Cafés in America

Listen to the customer's voice to improve performance

06.15.2021

—

Valentina Rizzati
Data Scientist
Yelp, Business Team

Opportunity

I am a Data Scientist for the Business team at Yelp. In light of the recent allegations about the Yelp Mafia (i.e. Yelp supposedly extorting small business owners for advertising fees in return for helping to improve reviews on their platform), I've been approached by a Product Manager who intends to build a new tool for businesses with the goal of demystifying the accusation of a corrupt system underlying Yelp reviews and, as a result, rebuild Yelp's reputation in the eyes of business owners.

Given the pandemic's impact on the hospitality industry, we intend to provide restaurants with a tool that will enable them to gather useful information from reviews and optimize their performance. In particular, I will focus on cafés because their relative low variance in pricing and customer preferences across regions will likely make the insights more applicable at a national scale.

As aligned with my business counterparts, I will first build the prototype based on all cafés in NY and then, depending on time, I will extend it to all of America.

Impact Hypothesis

If cafés owners have more visibility on what are the most valued elements in a café by a customer (e.g. decor, matcha latte), they will be able to invest in what really matters and align to customers' expectations. As a result, their performance and popularity will eventually improve.

Data

I will collect data from Yelp. This will be in the format of five json files (*business*, *review*, *user*, *tip*, and *checkin*) and for the purpose of this project I will focus on the *business* and *review* files.

I will use MongoDB to store the data locally and then reshape it in a tabular form in Python.

I will primarily focus on the text reviews and star ratings for every café in my scope. Depending on processing time and time availability, I might decide to filter the data by number of reviews (e.g. consider only the cafés that have at least 20 reviews).

Solution Path

I am planning to follow the following solution path:

- Unsupervised Learning model
    - Conduct data cleaning and initial EDA
    - Preprocess the text data
    - Perform dimensionality reduction and topic modeling to understand which are the elements of a cafe that customers value the most
    - Visualize the results with word clouds, scattertext or similar techniques
- Supervised Learning model (contingent on time)
    - Build a classification model to predict, based on a set of features (built on the topics identified in the unsupervised learning phase), if a café will be a *Top Yelp* (4 or 5 stars) or not (1 or 2 stars)
    - Run different models through cross-validation
    - Select optimal classification model
    - Build Tableau Dashboard to visualize the results

Tools

To store and query the data I will use MongoDB.

To handle the text data I will use processing libraries/tools such as NLTK and spaCy.

To manipulate the data and build the classification model I will use pandas, numpy, and scikit-learn.

For visualizations I am planning to use spaCy, seaborn, plotly and, depending on time, Tableau.

MVP Goal

As an MVP, I am planning to present a version of the topic modeling and a few relevant visualizations.