Page de Garde

Ecole : Efrei

Titre : Projet chatbot

Réalisé par : PIERSON Macéo et DOLLEZ—LOFFREDA Angelo

Groupe : PMP1

Année scolaire : 2023-2024 / promo 2028

Introduction au projet

Ce projet a pour but de créer un chatbot qui va pouvoir nous retourner des informations sur un fichier texte qu'on lui aura préalablement fourni.

Le projet est composé d'un menu et de plusieurs fonctions qui permette de faire fonctionner le menu.

L'objectif du chatbot sera donc de répondre au différentes demandent de l'utilisateur, donner grâce au menu, comme les noms des fichiers, ce dont ils sont composés, et les mots les plus prononcés dans les fichiers.

Ce projet comprend deux répertoires : “speeches” contenant 8 fichiers, chacun contenant un discours présidentiel, et “clean”.

functions.py contient les fonctions suivantes :

- **list_of_files(directory, extension : .txt)** : Prend un répertoire et une extension comme arguments et renvoie une liste des fichiers de l'extension donnée dans le répertoire. Le module **os** est utilisé.

- clean : Prend la liste des textes du répertoire “speeches” et supprime la ponctuation de chaque élément. Les textes nettoyés sont ensuite stockés dans le répertoire "clean", dans un sous-répertoire portant le nom de : “Nomination_(nom du président).txt” .

- **tf** : Prend un nom de répertoire et calcule le score TF pour l’occurrence de ce mot dans ce texte .
- **Idf** : Prend un nom de répertoire et calcule le score IDF pour chaque mot dans les fichiers du répertoire.
- **tfidf** : Prend un nom de répertoire et calcule le score TF-IDF pour chaque mot dans les fichiers.

Pour plus de précision sur le TF-IDF

Le TF mesure la fréquence d'un mot dans un document en le comparant au nombre total de mots du document. Il évalue ainsi l'importance relative d'un mot dans ce document. L'IDF, quant à lui, mesure l'importance d'un mot dans une collection de documents en mettant l'accent sur les mots moins courants, considérés comme plus pertinents. Ensemble, le TF-IDF est une méthode utilisée en traitement du langage naturel pour déterminer la pertinence d'un mot dans un document au sein d'une collection.

Ici , nous renvoyons un dictionnaire avec chaque mot comme clé et son nombre d’occurrence comme valeur.

Pour la partie 1, nous avons décidé de faire le menu ainsi que l’assignation des noms des présidents à leurs

Textes directement dans le main.py

Problèmes partie 1 :

- **Problèmes à partir des fonctions succédant TF-IDF :**

```
Pour afficher tout les noms des répertoires tapez 1
Pour afficher les noms des présidents tapez 2
Pour afficher les noms et prénoms des présidents tapez 3
Pour afficher tout le TF_IDF tapez 4
Pour afficher tout les mots important du TF_IDF tapez 5
Pour afficher tout les mots peu important du TF_IDF tapez 6
Pour afficher tout les mots les plus répéter par Chirac tapez 7
Pour afficher tout les mots en rapport avec la nation tapez 8
Pour afficher tout les mots en rapport avec l'écologie tapez 9
Pour arrêter le programme tapez 10
Tapez le numéro auquel vous voulez accéder :5
                        ^^^^^^^^^^^^^^^^^^
TypeError: object of type 'int' has no len()
```

functions_partie2.py si elle existait :

- **Token** : Renvoie une liste de mots à partir d'une chaîne de caractères.
- **Recherche (dico)** : Retourne une liste des mots communs entre une question et un dictionnaire.
- **vecteur** : Calcule un vecteur TF-IDF pour une chaîne de caractères donnée.
- **dictionnaire** : Calcule un dictionnaire de vecteurs TF-IDF pour une chaîne de caractères.

- **produit_scal** : Calcule le produit scalaire entre deux vecteurs.
- **norme**: Calcule la norme d'un vecteur.
- **similaire** : Mesure la similitude entre deux vecteurs.
- **comparer** : Trouve le fichier le plus proche d'un vecteur donné dans une liste de noms.
- **tfidf_max** : Identifie le mot avec le score TF-IDF le plus élevé dans une chaîne de caractères.
- **réponse** : Renvoie la première phrase contenant le mot avec le score TF-IDF le plus élevé dans une question.

main.py contient des instructions pour utiliser ces fonctions. Le programme principal interroge d'abord l'utilisateur sur ses préférences pour accéder aux fonctionnalités du projet, telles que le nettoyage des discours, la création d'un dictionnaire présidentiel, et l'analyse des mots les plus et moins importants selon le score TF-IDF.