Projet N°1 : Text Mining : Text Classification Problem

Description du problème :

Considérant un fournisseur de soins de santé de premier plan aux États-Unis qui fournit des soins de haute qualité en mettant l'accent sur des services de soins de santé de bout en bout. Les services de soins de santé vont des diagnostics médicaux de base aux services d'urgence critiques.

Le fournisseur suit un système de billetterie pour tous les appels téléphoniques reçus dans tous les départements.

Les appels au fournisseur peuvent concerner un nouveau rendez-vous (**New Appointment**), une annulation (**Cancellation**), des requêtes de laboratoire (**Lab Queries**), des renouvellements médicaux (**Medical Refills**), des assurances (**Insurance Related**), un avis médical général (**General Doctor Advise**), etc.

Les billets contiennent les détails du résumé de l'appel et la description des appels rédigés par divers membres du personnel sans directives textuelles standard.

Data : projet1_data.csv

L'objectif est de classer le **texte** 'SUMMARY' dans ses catégories et sous-catégories en suivant les étapes suivantes :

1. **Nettoyage des données**
   a. Supprimer les colonnes indésirables field_id et DATA et garder les colonnes - "summary","categories","sub_categories","previous_appointment","id".
   b. Supprimer la catégorie et la sous-catégorie indésirables **Junk**.
   c. Modifier les catégories et sous-catégories en minuscules puis supprimer les doublons.
   d. Analyser les données pour trouver les valeurs manquantes. Si une valeur manquante est trouvée, les lignes correspondantes sont supprimées.
   e. Écrivez une fonction nommée clean_text qui prend une chaîne contenant du texte et renvoie une version nettoyée en la mettant entièrement en minuscule, en supprimant tous les espaces du début ou de la fin et en supprimant tous les caractères qui ne sont ni des espaces ni des caractères alphabétiques.
   f. Appeler **clean_text(' This "1" wasn`t BAD!!!! ')** pour vérifier que votre fonction fonctionne correctement.
2. **Compréhension des données**
   g. Tracer la proportion de catégories et de sous-catégories dans l'ensemble de données.
   h. Trouver les mots fréquents et tracer l'histogramme en fonction du nombre de mots dans chaque catégorie.
3. **Construction Term document Matrix**
   i. Prétraiter le texte 'SUMMARY' en supprimant les chiffres, les ponctuations, les espaces, les mots vides, en les transformant en minuscules et en racinant le document (lemmes)
   j. Utiliser les poids TF-IDF et construire la matrice document terme.

k. Quelle est la dimensionnalité de la matrice et supprimer les termes épars (termes avec une valeur maximale de zéro)
l. Construisez le nuage de mots pour comprendre la distribution des termes fréquents.

4. Création des modèle

m. Diviser les données en train et teste
n. Utiliser la régression logistique pour construire le modèle de classification.
o. Choisir un autre algorithme de catégorisation et construire le nouveau modèle.

5. Validation du modèle

p. Utiliser le modèle de régression pour les données de test et afficher la matrice de confusion.
q. Calculer la précision du modèle.
r. Comparer la précision des deux modèles.

6. Question Ouverte

Proposer en 2 lignes une suggestion qui pourra améliorer les résultats obtenus.