précise.

5 Analyse de géolocalisation améliorée

Pour mieux comprendre la répartition géographique des discussions sur la sécurité alimentaire, nous utiliserons :
— Reconnaissance d'entités nommées (NER) : pour identifier les lieux mentionnés dans les textes et Hugging Face's Transformers (CamemBERT)
— Association d'emplacement au niveau de la phrase/du paragraphe.
— Géocodage : pour convertir les noms de lieux en coordonnées géographiques.

6 Analyse de séries chronologiques pour la dynamique temporelle

Nous intégrerons une dimension temporelle en extrayant les dates des articles et en analysant les tendances au fil du temps..

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. Traitement des Données Textuelles	Tokenization : 1 jour Lemmatization : 1 jour Suppression des mots vides : 1 jour Vectorisation (TF-IDF) : 1 jour	07/11 - 11/11
3. Filtrage de Pertinence	Configuration de Word2Vec et évaluation de pertinence : 1 jours	11/11
4. Modélisation des Sujets	Configuration et entraînement de LDA : 2 jours Vérification automatisée des sujets : 1 jour	11/11 - 13/11
5. Analyse de Géolocalisation Améliorée	Configuration de NER (CamemBERT) et géocodage : 2 jours Association d'emplacement au niveau de la phrase/du paragraphe : 1 jour	14/11 - 16/11
6. Analyse de Séries Chronologiques pour la Dynamique Temporelle	Extraction des dates et analyse temporelle : 1 jours	17/11
7. Développement de l'Indicateur TXT-FS	Calcul de la similarité sémantique : 1 jours Agrégation et analyse au niveau provincial : 1 jour	17/11