Data Mining Capstone Task 1: Exploration of Data Set

Introduction

The task given was to explore the Yelp data set to get a better understanding of common themes in restaurant reviews. In so doing, we needed to visualize the topics for all reviews (Task 1.1) and for two subsets of the data (Task 1.2).

I used the **entire set of restaurant reviews** for Task 1.1, and a subset of restaurants that could represent **North American cuisine** for Task 1.2. To explore the data set and model the topics, I used a combination of the **R programming language** and **Java**. To visualize the topic models, I used the **D3 JavaScript library**.

Distribution of Star Ratings
(Entire Restaurant Data Set)

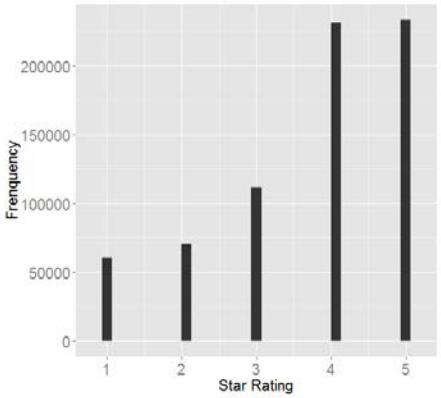

Figure 1: There are clearly a lot more positive reviews than negative ones.

Implementation

Exploring Data Set

First, to get to know the data, I imported the various files in R and used the "jsonlite" library to convert them to data frames. I then used functions such as "dim" and "str" to understand what type of data was available. Considering the tasks in mind, the "yelp_academic_dataset_review.json" and "yelp_academic_dataset_business.json" files quickly stood out as key files for this task.

To reduce memory constraints during analysis, I used the business data set to determine "Restaurant" business ids, and created a subset of reviews only for restaurants. Restaurant reviews totaled 706,646. To further reduce the object size, I only kept the "stars", "text", and "business_id" attributes.

Wanting to explore the distribution of stars and cuisines to decide on a topic for Task 1.2, I plotted a histogram of star frequencies (figure 1) and cuisine frequencies (figure 2). The cuisine frequencies used data from the "categories" attribute in the business data set.

Distribution of Restaurant Categories
(Top 28 Categories)

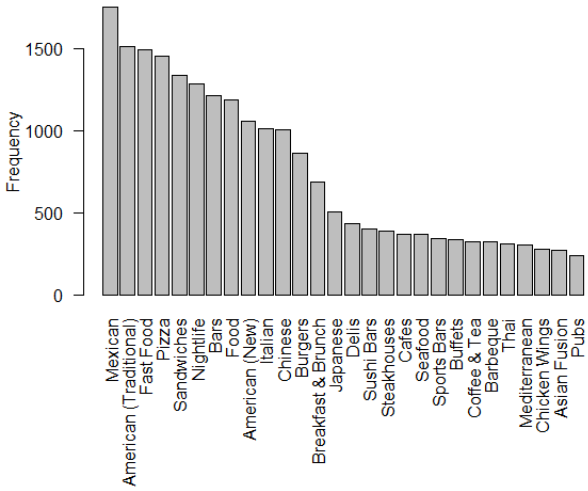

Figure 2: The data set appears to represent a wide variety of cuisines

Data Pre-processing

Before applying a topic model, I processed the reviews using several functions available in the R "tm" package:

1. First, all words in the reviews were converted to lower case using the "tolower" function.
2. Then, stop words such as "the", "a" were removed using the "removeWords" function. I decided to use the "SMART" dictionary available through the "stopwords" function, which is more comprehensive that the "en" dictionary.
3. After that, punctuation, numbers, and white spaces were removed using the "removePunctuation", "removeNumbers", and "stripWhitespace" functions, respectively. I also used "gsub" to remove white space at the beginning and end of each review.
4. Finally, I used the "stemDocument" function to stem words. This last step required installation and loading of the "SnowballC" library, and conversion of the reviews to a corpus using the "Corpus" function (tm package). Once stemming was finished, reviews were converted back to a data frame format, ready for the topic model.

Topic Model – All Reviews

I decided to use Latent Dirichlet Allocation (LDA) because R contains several packages that enable this type of topic modeling. Considering the size of the data set, I first created a sample of 10,000 reviews and tested the speed of each of the following packages: "lda", "topicmodels", and "mallet" (using similar parameters). The "mallet" package in R, which uses the Java-based MALLET package, performed much faster than the other two (roughly 3x faster than "topicmodels" and 3.5x faster than "lda") so this is the package I decided to use. Mallet uses Gibbs sampling.

Topic models were created for 10, 20, and 30 topics using the entire data set of reviews. For the visualization, I decided to use the model with 30 topics because it appeared to offer better segmentation. Hyper parameters were optimized every 200 iterations, after 400 burn-in iterations. Iterations for the modeling were set to 2000. The visualization of this model uses the top 7 words from each topic (see figure 5).

Topic Model – North American Cuisine

To focus on "North American" cuisine, I pulled out all reviews from restaurants found in the following categories: American (Traditional), Fast Food, Burgers, Delis, Sandwiches, and Steakhouses. The total number of reviews was 152,783. I then decided to divide this into two subsets: 5-star reviews and 1-star reviews. I did not include 2- to 4-star reviews because I wanted to get an idea of the extremes. There were 53,478 positive reviews (i.e. 5-star), and 12,361 negative reviews (i.e. 1-star) for the selected restaurant categories.

The mallet package in R was used to model topics for the above two subsets. In this case, I set the number of topics to 10, optimized hyper parameters every 20 iterations (after 50 burn-in iterations), and trained the model with 2000 iterations. The visualization of this model uses the top 9 words from each topic (see figure 6).

Visualizations

The D3 JavaScript library was used for visualizations for both tasks. In Task 1.1, black and blue are used alternatingly simply for aesthetic purposes, but do not highlight any difference between topics. The same was done for positive reviews in Task 1.2. For negative reviews, I decided to alternate between orange and red.

To show the relationship of term weights within a given topic, I decided to not only use a color gradient, but also use the radius length of each circle. I normalized the weights by setting the radius for the term with the highest weight to 4.5. Other radii were determined with a self-created formula[1]: the smaller the radius, the greater the difference between the frequency of that term and the term with the highest frequency in that topic. This method has the added benefit of enabling the color-blind to more easily visualize the weight relationships between same topic terms. Figures 3 and 4 show an example of differences between topics and how they are represented using the color gradient and different size radii.

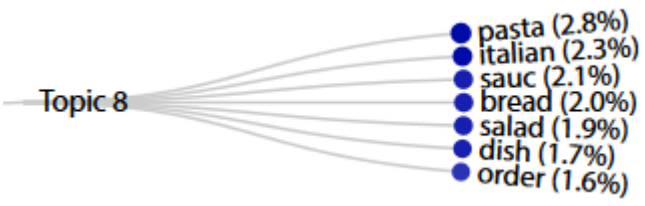

Figure 3: When term weight is similar, circles are relatively the same color and radius.

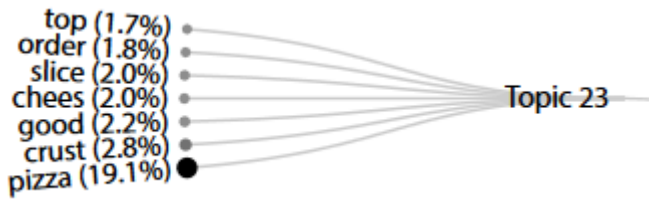

Figure 4: When there is a large difference between term weights, the color gradient becomes lighter and the radius becomes smaller.

1: Formula: `r = 4.5+1.1*log("term weight"/"top term weight".`

Conclusion

Word distributions in both visualizations appear to create relevant topics by grouping similar terms together. Figure 5, which shows a topic model for all reviews, separates types of cuisines very nicely. For instance, it's clear that topic 3 is about Japanese cuisine, topic 8 is Italian cuisine, topic 9 is about breakfast, topic 19 is Mexican cuisine, etc. Figure 6 does the same. For instance, the term "wait" appears in two different topics for the negative reviews model and only in one topic for the positive reviews model. Similarly, "good" appears in six different topics for the positive reviews, and in only one topic for the negative reviews. To verify these differences, I took a sample of 10,000 reviews from each subset and compared the frequency of the following terms found in the visualization that I felt might reflect a positive or negative sentiment: great, good, love, fresh, time, wait, busi, bad, dirti, and hair. The frequencies of these terms, as seen in figures 7 and 8, confirm the differences seen in figure 6. It seems safe to conclude that both tasks helped provide a high-level understanding of reviews, including an understanding of the different topics into which these could be classified.

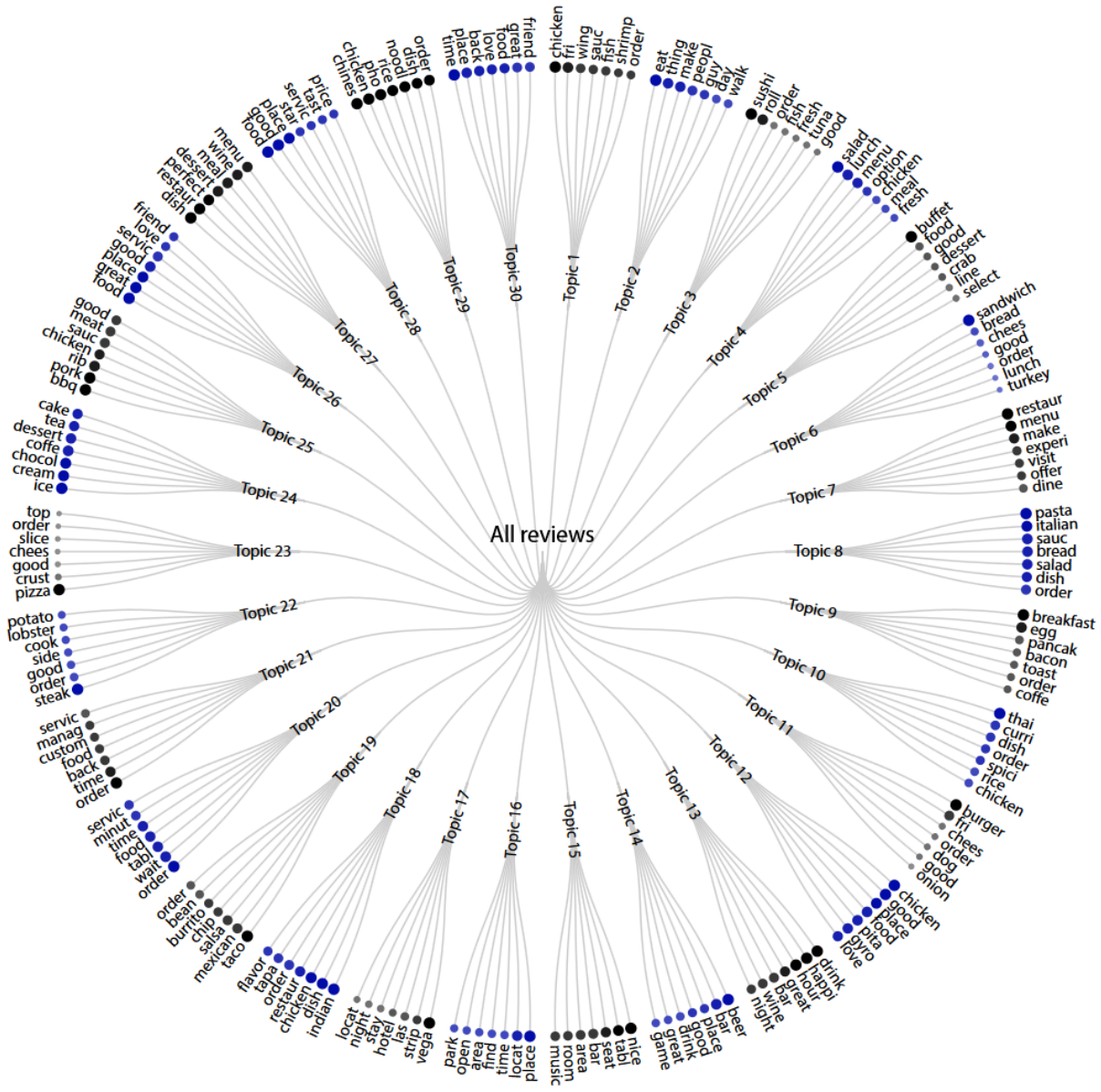

Figure 5: The above radial tree diagram represents over 700,000 reviews modelled into 30 topics using LDA.

Data Mining Capstone Task 1: Exploration of Data Set

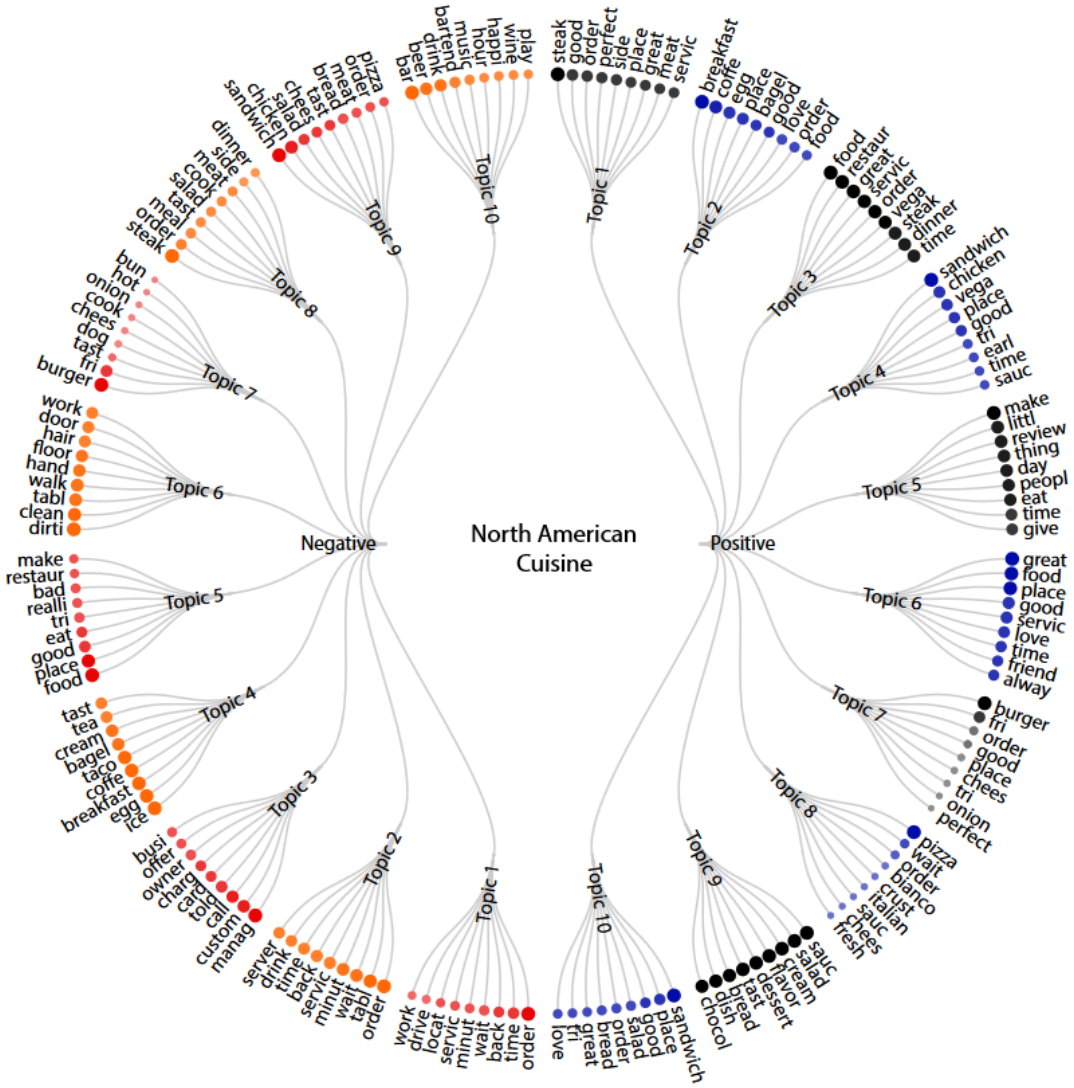

Figure 6: The above radial tree diagram represents over 50,000 positive reviews (5-star), and over 10,000 negative reviews (1-star) modelled in 10 topics each using LDA.

Distribution of Selected Words in Negative Reviews (n = 10,000)

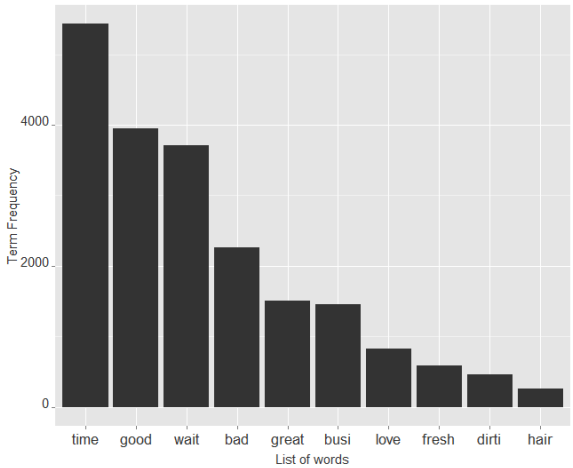

Figure 7: Bar plot showing the frequency of selected key words in 10,000 negative reviews of N.A. cuisines.

Distribution of Selected Words in Positive Reviews (n = 10,000)

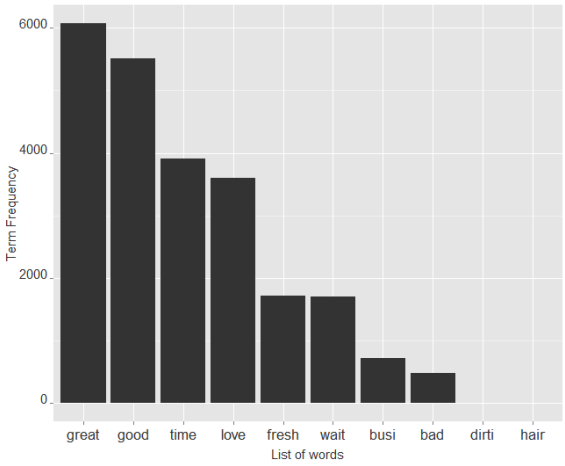

Figure 8: Bar plot showing the frequency of selected key words in 10,000 positive reviews of N.A. cuisines.