Yelp Healthy Restaurant Engine

CSE D 515 - Software Design For Data Science
March 2021

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

Project Introduction

Background. Describe the problem or area being addressed.

Background & Problem Addressed

- Increasing consideration of healthy lifestyle and diet
- Growing demand for healthy restaurants, but few search engines incorporate this dimension
- **Project goal:** provide a way for users to discover healthy restaurants within the existing Yelp directory
- **Methodology:**
  - Train text classification model from Nutritionix training data
  - Apply model to photo captions from Yelp dataset to determine if items are healthy
  - Generate health score for restaurants based on the model results
  - Surface results in app where eaters can search for nearby healthy restaurants

Data Sources

Data Sources:
- Yelp open dataset : Contains business information of 209k+ restaurants in 10 cities, 8M reviews and 200k pictures
- Nutritionix API : Contains nutrition information from chain restaurants across the US. We used this to train our classification model that assigns healthy or unhealthy labels.
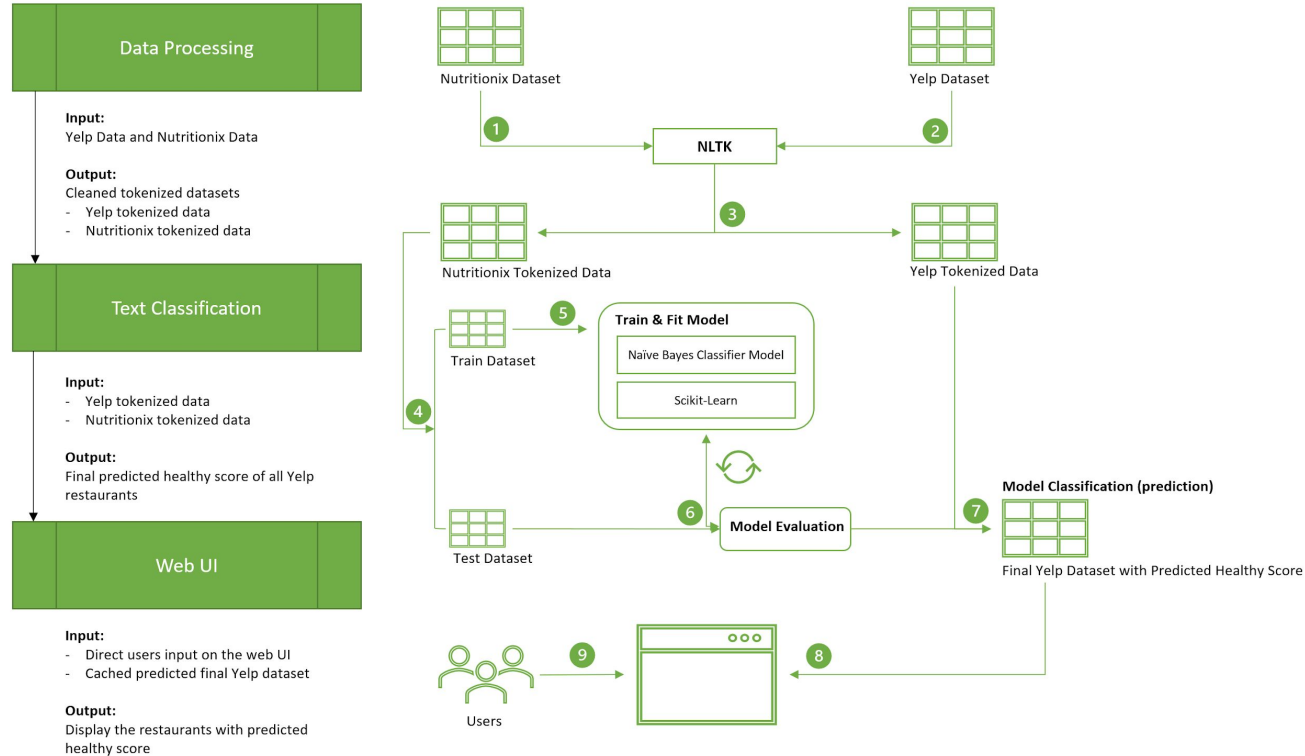
Limitations:
- Generated labels in training data based on nutritional fields; these were absent in the Yelp data
- Final predictions not necessarily accurate from a nutritional perspective

Use Cases

- **Use Case 1:** user wishes to find a healthy restaurant in close proximity by filtering on city or state

- **Use Case 2:** user wishes to find a healthy restaurant in a specific category, i.e. Greek food or a restaurant that serves vegetarian food

- **Use Case 3:** user wants to find healthy restaurants to have a health score of 70% or higher

Software Design

Data Processing

Input:
Yelp Data and Nutritionix Data

Output:
Cleaned tokenized datasets
- Yelp tokenized data
- Nutritionix tokenized data

Text Classification

Input:
- Yelp tokenized data
- Nutritionix tokenized data

Output:
Final predicted healthy score of all Yelp restaurants

Web UI

Input:
- Direct users input on the web UI
- Cached predicted final Yelp dataset

Output:
Display the restaurants with predicted healthy score

Nutritionix Dataset

Yelp Dataset

1

2

NLTK

3

Nutritionix Tokenized Data

Yelp Tokenized Data

Train Dataset

5

Train & Fit Model

Naïve Bayes Classifier Model

Scikit-Learn

4

6

Model Evaluation

7

Model Classification (prediction)

Test Dataset

Final Yelp Dataset with Predicted Healthy Score

Users

9

8

Project Structure

Github Repository

https://github.com/nhalfi/Yelp_Recommendations

```
.
└── Yelp_Recommendations-main
    ├── azure-pipelines.yml
    ├── docs
    │   ├── Final\ Project\ Draft\ Presentation.pptx.pdf
    │   └── Yelp_Healthy_Restaurant_Component_Specification.docx.pdf
    ├── examples
    │   └── README.md
    ├── LICENSE
    ├── README.md
    ├── setup.py
    └── Yelp
        ├── __init__.py
        ├── app.py
        ├── data
        │   ├── Yelp_Healthy_restaurant_predictions.csv
        │   ├── __init__.py
        │   ├── nutritionix_health.csv
        │   ├── nutritionix_tokenized.csv
        │   ├── restaurants_dao.py
        │   ├── yelp_business_clean.csv
        │   ├── yelp_final_tokenized.csv
        │   ├── yelp_joined_clean.csv
        │   └── yelp_photos_clean.csv
        ├── data_processing
        │   ├── __init__.py
        │   ├── data_processing.py
        │   └── download_nutritionix_data.py
        ├── tests
        │   ├── __init__.py
        │   ├── business.json
        │   ├── empty.csv
        │   ├── logic_test.csv
        │   ├── nutrition.json
        │   ├── photos.json
        │   ├── restaurant_sample.csv
        │   ├── restaurant_sample2.csv
        │   ├── sample_nutritionix_data.csv
        │   ├── sample_yelp_data.csv
        │   ├── test_data_processing.py
        │   ├── test_download_nutritionix_data.py
        │   ├── test_restaurants_dao.py
        │   └── test_text_classification.py
        ├── text_classification
        │   ├── __init__.py
        │   └── text_classification.py
        └── ui
            ├── __init__.py
            └── html_components.py
```

Demo

Lessons Learned and Future Work

Lessons

- Can be creative with choice of training data
- Proper package structure
- Test-driven development
- Setting up continuous integration
- Pull often and double-check before you push

Future Work

- Find additional training datasets closer to Yelp data
- Image classification of photos in addition to text classification

Q&A

Appendix

Data Diagram

- An operation is performed on these two separate data segments from Yelp into a comprehensive set for our model. Below is a data diagram of this Yelp joined dataset.

Yelp Business Data

Business ID
Name
Address
City
State
Postal Code
Latitude
Longitude
Starts
Review Count
Categories

Yelp Photo Data

Photo ID
Business ID
Caption
Label

Yelp Joined Data

Business ID
Photo ID
Caption
Label
Name
Address
City
State
Postal Code
Latitude
Longitude
Starts
Review Count
Categories

Yelp Data Structure

Training Dataset from Nutritionix

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