Data Mining Capstone Task 1

January 26, 2020

Asad Bin Imtiaz
CS598 Data Mining Capstone
MCS-DS UIUC
Email: aimtiaz2@illinois.edu

1 Introduction:

In this report, the approach for analyzing Yelp Review Academic Challenge Dataset described and the results for this analysis are presented. The analysis and findings are part of Task-1 for Data Mining Capstone course wherein the Yelp restaurant reviews were explored to uncover interesting patterns and knowledge such as frequent topic word distributions and cuisine landscape analysis. The goal of this task was to explore the data to understand the restaurant review characteristics such as main topics in the reviews and the distribution of star ratings.

2 Data exploration:

This section discusses the tools and approach used for the analysis and visualizations.

2.1 Tool and Libraries:

The processing of Yelp restaurant dataset was performed using Python 3. Several python libraries such as Spacy, Gensim and Sklearn were used to analyze the review texts and perform Natural Language Processing (NLP). For visualization, D3.js and python matplotlib were used.

2.2 Data Preprocessing

In order to get better results from NLP tasks, the data set was pre-processed to enhance distinctive features and to reduce noise such as stop-words and URLs. The following operations were performed in the pro-processing step:

- Any URLs present in the review texts were removed.
- Only reviews with category ‘Restaurants’ were retained.
- The review text was lemmatized, i.e. all tokens were brought in their base lemma form.
- Frequently occurring bi-grams and tri-grams were captured. These enabled a better of review topics which would be discussed in this document later.
- The stop-words were removed from the review test.
- Punctuation and symbols such as emojis were removed
- For better sentiment analysis, only parts of speech of type Nouns, Verbs, Adjectives, Adverbs, determiners and Adverb-participles were retained.

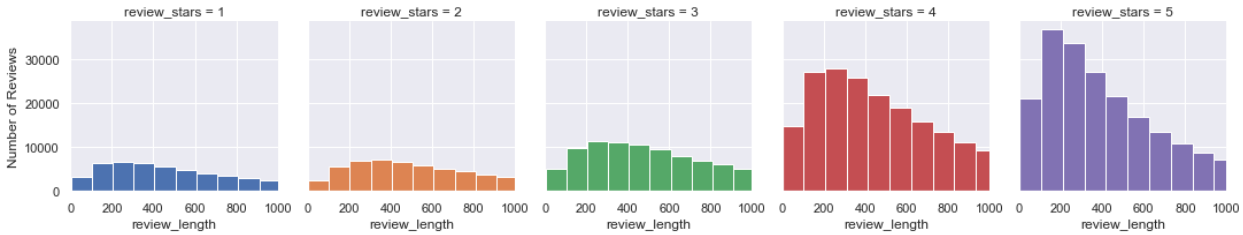
It is noteworthy that the stop-words were only removed after creating n-grams to retain words such as ‘not’ in phrases like ‘not bad’ for more accurate topic mining.

3 Understanding the data:

Before dining into the core tasks of Topic mining (chapter 4) and rating analysis (chapter 0), it was important to get a sense of data and its distribution.

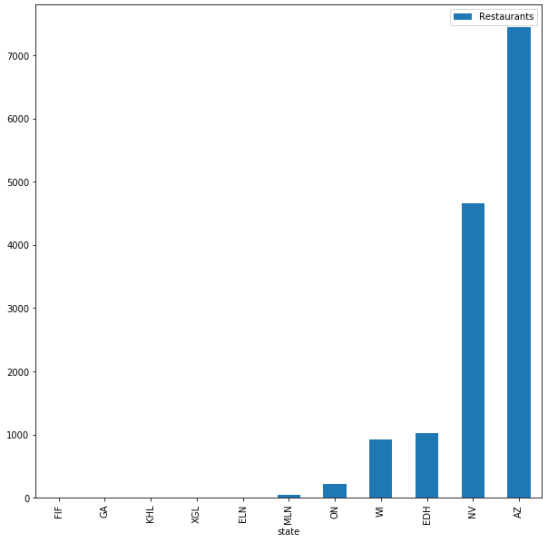
The first thing that was analyzed was the distribution of reviews with respect to the star ratings. It was discovered that the reviews were quite unevenly distributed among star ratings and were skewed towards the higher stars, as can be seen in the chart below:

One can conclude from this that users mostly refrained from writing reviews when they did not have much positive to say about the restaurants. This insight was corroborated by the fact that the length of reviews was also not evenly distributed and that the higher star rated reviews were more verbose than those with lower stars, as can be seen in the chart below:



This suggests that in general, positive reviews were lengthier and verbose to some extent than negative reviews.

Moreover, the reviews were also highly skewed when it comes to distribution per state, with the state of Arizona having more reviews than any other in the dataset.



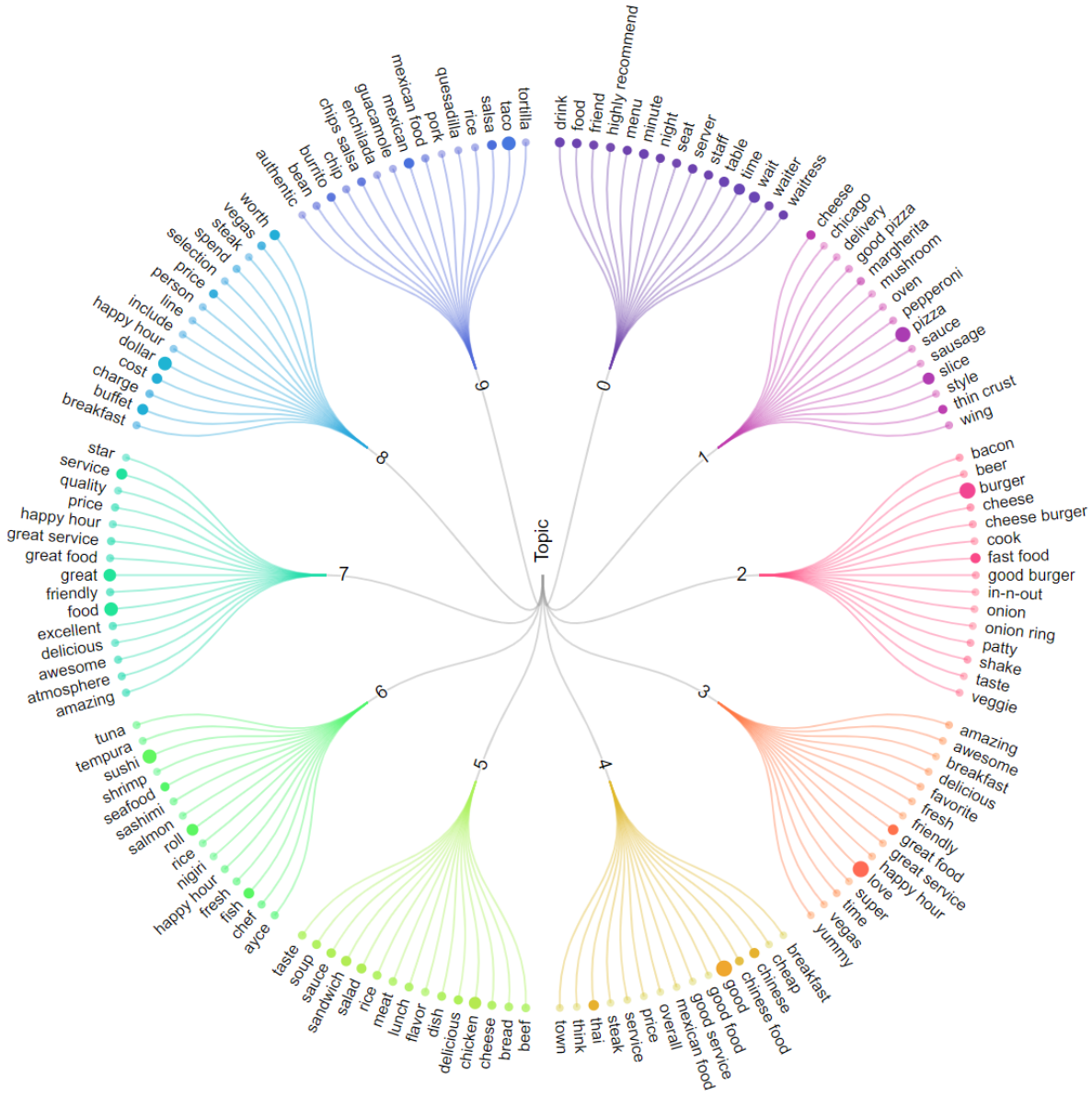
business_id	categories	city	review_count	name	stars
state					
AZ	7439	7439	7439	7439	7439	7439
NV	4655	4655	4655	4655	4655	4655
EDH	1016	1016	1016	1016	1016	1016
WI	916	916	916	916	916	916
ON	219	219	219	219	219	219
MLN	50	50	50	50	50	50
ELN	4	4	4	4	4	4
FIF	1	1	1	1	1	1
GA	1	1	1	1	1	1
KHL	1	1	1	1	1	1
XGL	1	1	1	1	1	1

4 Task 1.1: Topic Modeling

The entire pre-processed data set was used to mine topics as described in chapter 2.2. The topic mining was done in Python 3 with Gensim library using Latent Semantic Indexing (LSI) and Non-Negative Matrix Factorization (NMF) models. Ten topics models were extracted, and top 15 words for each topic were projected to describe the topic.

To extract the Topics, TFIDF-Vectorizers were instantiated using stop words from Spacy. A maximum features limit of 500,000 was used. The minimum document frequency was set to 10 while the minimum document length was 3 tokens.

The extracted Topics were visualized using D3.js. A radial dendrogram was generated for the resulting topics as below:



The diameter of a node associated with a word/phrase corresponds to its probability in the topic. Additionally, the opacity of a node also corresponds to its probability weight.

It can be seen that most of these topics were clearly distinctive, such as the words like ‘server’, ‘staff’, ‘wait’ & ‘time’ indicate a topic related to staff and service where as words such as ‘tuna’, ‘sushi’, ‘seafood’, ‘fish’ etc. in topic 6 indicate it is about sea-food.

To better analyze the extracted topics, word clouds of the top 15 words/phrases of each topic were generated using D3, as below:

Topic 0

Topic 1

Topic 2

Topic 3

Topic 4

Topic 5

Topic 6

Topic 7

Topic 8

Topic 9

Furthermore, 4 regional cuisines namely ‘Italian’, ‘Thai’, ‘Chinese’ and ‘Indian’ where also analyzed for word distributions. These cuisines were selected at random and have reasonable number of reviews in them. Every cuisine has some distinctive dishes, flavors and ingredients, which, if the topic generation is correct, should appear in the frequently in the topic of that particular cuisine. This was done by generating word-clouds of top 50 words of all reviews in 4 topics generated for each of the 4 selected cuisines.

The analysis was done to check the accuracy of the topic modelling, as one expects to see several distinct high frequency words for a particular type of cuisine. For example, among the most common words in Italian reviews for Italian cuisine were ‘Pizza’, ‘Salad’ and ‘Cheese’, whereas in Thai reviews were words like ‘Thai’, ‘soup’, ‘rice’ and ‘curry’ as in the charts below:

Italian

Thai

The word distribution is inline with what one would expect to find in these cuisines, indicating accuracy of the generated topics.

Similar analysis on the other 2 selected cuisines, namely ‘Chinese’ and ‘Indian’ also found similar results, as in the charts below:

Chinese

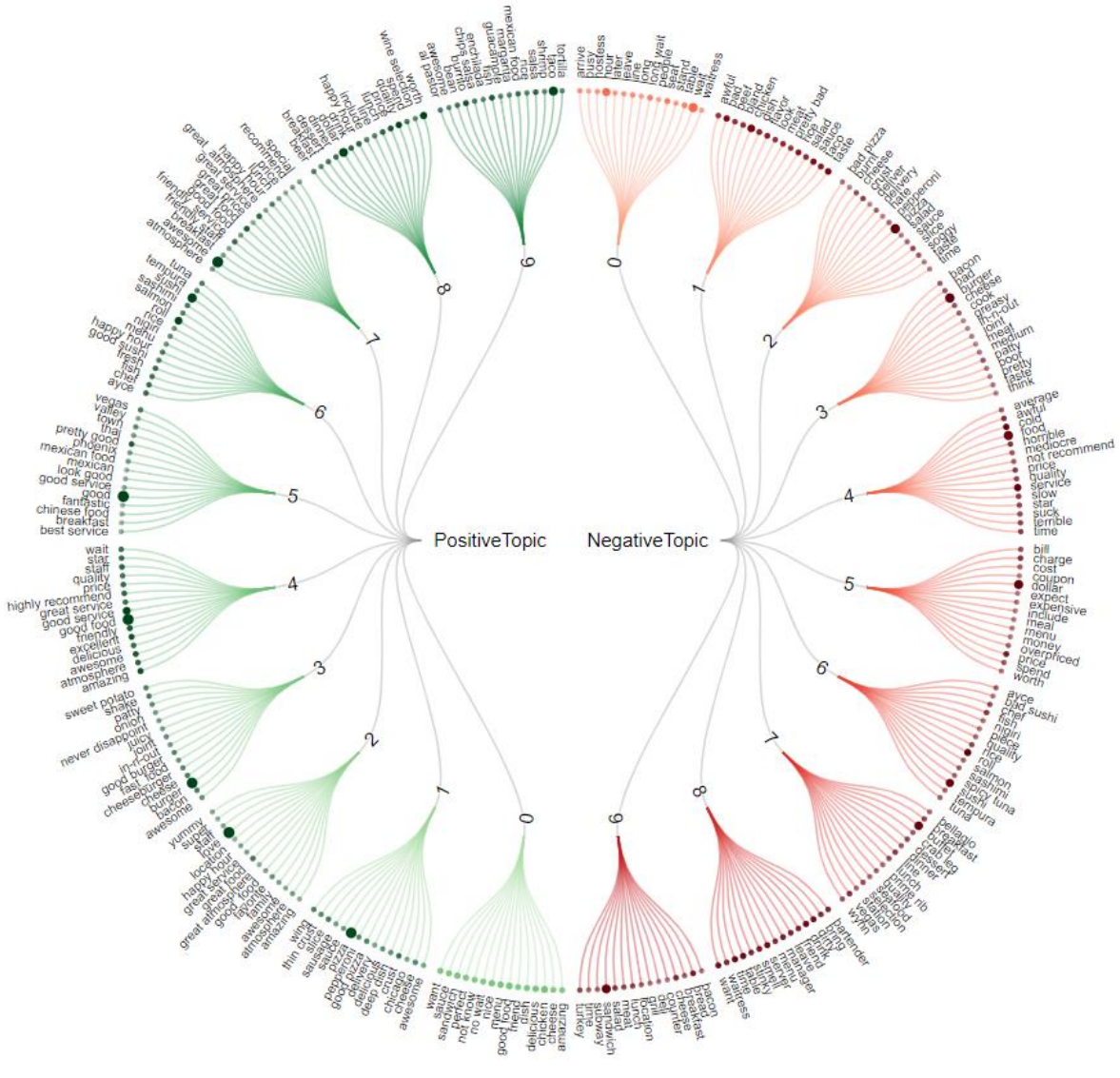
Indian

Here we find words such as ‘chicken’, ‘soup’, ‘noodle’, ‘roll’ and ‘dim sum’ among others, which are clearly associated with Chinese cuisine. On the other hand, among the frequent word in Indian cuisine where words such as ‘buffet’, ‘curry’, ‘naan’, ‘spicy’ and ‘tikka masala’, indicating correctness of the generated topics.

5 Task 1.2: Rating and Cuisine analysis

In this task Rating analysis of Yelp restaurant reviews was performed. The entire pre-processed data set was used to mine topics as described in chapter 2.2. The topic mining was done in Python 3 with Gensim library using Latent Semantic Indexing (LSI) and Non-Negative Matrix Factorization (NMF) models. For rating analysis, the start rating of above 3 was considered **Positive** while that of below 3 was considered **Negative**. Ten topics for each Positive and Negative were extracted, and top 15 words for each topic were projected to describe the topic.

To extract the Topics, TFIDF-Vectorizers were instantiated using stop words from Spacy. A maximum features limit of 500,000 was used. The minimum document frequency was set to 10 while the minimum document length was 3 tokens. The extracted Topics were visualized using D3.js. A radial dendrogram was generated for the resulting topics as below:



The diameter of a node associated with a word/phrase corresponds to its probability in the topic. Additionally, the opacity of a node also corresponds to its probability weight.

It can be seen that there are phrases and words such as ‘never disappoint’, great service’, ‘highly recommend’, ‘yummy’ etc. on the positive side, where as phrases and words such as ‘not

recommend', 'overpriced', 'horrible' etc. are present on the negative side, indicating correctness of the positive & negative split. Note that all topic words are in their base lemma form.

Additionally, an analysis on distribution of reviews per cuisine/food category was done and top 20 categories were visualized in the chart below:

Category	Review Count
American New	94183
Nightlife	87471
American Traditional	84895
Bars	82511
Mexican	70406
Breakfast & Brunch	61348
Italian	60976
Pizza	57500
Food	55668
Sandwiches	48716
Steakhouses	45338
Burgers	43431
Japanese	39553
Chinese	38716
Sushi Bars	38259
Buffets	32468
Seafood	31734
Fast Food	22291
Thai	20691
Asian Fusion	20084

Each bubble represents a category or cuisine in the top 20 categories while its diameter corresponds to the number of reviews in this category.

It can be seen that the top 5 categories with most reviews were 'American New', 'Nightlife', 'American Traditional', 'Bars' and 'Mexican' respectively. The word distribution of these 5 categories were analyzed for both negative and positive reviews to understand the top most aspects discussed for these categories.

This was done by generating word/phrase distributions and visualizing them in word clouds of 50 words. Two word-clouds, one for positive reviews and other for negative reviews, for each of the top 5 categories were generated using D3.js, as in the following:

American New (Positive)

American New (Negative)

Nightlife (Positive)

Nightlife (Negative)

American Traditional (Positive)

American Traditional (Negative)

Bars (Positive)

Bars (Negative)

Mexican (Positive)

Mexican (Negative)

From the word clouds, it can be seen that the words such as ‘wait’, ‘time’ and ‘service’ appeared quite frequently in almost all the negative reviews, while the positive reviews were more diverse and also discussed other aspects such as cuisine, service and value frequently. One can calculated from this that most of the complains were about the service in negative reviews, regardless of their category, while in positive reviews, no aspect dominated.

6 Conclusion

The major takeaways from the analysis were:

1 Cleansing, preprocessing and phrase mining of reviews before mining topics generated and reasonably accurate topics.
2 Selecting particular cuisines and generating topics resulted in word distributions which describe the cuisines reasonably well.
3 The fact that positive and negative reviews were skewed suggested that users are more like to write a review if they like the business/food than if they would not like it.
4 The distribution of the length of the review indicated that positive reviews were in general longer than negative reviews suggesting if user would not have much to say positively about the business, they would not write a lot in the review.
5 The negative reviews focused more on the service and staff than on other aspects.