Yelp – Healthy Restaurant Engine

Project Component Specification

DATA 515 Winter 2021

University of Washington

TEAM:

Sepideh Dibay Moghadam, Natasha Halfin, Harsha Koonaparaju, Xiaoyu Lin, Liem Luong, Divya Pandey

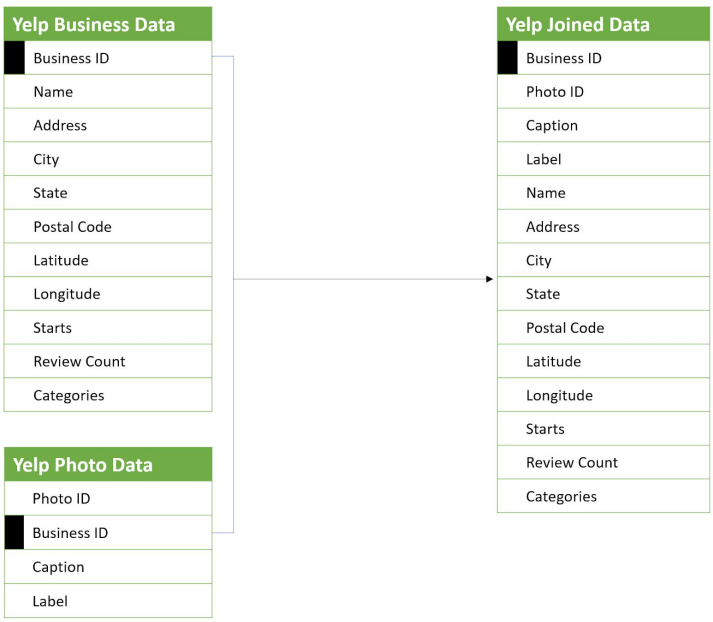
1. Introduction

The goal of this project is to create a Web App that allows users to search for healthy restaurants per their use case. The web app will provide healthy restaurant results within Yelp businesses on the basis of a health score generated by our classification model.

2. Datasets

Yelp Dataset - captures details of almost 209K Businesses across ~10 cities capturing ~ 8M reviews and ~200K pictures. Our project focuses on the classification of healthy or unhealthy restaurants based on the photo captions of the menu items in a restaurant.

a.) Link : https://www.yelp.com/dataset



Nutritionix Dataset - contains nutrition information from chain restaurants across the US. We are using it to train our classification model, which we will apply to assign healthy or unhealthy labels to restaurant photos in the Yelp dataset

b.) Link : https://www.nutritionix.com/business/api . Features we used are as following:

Restaurant Data
Item_name
brand name
upc
nf_ingredient_statement
nf_calories
nf_calories_from_fat
nf_total_fat
nf_saturated_fat
nf_trans_fatty_acid
nf_cholesterol
nf_sodium
nf_sugars
nf_protein
nf_serving_per container
Healthy score

3. Overall Framework

3.1 Logic for the Model

We leveraged nutritional science evidence based practices to generate the health flag labels for our training data, the Nutritionix database (this takes place during the "calculate_health" function in the data processing module). This logic is based on a 2,000-calorie average diet and three meals per day. The following features were considered:

Calories per meal: We consider a meal unhealthy if it provides more than 666 calorie (33% of the 2000 benchmark calories).

Overall fat:
Total fat allowance intake is 20% to 35% of daily calories, and we assume three meals per day. Thus, if a meal has more than 200 calorie or more than 20 gr fat then them meal gets categorized as unhealthy.

Saturated fat:
Saturated fat daily allowance is 10% or less of daily calories. So, if a meal provides more than 66 cal or 8 gr of saturated fat then it gets categorized as unhealthy.

Trans fatty: Any amount of trans fatty is considered carcinogenic and thus any meals with more than 0 grams of trans fatty acid is categorized as unhealthy.

Cholesterol:
The USDA recommends consuming no more than 300 milligrams(mg) of cholesterol a day. If a meal contains more than 100 mg cholesterol it is considered unhealthy.

Sodium:
The Daily Value for sodium is less than 2,300 mg per day. If a meal provides more than 766 mg sodium then it is categorized as unhealthy.

Sugar:
The American Heart Association suggests that the consumed amount of added-sugar should not be more than 24 g of sugar for women and no more than 36 g for men. For the purpose of our logic, we take an average and consider more than 30 g sugar in a meal as unhealthy.

If a meal does not meet any of the criteria defined above, it will be categorized as unhealthy. We will use the Nutritionix dataset to train our text classification model (discussed below). Then we will apply the trained model to the Yelp dataset.

3.2 Data Processing module- Data processing framework is designed to read and parse data from two datasets - Yelp and Nutritionix to generate clean and tokenized CSV files by utilizing NLTK and Numpy libraries. Features such as calories, sodium, cholesterol, sugar, fat etc for various menu items present in Nutritionix dataset are utilized to generate a final health score for each menu item/caption, which is then tokenized to create a training dataset for our classifier. Features such as Business Id, ratings, along with multiple photo captions of Yelp dataset, are cleaned and tokenized to prepare a final dataset that will be used for health classification.

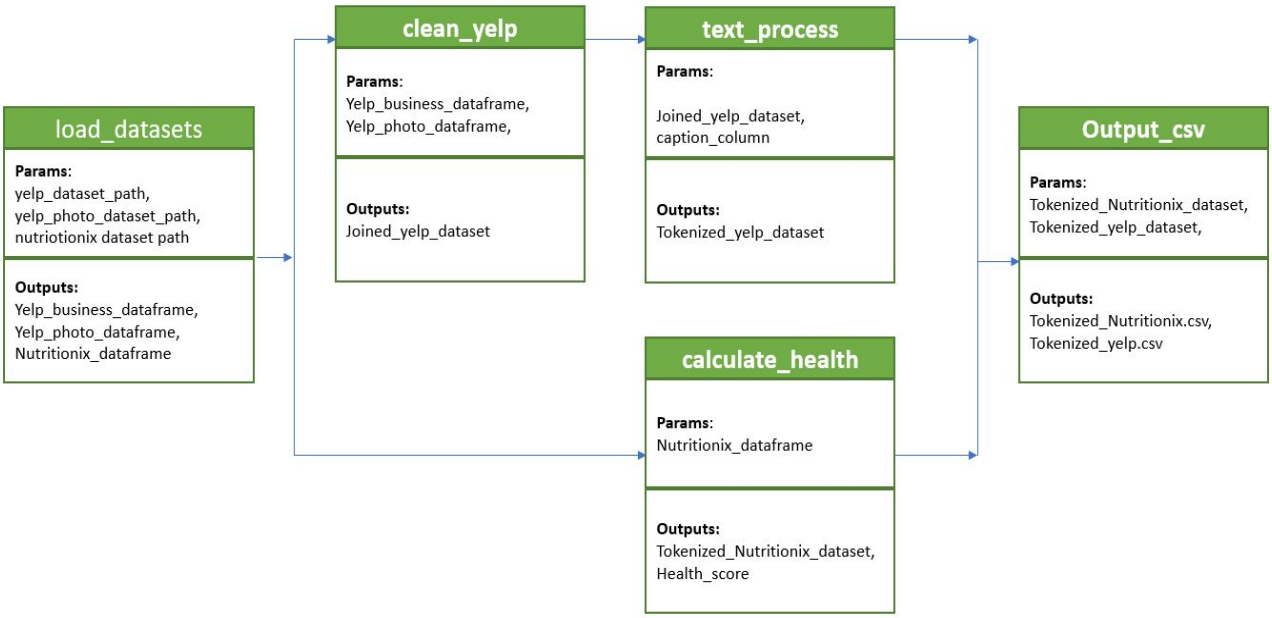


fig 1 : Data processing module

3.3 Text Classification module - The text classification module takes the tokenized Nutritionix dataset output from "Data Processing Module" and performs a test and train split using Scikit Learn library. These split datasets are then further input into the function that uses Scikit-learn (CountVectorizer) and generates a word frequency table, where sentences are scored and ranked based on word frequency. The Multinomial Naive Bayes model learns from this vectorized data and then predictions are calculated on the Nutritionix test dataset. The model trained on the Nutritionix dataset is then fit on the Yelp dataset to generate the healthy flag for each photo caption belonging to a business. Finally, the scores for all photo captions are averaged at a business Id level to provide a comprehensive health score per restaurant ranging from 0-100%.

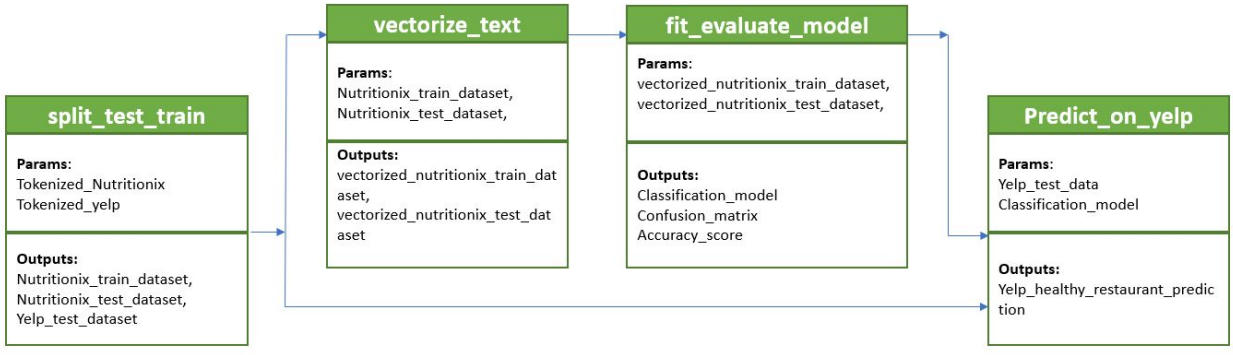


fig 2 : Text Classification module

3.4 Web UI module - Python Dash is used to build the map visual for the user interface web app which takes the input value from the users on the application interface (i.e. state or city and the food category). The cached model results are then exposed in the UI and the health scores (the propensity of healthy food offered by the restaurant) will be provided based on the user query.

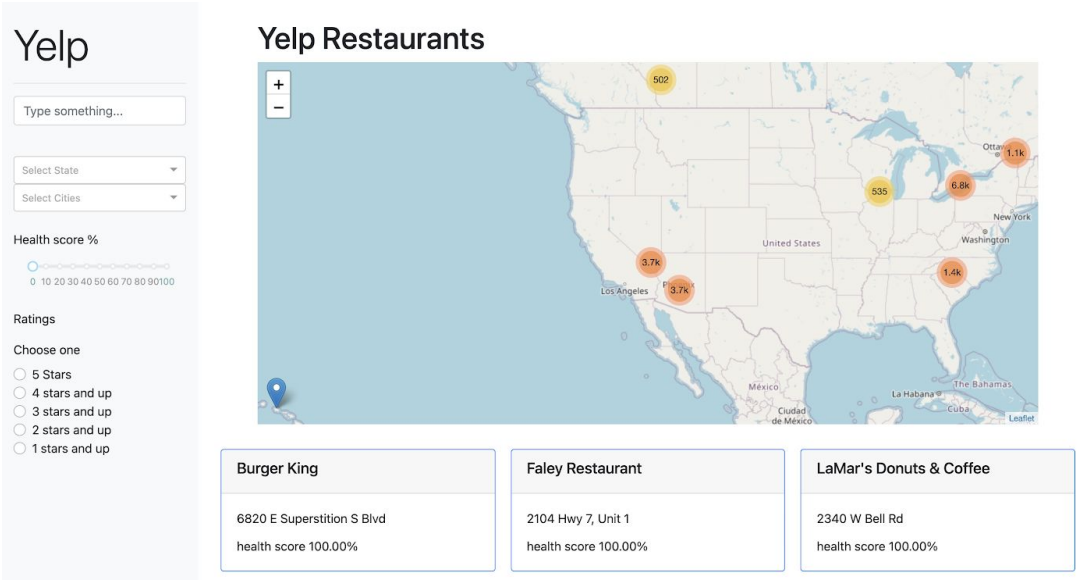


fig 3 : Web UI module

4. Interactions

4.1: Use case # 1

Description: Jack wants to look for a healthy restaurant in his area

Interaction:
Once the user locates the site, he can first enter his city or state to filter to restaurants in his area. The map on the right side of the screen will display the corresponding restaurant based on the search results. Secondly, Jack can zoom in on the map and target the restaurant he wants to visit. At the bottom of the map, the corresponding information about the restaurant is displayed, including the location, rating stars, review counts, categories and prediction score.

4.2: Use case # 2

Description: Jane wants to lookup an Asian vegetarian restaurant

Interaction:
The interactions in this case are very similar with the last one except that the user has the option to choose based on a specific ethnic cuisine and/or dietary practice. When Jane opens the webpage, first, she can enter the keywords of "Asian" or "vegetarian" in the search box on the left side of the screen. The map on the right side will display the restaurants corresponding to the search results. Second, Jane can zoom in on the map and locate the restaurants she is interested in. When she hovers over a specific restaurant, the corresponding restaurant information will be displayed at the bottom of the map.

4.3: Use case # 3

Description: Tyler wants to find a restaurant to has a minimum health score of 70%

Interaction:

Again, the interactions for this use case are similar except that the user can leverage the sliding scale on the left panel to filter results to restaurants that meet a minimum health score. After applying this filter, the user will only see results that have a health score of 70% or higher (in this scenario). The user can again use the map feature to zoom in on the restaurant results.

5. Preliminary Plan

The following describes the milestones that we plan to hit for each target date.

2/27:

- Final drafts of software design documentation
- Trained model based on labeled data and applied to Yelp data; performance metric assessment
- Basic UI stood up with dummy data
- Connect Travis CI to GitHub repository

3/6

- UI consuming and surfacing model results
- Custom packages and functions completed
- Draft version of repo requirements:
    - Readme
    - Setup.py
    - Examples folder
- Unit tests written
- Travis CI integration
- Review of coding practices (PyLint, PEP-8)

3/13

- Final tweaks to classification model
- Repo code complete and adhering to coding standards
- Draft slides created

3/15

- Final presentation
- No more changes to repo

References

- How to track saturated fat
- 2010 Dietary Guidelines for Americans
- How much sodium should I eat per day?
- Added Sugar in the Diet | The Nutrition Source | Harvard TH Chan School of Public Health