**Final Summary**

**Introduction**

Yelp is an Internet company that providing a platform for users to write reviews of businesses. It published a data set containing 8,635,403 reviews and the information of 160,585 businesses and 2,189,457 customers which covers multiple industries.

Our project will focus on Sandwich shops, and our target is to analyst the review data and providing data-driven suggestions to business owners. Specifically, we will categorize all the reviews by topic, and for each sandwich business, find out the reasons why they received good reviews and the reasons why they got bad reviews. We will also study consumer behavior and preferences in different seasons. Provide businesses with unique business strategies in different seasons.

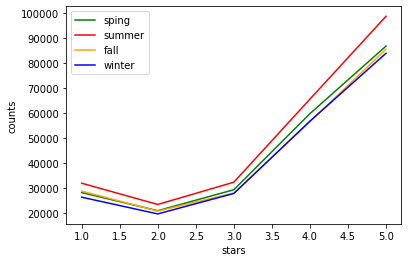
**Statistical Analysis**

**Data cleaning and overview**

Our research object is sandwich shops. First filter out the businesses that contain “sandwiches” and “burger” in the categories label. But we found 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.

**Seasonal effects**

According to the data of the review month, we divide the data into four seasons. Based on the data of the review month, we divide the data into four 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 there are some differences between seasons

**Distribution of review stars**

Then, we calculated the mean and standard deviation of review stars in different season groups.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Spring | Summer | Fall | Winter |
| Mean | 3.691 | 3.695 | 3.677 | 3.707 |
| Std | 1.387 | 1.394 | 1.405 | 1.382 |

**Statistics table**

Then we conducted 6 pairs of T-test to judge whether the differences are statistically significant.

**P=1.9\***

**P=1.5\***

**P=0.34**

**P=0.0029**

**P=0.001**

**Spring**

**Summer**

**P=0.009**

**Fall**

**Winter**

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 are no difference between Spring and Summer under 0.05 significant level.

**Fall**

**Spring**

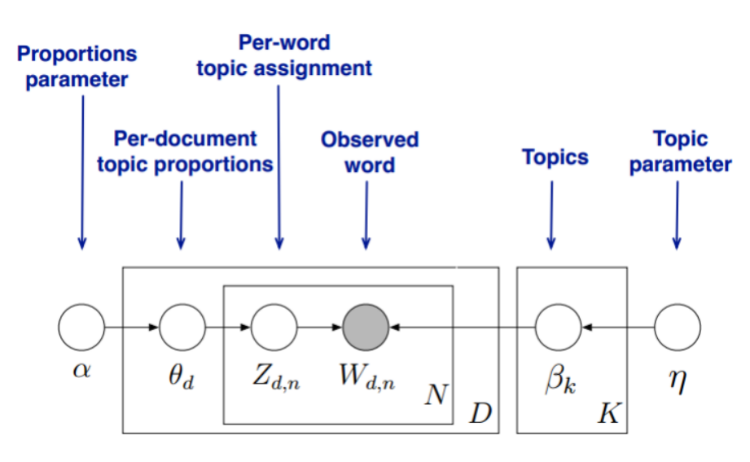
**Winter**

**Summer**

For business, it is more easily to earn 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, but should adopt a more conservative strategy in autumn such as strengthen the training of employees to avoid mistake.

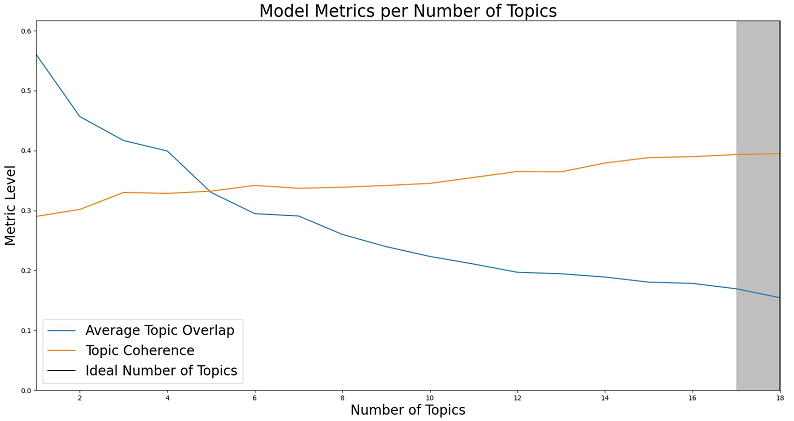
**NLP model**

**Model Selection**

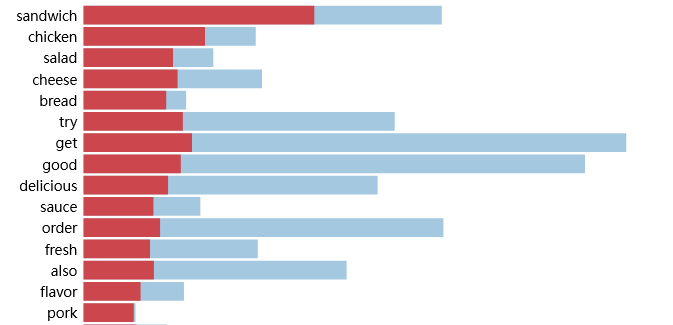
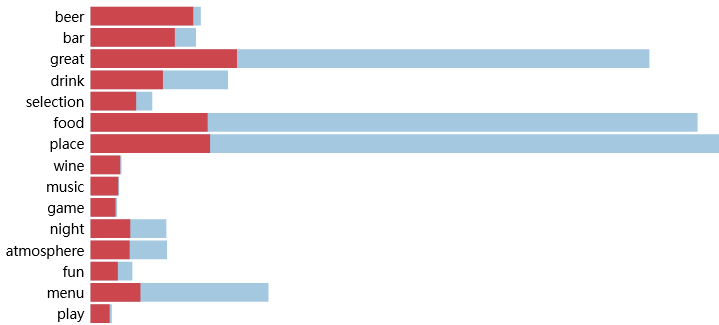
Topic modeling is a type of statistical modeling for discovering the abstract topics that occur in a collection of documents. LDA is one of the most popular topic models and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

**Principle of LDA model**

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

**ATO with different topics number**

**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 two topics from good reviews.

**words in topic 1 words in topic 2**

The first one on the left below is obviously talking about the taste of sandwiches, and the second one on the right below seems to be discussing the restaurant’s drinks and environment.

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

For good reviews:

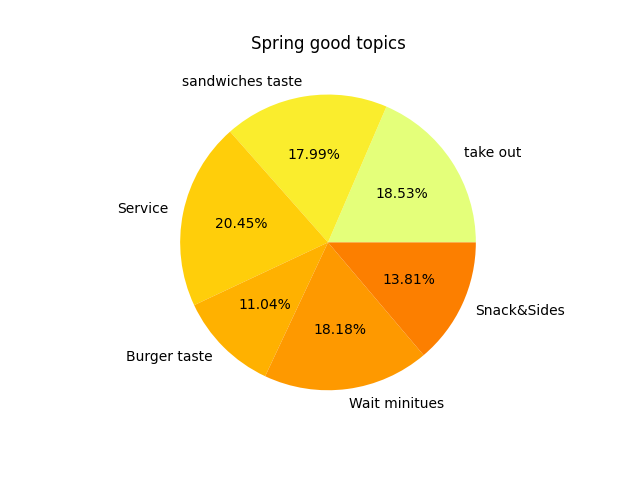
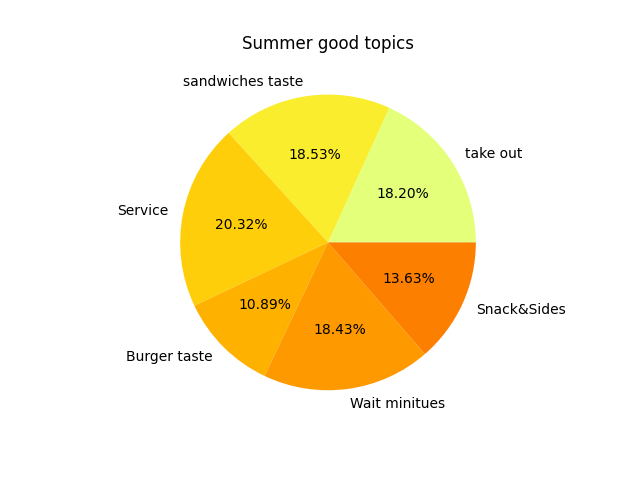
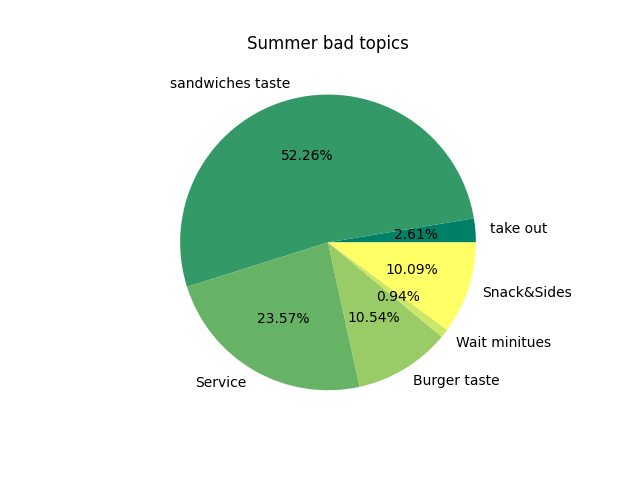
1. Sandwiches taste 2. Burger taste 3. Customer service 4. Drinks and atmosphere

5. Breakfast taste 6. Convenient

For bad reviews:

1. Sandwiches taste 2. Burger taste 3. Waiting minutes 4. Deliver & 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.

**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 fresh 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. Business should focus more on snack 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 is varied. But negative topics for almost all businesses focused on the taste of the sandwich, and followed by the customer service.

**Strength & Weakness**

The t-test simply and intuitively reflect the difference the of review scores between different seasons. With such a large sample, 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 large sample of reviews so that we can better understand and analyze the text. It also provide high-frequency words within each topics and we can summarize what the topic is. However, changing in hyperparameters (numbers of topics) will have a huge impact on the model. In addition, there are many influencing factors such as different years, different shop size etc. that have not been considered in the model.

**Conclusion**

From the analysis above, we would give such general suggestions to sandwiches businesses in different seasons. In the spring, open restaurant for breakfast. In the summer, offering better drinks and beer. Treating customers' problems carefully and take conservative strategy in the fall. And providing better experience to customers and be more creative in the winter. For each individual business, browsing our shiny app.

Shuren He wrote most part of the LDA model including selecting and training model.

Ziyue Zheng wrote the t-test part and the analysis of LDA model outcome.

Ouyang Xu wrote the Shiny app and helped Shuren with the LDA model.