Large language Models (ADIA94)

Département : Big Data & Machine Learning
Disciplines : Sciences des données et intelligence artificielle
Enseignant : Stefani EL KALAMOUNI - stefani.el-kalamouni@efrei.fr

---

TP2 - Instructions

1. Installez et importez les bibliothèques suivantes : `Nltk`, `numpy`, `random`, `string`, `bs4`, `re`, `urllib.request`, `TfidfVectorizer` de `sklearn.feature_extraction.text`, `cosine_similarity` de `sklearn.metrics.pairwise`.
2. Importez les données à partir du lien suivant : https://fr.wikipedia.org/wiki/Intelligence_artificielle et lisez-les.
3. Utilisez BeautifulSoup pour convertir les données au format `lxml`.
4. Chaque paragraphe sur Wikipédia est divisé par `<p>`. Utilisez `find_all()` pour obtenir tous les paragraphes.
5. Créez une fonction pour extraire le texte de vos paragraphes.
6. Assurez-vous que toutes vos données sont en minuscules.
7. Supprimez les balises, symboles et chiffres, et remplacez-les par des espaces.
8. **Tokenisation**
    - a. Tokenisez vos phrases.
    - b. Tokenisez vos mots.
9. **Lemmatisation**
    - a. Initialisez votre lemmatiseur.
    - b. Créez une fonction pour effectuer la lemmatisation.
    - c. Supprimez la ponctuation.
10. Définissez une fonction unique qui traite votre document en effectuant la lemmatisation, la tokenisation, la mise en minuscules, et la suppression de la ponctuation en utilisant les fonctions précédentes.
11. Créez 2 dictionnaires : `welcome_input` et `welcome_responses` avec les réponses et entrées possibles.
12. Créez une fonction `Welcome()` qui prend en entrée la réponse de l'utilisateur et, si elle est dans le dictionnaire créé, sélectionne une réponse aléatoire parmi les réponses possibles.
13. Créez une fonction `generateResponse()` qui prend l'entrée utilisateur, la traite et génère une réponse si elle existe dans la base de données.