Springboard Data Science Career Track

Capstone Project 2 Milestone Report 1

Topic Models for Yelp Reviews

Walker Page

2020

1. Introduction

Yelp is an online platform that, among other things, allows customers to post reviews and ratings of businesses to describe their experience with those businesses. This platform is a valuable tool for customers because it enables them to make more informed decisions about which businesses to use and which to avoid. Yelp has the potential to be invaluable for business owners as well, and this project makes progress towards developing one way that Yelp could be improved as a resource for business owners. The project focuses specifically on restaurants.

For restaurants to thrive as businesses it is important for them to know what their customers are saying about their experiences with the restaurant. Restaurant owners might acquire this knowledge in a number of ways, but one method is to read Yelp reviews that customers write about their experiences. A problem with this method, however, is that it involves conflicting incentives for restaurant owners. On the one hand, reviews are a valuable source of information about customer experiences, so restaurant owners should want more customers to write more reviews. On the other hand, reading and interpreting reviews can be time-consuming, and the more reviews there are, the more time it takes to gain value from them. Furthermore, it can be difficult to gain an overall sense of customer experiences simply by reading a set of reviews.

The goal of this project is to provide a more efficient means of summarizing the content of reviews. In particular, the goal has been to construct models that represent the dominant "topics" in reviews. Having models that can do this not only removes the problem of needing to read through each review just to get some sense of customer experiences, but it also provides a way to get a clearer view of overall customer experience. For example, after identifying the dominant topics in reviews, similar reviews could be grouped together and sorted by topic. If Yelp added a feature that enabled restaurant owners (and eventually any business owner) to easily identify the dominant topics of reviews as well as overall customer experience, the platform would be a much more useful resource for owners.

For this project, I have construed the business problem as an unsupervised problem, and I have developed and evaluated 24 models to generate topics in the reviews.[1] In what follows, I will describe my approach, summarize my findings, draw conclusions from those findings, and make recommendations for Yelp moving forward.[2]

2. Approach

2.1 Data Acquisition and Wrangling

The data I used for the project includes approximately 6.5 million reviews of a variety of different businesses.[3] I acquired the data directly from Yelp, but it needed to be manipulated in several ways to be useful for the present project.[4] First, since the dataset contained reviews for many different kinds of businesses, and the present project is concerned only with restaurants, I removed reviews for businesses that were not restaurants.

Next, I inspected how many rows were duplicated in the dataset, and how many values in the column of reviews were missing/null. Since it turned out that the percentage of duplicate rows and missing reviews relative to the entire dataset was miniscule I simply removed these.

Finally, after closer inspection, I discovered that some of the reviews were not written in English. Since analyzing non-English reviews was beyond the scope of the project, and the number of non-English reviews was negligible relative to the size of the dataset (only around 24,000 out of over 4 million total reviews), I removed them from the dataset as well.

---

[1] The models can be viewed here:
<https://nbviewer.jupyter.org/github/walkerpage/springboard/blob/master/Capstone_Projects/Capstone_Project_2/Code/Capstone_Project_2_Extended_Modeling.ipynb#topic=0&lambda=1&term=>

[2] All code for the project can be found in the following repository:
<https://github.com/walkerpage/springboard/tree/master/Capstone_Projects/Capstone_Project_2/Code>

[3] For the documentation for the dataset, see here: https://www.yelp.com/dataset/documentation/main.

[4] The dataset was acquired here: https://www.yelp.com/dataset.
The notebook in which I clean the data can be found here:
https://nbviewer.jupyter.org/github/walkerpage/springboard/blob/master/Capstone_Projects/Capstone_Project_2/Code/Capstone_Project_2_Data_Wrangling.ipynb

2.2 Storytelling and Inferential Statistics

After acquiring and cleaning the data, I explored a number of interesting features of the data.[5] In particular, since the goal has been to construct topic models that indicate the content of reviews, I inspected a number of word clouds using subsets of the data to gain some sense of the content in those subsets. The subsets of data included a sample of all reviews in the dataset, all reviews for a single restaurant with a large number of reviews, a sample of reviews of restaurants in Las Vegas, all reviews in the category 'Mexican, Restaurants', and all reviews for a single user. Let me comment on just three of the word clouds generated from these subsets.

First, it is interesting to note the similarity between the word cloud for reviews of all restaurants in the dataset (Figure 1) and the word cloud for reviews of all restaurants in Las Vegas in the dataset (Figure 2). They share a number of common words. I constructed a word cloud of Las Vegas restaurant reviews because Las Vegas is the most frequent city in the dataset. It turns out that the similarity of these word clouds is unsurprising since reviews of restaurants in Las Vegas make up approximately one fourth of the entire dataset. Second, I constructed a word cloud of all restaurants in the most frequent category in the dataset, which is 'Mexican, Restaurants' (Figure 3). It is interesting to see how easily one can tell that the word cloud represents reviews of mexican restaurants.

[5] The notebooks in which I explore an analyze the data can be found at these two links: https://nbviewer.jupyter.org/github/walkerpage/springboard/blob/master/Capstone_Projects/Capstone_Project_2/Code/Capstone_Project_2_Data_Storytelling.ipynb, and https://nbviewer.jupyter.org/github/walkerpage/springboard/blob/master/Capstone_Projects/Capstone_Project_2/Code/Capstone_Project_2_Statistical_Data_Analysis.ipynb

Figure 1

Figure 2

Figure 3

In addition to visually inspecting word clouds of these subsets of the review data, I also constructed several artificial "topics" (i.e. groups of words) to identify their prevalence in the data set. For example, I constructed two topics labeled 'positive sentiment' and 'negative sentiment' respectively. These topics were composed of the following words:

- positive sentiment: ‘good’, ‘delicious’, ‘yummy’, ‘tasty’, ‘superb’, ‘best’, ‘great’, ‘amazing’, ‘awesome’
- negative sentiment: ‘bad’, ‘disgusting’, ‘gross’, ‘nasty’, ‘terrible’, ‘worst’, ‘horrible’

With these topics, I identified two subsets of the dataset (one for each topic). Each subset was composed of all and only reviews that contained at least one of the words that composed the topic. Using this approach, I assumed that any of the individual words composing a topic could be used by itself to identify a review with that topic. For example, I assumed that any review with the word ‘great’ could be identified as having positive sentiment, and any review with the word ‘terrible’ could be identified as having negative sentiment.

I evaluated the performance of this approach by comparing the proportions of star ratings associated with the reviews in the entire dataset, the ‘positive sentiment’ dataset, and the ‘negative sentiment’ dataset (Table 1). I assumed that reviews associated with star ratings of 4 or 5 could reasonably be identified as having positive sentiment whereas those associated with star ratings of 1, 2, or 3 could reasonably be identified as having negative sentiment. Given this assumption, one would expect the proportions of 4 and 5 star ratings to be higher in the subset of reviews identified using the ‘positive sentiment’ topic than the proportion of these star ratings in the overall dataset. Similarly, one would expect the proportions of 1, 2, and 3 star ratings to be higher in the subset of reviews identified using the ‘negative sentiment’ topic than the proportion of these star ratings in the overall dataset. Interestingly, these are in fact the results I obtained (Table 1), which suggests that this approach successfully identified reviews having positive or negative sentiment. These results surprised me since the approach does not account for context when discriminating topics, but instead relies only on individual words.

Star Rating	Proportions for Entire Dataset	Proportions for 'Positive Sentiment' Dataset	Proportions for 'Negative Sentiment' Dataset
5.0	39.3%	43.3%	12%
4.0	26.1%	29.4%	15.6%
3.0	13.3%	13.7%	18.4%
2.0	9.4%	7.8%	19.1%
1.0	11.9%	5.9%	34.9%

Table 1

Finally, in order to practice applying statistical data analysis, I computed the mean star rating for two categories of restaurants (Sushi Bars and Mexican restaurants) and performed a t-test to evaluate whether the difference in mean star rating is statistically significant. The mean star rating for Sushi Bars was 3.78, and the mean star rating for Mexican restaurants was only 3.70. For the hypothesis test, the null hypothesis was that the mean star rating is the same for the two categories, and the alternative hypothesis was that the mean star rating is not the same for the two categories. The test produced a t-score of 13.33 with a p-value approximating 0, which is considerably smaller than the threshold I used of alpha = 0.05. The test thus provided significant reason to reject the null hypothesis and conclude that the mean star ratings for Sushi Bars and Mexican restaurants is different in the population. Sushi Bars are on average rated more highly than Mexican restaurants.