**Projet python**

**Sujet : analyse comparative de corpus**

Nom : MAHAMOUD ISMAN

Prénom : Abdoulrazack

**Introduction :**

Ce présent projet a pour objectif d’explorer les corpus volumineux en utilisant les outils que nous offrent les différentes bibliothèques de python. Dans ce sens, il est demandé d’étudier deux corpus : Reddit et Arxiv. D’une part, on fait une étude comparative de ces deux corpus et d’autre part, on s’intéresse à l’évolution d’un mot d’un point du temps.

Pour cela, il est nécessaire de commencer d’abord par accéder aux deux corpus pour extraire les données. Avec l’API praw pour Reddit et urllib, xmltodict pour Arxiv. Pour construire le programme de cette application, on a décidé d’utiliser comme éditeur de texte visual studio code et jupyter notetbook.

**Présentation deux corpus**

Reddit :

Le Corpus Reddit fait partie d'un référentiel d'ensembles de données conversationnelles comprenant des centaines de millions d'exemples et une procédure d'évaluation standardisée pour les modèles de sélection de réponses conversationnelles utilisant une "précision de 1 sur 100". Le corpus Reddit contient 726 millions de dialogues multi-tours provenant du forum Reddit. L’ensemble de données se compose de 3 848 330 articles d'une longueur moyenne de 270 mots pour le contenu et de 28 mots pour le résumé.

Pour accéder au corpus Reddit, il faut d’abord créer un compte utilisateur pour pouvoir utiliser l’API. Une fois le compte est crée, on peut créer une application pour pouvoir s’identifier par la suite avec la librairie « praw »,  acronyme de «Python Reddit API Wrapper». Lors de la création de l’application, on récupère : le client\_id, client\_secret, user\_agent. Ces identifiants sont demandées lors de l’identification du corpus.

reddit = praw.Reddit(**client\_id**='Q-hGVHp0E7hJfaBlGF4xKw',

**client\_secret**='DMbe1oti1MLFfTbv-oxn4-NxnAAfHg',

**user\_agent**='proj\_prog\_python')

Une fois on s’identifie pour accéder au corpus, on peut récupérer le type de sujet souhaité en précisant en paramètre de la fonction reddit.subreddit("politics par exemple "). Voici un exemple de la recuperation d’un sujet donné ( exemple politics) :

**subs = reddit.subreddit("all")**

**for submission in subs.hot(limit=None):**

**print(submission.title**)

Arxiv

ArXiv est une archive ouverte d'articles de recherche principalement dans les domaines de la physique, des mathématiques et de l'informatique. Cet archive offre un accès libre à environ 1,5 million de prépublications.

**Environnement :**

Concernant l’environnement de travail, j’ai commencé d’abord par récupérer les corpus aussi bien Reddit qu’Arxiv. Pour écrire les programmes, j’ai utilisé l’éditeur de texte Visual studio code et en même temps jupyter notebook pour installer les librairies nécessaires.

**Diagramme des classes :**



J’ai utilisé diverses classes qui interagissent entre elles. Il y a d’abord la classe Doc qui possède des attributs comme : titre, auteur, date, url, texte, etc. cette classe possède deux classes filles comme RedditDocument et ArxivDocument. Ces classes héritent les attributs et les méthodes de la classe mère en l’occurrence « Doc ». la classe dite corpus qui contient les caractéristiques propres au corpus Reddit et Arxiv . Dans cette classe on récupère les informations contenues dans les deux corpus qu’on étudie.