Group Member: Peiyilin Shen, Yuqi Shen, Jingxuan Zhou, Baixue Zheng, Jueyan Lin

# Final Project

1) **Background**

- Large amount of text data available on Yelp

Yelp claims itself a business directory service and crowd-sourced review forum. In the customers' view, Yelp provides many aspects information about shops, including restaurants, cleaners, movers, delivery and so on. Users can find the information about any shop listed. For example, users can find the basic information of a shop, and after visiting it, they can post reviews and give scores (the number of stars out of five) based on their experiences. Compared with scores, reviews provide more information and we try to find out the hidden information for text data.

- Information from text

The reviews data we use from Yelp are written by customers. Generally, if a review of a restaurant is posted, customers would likely to share their experiences about the food, the service, the environment and etc, and these reviews actually provide more information than scores. Therefore, it is necessary to figure out the information based on the large amount of text. For example, customer attitudes towards the service, the comments about the food, or even whether they would recommend this restaurant.

- Our goal

For the interpretation of the reviews data from Yelp, we basically focus on two aspects of the reviews. In the first part, we generated the topics which are more frequently mentioned by the customers. With the topics that we extract from the reviews, we try to figure out what kind of topics customers are more likely to mention when providing reviews on Yelp and which kind of topics they care more about. Based on the topics we generated, we find out the new scoring metrics based on the current one to make some improvement. Since customers provide more aspects of scores to a restaurant, other customers can get more information through the scalable scoring system while restaurant owners can also receive more clear and detailed information. In the second part, we combined sensitivity analysis and topic modeling. Instead of only focusing on topic, we try to figure out what do customers care more when they leave negative comments on these businesses, which can provide more insights for these business owners in specific aspects they can improve.

2) **Data Description**

In this project, we used the Yelp dataset to analyze reviews from customers. We hope to extract more features out of these reviews. The original "Review" data are in JSON format, with 5.6 GB in total. We used Python to read the JSON data and converted them into Python list for further analysis. Since it takes almost half an hour just to read the 5.6G data, we figured that we would only use parts of the data to represent the customer population. As a result, we randomly selected 10000 customer reviews as our targeting dataset to do topic modeling and sentiment analysis. For sentiment analysis, we only used negative reviews from the 10000 customers reviews to extract topics that matter most to the customers.

- Data cleaning: using Text_Normalization_Function notebook to clean data.
    - Expand stop words from NLTK by adding more words such as "mr", "mrs", "would", "also", and etc.

- Split each sentence from each customer review into a list of words through defining a function called “tokenize_text” to clean data word by word in each sentence.
- Defining a function called “expand_contractions”: we generated a dictionary for expand contracted words, such as converting “ain't” to “is not”.
- Annotating each token word with part-of-each tags through defining a function called “pos_tag_text”, which is aim to lower case for each word in each sentence.
- Defining a function called “lemmatize_text” to restore the original form for each word based on part-of-each tags.
- Removing any special characters such as quotation marks from text through defining a function called “remove_special_characters”.
- Removing stop words from text through defining a function called “remove_stopwords”.
- Removing all non-text characters such as numbers through defining a function called “keep_text_characters”.
- Cleaning up any HTML markups at the beginning the text and at the end of the text via defining function called “strip_html”.
- Removing any accents from characters through defining a function called “normalize_accented_characters”.
- Finally, putting all functions together through defining a function called “normalize_corpus”.

3) **Objectives for the Analysis**

- Business Insight

Yelp is widely used in our daily life to provide recommendations about restaurants, or any type of business. On the Yelp app, we observed that their recommendation system only has a general rating (stars), price assessment rating and customers reviews, which is lack of precise rating such as services, staff, and etc. Since Yelp stores so many reviews for each business, we can utilize these reviews to get some business insights to improve its recommendation system, such as adding more features. Through using topic modeling method on these 10000 customer reviews, we can find out which topics are most frequently mentioned in reviews, implying that these aspects are relatively cared by customers, then add these topics as additional features to the recommendation system, providing customers a diverse and precise perspective to make decisions. Specifically, we used word frequency and topic modeling techniques towards customers’ reviews to identify these valuable aspects. Finally, we extracted the most popular words from the Review dataset as our choices of features that are added into Yelp recommendation system.

On the other hand, because Yelp already owned millions of customer reviews, and these comments are either praising or expressing disappointment, thus Yelp could provide business insights to those business owners about what customers care about or are unsatisfied the most. For instance, restaurants who partner with Yelp can improve their services based on customers’ preference generated through analyzing customers reviews. We also used sentiment analysis to sort out positive and negative comments, then we only analyzed negative comments by word frequency and topic modelling techniques. By using this approach, we can provide feedbacks to business owners in general about which aspects they should improve to obtain more customers. Perhaps in the future analysis, we can provide personal diagnosis for improvement for each business owner.

By improving with the respect to users and business owners, Yelp is able to customize the recommendation system and reach out users more accurately, and it can also help business owners improve their products, which could provide another opportunity of generating revenue for Yelp.

- Application

In order to make Yelp more competitive in the saturation market, Yelp has to provide customized user experience for each customer. Yelp needs to go deeper to explore different customers' needs. In this case, Yelp can customize its service by upgrading the rating system.

For instance, in the past experience, customers can only give an overall evaluation for each restaurant in Yelp's existing rating system, and they can only evaluate the service, food quality and dining environment of these restaurants under the reviews section. By classifying the reviews by n topics, we can find the most common topics mentioned by customers and the perspectives they value most. Based on the topic analysis, Yelp could develop a new rating system that includes "Best food", "Best Service", "Least waiting time", etc. Customers can still rank the restaurants by the original overall ratings, but if they want to find restaurants with "Best food quality", they can refer to each characteristic ratings. Meanwhile, restaurants can also improve their service by knowing the current situations and customer preference from those specific ratings. Yelp could customize the service for restaurants as well. Without text mining, it's sometimes hard for restaurants to summarize the core aspects among all positive and negative reviews. After sentiment analysis, Yelp can send keywords of negative or positive reviews privately to restaurants if they want. Hence, we did topic selection again on negative reviews to find the most common topics that customers are unsatisfied about. Sentiment analysis can go further if Yelp wants to tease out fake reviews and make reviews more convincing.

4) **Methodology**

- **Frequent Topic Selection**
  - <u>Model Selection</u>

Based on the Yelp's review dataset, we found out topics that customers care about the most, and then added these features to Yelp's rating metrics. With large volumes of unlabeled Yelp's reviews, we chose topic model for discovering the abstract "topics" that occur in the reviews dataset. We will utilized Latent Dirichlet Allocation(LDA) as our main methodology.

Specifically, we first cleaned the reviews dataset and converted them to different corpus. Then we normalized the corpus and represented it with "bag-of-words" method. As for the parameter, we limited the number of features to 1000 and set 8-10 topics. By changing alpha and beta, we selected the optimal value of alpha and beta for the Latent Dirichlet Allocation model. Based on these 8 topics and the corresponding top keywords, we narrowed down the topics by comparing the results manually, and finally filtering most reasonable topics as our features to be added in the rating metrics.

Group Member: Peiyilin Shen, Yuqi Shen, Jingxuan Zhou, Baixue Zheng, Jueyan Lin

Overally, we chose 1000 features with 8 different topics and set both alpha and beta as 0.005. Based on our analysis(8 topics + 8 top keywords), we summarized 5 topics as our final features.

- Pros and Cons

Pros

According to Latent Dirichlet Allocation model, we can easily extract several topics among the unlabeled reviews from Yelp and better understand what aspects customers pay attention to. Besides, the elaboration of the topics helps us further define each feature clearly. Hence, it is handy to interpret our analysis by utilizing LDA method.

Cons

The main problem for Latent Dirichlet Allocation model is to accurately summarize the topics. Generally, with different values of the parameter, we will get different results. There is no certain rule for topic selection, that is, to well divide the reviews into several topics with no overlap. Therefore, in order to get the most reasonable topics, we need to run the models for several times and summarize the topics manually. Hence, the subjective problem may influence our conclusion.

- **Customer Sentiment Extraction**
  - Sentiment Analysis

Sentiment analysis is a text mining method to identify and extract different attitudes from authors in the text content. In Yelp's reviews, customers will show positive or negative attitudes towards different businesses. We would use sentiment analysis to identify keywords related to polarized attitudes and then find the key aspects that customers value most.

- Method Selection

We have learned several supervised machine learning methods including support vector machine, logit, Naive Bayes, and unsupervised machine learning model like lexicon. Because the raw data from Yelp is lack of positive or negative information in this dataset, we can not use supervised learning in this case. Hence, we chose to use Lexicon model to generate the scores of three sentiment level.

First, we cleaned and normalized data as we mentioned above. Then we generated the vander score by using Sentiment Intensity Analyzer function. We figured out the sentiment polarization with the compound score, that is, when the compound score is larger than/equal to 0, it reveals "positive review" and when the score is smaller than 0, it reveals "negative review". After we got sentiment for each review, we extracted the common topics mentioned on negative reviews. We used the same frequent topic selection model as above and then got the specific topics for negative reviews. We set up the features as 1000 and number of topics as 3.

- Pros and Cons

Pros:

We can have a clear view on sentiment analysis with Lexicon model even if we only got the unlabeled data, so it reduced the requirement of the datasets. Besides, a narrower topic analysis focusing on negative reviews will give both Yelp and business owners who collaborate with Yelp an intuitive insight about the existing problems and aspects which they should value in the future.

Cons:
Lexicon model is less accurate than supervised learning models because we don't have training and testing dataset to validate the model accuracy. As we tried in the class, the Lexicon model performs worse than supervised learning model on prediction accuracy and precision.

5) **Results and Discussion**
Since we had different goals for the two methods, we separated our discussion based on their results. For topic modeling, we mainly want to know what customers are interested in on a general level. However, for sentiment analysis, we specifically cared about customers who expressed negative attitudes in their reviews.

- **Topic Modeling**

The main goal for the topic modeling is to divide the reviews into several topics. Aftering adjusting various values of parameters, we found that 0.005 is the optimal value for both alpha and beta in the LDA model, since the topics can be divided neatly.

With the visualization of the model, we summarized the topics and keywords manually and narrowed down to 5 main topics that customers are mostly and generally interested in: Location, Food, Service, Order and Pay. We further collected keywords that appeared most frequently under each topic to better interpret these 5 topics/features. Yelp can add these five metrics to its rating system so that customers have a better understanding of each business in general.

The five metrics are defined as follows:

- <u>Location</u> (See Appendix 1)

The first metric "Location" measures how convenient it is for customers. Relevant keywords mentioned in Topic 1 are "Place", "Walk" and "Way". These words together provide a good clue to determine this new feature. For example, customers prefer a restaurant with more convenient transportation and a walk-distance location. Parking is also a crucial aspect to measure the location since it is sometimes a big concern for customers, especially in the urban area.

- <u>Food</u> (See Appendix 1)

The second valuable metric for a rating system is "Food", which is usually used in rating restaurants. The measurement includes two dimensions, food diversity and food quality. From topic 2 and topic 5, we observed that customers frequently mentioned diverse types of food such as "Beer", "Sushi", "Sandwich" and "Salad" in their reviews, and they cared a lot about food menu when evaluating restaurants. Besides, customers mentioned many emotional words such as  "Delicious", "Amazing", "Nice" which indicates the importance of the food quality. Obviously, the more delicious the food is, the better reviews customers may leave. Hence, we considered synthetically the topic and the keyword frequency here and generalized these findings into "Food" metric.

- Service (See Appendix 1)

Combining the result of topic 3 and 4, we summarized a third metric called "Service". As the fourth top keyword in topic 3, "Service" came up a lot in the customer reviews accompanied by other relevant keywords such as "Work", "Staff", "Experience" and "Friendly". For one thing, staff is the key to the high quality service. On the other hand, the service diversity is important as well. How to deliver an unforgettable customer experience and a comfortable environment to customers should be taken into account for all Yelp businesses.

- Order (See Appendix 1)

"Order" is another metric for Yelp's rating system. As shown in topic 5, it is the most popular keyword mentioned by customers. When it comes to "Order", order time and order method should be included. When we analyzed the result of topic 1, we evidently found that "Time" is the top one keyword with 26.8% of all the reviews. Some other keywords related to the "Time" such as "Wait" and "Minute" were also listed on the top 30 keywords. Thus, we took a deeper look at the customer's feeling about the time. Generally, customers care about the time cost when getting served, and as the keywords reflect, customers do not want to wait the food/service for a long time.

Additionally, order method refers to the way to get the food/service. We found that "Booking" is a frequent word mentioned by customers. Therefore, offering different ways to purchase is an essential aspect for evaluating businesses. In general, customers prefer a more flexible way to access to the service, such as booking or delivery.

- Pay (See Appendix 1)

The last metric is "Pay" which includes the price and the pay method. Different from the former metrics, customers did not mention price directly and we could rarely see this word shown among the top keywords. However, customers did mention "Pay" and "Card" in some topics which reflects that there are some price-sensitive customers in reality and they want a more flexible way for the payment. Hence, we generalized these into "Pay".

- **Sentiment Analysis**

Our goal in sentiment analysis is mainly filtering out all the negative reviews from the original 10,000 reviews and performing a topic modeling analysis to identify the major topics in these negative reviews. After reading in the 10,000 reviews, we analyzed the data using the Lexicon model. The polarity is an index in this model, with parameter "compound" indicating the positivity and negativity of comments. If the sentiment is positive, the compound score will be greater than 0, and vice versa.

Out of the 10,000 reviews, 1052 were negative reviews, then we performed a topic modeling analysis on these negative reviews to decide the main topics that customers complained about. Having adjusted for different values of number of topics, alpha and beta, we finally settled down on 3 topics, with alpha and beta being both 0.005. For $\lambda$, it is known that small values highlight potentially rare, but exclusive terms for the selected topic, whereas large values highlight frequently but not necessarily exclusive terms. Since we want to understand the story behind each topic, we chose a relatively small $\lambda$ (0.2) to get the unique terms. With the visualization panel, we summarized the 3 topics that describe customer's negative sentiments in general: food quality, services, appointment.

- Services (Appendix 2)

The first topic is "Services" (with 37.2% of overall tokens). Keywords such as "wait" "stand" "register" "tip" indicate the part of customers' dissatisfaction within the services area. This is consistent with what we found in topic modeling on the overall dataset. Finding a presence for services in both analyses indicate that it is a significant component that restaurants/other service companies should be aware of when building business models. This also complies with common knowledge in that Consumer Relationship Management has always been a great priority and success driver across all industries.

- Food Quality (Appendix 2)

The second topic is Food Quality (with 33.4% of overall tokens), which is mainly used in restaurants evaluations. The major keywords in this topic are names of different foods, negative adjectives such as "disgust", and food related things like menu. It's no surprise that food quality has been a main concern among customers' negative reviews. Yelp mostly constitutes of restaurants, and food quality problems are easily and likely included in customer dissatisfactions. Besides, it's also consistent with what we found in the general data (that food has been the major topic of overall reviews of the customers). Hence, it's important for restaurants to deal with reviews complaining food quality appropriately.

- Appointment (Appendix 2)

The third topic is Appointment (with 29.4% of overall tokens). Keywords include services such as "hair" "nail" "doctor" "tech" "schedule". Customers who having issues on their appointment also expressed dissatisfaction in their reviews. Yelp can use this information to optimize its recommending system. For example, placing service companies that have more negative comments about appointment at a lower rank tier, reducing customers' potential dissatisfaction.

6) **Conclusion**:

As a result, we generated our final conclusion in two parts. The first part is to provide advices to Yelp recommendation system, explaining our work and giving out some situations that our models and results can be used. In the second part, we indicated possible shortcomings of our work and how these shortcoming affect our modeling processes and results. Furthermore, we provided possible ways to overcome them.

- Advices
    - Yelp

The data provided by Yelp can be analyzed with respect to users and business owners. With the information given from text mining, Yelp can do assessments about their current business model, which can help them to get up-to-date information about different features from the reviews. Based on our analysis, it is obvious that users often focus on location, food, sevice, order and pay. Our recommendation is that Yelp can expand their current scoring metric or rebuild a new one, in order to provide precise aspects about the businesses, thus customers can receive information not just based on the overall score but also the specific parts, such as location, food quality, service. From the sentimental analysis, Yelp also can find out what customers care more and business owners will receive negative reviews from it.

- User

As we recommended above, Yelp could provide the revised scoring system, then users can search for a particular business, while they still can have a brief overview from star rating. Without going through others' review, customers can still receive detailed information about the business.

- Business owner

For business owners, our recommendation is that owners can take advantage of the topics generated from our analysis to figure out what customers care about. For example, they can have the diagnosis instead of reviewing each comment from their customers, which efficiently help them to enhance their competitive ability. With the topics generated from all these reviews, more information could be provided to the business owners.

- Possible shortcomings
  - Business owners focus on different topics

This might be one of the shortcomings in our analysis. The analysis that we generated are based on different markets. However, in reality, business owners from different industries have different target customers, which means that different business owners may have specific requirements to evaluate their businesses. However, we do not provide such specific information for them in our analysis since the dataset is too large to run. To deal with this problem, one possible approach is to use cloud technique to combine all data set, then categorize each type of business and generate different topics of each industry, and finally present results in the form of dashboard.

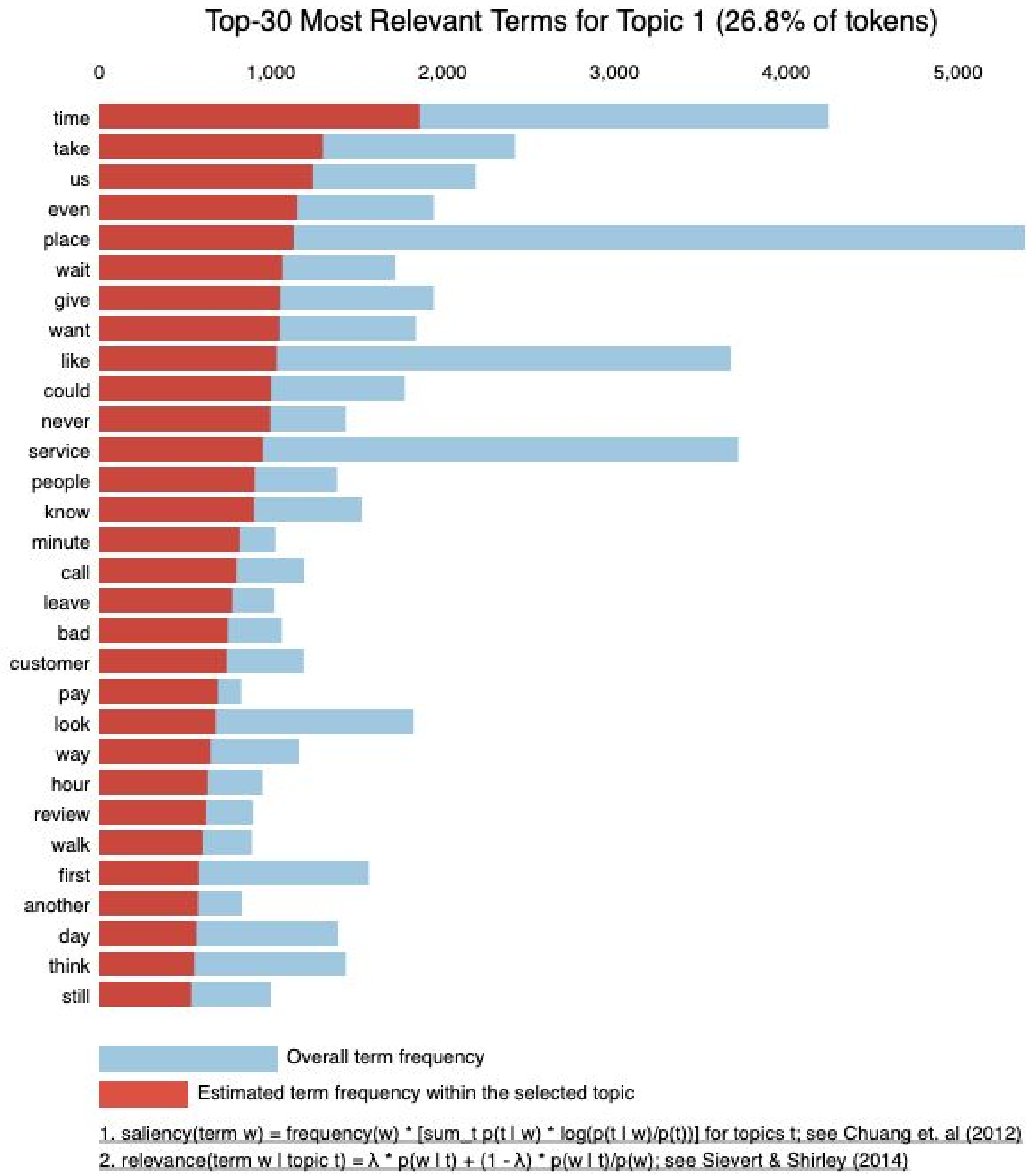
- More than two kinds of sentiments exist in reviews

If we go over all the reviews, we can find out some examples that the customer could express several feelings on one business, which may not reduce the precision of our modeling. But when designing the rating system, these multi-attitude review need to be reviewed carefully about how these reviews influence our rating system. One possible way to resolve this problem is that one review can explain different rating metric on the business. For example, a review can generate positive score on food quality but negative score on time consuming.

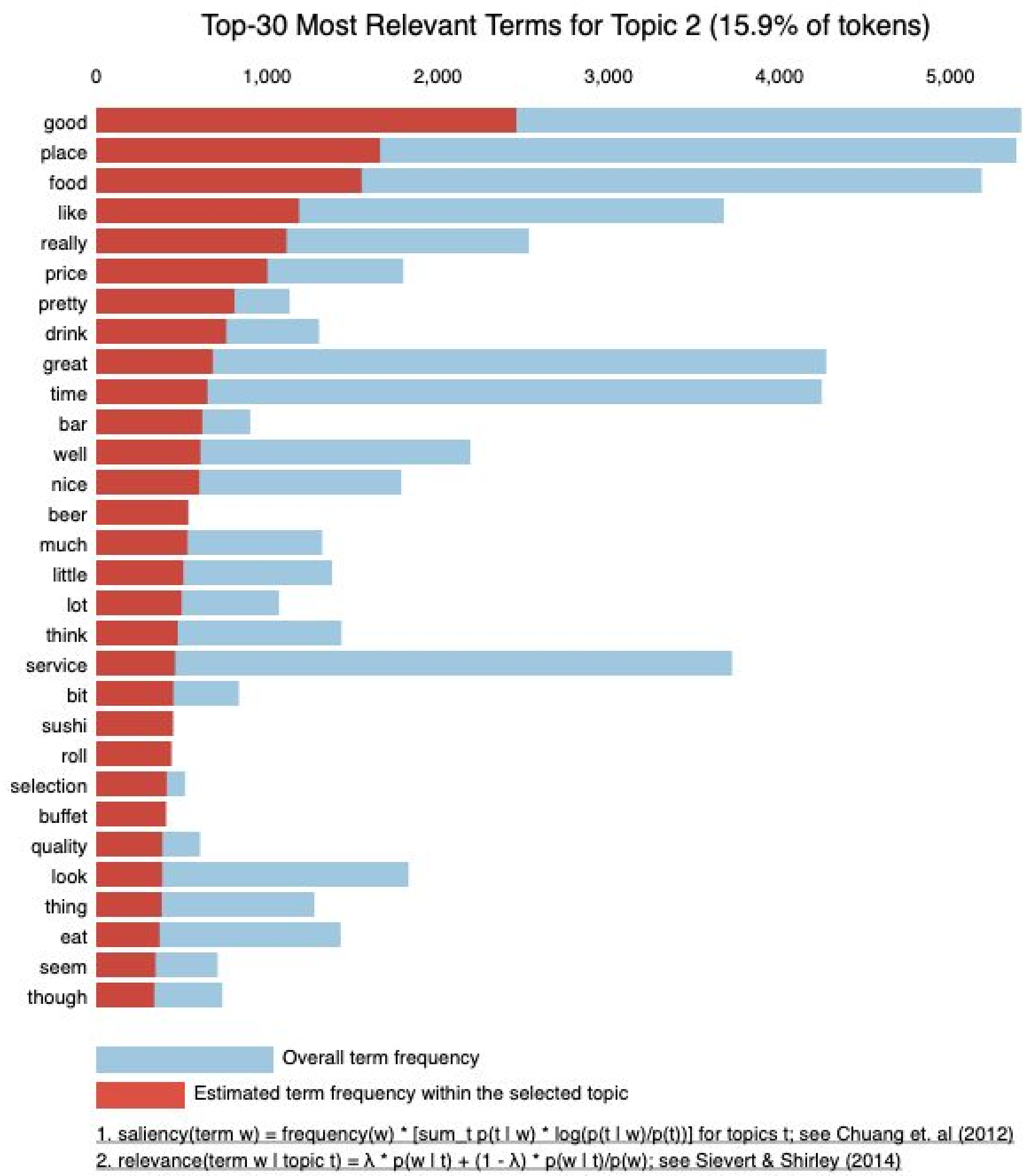
7) **Appendices**: tables and plots.

Group Member: Peiyilin Shen, Yuqi Shen, Jingxuan Zhou, Baixue Zheng, Jueyan Lin

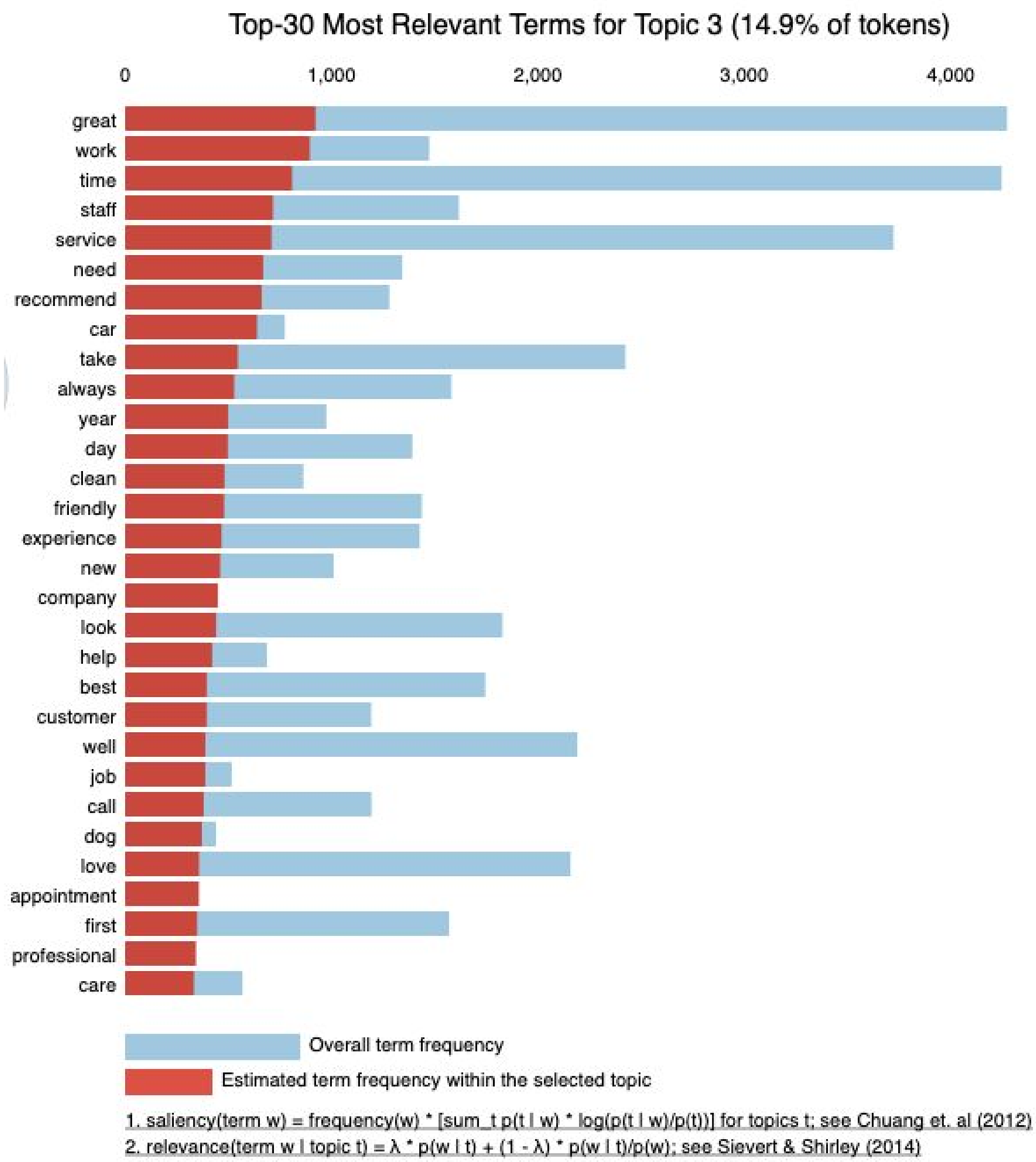
Appendix 1: Top-30 Most Relevant Terms for Topic 1-5



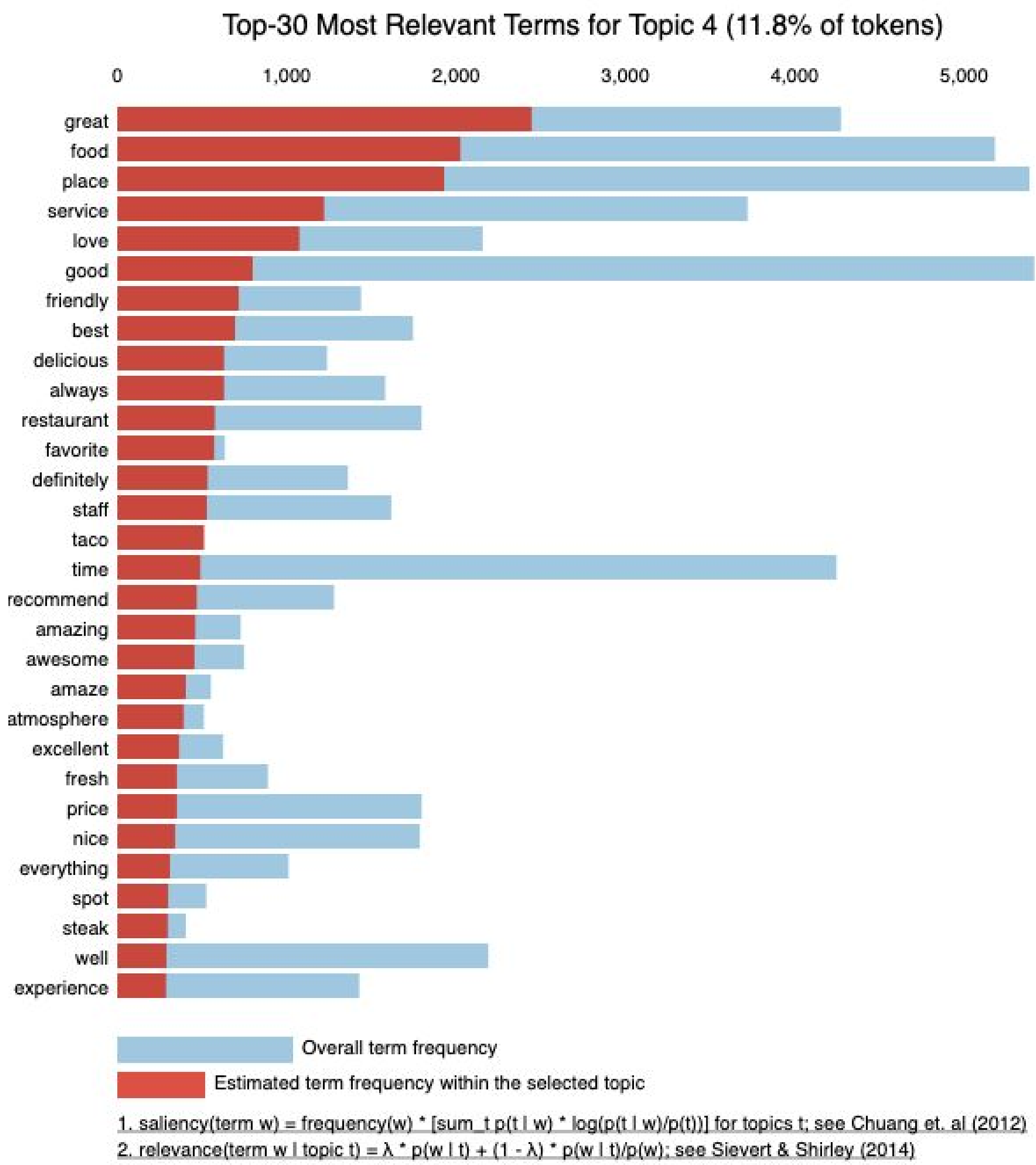
Group Member: Peiyilin Shen, Yuqi Shen, Jingxuan Zhou, Baixue Zheng, Jueyan Lin



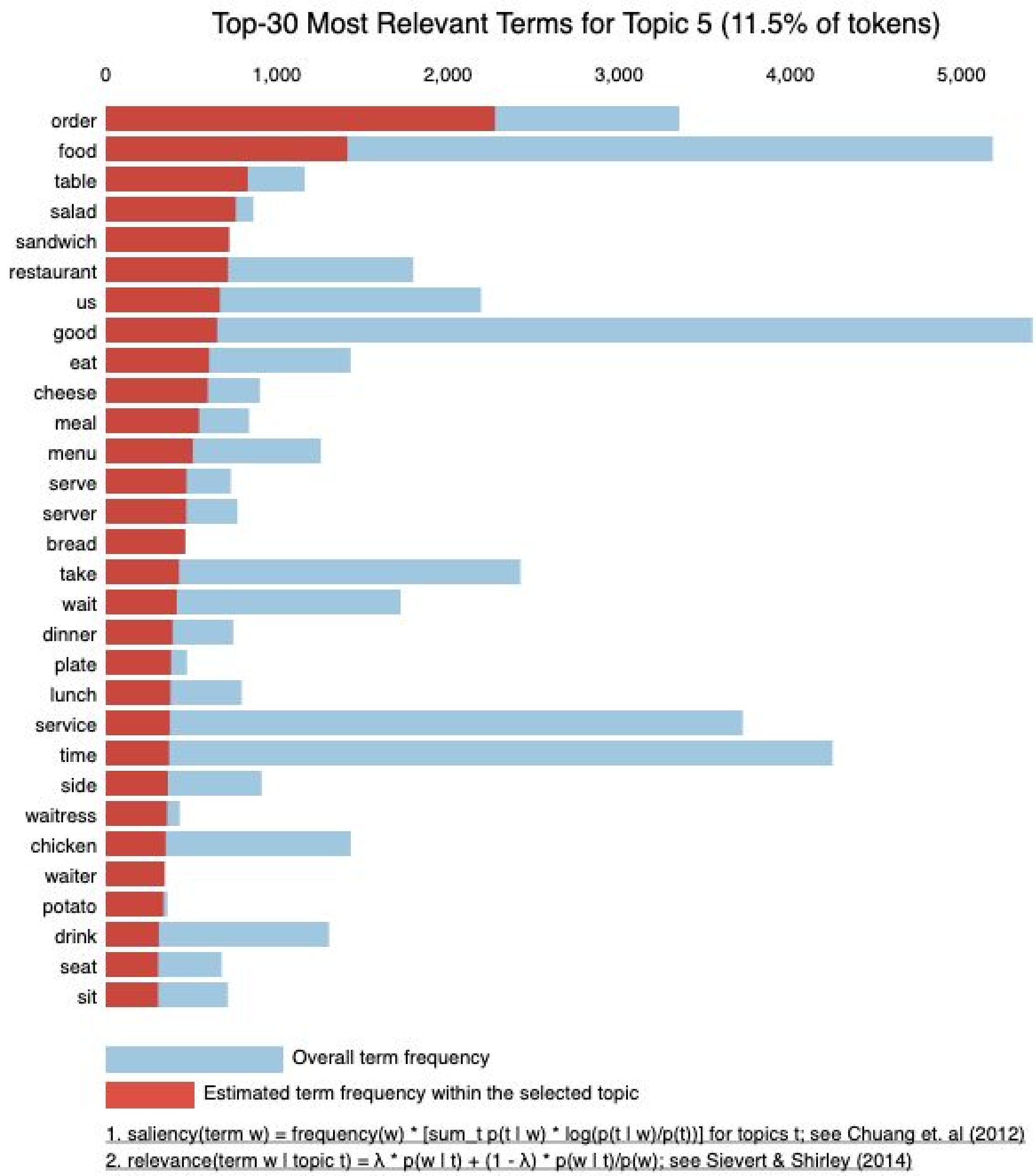
Group Member: Peiyilin Shen, Yuqi Shen, Jingxuan Zhou, Baixue Zheng, Jueyan Lin



Group Member: Peiyilin Shen, Yuqi Shen, Jingxuan Zhou, Baixue Zheng, Jueyan Lin

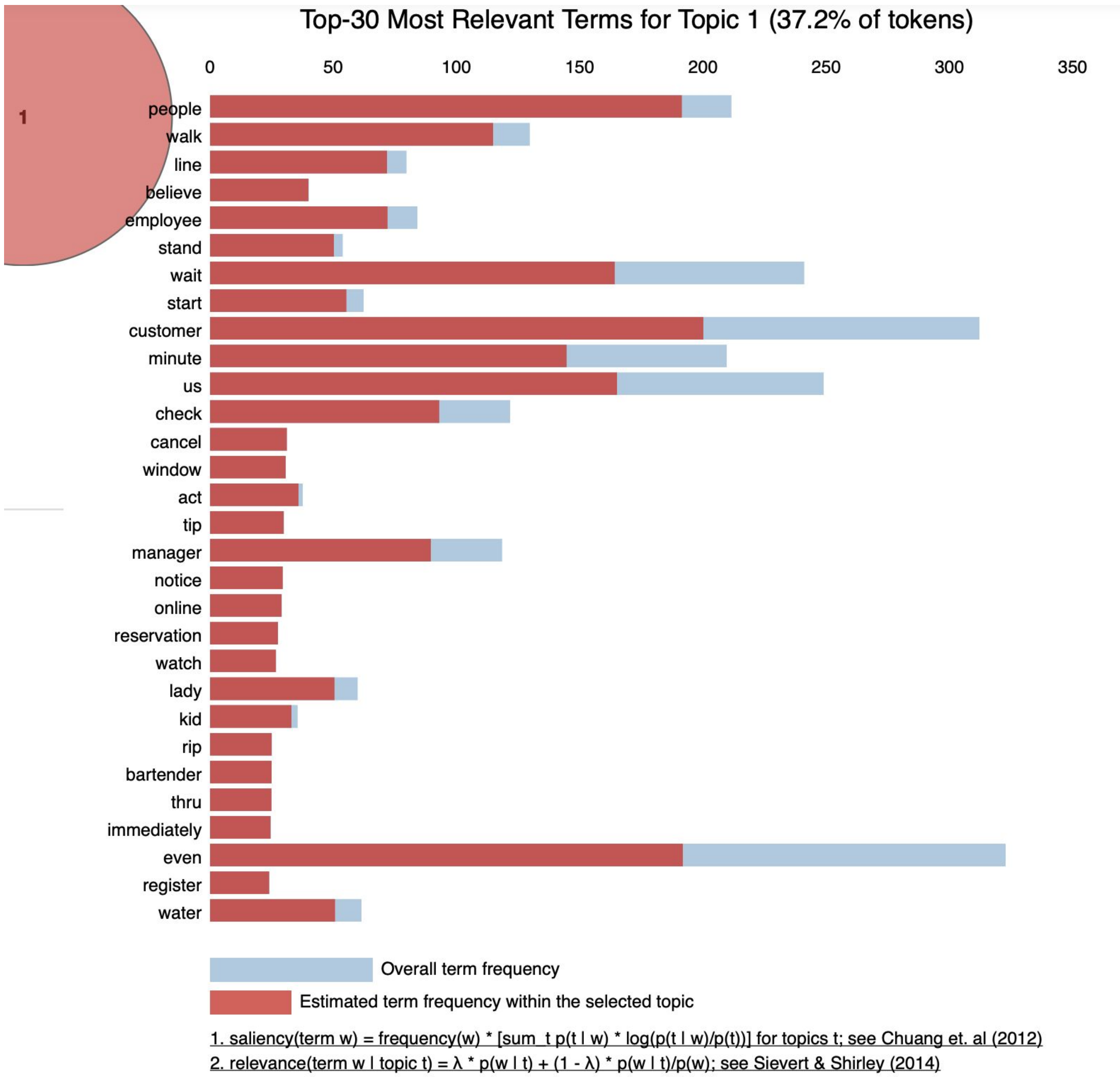


Group Member: Peiyilin Shen, Yuqi Shen, Jingxuan Zhou, Baixue Zheng, Jueyan Lin



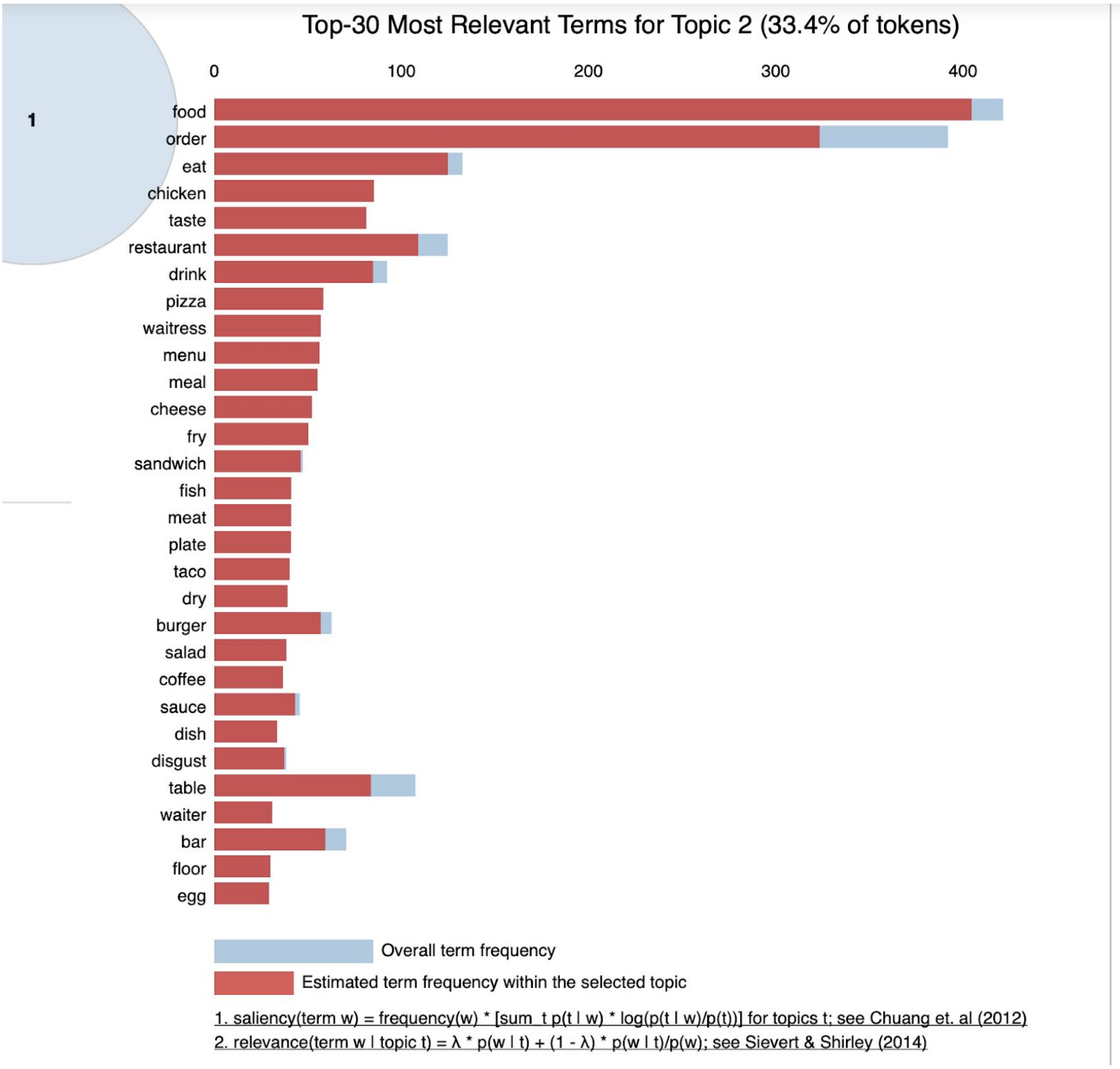
Group Member: Peiyilin Shen, Yuqi Shen, Jingxuan Zhou, Baixue Zheng, Jueyan Lin

Appendix 2: Sentiment Analysis Topic 1: Service



Group Member: Peiyilin Shen, Yuqi Shen, Jingxuan Zhou, Baixue Zheng, Jueyan Lin

Sentiment Analysis Topic 2: Food Quality



Group Member: Peiyilin Shen, Yuqi Shen, Jingxuan Zhou, Baixue Zheng, Jueyan Lin

Sentiment Analysis 3: Appointment