Thomas DANJOU et Rayan EID
Groupe D 2023/2024

Promo: 2028

Chatbot2000

Table des matières

Introduction

- Ce document est notre rapport du projet python du semestre 1.
- Ce projet consiste à développer un chat bot et plusieurs autre fonctionnalités associé à un corpus de texte de discours de président.
- Un chatbot est un programme simulant une conversation soit sur des thèmes généraux soit sur un domaine précis. Dans notre cas, le chatbot est consacrer à répondre aux questions qui sont en rapport avec le corpus de texte.

Fonctionnalité

Notre programme contient 7 fonctionnalités :

-1 : La première fonctionnalité sert à afficher la liste des mots les moins importants dans le corpus de documents. Un mot est dit non important, si son TD-IDF = 0 dans tous les fichiers.

-2 : la deuxième fonctionnalité afficher le(s) mot(s) ayant le score TD-IDF le plus élevé.

-3 : La troisième fonctionnalité affiche le(s) mot(s) le(s) plus répété(s) par le président Chirac hormis les mots dits « non importants ».

-4 : La quatrième fonctionnalité indique le(s) nom(s) du (des) président(s) qui a (ont) parlé de la « Nation » et celui qui l'a répété le plus de fois.

-5 : La cinquième fonctionnalité indique le premier président à parler du climat et/ou de l'écologie.

-6 : La sixième fonctionnalité cherche quel(s) est(sont) le(s) mot(s) que tous les présidents ont évoqués hormis les mots dits « non importants ».

-7 : La dernière fonctionnalité est le chatbot, il permet de répondre aux questions posées par l'utilisateur par rapport à l'ensemble des documents.

Qu'est-ce que le TF et l'IDF ?

Tout d'abord le TF c'est une mesure de la fréquence à laquelle un terme particulier apparaît dans un document. Elle est calculée en comptant le nombre d'occurrences d'un terme dans un document et en le divisant par le nombre total de termes dans le document. Elle vise à donner une idée de l'importance d'un terme dans un document spécifique.

L'IDF c'est une mesure de l'importance d'un terme dans l'ensemble de la collection de documents. Elle est calculée en prenant le logarithme inverse du nombre total de documents divisé par le nombre de documents contenant le terme. Elle vise à donner plus de poids aux termes rares qui sont susceptibles d'être plus informatifs.

En général : Le TF-IDF est une mesure statistique utilisée en traitement automatique du langage naturel et en recherche d'information pour évaluer l'importance d'un terme dans un document par rapport à une collection de documents.

Fichiers et algorithme

Le fichier main.py est fichier principal de notre code. Il relit les différentes fonctions et les fichiers pour permettre de faire fonctionner le programme.

Le dossier « speeches » est l'ensemble des textes avec ponctuations et accents.

Le fichier « fonction_de_base » comporte plusieurs fonctions dont :

- Une fonction list_of_files qui permet de parcourir la liste des fichiers d'une extension donnée et dans un répertoire donné.
- Une fonction nom_president qui extrait le nom des présidents dans le nom du discours associé.
- Une fonction prenom_president qui associe un prénom à chaque nom de président.
- Une fonction retourne_nom_president qui permet d'enlever les doublons car un président peut avoir plusieurs discours.
- Une fonction convertion_minuscule qui change tous les mots dans l'ensemble des discours en minuscule.
- Et enfin la fonction supression_ponctuation qui retire toute la ponctuation dans l'ensemble des discours.

Les documents après être modifié sont dans le dossier cleaned.

Le fichier « fonction_methodeTFIDF.py » contient aussi plusieurs fonctions :

- Une fonction TF qui permet d'associer un nombre à chaque mot dans les différents textes et ce nombre correspond à sa fréquence d'apparition dans les discours. On les stocks dans un dictionnaire. L'utilisation d'un dictionnaire permet de retrouver facilement les valeurs ainsi que le mot qui leur est associé comme il n'y a peu d'opération qui seront effectué en utilisant le TF l'utilisation d'un dictionnaire et le plus simple car les valeur n'aurons pas à être modifiée.
- Une fonction IDF qui fait le log décimal du nombre total de document diviser par le nombre de document contenant le terme. Tout comme pour la fonction TF l'utilisation du dictionnaire et le choix de la simplicité afin d'associé un IDF à chaque mot sans avoir à créer beaucoup de variable et la faible utilisation de l'IDF en fait le meilleur choix.
- Une fonction TF_IDF qui prend en paramètre le répertoire où se trouvent les fichiers à décortiquer et elle renvoi un tuple contenant la matrice TF-IDF et une liste contenant tous les mots du corpus. La fonction TF_IDF est très utilisé dans tous les autres fonctions développé dans ce projet c'est pour cela que l'utilisation d'un tuple contenant une liste de mot et une matrice de valeur et le choix le plus simple car la modification d'un dictionnaire n'est pas une tâche facile comparé à celle d'une matrice qui se modifie très facilement c'est pour cela que l'on a utilisé le tuple plutôt que le dictionnaire.

Le fichier « générateur_de_reponse.py » contient plusieurs fonctions :

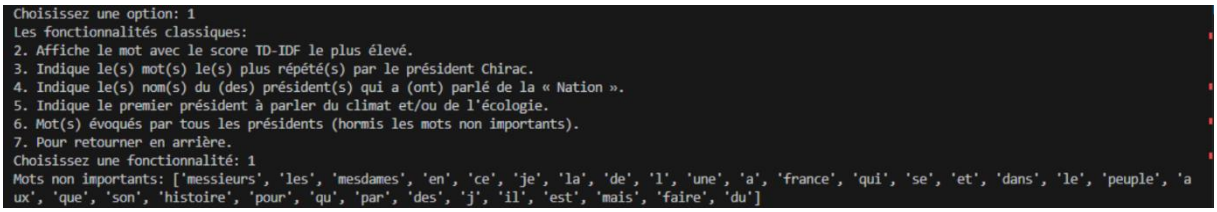
- Une fonction Tokenisation question qui prend tous les mots de la question et supprime la ponctuation et change les majuscules en minuscules. Elle retourne une liste de mot.

- Une fonction intersection_corpus_question qui prend une liste de mot qui sont présents dans les documents et qui ignorent les mots qui ne sont pas dans les documents car ils n'auront pas de valeurs TF-IDF associées.
- Une fonction vecteur_TF_IDF qui créée une matrice avec N nombre de lignes qui est égal au nombre de documents et M le nombre de colonnes égal au nombre de mots dans le corpus. Elle associe chaque mot de la question à un score TF puis sachant que les scores IDF de la question sont les même que ceux des mêmes mots dans le corpus, utilise le score IDF des mots de la question pour calculer le TF-IDF des mots de la question. On se trouve ici dans le même cas de la fonction TF_IDF calculé plus on est donc dans le même raisonnement vis-à-vis du choix de structure de donné qui est l'utilisation d'un tuple contenant la liste de mot du corpus et de la matrice possédant neuf lignes car nous avons dû faire une transposé de matrice pour pouvoir ensuite l'utiliser pour faire des calcules par rapport au document et non par rapport au mot comme précédemment.
- Une fonction produit scalaire qui calcul le produit scalaire de deux vecteurs qui sont enfaite des listes de nombres.
- Une fonction normes_vecteur qui calcul les normes des listes de nombres.
- Une fonction similarité qui est produit scalaire du premier vecteur avec le deuxième diviser par la norme du premier vecteur fois la norme du second.
- Une fonction document_pertinent qui calcule la similarité du vecteur de la question avec chacun des vecteurs du document puis retourne le nom du document correspondant à la valeur de similarité la plus élevée.
- Une fonction generation_reponse qui retourne une chaine de charactère et qui répond à la question de l'utilisateur grâce au différentes fonction défini auparavant. La question est bien évidement aussi une chaine de caractères
- Et enfin la fonction affine_reponse qui permet de rendre le chatbot plus humain en associant un début de phrase prédéfinit au question courante de l'utilisateur et qui permet aussi de faire une phrase de réponse lorsqu'il n'a pas la réponse.

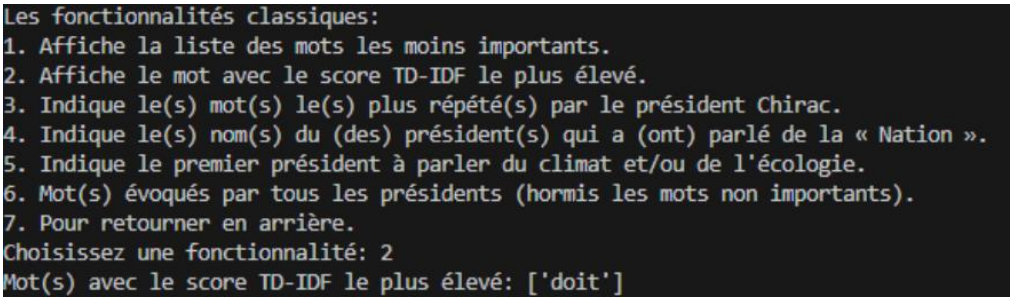
Le dernier fichier restant est le fichier fonction_menu.py est là où on a stocké le menu et défini les menu généraux.

Présentation des résultats

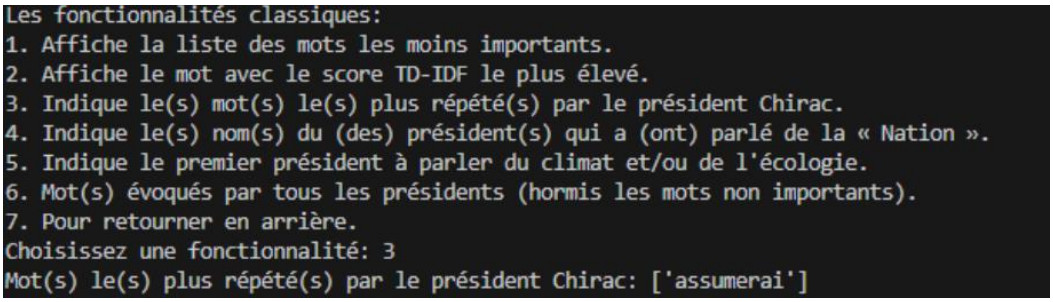
Fonctionnalité 1 :



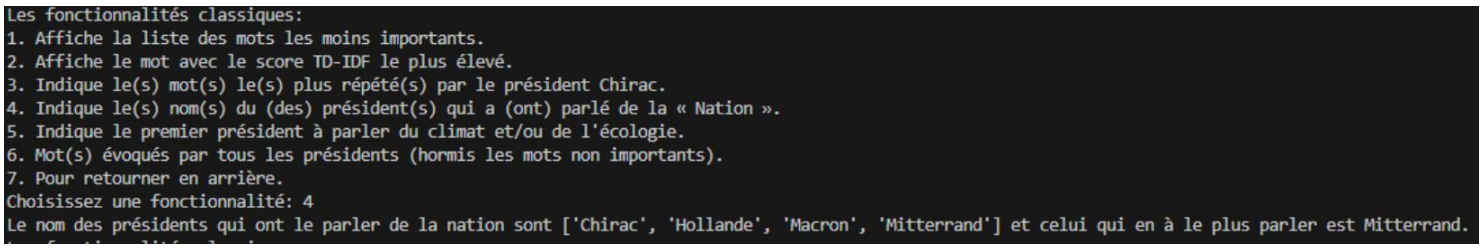
Fonctionnalité 2 :



Fonctionnalité 3 :



Fonctionnalité 4 :



Fonctionnalité 5 :

```
Les fonctionnalités classiques:
1. Affiche la liste des mots les moins importants.
2. Affiche le mot avec le score TD-IDF le plus élevé.
3. Indique le(s) mot(s) le(s) plus répété(s) par le président Chirac.
4. Indique le(s) nom(s) du (des) président(s) qui a (ont) parlé de la « Nation ».
5. Indique le premier président à parler du climat et/ou de l'écologie.
6. Mot(s) évoqués par tous les présidents (hormis les mots non importants).
7. Pour retourner en arrière.
Choisissez une fonctionnalité: 5
Le premier président qui a parlé de l'écologie est Hollande.
```

Fonctionnalité 6 :

```
Les fonctionnalités classiques:
1. Affiche la liste des mots les moins importants.
2. Affiche le mot avec le score TD-IDF le plus élevé.
3. Indique le(s) mot(s) le(s) plus répété(s) par le président Chirac.
4. Indique le(s) nom(s) du (des) président(s) qui a (ont) parlé de la « Nation ».
5. Indique le premier président à parler du climat et/ou de l'écologie.
6. Mot(s) évoqués par tous les présidents (hormis les mots non importants).
7. Pour retourner en arrière.
Choisissez une fonctionnalité: 6
['m', 'ont', 'leur', 'confiance', 'ai', 'chacun', 'à', 'hui', 'avec', 'tous', 's', 'libertés', 'notre', 'nation', 'toujours', 'sur', 'depuis', 'droits', 'mo
i', 'ceux', 'grande', 'emploi', 'comme', 'sont', 'tout', 'un', 'cette', 'sera', 'nouveau', 'monde', 'pays', 'bien', 'autres', 'femmes', 'au', 'rôle', 'dével
oppement', '', 'effort', 'intérêts', 'me', 'jour', 'n', 'pas', 'mai', 'affermie', 'politique', 'toutes', 'lui', 'mitterrand', 'avenir', 'conformément']
```

Fonctionnalité 7 (chatbot) plusieurs test :

Test1 :

```
PS D:\Documents\pychatbot-danjou-eid-d> & C:/Users/Thomas/AppData/Local/Programs/Python/Python39/python.exe d:/Documents/pychatbot-danjou-eid-d/main.py
1. Pour accéder aux fonctionnalités classiques.
2. Pour accéder au générateur de réponses automatiques.
3. Pour quitter.
Choisissez une option: 2
Les fonctionnalités du générateur de réponses automatiques:
1. Pour poser une question.
2. Pour retourner en arrière.
Choisissez une option: 1
Entez votre question:Es-ce que l'informatique est important dans le monde d'aujourd'hui?
Je ne connais pas cette question mais je dirais, le monde et l'Europe ont aujourd'hui, plus que jamais, besoin de la France.
```

Test2 :

```
PS D:\Documents\pychatbot-danjou-eid-d> & C:/Users/Thomas/AppData/Local/Programs/Python/Python39/python.exe d:/Documents/pychatbot-danjou-eid-d/main.py
1. Pour accéder aux fonctionnalités classiques.
2. Pour accéder au générateur de réponses automatiques.
3. Pour quitter.
Choisissez une option:
Choix invalide.
1. Pour accéder aux fonctionnalités classiques.
2. Pour accéder au générateur de réponses automatiques.
3. Pour quitter.
Choisissez une option: 2
Les fonctionnalités du générateur de réponses automatiques:
1. Pour poser une question.
2. Pour retourner en arrière.
Choisissez une option: 1
Entez votre question:Peux-tu me dire comment une nation peut-elle prendre soin du climat?
Oui, bien sûr!Et je songe bien sûr à François Hollande, faisant oeuvre de précurseur avec l'Accord de Paris sur le climat et protégeant les Français dans un
 monde frappé par le terrorisme.
```

Conclusion :

En général, ce projet fût une expérience très intéressante. Elle nous a appris beaucoup sur le plan technique et théorique. La gestion de temps pour la première parti était très bien gérer, la charge de travail et le temps donné était adéquat à la première partie mais pour la seconde se fût compliqué car la charge de travail attendu était bien trop élevé comparé au temps donné de la partie 2 a la partit 1. L'organisation du projet était bien faite entre nous et nous communiquions beaucoup sur nos avancé sachant que nous avions fait la plupart du projet en TP.