**Name**: Ajanthan Mathialagan

**Student Number**: 500768117

**Superviser**: Bilgehan Erdem

**Date**: Dec 3rd 2021

**Sentiment Analysis Of The Businesses On Yelp**

**Abstract**

Yelp is a crowd sourced local business review and social networking site. It is mainly for context on locations such as restaurants, schools, etc, where yelp users can submit a review of their products or services they received using their respected five star rating system. Accordingly, Yelp has released an open, all-purpose data set for learning. The dataset contains information of businesses, reviews and user data to inform government policy, academic research and form business strategies. The objective is to find out why users love to go to these top rated restaurants based on the common reviews and tips. I’ll use a NLP method known as topic modeling. Moreover I'll be using a specific topic modeling technique known as Latent Dirichlet Allocation (LDA), which is a type of topic model used to extract topics from a given set of text. It will be used to uncover hidden structures from the texts provided in the tips data. Find the positive and negative comments made by users on these businesses and the top common words and topics that show up in the reviews. The main tools I'll be using in this project will consist of Python using pandas and numpy for altering and working with the yelp dataset, creating LDA models and for the visualization packages such as seaborn, mathplotlib for making plots of the word counts, word clouds, etc.

**Literary Review**

Natural Language Processing is a field in machine learning where we grant the computer the ability to understand, analyze, manipulate and potentially generate the human language. An example of a company that uses NLP is google. In the case of information retrieval Google finds relevant and similar results when searching for a topic using their search engine. In addition, information extraction. Gmail, the online emailing service provided by google, extracts information from the emails received and structures events from the emails. I will be providing a sentiment analysis which is a subfield of Natural Language Processing. It is the procedure used to determine whether a chunk of text is either positive, negative or neutral. Sentiment analysis basically breaks the text down into topic chunks and then assigns the chunks to a sentiment score for each topic. Companies and brands often utilize this method to monitor brand reputation across social media platforms or the web as a whole. One of the most widely used applications for sentiment analysis is for monitoring call centers and omnichannel customer support performance.

LDA which is known as Latent Dirchlet allocation is a type of topic model where each topic will be represented by a set of words. The goal of LDA is to map all the reviews made by the yelp users to a topic in a way that the word in each topic is in the imaginary topic. Yelp is a social networking platform that lets users post reviews and rate businesses. When a customer goes to a new restaurant and has a good or bad experience, in most cases the customer will not keep it a secret. Using the data set yelp has provided which is a subset of businesses, reviews, and user data for use for personal, educational and academic purposes. Moreover, using the data set for performing sentiment analysis, finding customer trends and popular restaurants amongst the 8 metropolitan cities, finding what the top restaurants in each metropolitan city have in common based on the reviews presented. In addition what are the common tips users from yelp have given these restaurants for improving their business.Since Yelp’s data set is a common textual data set many use for educational purposes. There are a variety of projects that can be found through the web concerning the use of NLP on the yelp data set. For instance, analyzing the cultural impact of social commerce using the restaurants provided in the data set (Nakayama & Wan, 2015). This paper emphasizes the use of bilingual text mining software used to demonstrate that Japanese customers have a different sentiment distribution on the four basic attributes of dining experience which are food, price, service and quality.

In correspondence to topic modeling, there are many approaches used for studying abstract topics in a collection of text. LDA is a common unsupervised method to uncover hidden topics from the selected texts. For instance, capturing word choice patterns with LDA for fake review detections. (Lee, Han & Myaeng, 2016). This paper applies LDA to capture aspects of fake and truthful reviews by means of "topics" that are not necessarily subject areas but related to the word choice patterns reflecting behavioral and linguistic characteristics of the fake review writers. In addition, many have taken other approaches other than LDA to answer a similar business problem but using a different approach. For instance, identifying restaurant features using models based on Support Vector Machines (Yu, Zhou, Zhang, & Cao, 2015). This paper conducts a novel analysis on the restaurant features provided, the overall sentiment polarity that shows a preference on service in the reviews, which might allude that customers ‘self-select’ the food that they like. In addition, a sentiment analysis based on ratings (Xu, Wu & Wang, 2015). This paper emphasizes the use of Naive Bayes, Perceptron, and Multiclass SVM and compares the predictions with the actual ratings provided.

Using user recommendations to improve a business is a common problem that is solved not just in the food industry, but in many industries such as retail stores, healthcare clinics, etc. In my case, having the ability to visualize what recommendations the users have for businesses, enables those businesses to process these changes to attract more customers, evidently leading to more sales. In addition, finding what are the common trends that are captured through user tips for the top businesses. Moreover, it will tell us why the top businesses stand out compared to the businesses with average to below average ratings.

**Yelp Data Description**

The Yelp Data set is a subset of businesses, reviews, and user data for use in personal, educational and academic purposes. It is presented in JSON files using it to teach students about databases, learn NLP and used for sample production data while learning how to make mobile applications. The Dataset contains 8,635,403 reviews, 160,585 businesses, 200,000 pictures,

8 metropolitan areas, 1,162,119 tips made by 2,189,457 users. Over 1.2 million business attributes such as hours, parking availability and ambience aggregated check-ins over time for each of the 138,876 businesses.

The main data set is composed of 6 JSON files. The business.json, review.json, user.json and checkin.json, tip.json and photo.json. Since my project is based on text mining and sentiment analysis I’ll mainly be using the business.json and tip.json files.

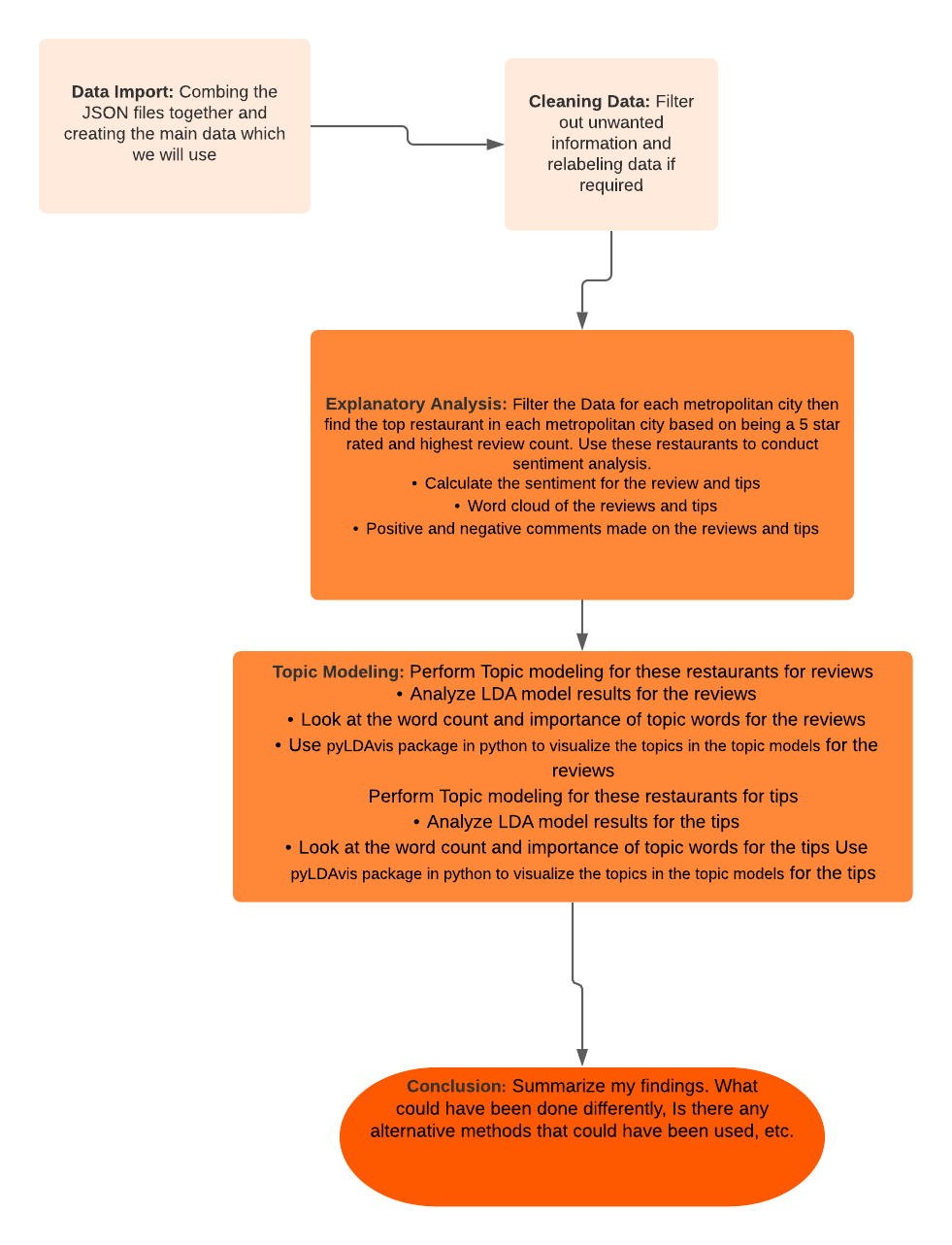
**Business.json** file contains business data, including the location data, attributes and categories.

| **Name** | **Description** | **Data type** |
| --- | --- | --- |
| Business\_id | 22 character unique string character. | string |
| Name | Name of the business | string |
| Address | Address of the business | string |
| City | City of the business, | string |
| Postal Code | Longitude coordinates | integer |
| Latitude | Latitude coordinates |  |
| Stars | Stars which is the 5 star rating system yelp uses | integer |
| Review\_count: | he review count of the total amount of reviews for the business, | integer |
| Is\_open | if the business is open or closed, | string |
| Attributes | if the restaurant provides takeout, business parking, etc, | string |
| Categories | business categories | string |
| Hours | The hours which the business functions Monday through friday | string |

**Tips.json** is tips written by users on a business. Tips are shorter than reviews and tend to convey quick suggestions.

| **Name** | **Description** | **Data Type** |
| --- | --- | --- |
| Text | Text of the tip | string |
| Date | The date the user has written the tip | string |
| Compliment\_count | How many compliments the text has | integer |
| Business\_id | 22 string characters business id maps to business in business.josn | string |
| user\_id | 22 character user id maps to users in user.json | string |

**Methodologies**

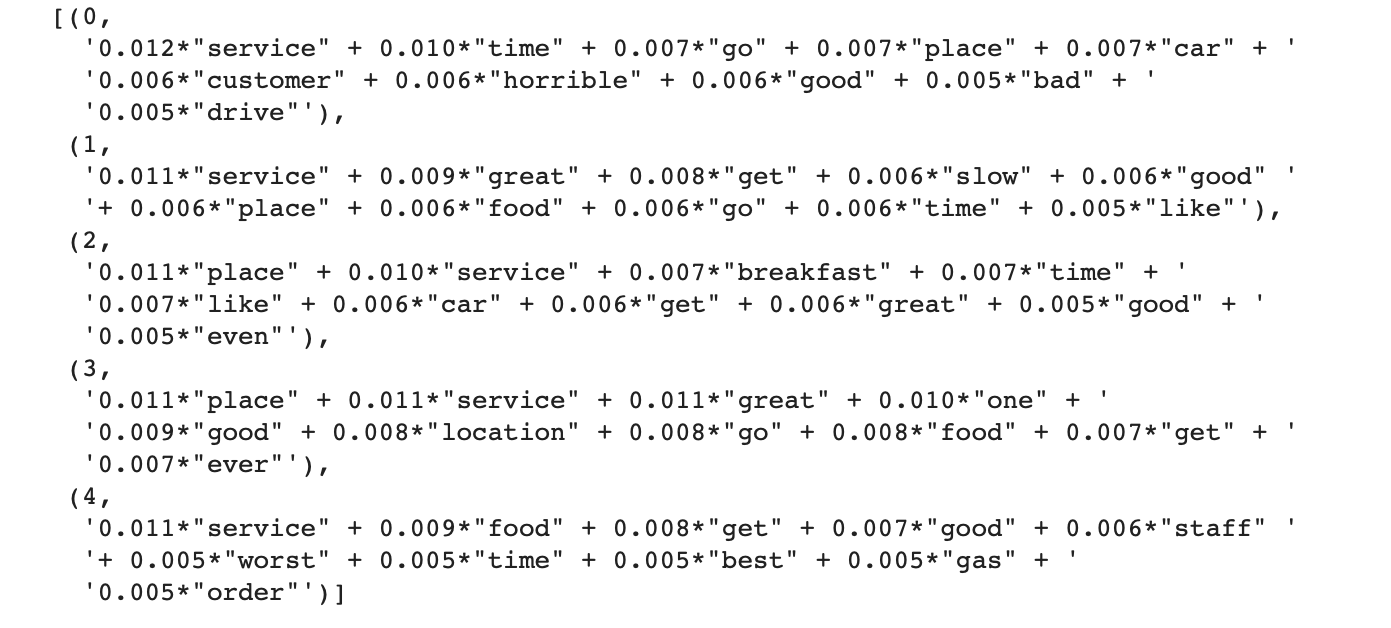
****

**LDA Model:**

I am going to be working with a LDA model. The LDA assumes that the documents are composed of words that help determine the topics and maps documents to a list of topics by assigning each word in the document to different topics. I am going to have 2 models to test my hypothesis one model is looking at the tips given by customers from austin texas to 5 star businesses. My other model will be looking at tips given by customers from austin texas to 2 star and below businesses.

**Figure 1: LDA Model Austin Texas 5 Star Businesses**

**Figure 2: LDA Model: Austin Texas 2 Stars and below Businesses**

****

**Model Evaluation:**

Ill be evaluating the models by looking at the coherence scores. The coherence score is basically assessing the quality of the learned topic for the LDA model. We set the default amount of topics to 5 so our initial coherence score for our model will be based on 5 topics.

**Coherence score LDA model: 5 star businesses in Austin Texas**

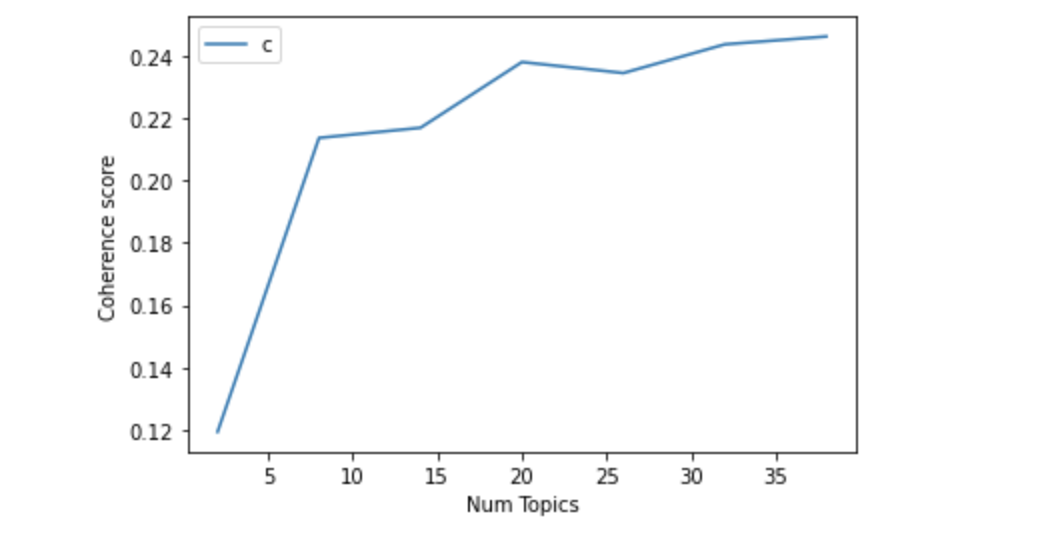


**Coherence score for LDA model: 2 stars and below businesses in Austin Texas**

****

We can see that in general the coherence score for both models is pretty bad. We will want to iterate through a number of topics to see which set amount of topics would be best for our LDA models.

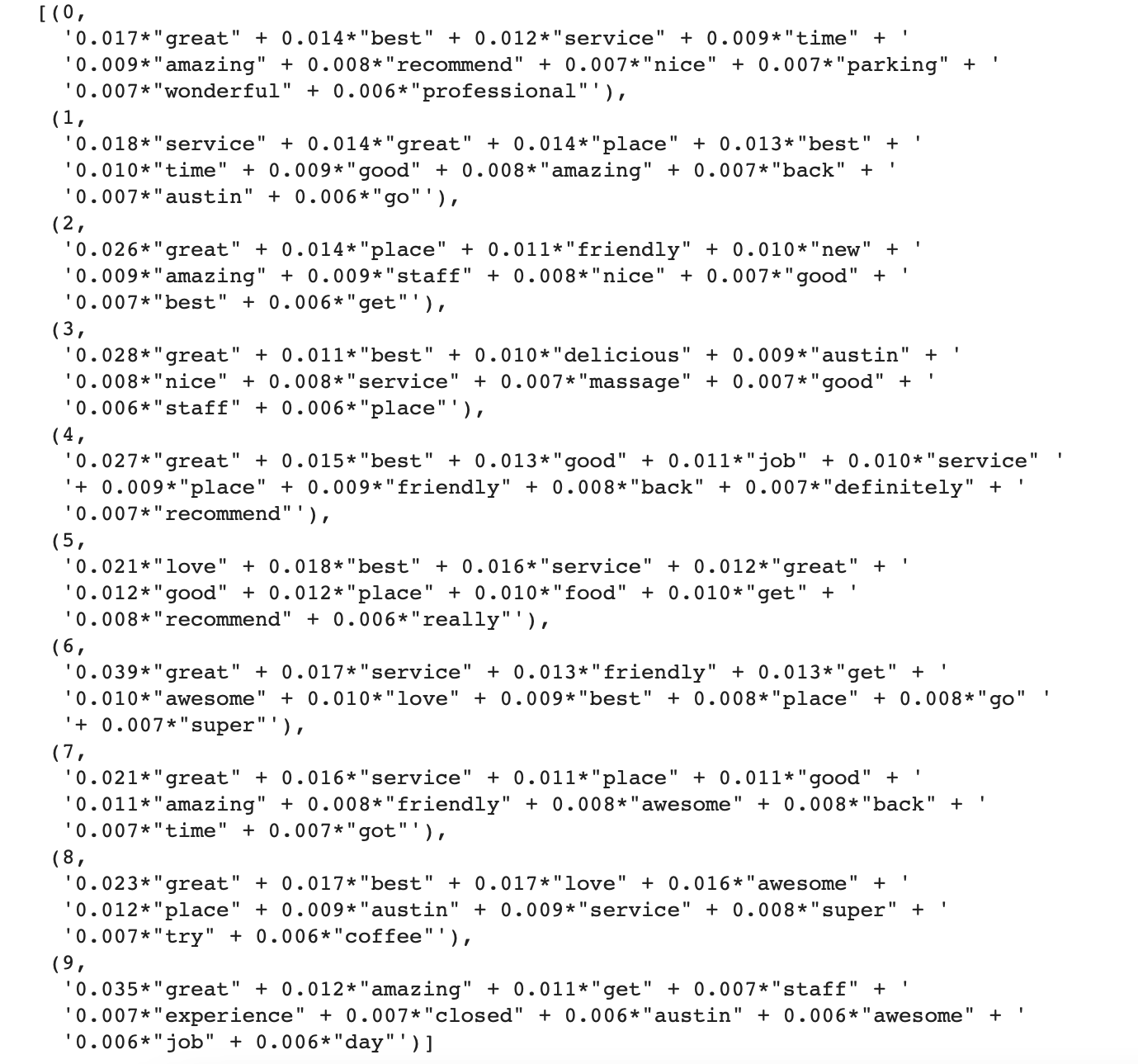
**Figure 3: Coherence Score For Multiple Topics**



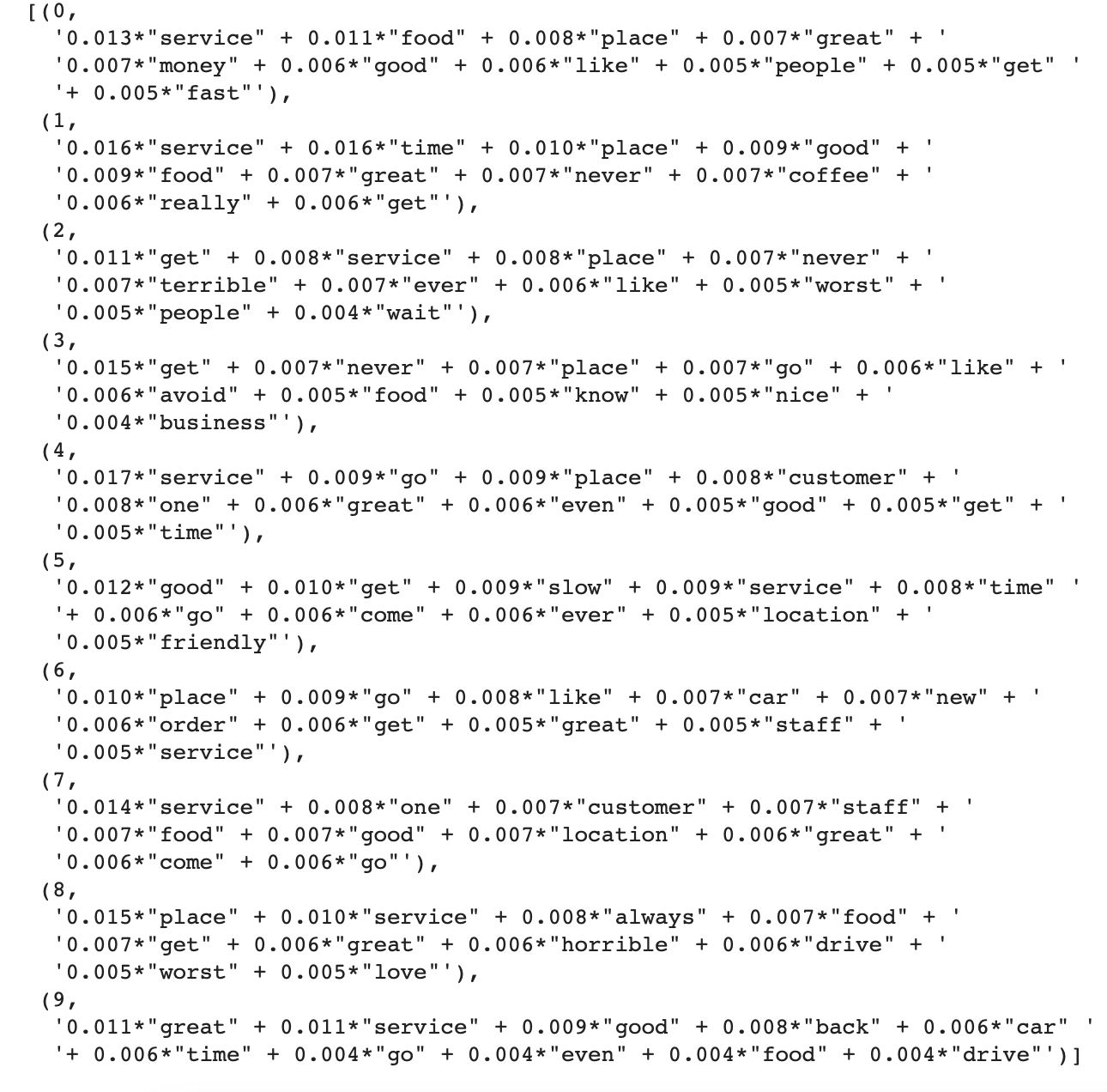
The plot in Figure 3 shows that coherence score increases with the number of topics, with a decline. Now, choosing the number of topics still depends on your requirement because topics more than 5 have good coherence scores but may have repeated keywords in the topic. Topic coherence gives you a good picture so that you can make better decisions. In our case the optimal amount of topics we will choose is 10 since the coherence score raised tremendously at 10 topics choosing more topics has a higher chance of repeating keywords in the topic which we don't want

**Re-evaluated Models:**

**Figure 4: LDA Model Austin Texas 5 Star Businesses**

****

**Figure 5: LDA Model Austin Texas 2 Stars and below Businesses**

****

**Results**

In LDA models, each document is composed of multiple topics. But, typically only one of the topics is dominant. The tables below shows the dominant topic for each sentence and shows the weight of the topic and the keywords in a nicely formatted output. This way, you will know which document belongs predominantly to which topic.

**Figure 6: Table of Dominant topics and words for 5 stared rated businesses in Austin**



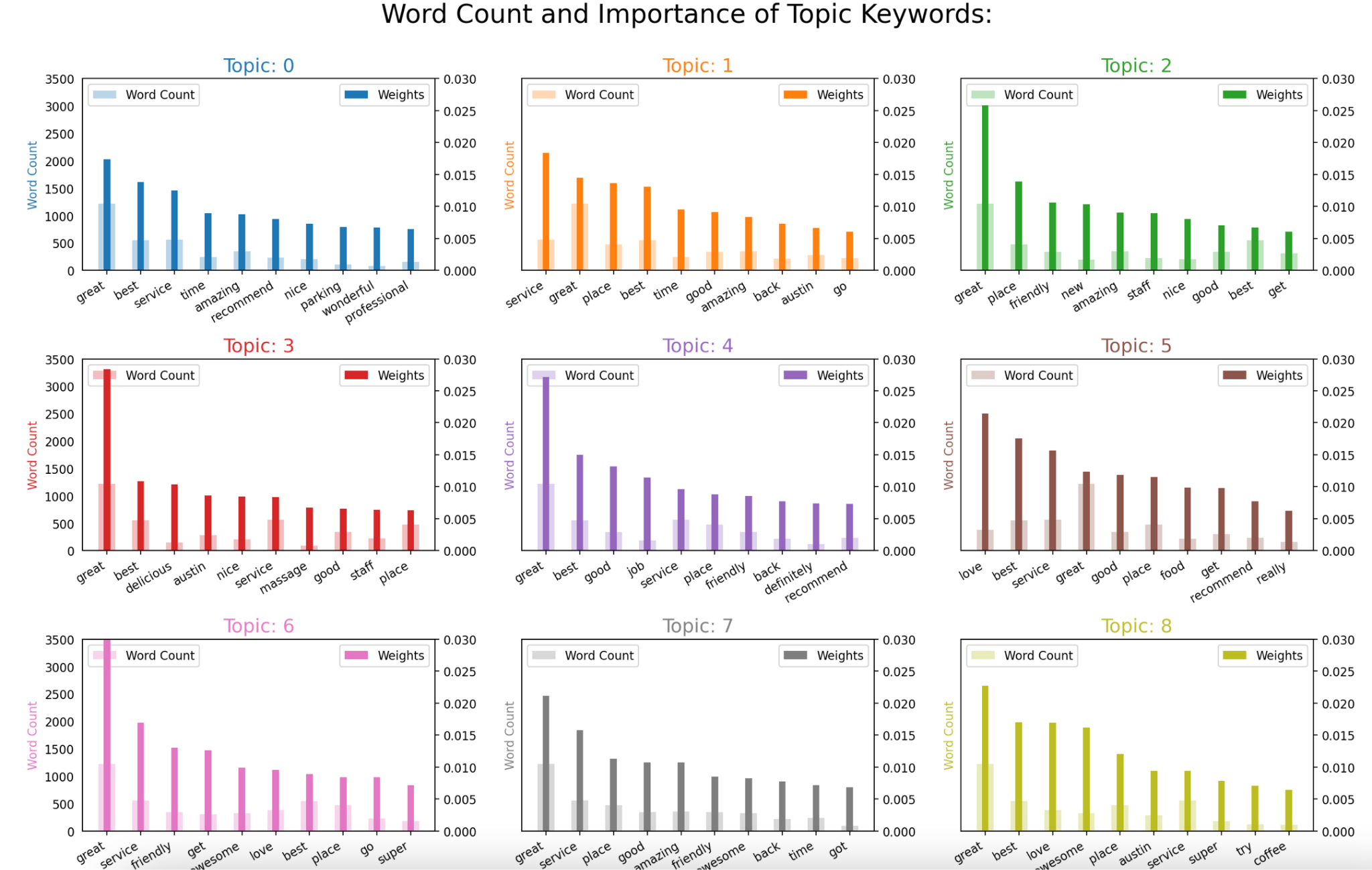
**Figure 7: Table of Dominant topics and words for 2 Star and lower rated businesses in Austin**



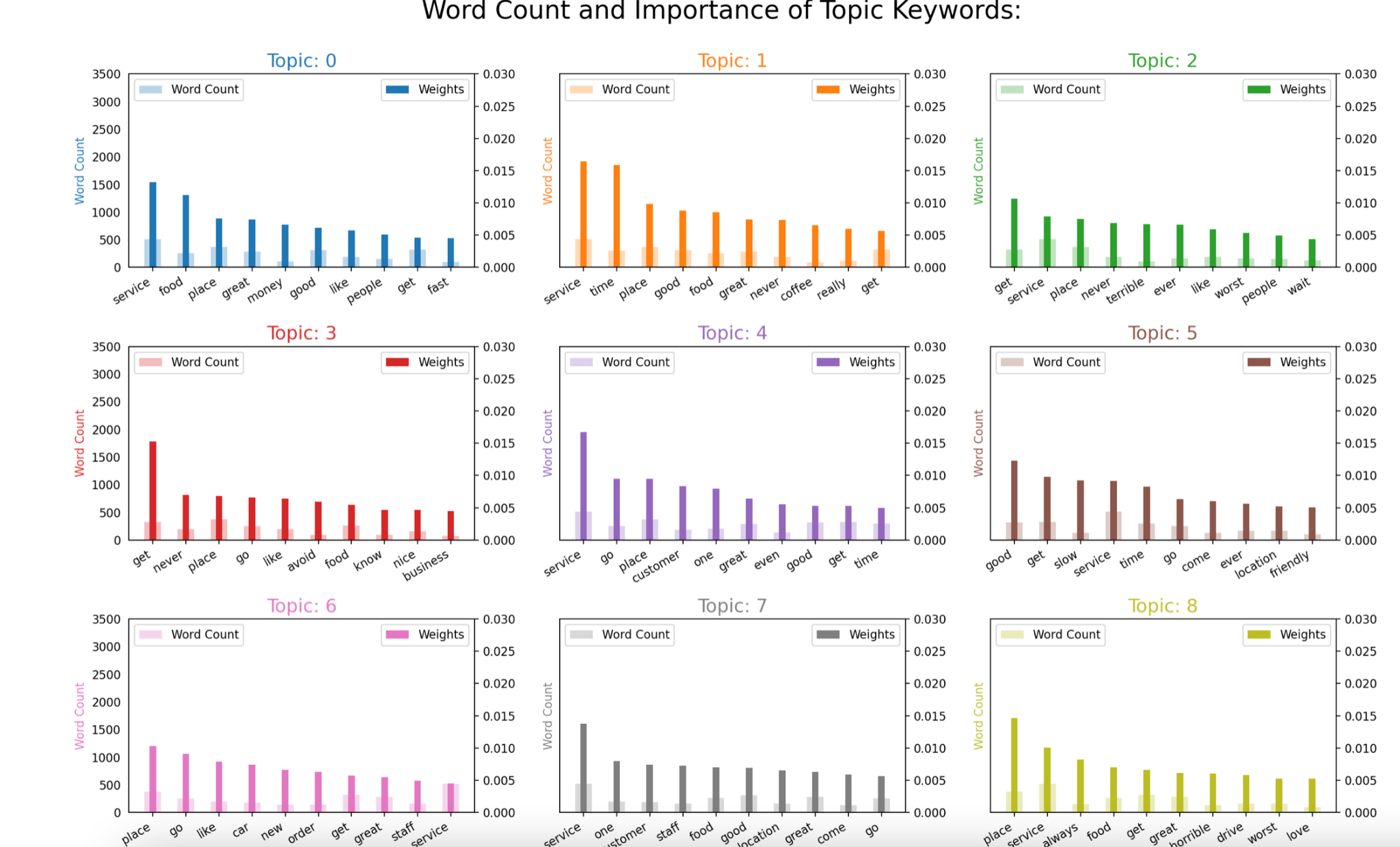
**Visualizing LDA Models:**

Looking at both models provided by the LDA model we can see that the 5 star rated businesses receive tips and compliments about having a positive experience. Words such as great as shown in figure 8 has a high weight over word count which shows that customers had a positive experience. Comparing it to the visualizations shown bas on the LDA model for 2 star rated business and below, shows that there is barely any positive comments in the topics and most of the words that come up involve negative experiences with employee or service provided by that business

**Figure 8: 5 Star Rated Businesses In Austin**

****

**Figure 9: 2 Star and Below Rated Businesses In Austin**

****

**Conclusion**

Looking at the results, through the analysis done by comparing customer tips/comments given to the 5 star rated business and 2 star rated business in Austin. We can see that 5 star rated businesses get more positive words that include positive feedback based on the service they provide to the customer, location of the business, pricing, items they purchased, etc.. The 2 star and lower rated businesses business gets a mixture of both negative and positive comments, but more negative comments regarding their staff, service, time, quality of item, quality of service provided etc. What the 2 star business can do to improve their quality of business is to take proper precaution and work on these complaints left by the customers. Regarding their service, items they are providing, etc, in order to improve customer experience. This will most likely give them a boost in ratings and increase daily customer activities for those respective businesses who are in the lower ends of the scale when looking at the star ratings.

**References**

Xu, Y., Wu, X., & Qinxia Wang, Q. (2015). Sentiment Analysis of Yelp‘s Ratings Based on Text Reviews. Retrieved 23 October 2021, from <http://cs229.stanford.edu/proj2014/Yun%20Xu,%20Xinhui%20Wu,%20Qinxia%20Wang,%20Sentiment%20Analysis%20of%20Yelp's%20Ratings%20Based%20on%20Text%20Reviews.pdf>

Lee, D., Han, K., & Myaeng, S. (2016). Capturing Word Choice Patterns with LDA for Fake Review Detection in Sentiment Analysis | Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics. Retrieved 23 October 2021, from <https://dl.acm.org/doi/abs/10.1145/2912845.2912868>

Yu, B., Zhou, J., Zhang, Y., & Cao, Y. Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews. Retrieved 23 October 2021, from <https://arxiv.org/pdf/1709.08698.pdf>

Nakayama, M., & Wan, Y. (2021). The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews. Retrieved 23 October 2021, from https://www.sciencedirect.com/science/article/abs/pii/S0378720617306225?via%3Dihub

Ma, E. (2018, August 18). *2 latent methods for dimension reduction and topic modeling*. Medium. https://towardsdatascience.com/2-latent-methods-for-dimension-reduction-and-topic-modeling-20ff6d7d547

A. (2018, March 9). *A Very Extensive Data Analysis of Yelp*. Kaggle. https://www.kaggle.com/ambarish/a-very-extensive-data-analysis-of-yelp/report

Yelp dataset. (n.d.). Retrieved September 26, 2021, from <https://www.yelp.com/dataset/documentation/main>.

*Yelp Dataset*. (n.d.). www.yelp.com. https://www.yelp.com/dataset/download