Polytech Paris-Sud

Traitement Automatique du Langage

Création d'un Chatbot Weeaboo

Weeabot

Auteurs :
Mohamed BELDI
Alban DESCOTTES
Camille FOSSIER

Superviseur :
Rachel BAWDEN

5 mai 2018

1 Introduction

1.1 Objectif du projet

L'objectif de ce projet a été la réalisation d'un chatbot fonctionnel, ainsi, nous avons décidé de réaliser un chatbot Weeaboo nommé Weeabot. Le mot Weeaboo est un mot désignant les admirateurs fous de tout ce qui concerne le Japon.
Cependant, nous avons choisi d'orienter l'aspect Weeaboo de notre bot sur les mangas, qui sont les bandes dessinées japonaises. Ainsi, le Weeabot sera susceptible de discuter avec l'utilisateur à propos de certains mangas et même de donner son avis dessus.
Cette idée nous est venu du fait qu'un membre de notre groupe aime beaucoup les mangas et nous nous sommes dit que cela pouvait être un sujet avec énormément de matière à travailler.

1.2 Mise en oeuvre du projet

Les modes 1 et 2 étant des modes génériques, nous allons nous pencher ici plus particulièrement sur le mode 3 qui est le mode le plus original et travaillé. Afin de nous simplifier la tâche concernant la base de données du Weeabot, nous avons utilisé la base de données du site `https://myanimelist.net/` qui est un site recensant la majorité des mangas et dessin animés japonais.
Il a été décidé que le Weeabot parlerait anglais étant donné que c'est une langue possédant des règles de grammaire et de conjugaison moins compliquées que le français, son implémentation serait donc plus simple et éviterait des erreurs syntaxiques.

2 Fonctionnement du Weeabot

2.1 Capacités du Weeabot

Le Weeabot possède 3 modes, le mode 1 étant le plus simple et le 3 le plus avancé. Nous avons aussi simulé le fait qu'une personne réelle soit en train de nous répondre en tapant sur son clavier à l'aide d'intervalles aléatoires.

- **Mode 1 :** Ce mode correspond à celui décrit dans l'énoncé, il répond seulement des onomatopées ou acquiesce à ce que l'utilisateur dit, sans y prêter attention.
  Il fait cependant attention à ne pas utiliser la même onomatopée ou le même mot à la suite.

- **Mode 2 :** Dans ce mode il existe une meilleure interaction entre l'utilisateur et le Weeabot, celui ci possède un panel de phrases prédéfinies correspondant à plusieurs thèmes. Il va piocher aléatoirement dans ses réponses possibles en prenant garde à ne pas choisir la même phrase deux fois de suite pour le même thème.
  Nous avons essayé de faire en sorte qu'il ait certaines réactions en fonction du comportement de l'utilisateur, par exemple si l'utilisateur répète plusieurs fois de suite la même phrase, il lui fera remarquer.

- **Mode 3 :** C'est dans ce mode que le Weeabot révèle son côté Weeaboo en parlant essentiellement de mangas. À l'aide de sa base de données qui contient plusieurs centaines de mangas et leurs informations provenant de `https://myanimelist.net/`, il peut nous renseigner sur le classement du manga sur ce site, son genre, le nombre de personnes l'ayant en favori, la note que lui ont mis les utilisateurs du site, son nombre de chapitres, sa date de publication, son ou ses auteurs et enfin ses personnages.
  Nous pouvons demander plusieurs informations à la suite sur le même manga grâce au fait qu'il garde toujours le dernier manga en mémoire. Ainsi nous avons la possibilité de comparer des mangas de demander le rang de l'un et de dire `what about [other manga title]?` afin que l'on nous donne l'information pour l'autre manga.
  Il est aussi possible de donner son avis sur un manga en disant si on l'aime ou non, et le Weeabot nous donnera son avis sur ce manga en fonction de critères se basant sur la popularité et la moyenne des mangas de sa base de données.
  Par exemple si nous lui disons `I like Naruto`, il nous répondra qu'il est d'accord avec nous en nous

donnant un critère aléatoire qu'il aime bien (*sa longueur, son rang, sa popularité...*).
Si nous voulons lire un type de manga en particulier nous pouvons demander conseil au Weeabot et il choisira le manga le mieux classé correspondant au genre que nous lui avons demandé.
Bien sûr nous pouvons demander un manga correspondant à plusieurs genres et s'il n'en trouve aucun dans sa liste, il nous le fera savoir. Par exemple nous pouvons lui dire `give me a manga with adventure genre` et il nous donnera le titre d'un manga par contre, si nous lui disons `give me a manga with adventure and action genre` il nous donnera le titre d'un autre manga.
Bien sûr s'il ne trouve aucune réponse possible il essaye de répondre en utilisant le mode 2 et si il ne trouve rien à dire il utilise une réponse du mode 1.

Voici quelques phrases que vous pouvez taper dans le mode 3 du Weeabot pour tester :

— `give me a drama manga`
— `I like Naruto`
— `which manga is the best ?`
— `who is the author ?`
— `can you talk about Berserk ?`
— `give me a drama and police manga`
— `which manga is the best ?`

2.2 Structure du Weeabot

Afin d'avoir une base de données de mangas sur laquelle travailler pour le mode 3, nous avons tout d'abord voulu nous pencher sur le module `BeautifulSoup`, mais après quelques recherches nous nous sommes rendus compte qu'il existait un module nommé `myanimelist` qui s'occupait d'extraire les informations de `myanimelist.net` pour nous, il nous a donc suffit de nous documenter sur ce module et d'enregistrer les données localement à l'aide de `pickle` afin de ne pas avoir à les retélécharger.

La structure du mode 3 du Weeabot est simple :

L'analyse de la structure de la phrase se fait en parsant plusieurs fois d'affilée les mots qui la composent afin d'identifier, si on parle d'un manga précis, d'un genre de manga précis, si on spécifie une action particulière, etc. Nous avons donc pour chacun de ces thèmes à vérifier un vocabulaire qui va être parcouru, afin de compter les mots qui sont le plus présents dans la phrase et essayer de comprendre le souhait de l'utilisateur. Plusieurs catégories peuvent contenir les mêmes mots de détection, nous avons donc mis en place un compteur qui indique lequel est le plus probable.
D'abord, nous cherchons donc un nom de manga. Si on en trouve un, on actualise les deux mangas actuels de façon à garder en mémoire le manga dont on parle.
Exemple : Si auparavant nous avions parlé de Naruto et de Dragon Ball, puis que nous mentionnons Berserk, Dragon Ball sera supprimé de la liste, Naruto passera en deuxième position et Berserk en première. Nous avons choisi d'en garder deux de façon à pouvoir effectuer des comparaisons entre les deux en mémoire.

Ensuite, si on trouve un ou plusieurs genres, une fonction s'occupe de trouver un manga qui contient l'intégralité des genres mentionnés par l'utilisateur. S'il n'en trouve pas il cherche un manga qui en contient au moins un. Puis il garde en mémoire le résultat qu'il a fourni.
Exemple : Si je demande un manga d'action et d'aventure, le Weeabot me proposera par exemple Naruto, mais si je lui redemande un manga uniquement d'action, il ne me proposera plus Naruto qu'il m'a déjà donné avant.
De plus si je demande un manga qui se passe dans l'espace, d'amour et qui soit en plus d'action et de comédie, il est probable que le Weeabot ne me fournisse qu'un manga se passant dans l'espace.

Pour détecter l'action souhaitée par l'utilisateur on fonctionne à nouveau sur un système de mots clefs qui permettent de savoir si on veut le meilleur manga dans une catégorie, comparer deux mangas, ou juste obtenir des informations. Une fois cela détecté, nous avons une simple fonction qui génère des phrases en fonction de ce que recherche l'utilisateur en concaténant des structures de phrases prédéfinies avec les champs du manga correspondant.

Une de ces actions est de demande au bot s'il apprécie le manga ou nom et s'il est d'accord avec nous. Il se base sur la note moyenne des mangas présents dans la base de données, et prend toujours parti de la population générale.
Exemple : Si un manga est bien noté mais que l'utilisateur dit ne pas l'aimer, alors le Weeabot dira qu'il n'est pas d'accord avec nous et se justifiera.
La justification du Weeabot quant à la qualité d'un manga se fait en partie par des critères chiffrés. Pour le manga à évaluer il regarde celui qui s'éloigne le plus de la moyenne générale pour fonder son argument. Pour simplifier il nous dit par quoi le manga se démarque.

Exemple : Si on considère un manga dont la note est très proche de la note moyenne globale, dont le nombre de chapitres est proche du nombre moyen global, etc. Mais que ce manga est très vieux comparé aux autres, alors il dira qu'il apprécie le manga car il est plus ancien que les autres.
Il y a également un facteur aléatoire qui fait que le bot peut se justifier par des critères différents, par exemple qu'il aime l'auteur, ou le titre du manga, etc. Ce facteur aléatoire a été incorporé à plusieurs endroits pour diversifier la conversation, par exemple pour éviter que le bot ne répète constamment la même information quand l'utilisateur ne montre pas de volonté précise d'informations dans une phrase, et cela en appliquant au hasard des thèmes de discussion, ou même en appelant le système de réponse du mode 2.

2.3 Limites du Weeabot

Malheureusement, le Weeabot n'est pas parfait... Le vocabulaire utilisé pour détecter le souhait de l'utilisateur n'est peut-être pas assez développé dans le mode 2 et il y a probablement des risques de collisions de vocabulaire de déclenchement pour ces modes. De plus il arrive que la réponse à la question soit `None`, ce qui provient de cas (rares) de non information concernant un sujet précis sur un certain manga, et qui n'ont pas été traités lors de la rédaction des phrases réponses.
Par exemple on ne connait pas le nombre de chapitres d'un manga qui n'est pas fini, par conséquent quand on demandera son nombre de chapitres, le Weeabot nous répondra `None`, et même en essayant d'y remédier nous n'y sommes pas encore parvenu.

Il subsiste malheureusement quelques bugs qui concerneraient la liaison entre les différents modes qui peut ne pas fonctionner dans certains cas. Il se trouve que notre manière de détecter le vocabulaire dans le mode 3 n'est pas forcément la meilleure et on a donc du quelque fois du mal à obtenir la réponse à la question que l'on a posé si celle ci est mal formulée.
Un autre problème que nous avons détecter mais pour lequel nous n'avons pas de solution est le fait que le nom d'un manga contienne le début du nom d'un autre manga, c'est le cas entre `One` et `One Piece`, ainsi dans certains cas le Weeabot confondra les deux.

3 Bilan du projet

3.1 Répartition des tâches

Concernant la distribution des tâches nous avons essayé de tous travailler de manière équitable. Le sujet a rapidement été trouvé donc nous avons rapidement pu nous mettre au travail.
Le taux de participation sur le github n'est pas forcément représentatif du travail fourni par chacun (bien que Camille ait fourni une grande quantité de travail), en effet, nous avons plusieurs fois codé à plusieurs en nous échangeant des bouts de code et malgré le fait que Mohamed utilisait le shell de son ordinateur pour commit, il n'apparait pas parmis les contributeurs.

Pour résumer :

- **Camille :** Réalisation des modes 1 et 2 et de plusieurs fonctions du mode 3.

- **Alban :** Réalisation de plusieurs fonctions des mode 2 et 3.
- **Mohamed :** Plusieurs recherches et fonctions pour le mode 3 ainsi que le téléchargement et le stockage des données.

Nous avons tous contribué à la rédaction de ce rapport.

3.2 Améliorations à apporter

La première amélioration à apporter serait évidemment plus de stabilité ainsi que corriger tous les bugs enoncés précédemment. Il serait intéressant d'agrandir la base de données de manga et de faire en sorte que celle-ci soit connectée en temps réel au site web afin d'avoir toutes les informations sur le manga mises à jour, ou encore l'étendre au domaine de l'animation japonaise.
Ainsi, on aurait accès au nombre de chapitres d'un manga en cours de parution. Concernant le bot, il faudrait lui donner un peu plus de vie, en lui permettant d'avoir plus de phrases de réponses, qu'il pourrait construire lui-même, dans le but d'avoir une interaction qui nous semble plus réelle.
Il serait aussi original que le bot utilise certaines mimiques et expressions propres aux weeaboos, qui adore inclure des mots japonais dans leur phrase, cela lui donnerait une certaine personnalité et renforcerait son identité de Weeabot.
Enfin, il serait judicieux d'inclure du machine learning à notre programme afin d'apprendre à la machine à construire des phrases, ou même à mieux analyser celles de l'utilisateur.

3.3 Conclusion

Nous avons trouver ce projet très intéressant et enrichissant, en effet, il n'était pas orienté seulement sur le code, il y avait une partie linguistique qui était très importante.
Même si nous avons eu quelques difficultés et échecs dans la réalisation de ce projet, nous pensons l'avoir mené à bien et avoir beaucoup appris.
Ainsi, si nous devions le recommencer, nous aurions une meilleure idée de comment le réaliser, où commencer mais aussi, comment le structurer.