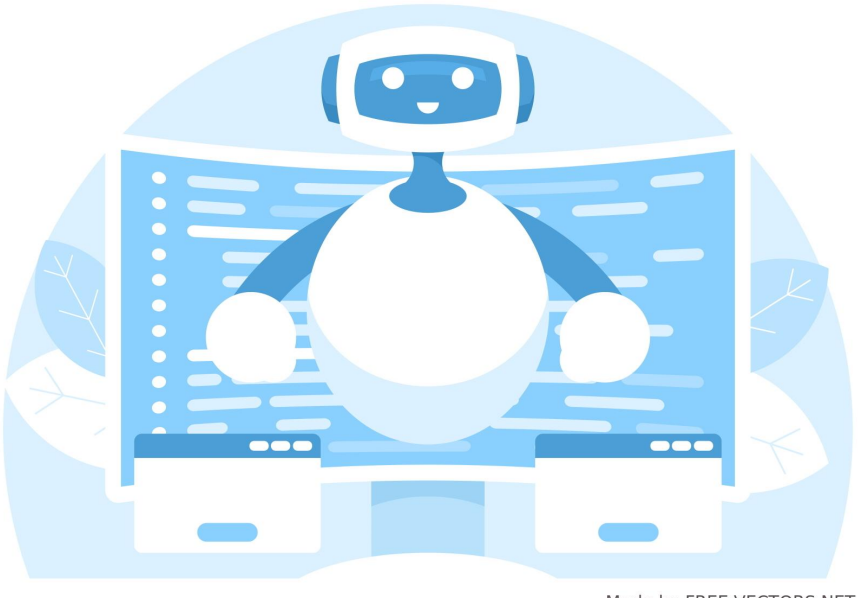
PARIS PANTHÉON-ASSAS UNIVERSITÉ

PROJET PYTHON:
My first ChatBot



Ali IBNOUZAHIR
Lindsay ELLEPO

L1 - Groupe F - Promo 2028

Année 2023

SOMMAIRE

Introduction

Notre projet repose sur une exploration approfondie des concepts fondamentaux du traitement de texte, avec un accent particulier sur les méthodologies employées dans le développement de chatbots et d'intelligences artificielles génératives, à l'image de chatGPT. Évitant la complexité des réseaux de neurones, notre approche se démarque par l'utilisation d'une méthodologie axée sur le nombre d'occurrences des mots, offrant ainsi la possibilité de générer des réponses intelligentes à partir d'un corpus de textes.

Les objectifs de ce projet sont multiples. Tout d'abord, il vise à concrétiser une compréhension approfondie des concepts fondamentaux du traitement de texte. Ensuite, il s'agit de mettre en pratique une méthodologie spécifique à travers le développement d'une application fonctionnelle. Cette dernière aura pour mission de répondre à des questions en se basant sur la fréquence des mots dans le corpus, en suivant des étapes clés telles que le prétraitement des données, la création d'une matrice TF-IDF, la représentation des questions, le calcul de la similarité, et la fourniture de réponses pertinentes.

À travers cette démarche, le projet cherche à analyser les résultats obtenus, à identifier les forces et les éventuelles limites de l'approche adoptée. La communication claire et structurée, manifeste dans le rapport final, permettra de présenter de manière complète et compréhensible les différentes phases du projet, tout en offrant une expérience pratique dans le domaine du traitement de texte et du développement d'applications d'intelligence artificielle.

Présentation fonctionnelle du projet

Objectif du Programme :

Le programme a pour objectif d'analyser un corpus de discours de présidents français et de fournir des fonctionnalités telles que l'extraction de données, le nettoyage de texte, le calcul de la matrice TF-IDF, et la recherche d'informations spécifiques dans le corpus en réponse à des questions.

Fonctionnalités Principales :

Partie 1: Analyse du Corpus.

Extraction de Noms de Fichiers :

- La fonction `list_of_files(directory, extension)` permet d'extraire une liste de noms de fichiers dans un répertoire spécifié avec une extension donnée.

Nettoyage de Texte :

- La fonction `clean_text(text)` convertit les textes en minuscules, supprime la ponctuation, et normalise certains caractères spéciaux.

Calcul de la Matrice TF-IDF :

- La fonction `calculer_tf_idf(dossier)` calcule la matrice TF-IDF pour les documents textuels présents dans un dossier.

Recherche de Mots Répétés par un Président :

- La fonction `mots_plus_repetes_chirac(dossier_corpus)` trouve le mot le plus répété par un président spécifique (Chirac dans cet exemple).

Analyse des Discours sur le Thème de la "Nation" :

- La fonction `president_parle_de_nation(dossier_corpus)` identifie le(s) président(s) qui ont parlé de la "Nation" et affiche celui qui l'a fait le plus fréquemment.

Identification du Premier Président à Parler du Climat et de l'Écologie:
- La fonction `premier_president_climat_ecologie(dossier_corpus)` cherche le premier président à aborder les thèmes du climat et de l'écologie.

Mots Évoqués par Tous les Présidents :
- La fonction `mots_evoques_par_tous(dossier_corpus)` identifie les mots évoqués par tous les présidents dans leurs discours.

Partie 2: Mode Chatbot
Tokenisation de Questions et Recherche de Mots Pertinents :
- Les fonctions `tokenisation(question)` et `recherche_mot(question, dossier_corpus)` permettent de tokeniser une question et de trouver les mots pertinents dans le corpus.

Calcul de Similarité Cosinus entre une Question et le Corpus :
- La fonction `similarite_cosinus(question_vector, document_vector)` mesure la similarité cosinus entre un vecteur de question et les vecteurs des documents du corpus.

Génération de Réponses en Fonction des Résultats :
- Les fonctions `document_plus_pertinent`, `mot_max_tfidf`, et `generer_reponse` sont utilisées pour générer des réponses basées sur la similarité cosinus et les scores TF-IDF.

Affinage de la Réponse en Fonction de la Forme de la Question :
- La fonction `affiner_reponse(question, reponse)` affine la réponse en ajoutant des détails en fonction de la forme de la question.

Utilisation :
L'utilisateur peut poser des questions sur le corpus, et le programme répond en extrayant des informations pertinentes à partir des discours des présidents français. Le nettoyage de texte, le calcul de la matrice TF-IDF, et la recherche de mots spécifiques permettent d'analyser et de comprendre le contenu des discours.

Présentation technique du projet

Description des Principaux Algorithmes :
- **`list_of_files(directory, extension)` :** Utilise une boucle pour parcourir le répertoire spécifié, filtrant les fichiers avec l'extension donnée et les ajoutant à une liste.
- **`extraire(nom_fichier)` :** Lit le contenu du fichier, divise le nom du fichier en parties, et retourne le nom du président.
- **`associer(noms_presidents)` :** Utiliser un dictionnaire pour associer les noms de président à leurs prénoms correspondants.
- **`convertir(car)` et `ponctuation(car)` :** Implémentation simple pour convertir les caractères en minuscules et vérifier la ponctuation.
- **`clean_text(text)` :** Convertit le texte en minuscules, remplace certains caractères spéciaux, supprime la ponctuation et les espaces en double.
- **`process_file(file_path, output_folder)` :** Lit le fichier, nettoie le texte en utilisant `clean_text`, puis écrit le texte nettoyé dans un nouveau fichier.

Justification des Choix de Structures de Données :
- **Listes :** Utilisées pour stocker les noms de fichiers, car l'ordre d'apparition est important.
- **Dictionnaire (prenoms_presidents) :** Permet une recherche rapide et

efficace du prénom associé à chaque nom de président.
- **Set (dans la fonction `recherche_mot`) :** Utilisé pour stocker les mots trouvés dans le corpus sans doublons, facilitant la recherche d'occurrences.

Difficultés Techniques et Solutions Apportées :
- **Encodage des fichiers :** Les fichiers peuvent avoir différents encodages. Utilisation du paramètre d'encodage lors de la lecture/écriture des fichiers pour éviter des problèmes d'encodage.
- **Traitement des caractères spéciaux :**Certains caractères spéciaux peuvent causer des problèmes lors de l'analyse du texte. Utilisation de la fonction `clean_text` pour remplacer les caractères spéciaux par des équivalents normaux.
- **Gestion des Extensions des Fichiers :** Les fonctions qui manipulent les noms de fichiers doivent prendre en compte différentes extensions. Utilisation de la fonction `endswith` pour filtrer les fichiers par extension.

- Variables => ‘dossier_corpus’ : Pour accéder au dossier "cleaned" du projet Python, nous avons rencontré des difficultés car la variables désiré n’était pas implémenter. Alors ,nous avons modifié ‘dossier_corpus’ pour qu’elle utilise ‘os.path.dirname(os.path.realpath(__file__))’ pour obtenir le répertoire du script Python en cours d'exécution, puis utilise os.path.join() pour construire le chemin complet vers le dossier "cleaned".

- **variables** => **`scores_idf`**: cette variable n'était pas passée en tant qu'argument à la fonction(TFIDFQuestion), alors nous l’avons passée en argument à la fonction pour que cela fonctionne correctement.

Cette approche garantit une gestion efficace des fichiers, une association rapide des prénoms aux noms des présidents et un nettoyage de texte précis pour une analyse ultérieure. Les structures de données ont été choisies en fonction de leur pertinence pour chaque tâche spécifique, et des

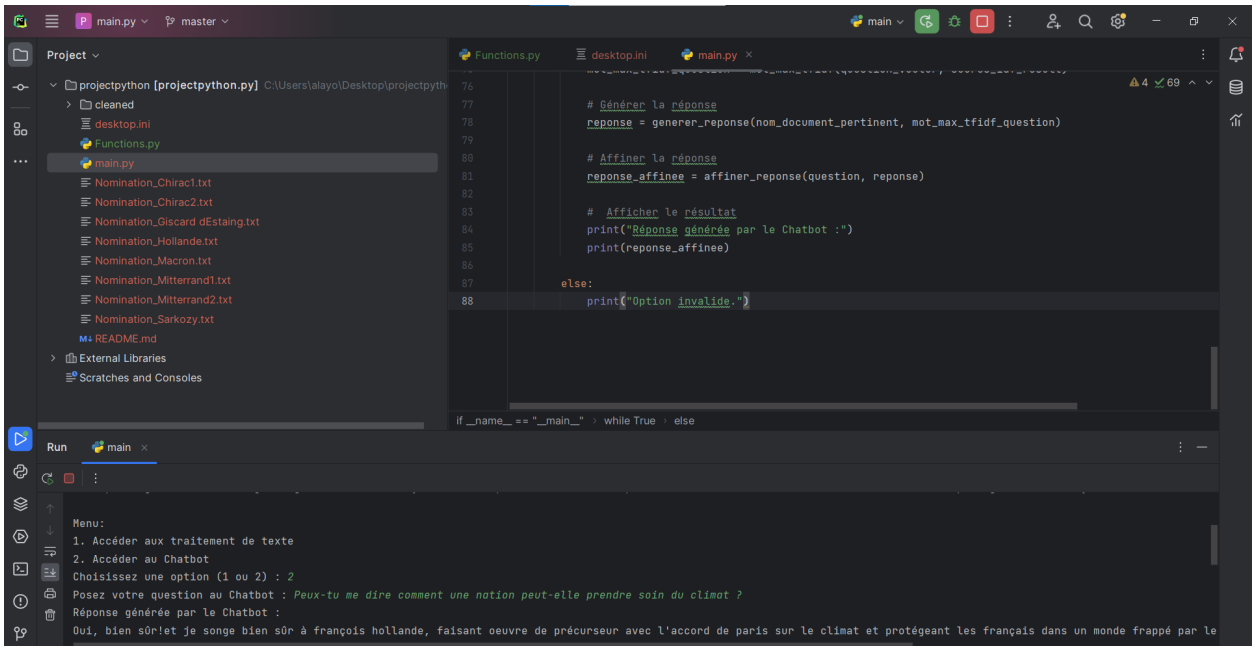
solutions ont été apportées pour résoudre les problèmes techniques potentiels.

Conclusion

En conclusion, notre projet a accompli avec succès l'exploration des concepts clés du traitement de texte, en mettant l'accent sur le développement de chatbots et d'intelligences artificielles génératives. En adoptant une approche novatrice axée sur la fréquence des mots plutôt que sur la complexité des réseaux de neurones, nous avons créé une application fonctionnelle offrant des réponses intelligentes à partir d'un corpus de discours présidentiels français.

Les objectifs variés de notre projet, de la compréhension approfondie des concepts du traitement de texte à la mise en œuvre pratique d'une méthodologie spécifique, ont été atteints. La présentation fonctionnelle du programme a démontré sa capacité à extraire des informations pertinentes du corpus, à effectuer le nettoyage du texte, à calculer la matrice TF-IDF, et à répondre de manière intelligente à des questions spécifiques. De plus, le mode chatbot offre une expérience interactive, permettant aux utilisateurs d'explorer le corpus et d'obtenir des réponses contextualisées.En résumé, ce projet a enrichi notre compréhension du traitement de texte et du développement d'applications d'intelligence artificielle.En résumé, ce projet a enrichi notre compréhension du traitement de texte et du développement d'applications d'intelligence artificielle.

Résultat :


main.py
master
main
Project
projectpython [projectpython.py] C:\Users\alayo\Desktop\projectpyth
cleaned
desktop.ini
Functions.py
main.py
Nomination_Chirac1.txt
Nomination_Chirac2.txt
Nomination_Giscard dEstaing.txt
Nomination_Hollande.txt
Nomination_Macron.txt
Nomination_Mitterrand1.txt
Nomination_Mitterrand2.txt
Nomination_Sarkozy.txt
README.md
External Libraries
Scratches and Consoles
Functions.py
desktop.ini
main.py
# Générer la réponse
reponse = generer_reponse(nom_document_pertinent, mot_max_tfidf_question)
# Affiner la réponse
reponse_affinee = affiner_reponse(question, reponse)
# Afficher le résultat
print("Réponse générée par le Chatbot :")
print(reponse_affinee)
else:
print("Option invalide.")
if __name__ == "__main__" › while True › else
Run
main
Menu:
1. Accéder aux traitement de texte
2. Accéder au Chatbot
Choisissez une option (1 ou 2) : 2
Posez votre question au Chatbot : Peux-tu me dire comment une nation peut-elle prendre soin du climat ?
Réponse générée par le Chatbot :
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