Yelp – Healthy Food Recommendation

Project Functional Specification

University of Washington

Winter 2021

DATA 515

TEAM:

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

Background

Food is essential to life, and eating healthy has become a priority for more and more people. According to the survey from the US Department of Agriculture (USDA), more consumers care about their eating behavior in connecting to their health. With 71% of people who believe they could eat healthier from the survey, it is clear that there is growing awareness around this topic. The challenge for eaters is finding a restaurant that serves healthy food. Our project aims to solve these challenges by providing a platform where eaters can find nearby restaurants that serve healthy food. To accomplish this, we are leveraging the Yelp open dataset as the base and building a machine learning model on top that analyzes the captions of restaurants' food photos to assess if the item is healthy or not. Through this manner, we can help eaters make more health-conscious decisions when they pick where to dine.

Goals

- Extract the valuable information from the photo captions on the Yelp dataset to build a model that classifies healthy and unhealthy food.

- Update the business profile with the healthy food label.

- Build the location-based lookup tool where eaters can search for a recommended nearby restaurant that serves healthy food by zip code, and cities.

User Profile

Organic food, fruit & veggies, plant-based food, low sodium snacks, less sugary drinks, etc. are some of the trending words in the food industry over the last few years. According to Gallup research, consumers within the age range of 18 to 50 exhibit greater adoption of vegetarian or vegan eating. Thus, this group is included within our target consumers that share characteristics fitting the healthy food consumption profile.

Young working professionals who are single represent a growing demographic for large cities and suburbs. This group leverages the use of technology in every aspect of life; healthy fitness and healthy eating are some of the characteristics of this millennial group. Jack is our user profile in this category. The second group is geared toward the married working families. They are more health-conscious in providing healthy nutrition to all family members. Jane represents this category of working parents.

AGE: 32

STATUS: Single

EDUCATION: MS Data Science from UW

OCCUPATION: Software Developer

TECH SAVVY:
Strong background in technology and enjoying explore local places with apps on smartphone

NOTABLE POINT:
a young tech with healthy lifestyle in mind, love checking out the healthy local eat with friend.

JACK

AGE: 36

STATUS: Married

EDUCATION: BS Economics from UW

OCCUPATION: Financial Investment Manager

TECH SAVVY:
Use desktop more than a smartphone for day-to-day job and family planning on laptop
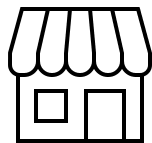
NOTABLE POINT:
a mother with a great sense of care for the family, love planning weekly family healthy eat-out.

JANE

Data Source

The main dataset for our project is the public Yelp Open Dataset. This dataset is available to the public in JSON files. A little note about the richness of this data source:
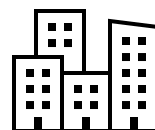
~ 8M reviews   ~ 209K business   200K pictures   10 cities

The Yelp dataset is structured into six major segments:

Segments	Description
Business	This segment contains business data: location, attributes, and categories, etc.
Review	This segment contains the review from users for a particular business
User	This segment contains the metadata of users
Checkin	This segment contains the detail of check-in on a business (date and time)
Tip	This segment contains the tip written by a user on a business
Photo	This segment contains the photo (include the caption) and the classification of the photo

The second dataset is the Nutritionix dataset which contains nutrition information from chain restaurants across the US. We are using it to train our classification model that will assign healthy or unhealthy labels.

Field	Description
item_name	Name of menu item
brand_name	Restaurant/brand name
upc	UPC if valid
nf_ingredient_statement	Ingredients if available
nf_calories	Caloric value
nf_calories_from_fat	Calories from fat
nf_total_fat	Total fat in grams
nf_saturated_fat	Saturated fat in grams
nf_trans_fatty_acid	Trans fatty acids in grams
nf_cholesterol	Cholesterol in grams

nf_sodium	Sodium in milligrams
nf_sugars	Sugar in grams
nf_protein	Protein in grams
nf_serving_per container	Servings per container

Based on the values of the nutritional fields, we have assigned a binary healthiness score (0 if unhealthy, 1 if healthy). The goal is to correlate the text to the score and use this trained model when classifying the food items in the Yelp dataset.

Use Cases

Below are two use cases tailored to our users (Jack and Jane).

1. Jack wants to check out the restaurant with his friends after work in the neighborhood. He plans the event by using his smartphone to look up a recommendation of a healthy restaurant before sharing it with his friends.
2. Jane has realized that her family consumes a lot of meat this week. She wants to change the routine with something light and healthy. While planning for the weekend grocery trip on her desktop, she also looks for a vegetarian restaurant on Sunday.
3. Tyler wants to find a restaurant that meets a minimum health criteria.

Use Case # 1

Jack inputs the location of Chicago as the desired city. After the input is supplied to the application user interface on his smartphone, a list of healthy restaurants are shown to Jack on the screen that fit his filters. Each restaurant's results include three basic meta-data:

- Business Name
- Location
- Serving Healthy Food (% of healthy out of 100%)

The application interface also includes a map for the user interaction (zoom in, zoom out, pan the map)
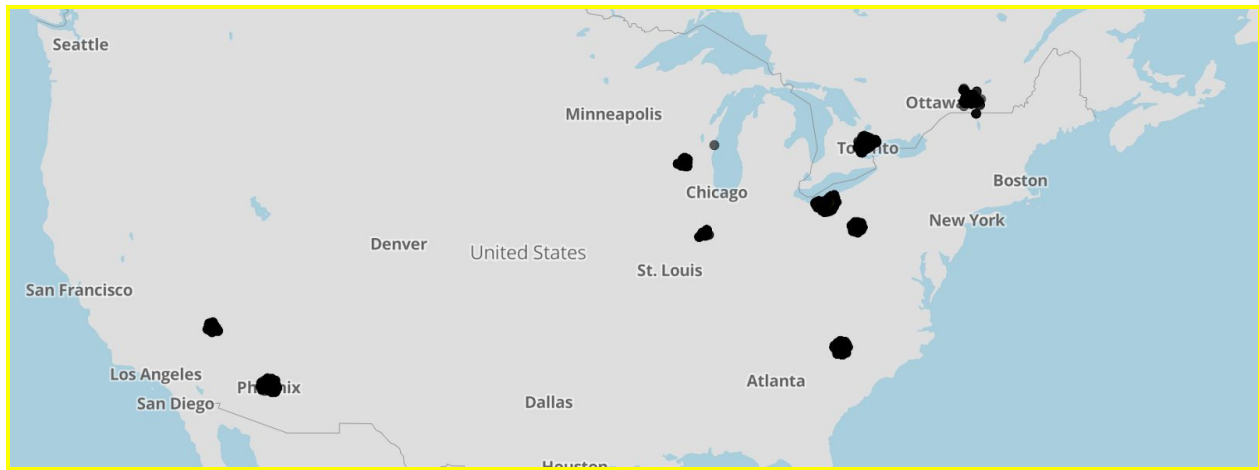
Figure 1 - Sample of application lookup

Use Case # 2

Jane opens the web browser on her computer and searches for a vegetarian restaurant in Phoenix, AZ. Vegetarian is the main category Jane uses for the filter.

After the inputs are supplied to the application user interface on her computer, a list of vegetarian restaurants are displayed on the screen. The same metadata listed in Use Case #1 are again included.

The application interface also accompanies the result with a map for the user interaction (zoom in, zoom out, pan the map). Jane also decides to add one more filter for the business category "Asian Vegetarian" to filter down the result list further to her desire.
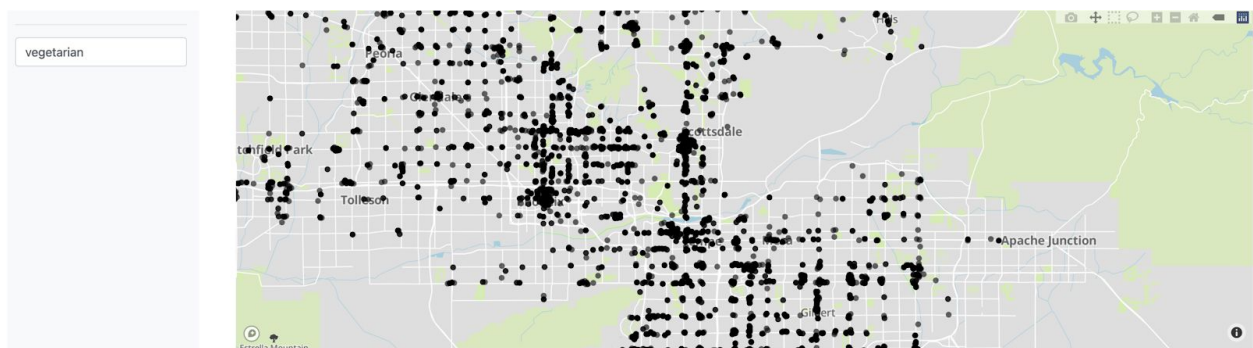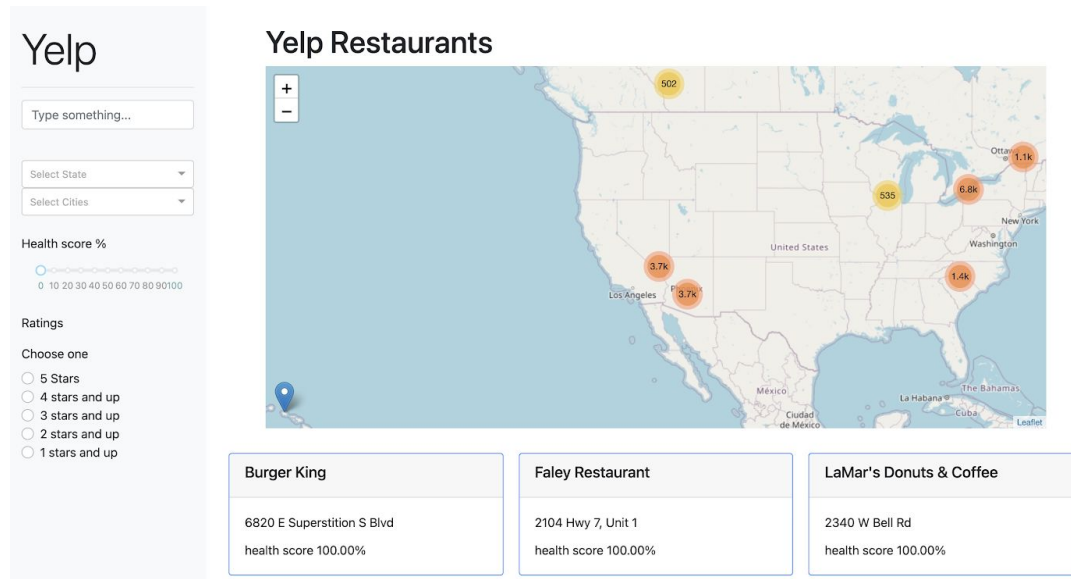

Figure 2 - Sample of application lookup for vegetarian restaurant in Phoenix, AZ

Use Case #3

Tyler manipulates the sliding scale feature on the left to filter results to those that meet a minimum health score criteria.

Reference

US Department of Agriculture (USDA) survey about Eating Behavior in the U.S.
- Published by Statista: U.S. Eating Behavior - Statistics & Facts | Statista

Gallup Research on "Who Are America's Vegans & Vegetarians?" in 2019
- Published by Statista: Chart: Who Are America's Vegans & Vegetarians? | Statista

Yelp Open Dataset
- An all-purpose dataset is available to the public: Yelp Dataset

Nutritionix Dataset
- A comprehensive nutrition database: Nutritionix