Summary of Yelp Data Analysis Project - Group 5

1 Introduction

Yelp is an Internet company that provides a platform for users to write reviews of businesses, which publishes a data set containing 8,635,403 reviews and the information of 160,585 businesses and 2,189,457 customers which covers multiple industries.
This project will focus on Sandwich restaurants, and the target is to analyze the reviews and provide data-driven suggestions to business owners. Specifically, it will categorize all the reviews by topic, and for each sandwich business, find out the reasons why they have received good reviews or bad reviews. Consumer behaviors and preferences are discussed in this project in order to provide businesses with unique strategies in different seasons.

2 Statistical Analysis

2.1 Data cleaning and overview

The research object is sandwich shops. Firstly we filter out the businesses that contain “sandwiches” and “burgers” in the categories label. But it turns out that many pizzerias and barbeques also include sandwich tags (such as Domino’s pizza Franklin Barbecue), so businesses that contain such tags are removed. We match businesses and reviews by inner-joining with business ID. After cleaning, we get 912,573 reviews of 4550 businesses.

2.2 Seasonal effects

Based on the data of the review month, we divide the data into 4 seasons. Spring is from March to May, summer is from June to August, autumn is from September to November, and winter is from December to February.
First, we took a look at the distributions. The distribution plot shows the overall shapes are quite similar between seasons.Due to the large scale, though, differences really exist.

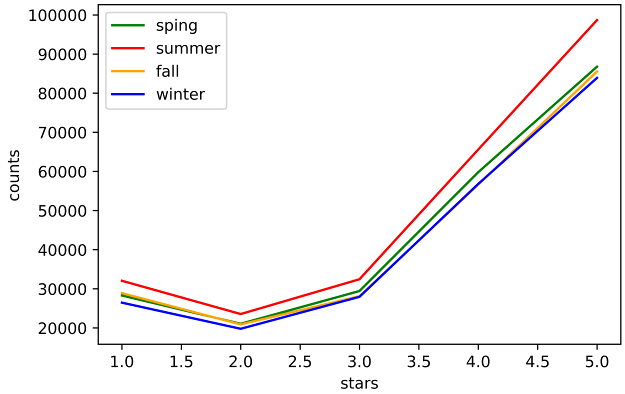


Figure 1: Distribution of review stars

Then, we calculated the mean and standard deviation of review stars in different season groups. 6 pairs of T-tests are conducted to judge whether the differences are statistically significant.

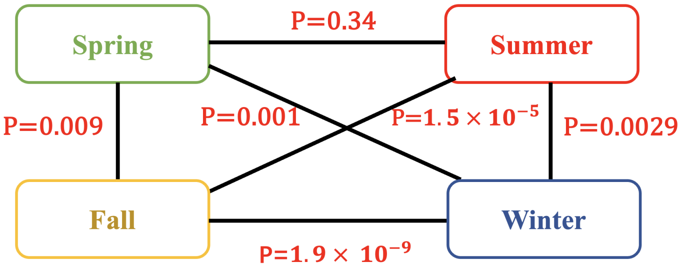


Figure 2: p-values of the seasons

The mean value of review scores is significantly different in different seasons. The T-test result shows that Winter has the highest average scores and Fall has the lowest. However, there is no difference between Spring and Summer under 0.05 significant level.

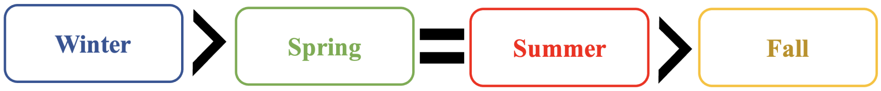


Figure 3: relationship between different seasons

For business, it is easier to earn a high score in winter and get bad feedback in fall. Therefore, from the perspective of business strategy, we recommend that businesses can try more proactive strategies in winter, such as launching new dishes, providing more service, while adopt a more conservative strategy in autumn such as strengthening the training of employees to avoid mistakes.

3 Natural Language Processing

3.1 Model Selection

Topic modeling is a type of statistical modeling for discovering the abstract topics that occur in a collection of documents. In a practical and more intuitively, you can think of it as a task of Dimensionality Reduction and Unsupervised Learning. LDA is one of the most popular topic models and is used to classify text in a document to particular topics. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

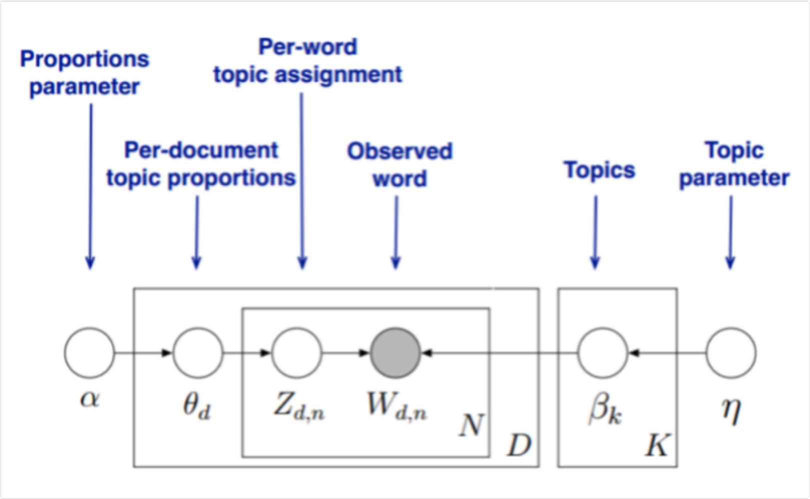


Figure 4: Distribution of review stars

The first step is to clean the data of reviews. We split the sentence into individual words, and then delete all the stop words like “and/or/but/so” and punctuation. Then we apply lemmatisation on our text data. The Average Topic Overlap (ATO) is an indicator to measure the quality of the LDA model. We have trained a series of models with different numbers of topics. We can see there is a sharp decrease of ATO when the number of topics is 6 and become much flatter after that. In addition, an excessive number of topics will make the model impossible to interpret. So, we choose 6 as the topic number for our model.

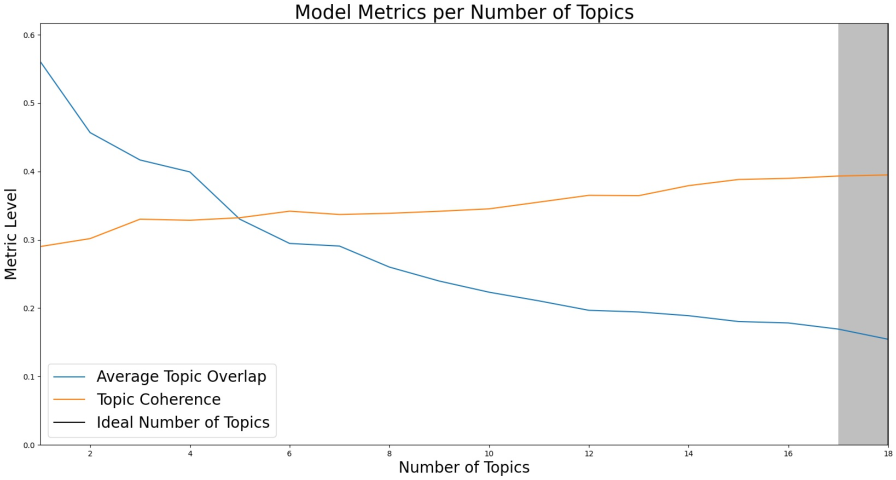


Figure 5: Metrics to choose the parameter

3.2 Outcome and Analysis

Finally, we trained a 6-topic model for positive and negative reviews respectively. By analyzing the high-frequency words in each topic, we will conclude what each topic is talking about. For example, here are the most relevant terms in a topic within good reviews.

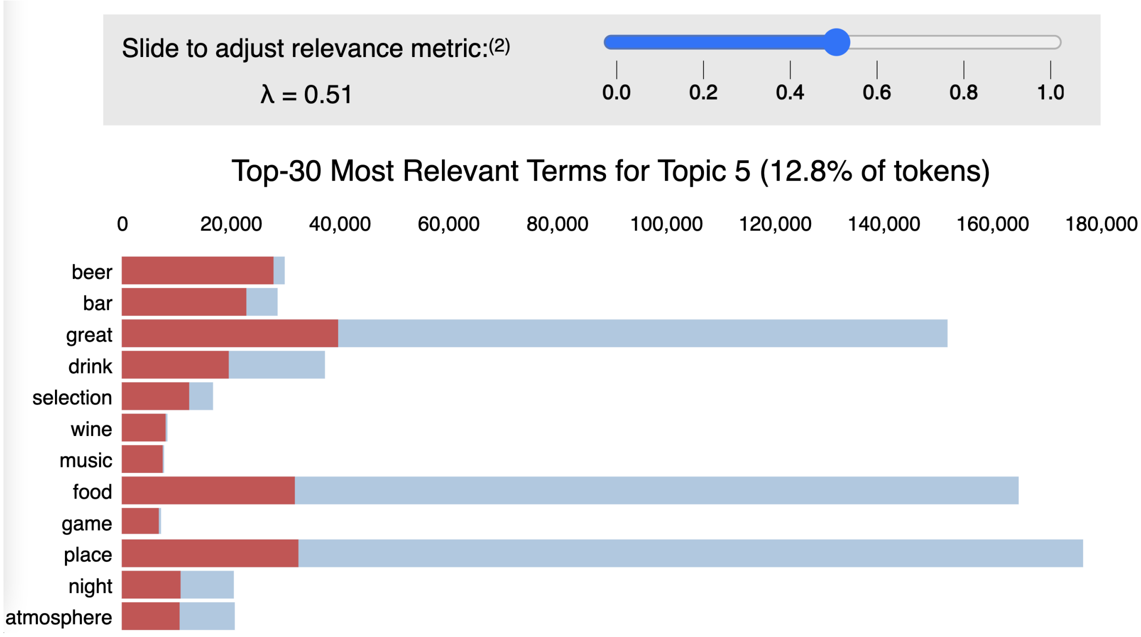


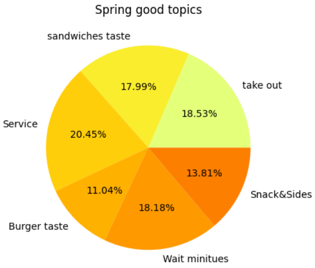
Figure 6: Top Most Relevant Terms for a certain Topic within good reviews

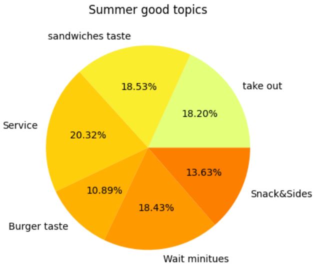
This topic is talking about that the atmosphere can affect the experience of eating sandwiches here. If the drink is better or the music meets the appetite of customers, the restaurant tends to get a higher star. Also there are more topics in the HTML files and it turns out, for example, that there is a topic talking about the taste of sandwiches, and another discussing the restaurant's diversity and convenience.

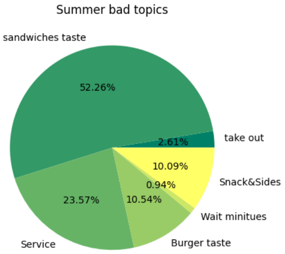
After careful analysis of high-frequency words, we concluded the topics in reviews are as below:

For good reviews: 1. Sandwiches taste 2. Burger taste 3. Customer service 4. Drinks and atmosphere 5. Breakfast taste 6. Convenience

For bad reviews: 1. Sandwiches taste 2. Burger taste 3. Waiting minutes 4. Delivery Take-out order 5. Customer service 6. Snack Sides
To measure the seasonal impact, we calculate the topic distribution of positive and negative reviews in each season. Below are examples of spring and summer:







We can see differences between seasons are not so significant. The reasons why consumers give good and bad reviews in different seasons are basically the same. But remember this is more than one million data, even a change of 1% means the change of tens of thousands of reviews. For example, there is a certain difference between the good reviews in spring and summer in terms of drinks and breakfast. This shows that in the spring people are more willing to have breakfast in the sandwich shop, and in the summer, people prefer to drink a glass of beer in the shop.

3.3 Suggestion for businesses

For bad reviews, sandwiches taste and customer service take more than 75%. Therefore, customers usually only give bad reviews when they eat bad or stale food, or when they are treated badly by service. Surprisingly, waiting time is the topic with the least proportion of negative reviews. People usually don't care much about waiting time. Snack sides is a part that cannot be ignored. Businesses should focus more on snacks and sides.
Based on the reviews of each sandwiches shop, we also calculated the topics with the most favorable reviews and the most negative reviews for each of them. The result will be shown in the Shiny App. Basically, we found the positive topics of every business are varied. But negative topics for almost all businesses focused on the taste of the sandwich and followed by the customer service.

4 Model Strengths and Weaknesses

The t-test simply and intuitively reflects the difference in review scores between different seasons. With such a large sample, the t-test is somehow effective. However, the data do not follow normal distribution because it only has 5 level and 5-star reviews account for the vast majority which means the data is skewed.
The topics model is a kind of unsupervised machine learning model. It effectively categorizes a large sample of reviews so that we can better understand and analyze the text. It also provides high-frequency words within each topic and we can summarize what the topic is. However, changes in numbers of topics will have a huge impact on the model. In addition, there are many influencing factors such as different years, different shop sizes, etc. that have not been considered in the model.

5 Conclusion

From the analysis above, we would give such general suggestions to sandwiches businesses in different seasons. In the spring, provide breakfast for customers. In the summer, offer better drinks and beer. Treating customers' problems carefully and taking conservative strategy in the fall. And providing a better experience to customers and being more creative in the winter. For each business in details, browse our shiny app.

References

Blei, David, Andrew Ng, and Michael Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, 05 2003, *3*, 993–1022.

Contributions

Shuren He: Contribute to most part of the LDA model including selecting and training model.
Ziyue Zheng: Contribute to the t-test part, the analysis of the LDA model outcome, R Shiny app, report writing.
Ouyang Xu: Contribute to part of the LDA model, R Shiny app, report writing.