

2023-2024
Groupe A

My First Chatbot



Un projet de Maël Brosset et Julien Casamian

Introduction :

Ce projet est de créer un algorithme de traitement de texte naturelle qui par la suite pourras répondre à des questions posées par un utilisateur.

Il se compose de deux parties, la première est une partie à but purement technique. Notre algorithme va pouvoir analyser tout un corpus de document, il va pouvoir comparer les textes regarder les mots non importants et ceux qui au contraire sont très utile.

La deuxième partie quant à elle reste technique mais est plus interactive avec l'utilisateur. Dans cette partie du projet, un utilisateur va pouvoir poser des questions au chatbot, celui-ci va alors analyser ces questions et va pouvoir y répondre de la manière la plus pertinente possible avec le corpus de document qu'il à sa disposition.

Les objectifs du projet :

- Méthode de traitement du texte naturelle :
 - Calcul du coefficient TF-IDF d'un corpus de documents
 - Possibilité de rajouté facilement de nouveaux documents dans ce corpus
- Mise en valeur des possibilités du score TF-IDF :
 - Diverses fonctionnalités de calcul basique sur le score TF-IDF
 - Comparaison de documents du corpus grâce à ce score
 - Identification de texte à partir de mots
- Mise en place du Chat bot :
 - Traitement de question avec un score TF-IDF
 - Comparaison des questions et des documents avec la similarité cosinus
 - Réponse du bot en se basant des phrases du texte

Fonctionnalités :

En plus de la fonctionnalité principale du chat bot, qui est de répondre à l'aide du corpus de document à l'utilisateur, nous avons aussi développés des fonctionnalités de base qui sont : l'affichage de la matrice TF-IDF, la liste des mots non-importants du corpus de document, l'affichage du mot avec le score TF-IDF le plus élevé et du mot le plus répété par Chirac qui n'est pas un mot non important, le président qui parle le plus de nation dans ces discours et combien de fois il en parle, la liste des présidents parlant d'écologie et les mots qui ont été dit par tous les présidents mais pas dans tous les discours.

Fonctionnement de l'algorithme :

Le chat bot fonctionne grâce à 5 algorithmes principaux :

- Traitement des textes sous forme d'une liste de mots
- Création de la matrice TF-IDF du corpus de document
- Traitement de la question et attribution d'un vecteur TF-IDF
- Comparaison de la question aux différents documents
- Génération de la réponse

Pour tout le stockage de données des scores TF-IDF nous avons opté pour des dictionnaires, ce qui nous permet d'avoir accès rapidement au score d'un mot.

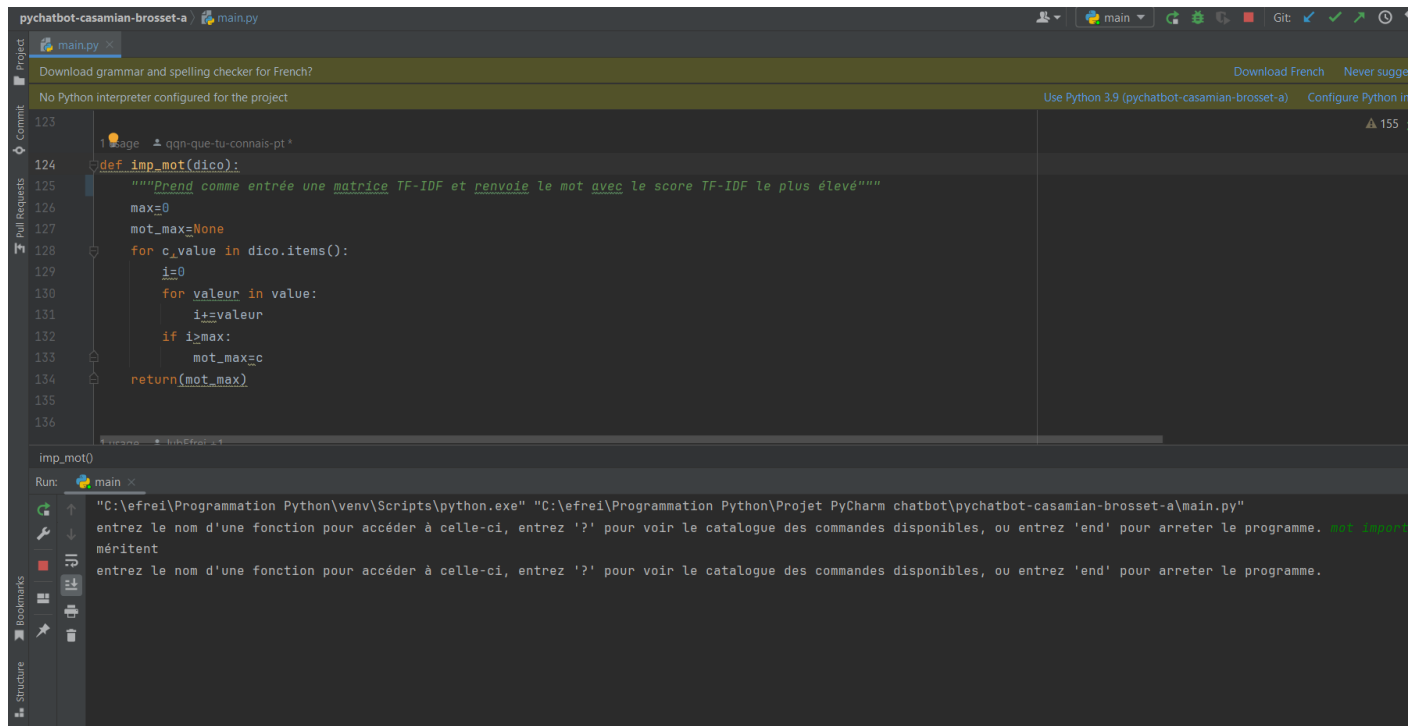
Pour la forme finale de la matrice nous avons donc un dictionnaire de dictionnaire, avec le premier dictionnaire un dictionnaire au même nombre d'entrées que de document avec une valeur par document. Et chaque valeur est elle-même un dictionnaire avec chaque mot en entrée et son score TF-IDF en valeur.

Lors de ce projet nous avons fait face à plusieurs difficultés, tout d'abord pour répondre à la question qui nous demande de trouver les mots dit par tous les présidents mais qui ne sont pas dit mots non importants nous avons dû créer un second corpus de document à cause de l'organisation de notre matrice.

Aussi, il y avait le vocabulaire employé qui n'était pas forcément compréhensible par moment (la notion de vecteur dans le texte)

Enfin nous avons remarqué que le temps nous manquait et avons dû nous précipiter pour la seconde partie.

Présentation des résultats



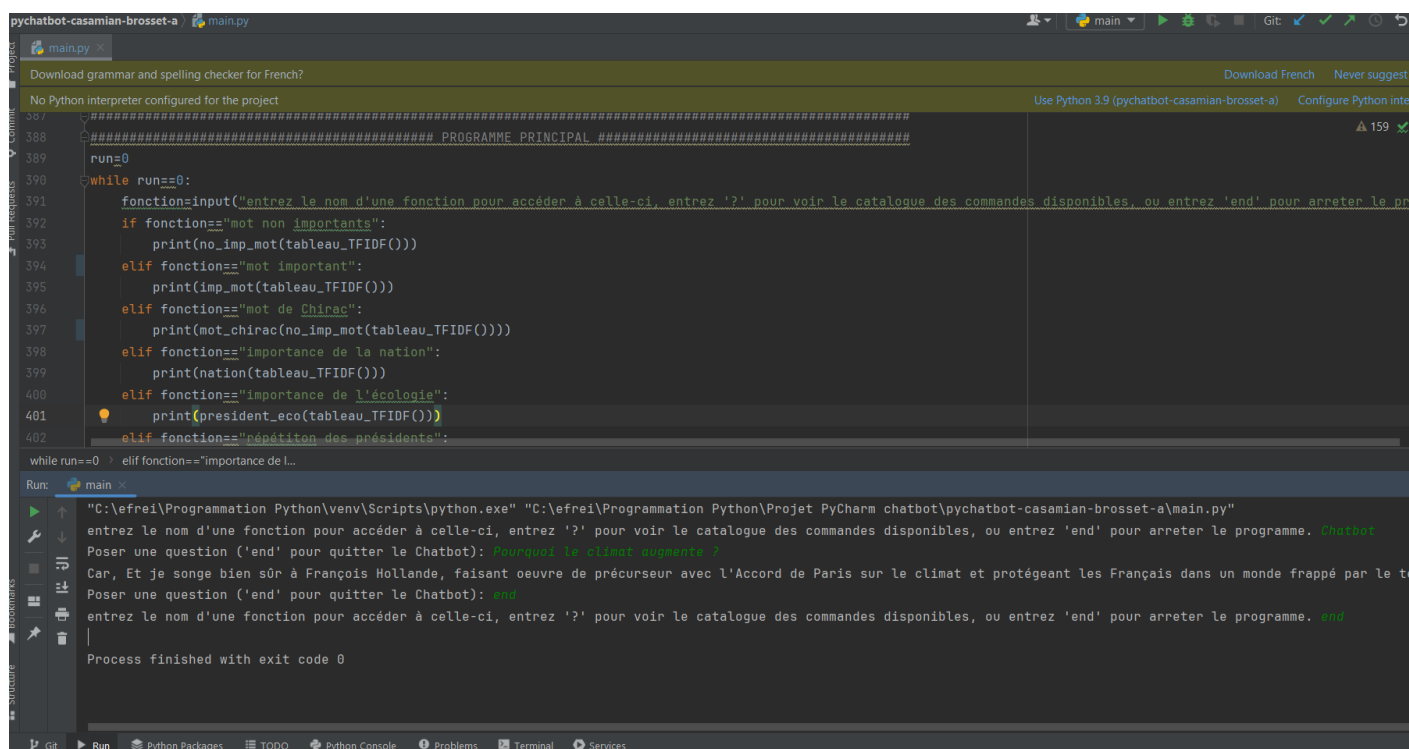
The screenshot shows the PyCharm IDE interface. The main editor displays a Python script named `main.py` with the following code:

```
123
124 def imp_mot(dico):
125     """Prend comme entrée une matrice TF-IDF et renvoie le mot avec le score TF-IDF le plus élevé"""
126     max=0
127     mot_max=None
128     for c_value in dico.items():
129         i=0
130         for valeur in value:
131             i+=valeur
132             if i>max:
133                 mot_max=c
134     return(mot_max)
135
136
```

Below the editor, the Run console shows the output of the program:

```
Run: "C:\efrei\Programmation Python\venv\Scripts\python.exe" "C:\efrei\Programmation Python\Projet PyCharm chatbot\pychatbot-casamian-brosset-a\main.py"
entrez le nom d'une fonction pour accéder à celle-ci, entrez '?' pour voir le catalogue des commandes disponibles, ou entrez 'end' pour arreter le programme. mot important
méritent
entrez le nom d'une fonction pour accéder à celle-ci, entrez '?' pour voir le catalogue des commandes disponibles, ou entrez 'end' pour arreter le programme.
```

Le mot le plus important est “méritent”



The screenshot shows the PyCharm IDE interface. The top toolbar includes icons for Run, Debug, and other development tools. The main editor window displays a Python file named `main.py` with the following code:

```
##### PROGRAMME PRINCIPAL #####
run=0
while run==0:
    fonction=input("entrez le nom d'une fonction pour accéder à celle-ci, entrez '?' pour voir le catalogue des commandes disponibles, ou entrez 'end' pour arreter le programme. ")
    if fonction=="mot non importants":
        print(no_imp_mot(tableau_TFIDF()))
    elif fonction=="mot important":
        print(imp_mot(tableau_TFIDF()))
    elif fonction=="mot de Chirac":
        print(mot_chirac(no_imp_mot(tableau_TFIDF())))
    elif fonction=="importance de la nation":
        print(nation(tableau_TFIDF()))
    elif fonction=="importance de l'écologie":
        print(president_eco(tableau_TFIDF()))
    elif fonction=="répétition des présidents":
```

The bottom pane shows the Run console output:

```
"C:\efrei\Programmation Python\venv\Scripts\python.exe" "C:\efrei\Programmation Python\Projet PyCharm chatbot\pychatbot-casamian-brosset-a\main.py"
entrez le nom d'une fonction pour accéder à celle-ci, entrez '?' pour voir le catalogue des commandes disponibles, ou entrez 'end' pour arreter le programme. Chatbot
Poser une question ('end' pour quitter le Chatbot): Pourquoi le climat augmente ?
Car, 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 t
Poser une question ('end' pour quitter le Chatbot): end
entrez le nom d'une fonction pour accéder à celle-ci, entrez '?' pour voir le catalogue des commandes disponibles, ou entrez 'end' pour arreter le programme. end
Process finished with exit code 0
```

Test du Chatbot avec une question de type ‘pourquoi... ?’

Lors de ce projet nous avons fait face à plusieurs difficultés, tout d’abord pour répondre à la question qui nous demande de trouver les mots dit par tous les présidents mais qui ne sont pas dit mots non importants nous avons dû créer un second corpus de document à cause de l’organisation de notre matrice.

Aussi, il y avait le vocabulaire employé qui n’était pas forcément compréhensible par moment (la notion de vecteur dans le texte)

Enfin nous avons remarqué que le temps nous manquait et avons dû nous précipiter par moment.

Conclusion

Ce projet nous a permis de mieux comprendre l'analyse et le traitement de texte qui au départ pouvait sembler assez flou. Il nous a aussi permis de découvrir des nouveaux modules et à nous débrouiller sans fonctions prédéfinies. A l'aide de ce projet on a pu approfondir notre maîtrise des fichiers ainsi que des dictionnaires.

Nous avons donc appris une méthode permettant l'analyse, le traitement et la comparaison de texte naturelle. Nous avons aussi pu avoir une meilleur idée du fonctionnement des matrices et des calculs de vecteurs de ces matrice.

Ce projet nous as aussi appris la collaboration sur git qui nous à permis de pouvoir travailler ensemble sur le même fichier sans déranger l'autre. Cela nous à permis de mieux gérer notre temps et de mieux collaborer ensemble.