NLP Analysis of Yelp Restaurant Reviews
Ankur Vishwakarma
Metis SF Winter 2018

Project Goal - Analyze all restaurant reviews via topic modeling to determine main topics for positive and negative reviews. With this information for all reviews, each restaurant was assigned topic weights to allow for comparison between restaurants to identify their positive attributes and areas of improvement.

1. Exploratory Data Analysis

The dataset came from Yelp and included 5.2 million reviews in total. Of those, 130,000 reviews of restaurants in the state of Ohio were kept for analysis. 1 and 2-star reviews were “negative” and 4 and 5-star reviews were “positive”.

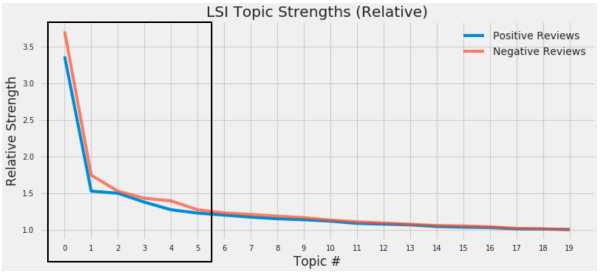
2. Tokenization Process

Documents (reviews) were tokenized using TFIDF and 1-grams, which gave the most consistent results and distinct topics.

For topic modeling, the following approaches were used:

	Notes
LSI	Negative word weights in resulting topics were not easily interpretable for this application. Topic strengths from LSI were used to determine # of topics.
LDA	Topics had lots of overlap in meaning among each other, which would be the result of relatively short document size.
NMF	Gave most interpretable results and separable topics. Rest of the analysis was done using NMF.

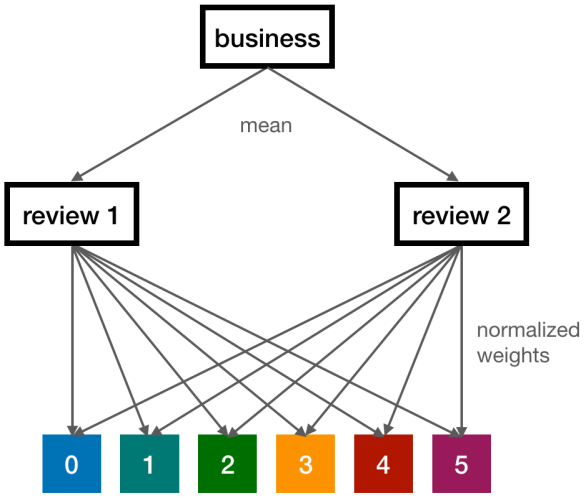
There were 1 set of topics for positive reviews and 1 set of topics for negative reviews. The elbow plot below shows the relative strengths of the topics and that 6 topics were chosen for the analysis.



3. Mapping All Reviews

The process here was:

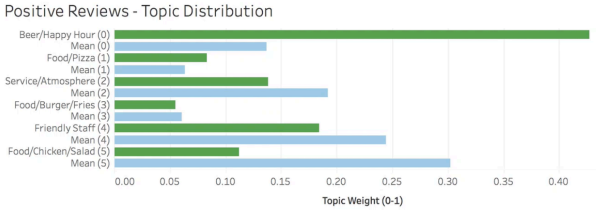
1. Map all reviews to topics.
2. Normalize topic weights to sum to 1.
3. Average topic distributions for all reviews of a restaurant to map that restaurant to the topic space (shown below).



4. Save information to CSV files.
5. Visualize data using Tableau.

3. Visualizing The Results

The data can be used to show topic strengths for positive & negative reviews and how they compare against other restaurants in Ohio.



The interactive visualization on Tableau public can be reached at this link.