PARIS PANTHÉON-ASSAS UNIVERSITÉ

RAPPORT DE STAGE

Projet Python S1 : “My First ChatBot”

Rendu : 19/12

Roch Triomphe / Rayane Karaouzene
Promotion 2028

Lien direct vers le repository Git

Année scolaire 2023-2024

=

SOMMAIRE

I- PARTIE 1 : My First ChatBot !

A. Présentation sommaire du projet

Le projet que nous avons eu à faire ce semestre dans le module Python portait sur l'analyse et la manipulation de textes. Plus précisément, il s'agissait d'une manipulation plutôt basiques des différentes méthodes de traitement de texte afin de fournir une réponse à une question donnée.

Ainsi, nous avons pu fournir, en nous basant sur la fréquence d'un terme dans le document, une réponse convenable à une question posée.

Cette manipulation pourtant très basique et très simple permet néanmoins d'apréhender dans une moindre mesure le fonctionnement de LLM et de réseaux neuronaux tels que ChatGPT, Gemini et consorts. Cependant évidemment de façon bien moins complexe.

B. Présentation technique

Le projet repose intégralement sur l'utilisation de matrices TF-IDF (**T**erm **F**requency-**I**nverse **D**ocument **F**requency), permettant ainsi de travailler efficacement sur les documents présentés.

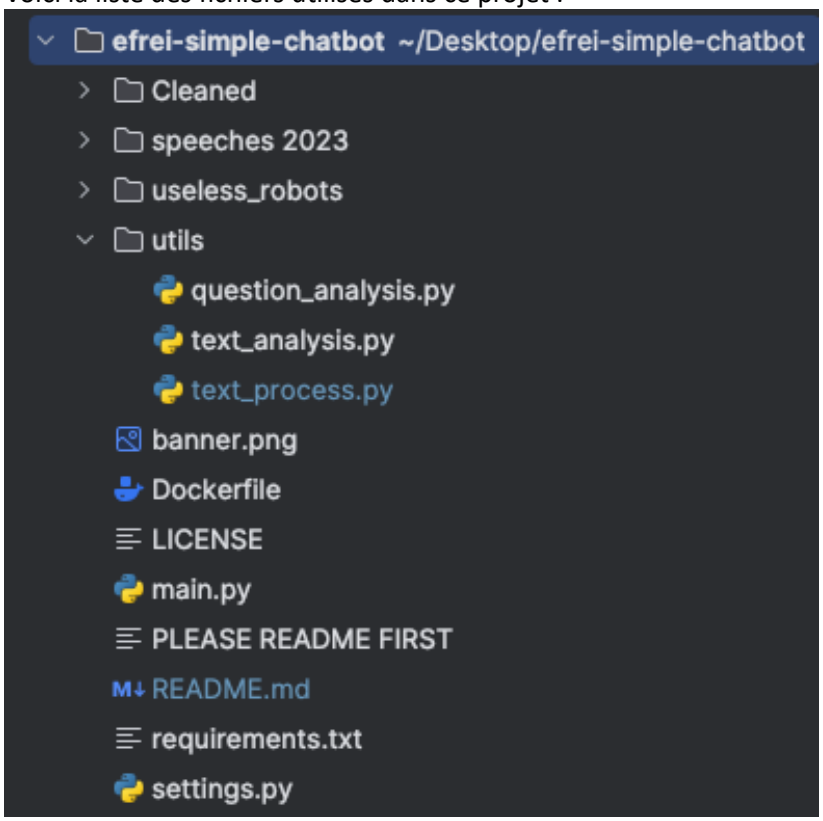
Tout ce système repose, du moins de notre côté, sur une grande utilisation des dictionnaires afin de permettre une plus grande flexibilité et de meilleurs résultats.

Pour expliquer simplement, lorsqu'une question est posée, le code détermine le mot le plus important de la question => vérifie sa présence puis son importance dans le texte => Renvoie un résultat, c'est-à-dire une réponse à la question posée.

D'un point de vue purement technique, notre code est assez concis (+- 600 lignes probablement optimisables) et utilise quelques fichiers :

- ⇨ Fichier main, qui gère le menu et l'execution des fonctions sous-jacentes
- ⇨ Un dossier utils dans lequel on retrouve les fonctions principales :
  - fichier text_analysis afin d'analyser le texte
  - fichier text_process afin d'optimiser et traiter le texte
  - fichier question_analysis afin de traiter et d'optimiser la question.
- ⇨ Un fichier settings contentant les réglages pouvant être modifiés.
- ⇨ Et enfin un certain nombres de fichier texte (les texte fournis), les fichiers readme etc.

Voici la liste des fichiers utilisés dans ce projet :



Les principales fonctions sont :

question analysis = texte => matrice tf-idf => comparaison => renvoi résultat
tf_calculator = reçoit un dict => calcule le tf de chaque mot => renvoie un dict
idf_calculator = reçoit un dict => calcule le idf de chaque mot => renvoie un dict
tf-idf calculator = calcule la matrice tf-idf et renvoie un dictionnaire.

Les dictionnaires nous offrant une plus grande flexibilité, et surtout la possibilité d'attribuer à un mot une valeur, nous avons donc opté pour ceux-ci.

Cependant, il a fallu nous accorder afin de produire les mêmes formats de sortie des valeurs afin que la coordination des deux soit bien effectuée.

C. Présentation fonctionnelle

Nous avons fait le projet (du moins les parties 1 et 2) dans son intégralité, c'est-à-dire que nous nous sommes assurés, avant d'aller plus loin (pourquoi pas une hypothétique partie 3 par exemple), que le programme fonctionnait parfaitement et surtout répondait à toutes les situations et à tous les cas de figure qui pourraient lui être proposés. Ainsi, la partie « chatbot » fonctionne ainsi que les questions de base telles que « trouver le mot avec le TF-IDF le plus faible » ou encore « quel président a dit le plus de fois nation ».

Ces fonctions, basiques, sont réalisées, et bien réalisées. Cependant, étant donné que la partie 3 était en bonus, et surtout que nous n'avions pas forcément le temps, nous n'avons pas jugé opportun de la traiter.

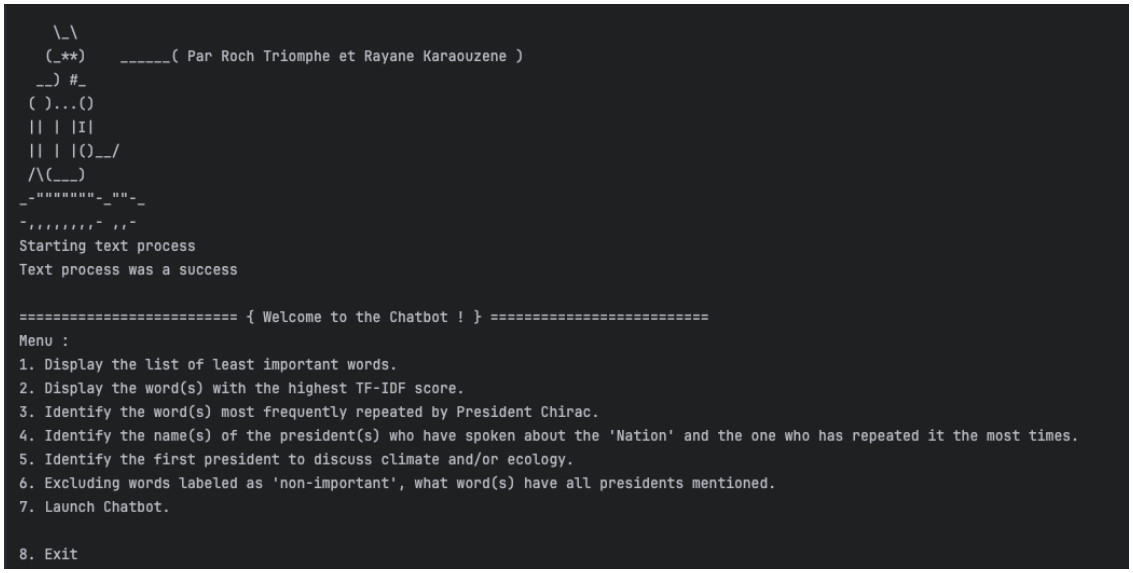
Nous n'avons donc pas rajouté de fonctionnalité en plus de ce qui était demandé initialement. (à part l'ajout d'un Dockerfile et la création d'un readme propre.) Cependant, une nuit, pris d'insomnie, Roch a décidé de modifier le code du chatbot sur une branche secondaire appelée BETA-DEVELOPEMENT (que vous pourrez retrouver sur github) avec une feature en plus. Si le chatbot ne trouve aucune correspondance (ça peut arriver, si vous lui donnez un mot absent du corpus fourni auparavant…) au lieu de simplement vous dire qu'il n'a pas trouvé, il vous proposera une réponse tout de même, et effectuera une requête à OpenAI afin de fournir coûte que coûte une réponse.

Cependant, n'étant pas le sujet initial, nous n'avons pas jugé opportun de mettre cette feature dans le code de base, mais nous avons préféré garde celle-ci à part.

(nous avons également ajouté de l'ASCIIart car c'est inutile mais joli… nous étions fatigués à ce moment-là.)

Vous trouverez ci-joint des captures d'écran des fonctionnalités présentes :

Menu :



Le mot avec le TF-IDF le plus élevé :

```
The words with highest score :
doit,
```

Le mot le plus répété par Chirac :

```
3
The most repeated words by Chirac are :
de,
```

Les présidents qui ont parlé de Nation, et celui qui l'a répété le plus de fois (ainsi que le nombre d'occurrences) :

```
4
Presidents mentioning nation :
Chirac, Mitterrand, Sarkozy, Hollande, Macron,
And it was repeated 15 times by Chirac
```

Le premier président à mentionner l'écologie, et le premier pour le climat :

```
5
The first president to mention the word éco is: Mitterrand
The first president to mention the word climat is: Sarkozy
```

Enfin, le chatbot en action.

```
What is your question? : Parle-moi de la nation
Avec plaisir ! il ne peut pas y avoir des sacrifices pour les uns, toujours plus nombreux, et des privilèges pour les autres, sans cesse moins nombreux.
Would you like to go back to menu, or ask another question ?
1. Go back to menu.
2. Ask another question.
```

II- PARTIE 2 : Seuls face au monde.

Évidemment, nous avons rencontré des problèmes durant la création de ce programme. En voici un petit panorama, ainsi que les solutions mises en place afin de les contourner, voir les surmonter.

A. La méthode TF-IDF

Un des obstacles a été l'apprentissage et la compréhension de ce qu'était le TF-IDF, et Son utilité qui nous paraissait somme toute obscure. La compréhension de cette notion pourtant cruciale a donc été un frein. Également, il a fallu comprendre la méthode de calcul des vecteurs et nous replonger dans ce chapitre de mathématiques, ce qui n'a pas forcément été une partie de plaisir.

B. L'utilisation de GIT

Contrairement à ce que l'on aurait pu penser, l'utilisation de Git ne nous a absolument pas gênés, et au contraire, a facilité notre travail et surtout la collaboration. Certes au début il a fallu mettre cela en place en expliquant sommairement à mon camarade comment utiliser ce logiciel, mais une fois les routines mises en place, il nous paraissait normal de communiquer notre code et notre avancement par ce biais. Les messages de commit étaient clairs, les best practices plus ou moins respectées, bref ce fut une très bonne expérience. De fait, notre repo totalise pas moins de 78 commits tout au long du projet, ce qui est, je trouve, franchement agréable lorsqu'on travaille en groupe.

C. Répartition des tâches

Rome avait son empereur, et un village a son maire. Tout projet ou institution quel qu'il Se doit d'avoir un leader afin de mener cette entreprise à bien. De fait, dans ce projet, instinctivement Rayane a pris la tête du projet, chose qui ne me déplaisait pas, loin de là. Nous avions donc un chef et une direction, tout ne pouvait que bien se passer. Après cela, nous avons codé plus ou moins équitablement afin de parvenir au rendu final. (Je pense tout de même que Rayane a écrit plus de code que moi…)

Cela dit, nous avons, d'un accord tacite, désigné deux responsables : Rayane, le responsable développement pur, et moi, le responsable git, qui me suis chargé de veiller sur le repo, et surtout donner les directives concernant les pull, merge etc.

CONCLUSION

Je pense que ce projet nous a beaucoup appris, peut-être pas forcément d'un point de vue technique étant donné que ce projet n'était pas non plus extrêmement dur (même si un peu, évidemment) mais il nous a surtout appris énormément dans le management d'un projet, et tout ce qui vient avec cela. Nous n'avons pas énormément été pressés par le temps, étant donné que nous avons choisi de nous focaliser sur la qualité du travail rendu, et non sa quantité. Avec cettre « contrainte » de temps en moins, nous avons probablement été plus efficaces, et avons mieux travaillé.