Projet
My first chatBot

Réalisé par Farida BEDEWY, Dissi GAM et Ines SERIDJ

SOMMAIRE :

Introduction

Le projet que nous avons à réaliser durant ce premier semestre dans la matière programmation python est centrée sur un sujet très actuel : les chatbot. En effet, il nous permet de comprendre et de développer les concepts de base de traitement et d'analyse de texte utilisé pour générer des réponses intelligentes. Notre objectif est ici de créer un système, qui, grâce à un ensemble de textes donnés, peut répondre à des questions simples en fonction de la fréquence des mots dans ces derniers.

Pour se faire, l'algorithme passe par six grandes étapes. Dans un premier temps, le programme prend connaissance des données fournit par les textes et les prétraitent afin d'en extraire le sujet principal et d'utiliser ce vocabulaire dans la réponse finale. Ensuite, en utilisant la technique TF-IDF, mesurant la fréquence d'un terme et son importance dans le texte, un vecteur de même dimension que le nombre de textes est créé pour chaque mot de ces derniers, formant ainsi une matrice TF-IDF (où les lignes sont les mots et les colonnes le nombre de texte). Pour traiter les questions posées, le système calcule des vecteurs TF-IDF aux mêmes dimensions que ceux des mots déjà référenciés. Ces longueurs sont ensuite comparées afin de reconnaitre les mots similaires. Enfin, il sélectionne la réponse avec le plus grand nombre de mots similaire et répond à la question posée.

Ce projet nous permet de mettre en pratique les notions de liste 1D, liste 2D, fonctions, fichiers et collections, étudiées tout au long du semestre.

Le projet est divisé en trois parties dont une facultative: la première servant à créer et développer les fonctions et fichiers de base, la deuxième permettant le calcul de la matrice et la génération de réponses simples et la dernière pour une généralisation et ouverture du programme.

I. Présentation fonctionnelle du projet

Dans la première partie de notre projet, nous nous concentrons sur le traitement préliminaire des données textuelles. Cela implique la collecte et la préparation des documents de notre corpus, notamment les discours présidentiels. Ces étapes sont cruciales pour préparer les données en vue d'analyses plus poussées et de génération de réponses dans la suite du projet.

- **FONCTION 1 → list_of_files(directory, extension) :** cette fonction permet de parcourir la liste des fichiers d'une extension ou dans un répertoire donné
- **FONCTION 2 → print_list(file_list)** : cette fonction permet d'afficher la liste des noms des fichiers des discours de présidents
- **FONCTION 3 → extract_presidents(files_names) :** cette fonction permet d'extraire les noms des présidents à partir des noms des fichiers texte fournis
- **FONCTION 4 → associate_presidents_names():** cette fonction permet de créer et de retourner un dictionnaire qui associent les noms de présidents à leurs prénoms
- **FONCTION 5 → convert_text(directory, cleaned_directory) :** cette fonction permet de nettoyer le contenu textuel (lit, convertit en minuscules, supprime la ponctuation, ajuste les caractères spéciaux) des fichiers situés dans un répertoire donné
- **FONCTION 6 → counter(text) :** cette fonction permet de calculer la fréquence des mots. Cette fréquence est appelée score TF
- **FONCTION 7 → calculate_idf (directory):** cette fonction permet de calculer la fréquence IDF. Un score IDF élevé indique qu'un mot est important car il apparaît dans moins de documents. Elle affiche donc la fréquence inverse des mots dans l'ensemble des documents textuels
- **FONCTION 8 → calculate_tf_idf(directory) :** cette fonction permet de calculer la matrice TF-IDF pour un ensemble de documents textuels
- **FONCTION 9 → non_important_words (tfidf_matrix, unique_words):** cette fonction permet d'identifier et d'afficher les mots les moins important c'est-à-dire ceux avec un score TF-IDF de 0
- **FONCTION 10 → highest_tfidf_words(corpus_directory) :** cette fonction permet d'identifier et d'afficher les mots ayant le score TF-IDF le plus élevé
- **FONCTION 11 → get_words_from_indices(indices, unique_words) :** cette fonction permet de récupérer les mots correspondants à une liste d'indices donnés dans la liste unique_words pour convertir des indices en mots réels, facilitant l'interprétation des résultats de l'analyse de texte
- **FONCTION 12 → words_nation(corpus_directory, president_names) :** cette fonction permet d'analyser la fréquence du mot nation dans un ensemble de discours du président
- **FONCTION 13→ word_environment (corpus_directory):** cette fonction d'analyser la fréquence des mots liés à l'environnement dans un ensemble de discours des présidents pour étudier comment et combien de fois les problématiques environnementales sont abordées dans les discours présidentiels

Nous entrons maintenant dans la phase cruciale de génération automatique de réponses aux questions posées. Pour cela, notre approche consiste d'abord à calculer le TF-IDF de la question. Ensuite, nous identifions le document le plus pertinent dans notre corpus, celui ayant le plus de termes en commun avec la question. Cela augmente les chances de trouver une réponse pertinente dans ce document. Finalement, nous extrayons et raffinons la réponse à partir de ce document.

- **FONCTION 14 → tokenize_question(question) :** cette fonction permet de préparer une question pour l'analyse en la divisant en mots individuels
- **FONCTION 15 → find_terms_in_corpus(question_tokens, idf_scores):** cette fonction permet de filtrer les mots d'une question en conservant uniquement ceux qui sont présents dans les scores IDF
- **FONCTION 16 → calculate_question_tf_idf(question_tokens, unique_words,ids_scores) :** cette fonction permet de calcule le vecteur TF-IDF pour une question donnée en fonction du nombre de textes
- **FONCTION 17 → dot_product(vector_a, vector_b):** cette fonction permet de calculer le produit scalaire (somme des produits des éléments correspondants à deux vecteurs) entre vector_a et vector_b
- **FONCTION 18 → vecteur_norm(vector) :** cette fonction permet de calculer la norme (la longueur) de chaque vecteur. On la calcule grâce à la formule racine carrée de la somme des carrés de ses éléments
- **FONCTION 19 → cosinus_similarite(vector_a, vector_b) :** cette fonction permet de la similarité de cosinus entre deux vecteurs grâce à la formule produit scalaire de a et b divisé par le produit des normes de ces deux vecteurs
- **FONCTION 20 → find_most_revelant_document(tfidf_matrix, question_vector) :** cette fonction permet de déterminer le document le plus pertinent dans un ensemble de texte par rapport à une question donnée en comparant la similarité cosinus
- **FONCTION 21 → highest_tf_idf_word(question_tf_idf, unique_words) :** cette fonction permet d'identifier le mot ayant le score TF-IDF le plus élevé dans une question donnée
- **FONCTION 22 → find_sentence_with_words(text, word) :** cette fonction permet d'extraire et de renvoyer la première phrase contenant un mot spécifique à partir d'un texte donné
- **FONCTION 23 → refine_response(question, answer) :** cette fonction permet de filtrer une réponse générée en fonction du début de la question posée

II. Présentation technique

Maintenant que nous avons expliqué les différentes fonctions formant l'algorithme, nous allons explorer et détailler les principales étapes du fonctionnement de notre application. En effet, la première étape est le nettoyage des données: cette fonctionnalité utilise convert_text() pour transformer le texte en minuscules et supprimer la ponctuation. Cela uniformise le texte.

La deuxième boucle principale concerne IDF et TF-IDF.

- IDF: Implémenté dans calculate_idf(), cette fonction évalue la rareté d'un mot dans l'ensemble des textes

- TF-IDF: calculate_tf_idf() combine la fréquence des termes (TF) et l'IDF pour chaque document..

La dernière partie importante est la génération de réponses.

- Identification des documents pertinents : find_most_relevant_document() utilise la similarité cosinus pour identifier le document le plus pertinent par rapport à une question.

- Extraction de réponses : find_sentence_with_word() et refine_response() cherchent et filtre la réponse de renvoyer la réponse la plus adéquate et précise.

A travers ce projet, nous avons développé des algorithmes qui demandaient une sélection minutieuse de structures de données, telles que l'utilisation de dictionnaires pour le stockage efficace des fréquences et de listes pour les matrices TF-IDF. Nous avons également accordé une importance particulière à l'optimisation des performances et à la pertinence des résultats.

On retrouve donc, listé ci-dessous, plusieurs structures primordiales à la réussite de notre projet.

- D'abord, les listes : elles sont utilisées pour stocker une série d'éléments ordonnés notamment comme dans la foncion **list_of_files**. Les listes sont idéales pour conserver l'ordre des fichiers et permettent un accès itératif efficace.

- Ensuite on peut noter les Sets : celles-ci sont en particulier employés dans la fonction **extract_presidents** pour éliminer les doublons. Le set est rapide pour vérifier l'existence d'un élément et garantit l'unicité, ce qui est crucial pour lister les présidents sans répétition.
- Dans les dictionnaires, on retrouve la fonction **associate_president_names.** : elle est utilisée pour associer des clés à des valeurs. Ils fournissent un accès rapide et efficace aux données via des clés, ce qui est idéal pour récupérer le prénom d'un président à partir de son nom.

- Et enfin, on a la structure Defaultdict : il s'agit ici d'une variante du dictionnaire utilisée dans la fonction **counter** et **calculate_idf** pour gérer des valeurs par défaut ce qui permet de simplifier la gestion des mots non présents initialement dans le dictionnaire.

Cependant, nous avons tout de même rencontré quelques difficultés techniques. En effet, pour la fonction **def_words_environment** qui est en relation avec le climat, le nom des présidents s'affiche en double. Nous ne parvenons pas à afficher le nom des présidents sans doublons car lorsque l'on appelle la fonction **president_name,** cette dernière ne fonctionne pas et nous avons le mauvais affichage. Par conséquent, le nom du président s'affiche sous format « txt ».

Par ailleurs, l'utilisation de git nous a aussi causé des soucis. Etant donné qu'il s'agissait de notre première utilisation et que nous n'avons pas été suivi quant aux différentes fonctionnalités de ce nouveau logiciel comme nous avions pu l'être pour coding rooms par exemple, les débuts ont été difficiles.

III. Présentation des résultats

Lorsque nous lançons le programme, un menu s'affiche et propose deux options à l'utilisateur. Prenons ici le cas où nous saisissons le mode fonctionnalité de la partie 1. Nous avons ensuite, la possibilité de choisir entre 12 différentes options.

```
/Users/macbook/PycharmProjects/pythonProject3/.venv/bin/python /Users/macbook/PycharmProjects/pythonProject3/MyChatBot4/main.py

Bienvenue dans le programme principal!

Menu Principal
1. Fonctionnalités de la partie I
2. Mode Chatbot
0. Quitter

Choisissez un mode (1-2) ou quittez (0): 1
        1. Afficher les noms de fichiers
        2. Extraire les noms de présidents
        3. Nettoyer les fichiers
        4. Afficher la fréquence des termes
        5. Afficher la fréquence inverse du document
        6. Afficher la matrice TF-IDF
        7. Afficher les mots les moins importants
        8. Afficher les mots avec le score TF-IDF le plus élevé
        9. Afficher le(s) mot(s) le(s) plus répété(s) par Chirac
        10. Afficher 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
        11. Afficher le(s) nom(s) président(s) qui a (ont) parlé de la 'Nation'  et celui qui l'a répété le plus de fois
        12. Afficher le(s) noms(s) du (des) président(s) qui a (ont) parlé du climat et de l'écologie
        0. Revenir au menu principal
Veuillez sélectionner une option (0-12) :
```

Lorsque nous choisissons l'option 2, les prénoms des présidents s'affichent sans doublon.

```
Veuillez sélectionner une option (0-12) : 2
//*** Prénoms associer au nom de presidents sans doublon ***//

Giscard dEstaing : Valéry
Mitterand : François
Chirac : Jacque
Sarkozy : Nicolas
Hollande : François
Macron : Emmanuel
```

L'option 6 affiche la matrice TF-IDF.

```
Veuillez sélectionner une option (0-12) : 6
//*** Matrice TF-IDF ***//

[0.2, 0.9, 0.9, 0.9, 0.6, 0.0, 0.9, 0.0, 0.9, 0.43, 0.0, 0.2, 0.0, 0.12, 0.06, 0.43, 0.06, 0.9, 0.6, 0.9, 0.9, 0.43, 0.2, 0.9, 0.3, 0.9, 0.2, 0.0, 0.6, 0.9, 0.
[0.2, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.28, 0.0, 0.2, 0.0, 0.12, 0.23, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.28, 0.41, 0.0, 0.3, 0.0, 1.02, 0.0, 0.0, 0.0, 0.
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.2, 0.0, 0.0, 0.0, 0.0, 0.06, 0.0, 0.0, 0.0, 0.0, 0.0, 0.41, 0.0, 0.9, 0.0, 0.41, 0.0, 0.0, 0.0, 0.0,
[0.61, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.5, 0.17, 0.0, 0.12, 0.0, 0.6, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
[0.2, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.12, 0.06, 0.0, 0.23, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.6, 0.0, 0.0,
[0.41, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.2, 0.0, 0.37, 0.12, 0.43, 0.23, 0.0, 0.0, 0.0, 0.0, 0.0, 0.85, 0.41, 0.0, 0.0, 0.0, 0.2, 0.0, 0.0, 0.0, (
[0.0, 0.0, 0.0, 0.0, 1.2, 0.0, 0.0, 0.0, 0.0, 0.43, 0.0, 0.0, 0.0, 0.25, 0.06, 0.43, 0.12, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.82, 0.0, 1.2, 0.0, 0.41, 0.0, 0.0, 0.0, (
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.2, 0.0, 0.0, 0.06, 0.0, 0.12, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, (
```

L’option 8 affiche les mots avec le score TD-IDF le plus élevé.

```
Veuillez sélectionner une option (0-12) : 8
//*** Mots ayant un score TD-IDF le plus élevé

Mots avec le score TF-IDF le plus élevé : ['voudrais', 'doit', 'faut', 'lobligation', 'pense', 'ville', 'avant', 'nous', 'president']
```

L’option 9 affiche le mot le plus répété dans le discours de Jacques Chirac.

```
Veuillez sélectionner une option (0-12) : 9
//*** Mots les plus répétés par Chirac ***//

Le mot le plus fréquent pour Chirac (hors mots non importants) est : doit
```

L’option 12 affiche les présidents ayant parlé du climat et de l’écologie. Comme mentionné précédemment, l’affichage final n’est pas celui souhaité à la suite d’un disfonctionnement dans la fonction liée.

```
Veuillez sélectionner une option (0-12) : 12
//*** Présidents parlant du climat et de l'écologie ***//

Le(s) président(s) qui a(ont) parlé du climat et/ou de l’écologie : ['Macron.txt']
```

On revient maintenant au menu principal pour accéder au mode 2, soit le mode chatbot. En posant la question “Comment une nation peut-elle prendre soin du climat ?”, le document le plus adéquat pour donner une réponse s’affiche. Cependant, il a un bug: le mot le plus pertinent est toujours “m”. Nous obtenons tous de même les deux réponses souhaitées : la réponse générée et la réponse raffinée.

```
Menu Principal
1. Fonctionnalités de la partie I
2. Mode Chatbot
0. Quitter

Choisissez un mode (1-2) ou quittez (0): 2

Posez votre question:
```

```
Posez votre question: Comment une nation peut-elle prendre soin du climat ?
Document pertinent: Nomination_Giscard dEstaing.txt

Mot le plus pertinent dans la question: m

Réponse générée: messieurs les presidents mesdames mesdemoiselles messieurs de ce jour date une ere nouvelle de la politique francaise
ceci nest pas seulement du monsieur le president du conseilconstitutionnel a la proclamation du resultat que vous venez de rappeler et dont par respect pour la frar
ceci nest pas seulement du aux 13396203 femmes et hommes qui mont fait la confiance de me designer pour devenir le vingtieme president de la republique francaise
ceci est du en realite a la totalite des suffrages du 19 mai 1974
ces suffrages egaux selon la regle democratique quil sagisse de ceux des femmes et des hommes des jeunes et des moins jeunes des travailleurs et des inactifs et qui
jadresse le premier salut du nouveau president de la republique a ceux qui dans cette competition aspiraient a le devenir et qui avaient la capacite de le faire et
francois mitterrand et m jacques chabandelmas
ainsi cest moi qui conduirai le changement mais je ne le conduirai pas seul
si jentends assumer pleinement la tache de president et si jaccepte a cet egard les responsabilites quune telle attitude implique laction a entreprendre associera l
je ne le conduirai pas seul parce que jecoute et que jentends encore limmense rumeur du peuple francais qui nous a demande le changement
nous ferons ce changement avec lui pour lui tel quil est dans son nombre et dans sa diversite et nous le conduirons enparticulier avec sa jeunesse qui porte comme c
messieurs les presidents mesdames mesdemoiselles messieurs voici que souvre le livre du temps avec le vertige de ses pages blanches
ensemble comme un grand peuple uni et fraternel abordons lere nouvelle de la politique francaise.

Réponse raffinée: Après analyse, Messieurs les presidents mesdames mesdemoiselles messieurs de ce jour date une ere nouvelle de la politique francaise
ceci nest pas seulement du monsieur le president du conseilconstitutionnel a la proclamation du resultat que vous venez de rappeler et dont par respect pour la frar
ceci nest pas seulement du aux 13396203 femmes et hommes qui mont fait la confiance de me designer pour devenir le vingtieme president de la republique francaise
ceci est du en realite a la totalite des suffrages du 19 mai 1974
ces suffrages egaux selon la regle democratique quil sagisse de ceux des femmes et des hommes des jeunes et des moins jeunes des travailleurs et des inactifs et qui
jadresse le premier salut du nouveau president de la republique a ceux qui dans cette competition aspiraient a le devenir et qui avaient la capacite de le faire et
francois mitterrand et m jacques chabandelmas
ainsi cest moi qui conduirai le changement mais je ne le conduirai pas seul
si jentends assumer pleinement la tache de president et si jaccepte a cet egard les responsabilites quune telle attitude implique laction a entreprendre associera l
je ne le conduirai pas seul parce que jecoute et que jentends encore limmense rumeur du peuple francais qui nous a demande le changement
nous ferons ce changement avec lui pour lui tel quil est dans son nombre et dans sa diversite et nous le conduirons enparticulier avec sa jeunesse qui porte comme c
messieurs les presidents mesdames mesdemoiselles messieurs voici que souvre le livre du temps avec le vertige de ses pages blanches
```

Conclusion

La réalisation du projet de programmation Python au cours de l'année académique 2023-2024 a été une expérience riche en enseignements. Sur le plan technique, nous avons consolidé nos compétences en programmation Python notamment grâce à la manipualtion de liste 1D, liste 2D, fonctions, fichiers et collections ainsi que leur utilisation dans des cas pratique. Parallèlement, la collaboration au sein de l'équipe a mis en lumière l'importance de la communication et de l'organisation du travail pour atteindre nos objectifs. De plus, malgré notre emploi du temps très chargé et un délai qui nous a semblé plutôt court, le projet a été une opportunité d'apprendre à gérer efficacement notre temps, soulignant l'importance de la planification. En résumé, cette expérience a été bien plus qu'un simple exercice académique, elle a constitué une étape cruciale dans notre développement, nous préparant à relever avec succès des défis futurs, tant sur le plan technique que professionnel.