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

---

Objectif du TP

Manipuler un modèle de langage pré-entraîné pour comprendre comment il génère du texte et explorer le processus de tokenisation et génération de texte.

TP1 -

1- Installez et importez les bibliothèques suivantes : Nltk, random, string, re, unicodedata, wordnet de nltk.corpus, WordNetLemmatizer de nltk.stem.wordnet, wikipedia, defaultdict de collections, warnings, sklearn TfidfVectorizer de sklearn.feature_extraction.text cosine_similarity, Linear_kernel de sklearn.metrics.pairwise.

2- Importez les données HR.txt sur moodle et lisez-les

3- Assurez-vous que toutes vos données sont en minuscules

4- Afficher les 1000 derniers caractères de vos données

5- Préparation du corpus :

   a. Tokenisez vos données : en utilisant ntlk, tokenisez vos données brutes en phrases tokenisées (nltk.sent_tokenize())

   b. Normalisez vos données : créez une fonction qui :

      i. Supprime la ponctuation

      ii. Supprime le décodage ascii et utf-8

      iii. Supprime les symboles indésirables

6- Définissez votre lemmatiseur et remplissez-le avec les données/jetons que vous avez nettoyés ci-dessus (lemmatize())

BONUS

7- Créez 2 dictionnaires l'entrée de bienvenue et les réponses de bienvenue avec les réponses et entrées possibles

8- Créez une fonction Welcome () qui prend en entrée la réponse de l'utilisateur et si elle se trouve dans le dictionnaire que vous avez créé, elle doit choisir une réponse aléatoire parmi les réponses saisies.

9- Créez une fonction generateResponse() qui répond à la question de l'utilisateur,le traite et génère une réponse s'il y en a une dans la base de données.

10- Créez l'interface utilisateur qui démarre le chat avec « Bonjour, je suis un chatbot RH, comment puis-je vous aider ? » et répond aux questions.