CS410 Project Topic Proposal, Fall 2022

Team: Tetra
Members: Dillon Harding (dth3), Kowshika Sarker (ksarker2), Nikhil Garg (nikhilg4)
Captain: Vaibhav Karanam (karanam5)

Proposed Topic

Our project is identifying food items beneficial or detrimental for certain disease conditions. By collecting textual diet recommendations of different disease conditions from reputable web sources, we plan to compare that with free text nutrient descriptions of food items collected from web sources to determine association of food items with disease conditions.

Expected Outcome

The output of the project will be - if the user is interested in a particular disease, our system will provide a list of association of food items as recommended, detrimental, or neutral in context of that disease.

Importance

Our proposed system can be integrated with grocery shopping applications to help users better plan their diet intake. User disease profile can be stored in the account information and when user adds any new food item to the cart, the system may provide a pop up notifying the association of the food item with respect to the user's disease profile.

Project Description and Planned Approach

Web Crawler

The project will gain data via a custom made web scraper. We are using the scraper to amass text data from credible web sources. The web scraper's design will be dependent on the URL structure of a source website of choice. In order to properly scrape the site, it will be important to know which query strings and URL stubs to navigate through in order to gain access to the data on all pages. As most sources with dynamic information will likely be contained within a tabular HTML structure, we will utilize the Python library Pandas to read tabular site data. Our group will ensure we are within the terms of use of each website with regards to their website scraping policies.

Sentiment Analysis

The project will require the use of sentiment analysis to primarily establish the link between the different diseases and food items that have been gathered with the web crawler. The analysis will be based on how the food item has been linked to the disease, does the food item prevent the disease or provides healthy addition to the diet to protect from diseases. We will be looking for terms such as 'diet recommendations' or 'foods to avoid' sections from the web pages used

by the crawler. This will also provide a way to map food items based on its positive and negative associations with the disease.

Website

For our project, a user will be able to interact with a website where they can input a medical condition or disease as a query to search with. For example, a user looking to prevent or reduce the symptoms of cardiovascular disease can use a term such as “cardiovascular disease” as a query on the website and see a list of foods that would be beneficial for someone with that condition or someone who is trying to prevent its onset. This website will be the outward facing user interface that any user would interact with, and their input would be used for searching through the data from the web crawler in order to find the most relevant information and deliver the best results.

Programming Languages

Python will be used as our primary language to build the web page crawler and perform sentiment analysis. Javascript will be used to build the front end of the tool’s website

Datasets

We will be looking at web pages such as Wikipedia, Mayo Clinic and USDA food database.

Evaluation

Our project will consist of two components - web crawler and sentiment analysis of food items and diseases. We will evaluate both of these components to make sure the tool is performing correctly and producing the right results. For the web crawler, we will sample a set of web pages that have been crawled to manually highlight the data that should be captured and is important. We will then compare our highlights with the data gathered from the web crawler to see what percentage of the data is similar. These will give us a threshold of the crawler’s accuracy and a benchmark. For the sentiment analysis, the predictions from the analysis will be compared with content from a reputable website such as Mayo Clinic. The team will then see what is the similarity to score between the predictions and expert knowledge and if the analysis was able to capture the positive and negative diet recommendations for the diseases sampled by our tool.

Workload

Our project has three main tasks - web page crawling, analysis of disease and food text descriptions, and website development. We are hoping each of the first and last tasks will take around 20 hours and the second task will take around 40 hours. The second task will be handled by two members.