ET4-info

INTRODUCTION AU TAL

POLYTECH PARIS-SUD

RAPPORT DE PROJET

1. Spécifications

- Sujet : Chatbot
- Objectif : Créer un bot capable de comprendre contextuellement l'interlocuteur. Afin de rendre le projet réalisable dans les temps on se focalise sur des données liées à des personnalités populaires. Sujets abordables concernant la personnalité : - Famille (parents, enfants, siblings) - Age (date de naissance, âge actuel) - Lieu de naissance - Mensurations (poids, taille) - Lieu de résidence
- Conversation type :
  - Exemple 1 :
    Humain : Salut ! Je cherche quelqu'un qui a les cheveux blonds. Vous avez une idée ?
    Bot : Bien-sûr, pouvez-vous être plus précis ?
    Humain : Cette personne a entre 30 et 50 ans et est un garçon.
    Bot : Je pense à Chris Evans et Ryan Gosling, la personne que vous cherchez est-elle parmi elles ?
    Humain : Oui c'était Chris Evans, merci !
    Bot : De rien.
  - Exemple 2 :
    Humain : Bonjour, savez vous qui est Obama ?
    Bot : Vous parlez peut-être de Barrack Obama ?
    Humain : Oui !
    Humain : Combien d'enfants a-t-il ?
    Bot : Il a 2 enfants.
    Humain : D'accord et il a quel âge ?
    Bot : il a 55 ans.

Humain : Et sa fille la plus âgée mesure-t-elle plus de 1,70 mètres ?
Bot : Oui, elle mesure 1,74 mètres.
Humain : D'accord merci, au revoir !
Bot : Au revoir

— Modules :
  — Analyse lexicale
  — Recherche de matchs selon des critères donnés
  — Création du corpus
  — Synthèse lexicale
  — Compréhension contextuelle (garder en mémoire le reste de la conversation)

2. **Utilisation de la librairie NTLK (Natural Language Toolkit)**

Afin de faciliter notre travail et de nous permettre de nous focaliser sur l'essentiel on a décidé d'utiliser la librairie NTLK. Celle-ci nous permet de travailler des phrases et d'en ressortir des informations qui nous intéressent.

Dans le cadre de notre projet on s'est limité à une utilisation basique de la librairie. En effet celle-ci pourrait nous permettre de créer des arbres donnant la logique d'une phrase avec, pour chaque mot, sa signification grammaticale. De plus il existe des packages plus ou moins variés permettant toute sorte d'opération comme par hasard comprendre si une phrase est plutôt positive ou négative. Une limitation de la librairie est le langage utilisé. Ici on choisi de travailler sur du français, ainsi il y a un certain nombre de possibilités qui nous sont plus difficilement accessibles.

On a décidé d'utiliser deux fonctions de cette librairie. La première est $sent_tokenize()$, une fonction permettant de séparer un texte composé de phrases en plusieurs phrases, ceci facilitant la compréhension de ce texte. La deuxième est $word_tokenize()$, cette fois-ci permettant de créer correctement tous les tokens de la phrase. C'est tout.

Une autre piste que l'on a évalué est la possibilité de taguer chaque mot en son sens grammatical (verbe, nom commun,

complément d'objet, etc.). Par défaut nltk ne le permet pas avec ses packages pour la langue française, cependant en creusant le sujet on a pu découvrir un travail réalisé à Stanford avec le Stanford Log-linear Part-Of-Speech Tagger, un bout de programme écrit en Java permettant l'assignation du sens grammatical des mots en français. De plus Nitin Madnani a écrit une interface python-java pour ce bout de programme permettant de faire le lien avec NTLK. Ainsi, d'une manière détournée, on aurait pu tagguer automatiquement les mots et, pourquoi pas, en créer des arbres. Cependant on a souhaité ne pas se mâcher tout le travail même si cela reste très intéressant du point de vue praticité. Donc on s'est limité aux deux premières fonctions citées plus haut.

Tableau des Tags <-> signification <-> exemples)

Tag	Signification	Examples
ADJ	adjectif	nouveau, bien, haut, spécial, grand, local
ADP	préposition	sur, de, avec, dans, sous
ADV	vraiment	déjà, encore, tôt, maintenant, grand, local
CONJ	conjonction	et, ou, mais, si, grand, tant que, bien que
DET	determinant, article	le, un, quelque, plupart, chaque, aucun
NOUN	nom	année, maison, coûts, temps, Afrique
NUM	numeros	vingt-quatre, quatrième, 1991, 14 :24
PRT	article	à, sur, dehors, dessus, avec
PRON	pronom	il, leur, sa, son, mon, nous
VERB	verbe	est, dire, donné, jouer, voudrait
.	ponctuation	., ; !
X	autre	dunno, esprit, university, gr8
PN	nom propre	Obama, Valentin, Paris, Mulhouse
IADJ	adjectif interrogatif	c'est, quel, qui, quand, quoi

3. **Contribution de chacun des membres du groupe**
   Théo :

— codage d'un premier tokenisateur (rendu inutile par l'utilisation de la librairie NLTK)
— Recherche de solutions accessibles à notre niveaux afin de générer les réponses du bot de manière contextuelle
— Complétion du dictionnaire d'équivalence token <-> tag

Raihane :

— Recherche d'informations sur les célébrités et mise en place du fichier contenant la base de données sur ces personnages.
— Recherche, compréhension et apprentissage de LaTeX
— Ecriture du rapport en LaTeX

Benoît :

— Moteur du programme ChatBotEngine
— Tokenisation

4. **Objectifs réalisés**
   — Tokenisation
   — Traduction en TAG
   — Fonctions de recherche dans la base de données

5. **Limitations du programme**
   — Génération de réponses contextuelles