Projet : web scraping sur The Canadian Encyclopedia

L'objectif de ce travail est de réaliser une analyse de contenus par web scraping de pages de l'encyclopédie Canadienne. On analysera les pages en anglais de cette encyclopédie.

Organisation du projet

De manière similaire au TD2, ce projet sera décomposé en :

- une partie scraping avec l'extraction d'informations dans un ensemble de pages web ;
- la sérialisation du contenu extrait sous forme d'un fichier .pick
- une partie analyse avec la construction d'une DTM et d'outils de requêtes sur le corpus.

Corpus analysé

L'objectif est d'analyser les articles d'une catégorie de l'encyclopédie en ligne The Canadian Encyclopedia. Chaque étudiant doit focaliser son travail **sur une catégorie donnée**, qui a été attribuée par tirage au sort. La liste des sujets tirés au sort est disponible sur Cursus.

Les catégories d'articles sont accessibles à cette adresse :
https://www.thecanadianencyclopedia.ca/en/browse/things
Les catégories sont accessibles par le menu situé à gauche de cette page. Par exemple, un étudiant qui aurait reçu le sujet Arts & Culture / Architecture doit se concentrer sur cette rubrique dans le menu de gauche.

Une rubrique fait référence à différents contenus. On pourra appliquer un filtre avec 'filter content by type' pour se focaliser uniquement sur les **articles**. En général, les articles référencés ne sont pas tous visibles sur une seule page. Il faudra donc balayer les différentes pages, en s'adaptant **automatiquement** au nombre de pages de résultats (ce nombre est visible en bas).

Dans notre exemple, l'étudiant qui travaillerait sur Arts & Culture / Architecture devrait ainsi analyser 24 articles, accessibles par https://www.thecanadianencyclopedia.ca/en/browse/things/arts-culture/architecture?type=article et https://www.thecanadianencyclopedia.ca/en/browse/things/arts-culture/architecture?type=article&page=2

Travail d'analyse

On pourra conserver les méthodes d'analyses vues lors du TD2. Pour le projet, il est demandé explicitement les fonctionnalités suivantes :

1. Une méthode **queryScore(chaine, N)** qui prend en entrée une requête sous la forme d'une chaîne de caractères, et un entier N, et qui renvoie les urls des N documents les plus pertinents pour cette requête,
2. Une méthode **wordCloud(numDoc)** qui affiche un nuage de tags pour le document d'indice numdoc. L'importance des mots sur le nuage est donné par la mesure tf-idf. On utilisera le module Python « wordcloud » pour la représentation des nuages de mots.

   Voici un exemple de nuage de mots construit pour le premier document du thème Arts & Culture / Architecture, qui porte sur l'architecture des aéroports.

3. Une méthode **nMostSimilar(numDoc, N)** qui prend en entrée un numéro de document et qui renvoie les titres des N documents les plus similaires à ce document. Pour calculer la similarité, on utilisera une similarité cosinus entre les vecteurs td-idf des documents.

Note : le code devra être clair, avec des exemples d'appels pertinents aux trois fonctionnalités.

Rendu du projet

Le travail devra être rendu sur Cursus pour le **mardi 22 mars à 14h.**
Il devra contenir :

- le code python du scrapping
- le fichier .pick des données extraites
- le code python de l'analyse.

La note comptera pour l'évaluation du contrôle continu. Ce projet est **strictement individuel.** Les fichiers sources seront comparés avec un logiciel d'étude de similarités, pour détecter les fraudes.