Application de méthodes de traitement de langage naturel pour l'extraction automatisée de données web pour la création d'une base de données pour les produits de santé naturels

Session	Hiver 2025
Nombre de personne	1 à 3 personnes (3 crédits : 135 heures ou 6 crédits : 270 heures)
Lieu	En personne à la faculté des sciences et la faculté de médecine En télétravail
Supervision	Pre Christina Khnaisser, christina.khnaisser@USherbrooke.ca Faculté de médecine et Faculté des sciences
Responsable	Pr Yohann Chiu, Faculté de médecine, yohann.chiu@usherbrooke.ca Pr Benoit Cossette, Faculté de médecine, benoit.cossette@usherbrooke.ca

Description du projet

L'utilisation de produits de santé naturels (PSN) est répandue au Canada chez les ainés, les données de l'Étude longitudinale canadienne sur le vieillissement (ÉLCV) suggérant que 86 % des participants ont consommé au moins un PSN entre 2011 et 2015, avec une moyenne de 5 PSN par personne. Les informations sur les PSN sont dans de multiples sources de données, y compris les sites Web, les médias sociaux, les bases de données locales et les questionnaires. Ce projet d'études s'insère dans un projet plus large qui vise à classifier les PSN de façon structurée, afin de les rendre accessibles pour la recherche en lien avec les maladies chroniques. Dans un premier temps, nous allons nous intéresser à l'extraction des données de plusieurs sites web. Le *Web Scraping* est une solution alternative à la collecte manuelle de données qui consiste à rassembler les informations disponibles sur différents sites web. Par exemple, l'outil (ex. Scrapy, Octoparse, etc.) navigue sur chaque page du site web WebMD pour extraire les informations pertinentes sur les produits (par ex, « Acides gras oméga-3 — huile de poisson 300 mg - 1 000 mg capsule ») tels que : le nom (« Acides gras oméga-3 — huile de poisson »), le dosage (« 1000 mg »), l'utilisation (par exemple, « Prendre ce produit par voie orale »), la liste des ingrédients (« huile de poisson », « Acide eicospentaénoïque 300 mg », « Acide docosahexaénoïque 200 mg »). Ensuite, nous utiliserons des outils de traitement de langage naturel (ex. spaCy, BERT, etc.) combinés pour extraire des informations à partir de données non structurées. Après avoir recueilli les données des sources de données cibles, nous devons les structurer et les sauvegarder dans une base de données pour pouvoir les mettre à jour et les rendre disponibles aux analystes de données.

Le projet consiste à :

- Définir les besoins de collectes de données ;
- Faire une étude comparative des outils de *Web Scraping* pour choisir celui qui correspond le mieux aux besoins ;
- Élaborer une stratégie de collecte de données en utilisant un outil de *Web Scraping* ;
- Élaborer une base de données relationnelle pour stocker les données recueillies ;
- Développer des requêtes pour visualiser les données.

Compétences cibles	Modélisation de bases de données Conception logicielle Développement logiciel Développement de bases de données relationnelles Vérification et validation logicielles
Langages de programmation et technologies	Python, SQL, Java, Bash Pycharm, IntelliJ, PostgreSQL Asciidoc, Git/Gitlab
Atout	Être inscrit ou avoir réussi le cours IFT599 ou IFT607