An NLP Project Report on Analyzing Sentiments of Restaurant Reviews provided by TripAdvisor - London City

1. Introduction

People must eat to live, and for that people who love to dine at a restaurant rely on restaurant reviews. In this modern era of food industry and internet, reviews left on the internet for a restaurant has a significant impact on their business, but also opens the opportunity to find hidden information on new opportunities for consumer business. TripAdvisor has a rating-based review system which helps the visitors get an idea on the review on restaurants.

Today, I am discussing the Rating Reviews obtained 1827 Restaurants from London City.

2. Problem Statement

The goal of this project is to explore and provide solutions to the following key questions:

- What factors contribute most to restaurant reviews?
- Can we predict the sentiment of a review based on common patterns or trends?
- What are the key words or phrases that help identify the sentiment in a review?

The aim is to analyze both positive and negative reviews, uncover the factors that drive these sentiments, and develop an action plan to categorize the keywords most strongly associated with each sentiment.

The London dataset, containing 196,134 reviews across 13 columns, includes detailed information such as restaurant names, ratings (1 to 5), review sentiments (positive or negative), full reviews, dates, and other relevant metadata. Each row represents a single customer review.

Our primary focus is the Sentiment Analysis of these reviews, using the "sample" column, which indicates whether a review is positive or negative, as the target variable for classification.

Since this is an NLP-based classification problem, we will employ Topic Modeling to identify clusters of related keywords and explore their association with sentiment. This approach will help us determine how different topics contribute to positive or negative reviews, allowing for a deeper understanding of customer feedback.

3. Data Collection and Preprocessing

The data for this analysis was sourced from a **Kaggle dataset**, which contains six tables representing reviews from six cities: London, New York, New Delhi, Paris, Barcelona, and Madrid, all in CSV format. For this project, I focused on the **London** dataset.

The preprocessing steps included the following:

- **No Missing Values**: During the web scraping process, there were no missing values. Any rows with null values were dropped from the dataset to ensure data quality.
- **Removal of Unnecessary Columns**: To prepare the dataset for NLP analysis, I removed irrelevant columns that would not contribute to the corpus. These columns included: "Unnamed:0", "parse_count", "review_id", "url_restaurant", "author_id", "date", "title_review", "review_preview", and "city".
- **Converting Data Types**: The rating_review column, which was initially in string format, was converted to float and integer types to facilitate numerical analysis.
- **Text Preprocessing**: To prepare the textual data in the reviews_full column for NLP analysis, I applied the following preprocessing steps:
    - Converted all text to lowercase.
    - Removed punctuation.
    - Set stopwords using the **NLTK stopwords** corpus, with additional custom stopwords including: 'food', 'place', 'restaurant', 'go', and 'london'.
    - Applied Parts of Speech (POS) tagging.
    - Lemmatized the text to reduce words to their base forms.
    - Tokenized the text into individual words.
    - Created bigrams, but avoided trigrams due to computational constraints.
- **Sampling the Dataset**: Due to computational limitations, I selected a balanced sample of 20,000 reviews, with an equal split between positive and negative reviews. This balanced dataset allowed for more accurate classification during topic identification. I further filtered the **Bag of Words** by removing words that appeared in fewer than 10 documents or in more than 90% of the documents, ensuring that only meaningful terms were retained for analysis.

4. Methodology:

This section outlines the approaches, algorithms, and models used to solve the NLP problem, along with the tools and libraries employed.

To address dimensionality reduction and topic discovery, I applied the **Latent Dirichlet Allocation (LDA)** approach for **Topic Modeling**. LDA was chosen over other techniques like **PCA** due to its ability to uncover meaningful associations between topics and sentiments within the text data. The goal was not merely to reduce dimensionality but to extract dominant topics that could be used as features for **Sentiment Prediction**.

By modeling the text corpus into topics, LDA helped reduce the complexity of the dataset while maintaining the interpretability of the topics. These dominant topics were then used as inputs for the sentiment analysis model, allowing for the prediction of sentiment based on the discovered topics.

The tools and libraries used for this task included Python's **SpaCy**, **Gensim** for LDA topic modeling, along with **NLTK** and **Scikit-learn** for preprocessing and sentiment prediction.

5. Model Training and Evaluation

Discuss the model training process, including hyperparameter tuning and performance metrics. Include the evaluation results.

To classify the sentiment of the reviews (positive or negative), I trained a machine learning model using the processed text data from the **London** dataset. The following steps outline the model training process, hyperparameter tuning, and evaluation metrics:

1. Model Selection

Given the nature of this problem, I experimented with several classification algorithms, including:

- **kNN**
- **Logistic Regression**
- **Decision Tree**
- **Random Forest**
- **Support Vector Machines (SVM)**

After initial testing, **Logistic Regression** and **SVM** performed the best in terms of both accuracy and computational efficiency for the text classification task.

2. Hyperparameter Tuning

For both the Logistic Regression and Naive Bayes models, hyperparameter tuning was performed using **GridSearchCV** to find the optimal values:

- **Logistic Regression**: Tuned the regularization parameter C to control overfitting.

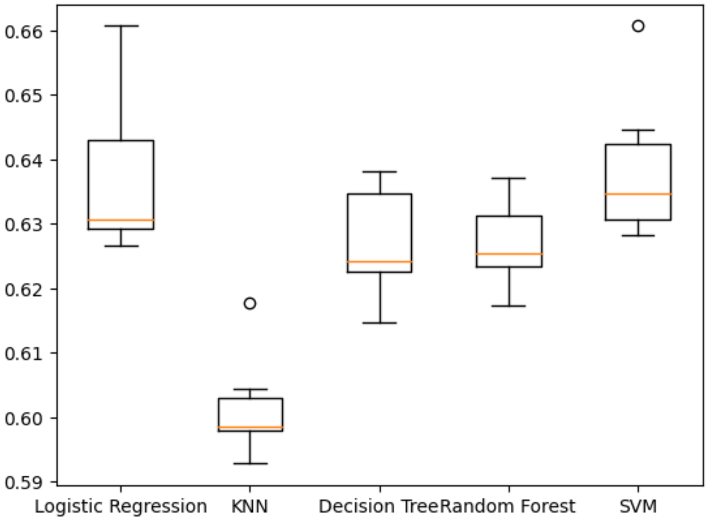
The best parameters found were:

- **Logistic Regression**: C = 0.1

3. Model Training

I split the dataset into **80% training** and **20% testing** sets to train the model. The **Bag of Words (BoW)** and **TF-IDF** (Term Frequency-Inverse Document Frequency) vectorization techniques were used to convert the textual data into numerical form, allowing the algorithms to process it.

The training process involved feeding the vectorized review texts to the chosen classifiers. I used **6-fold cross-validation** to ensure that the model generalizes well across unseen data and to prevent overfitting.



4. Evaluation Metrics

To evaluate the performance of the models, the following metrics were used:

- **Accuracy**: Proportion of correctly classified reviews.
- **Precision**: Accuracy of the positive predictions.
- **Recall**: Ability of the model to identify positive reviews.
- **F1-Score**: Harmonic mean of precision and recall, providing a balanced measure of model performance.

5. Results

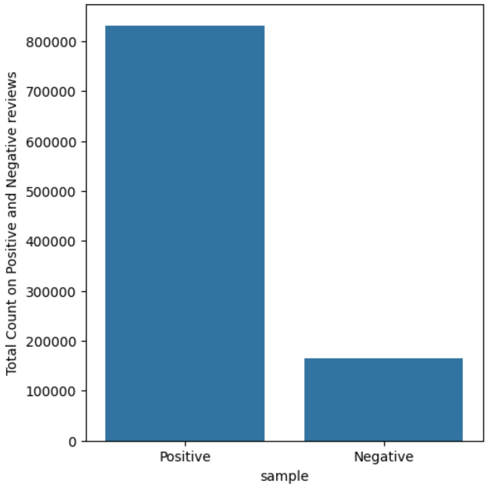
After training and evaluation, the following results were obtained on the test set:

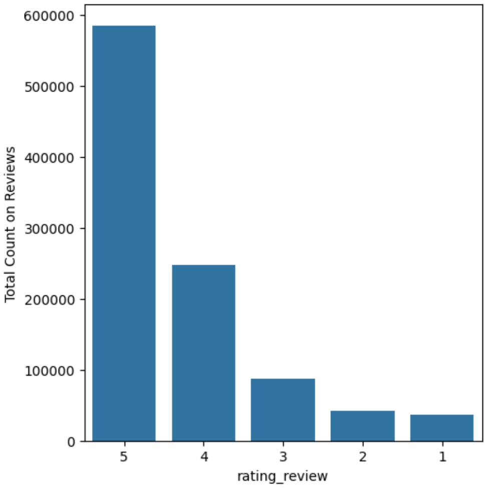
Model	Accuracy	Precision (Pos)	Recall (Pos)	F1-Score (Pos)
Logistic Regression	64%	0.41	0.75	0.53

The **Logistic Regression model** slightly outperformed the Naive Bayes model in terms of accuracy and precision. Therefore, it was selected as the final model for predicting the sentiment of restaurant reviews.

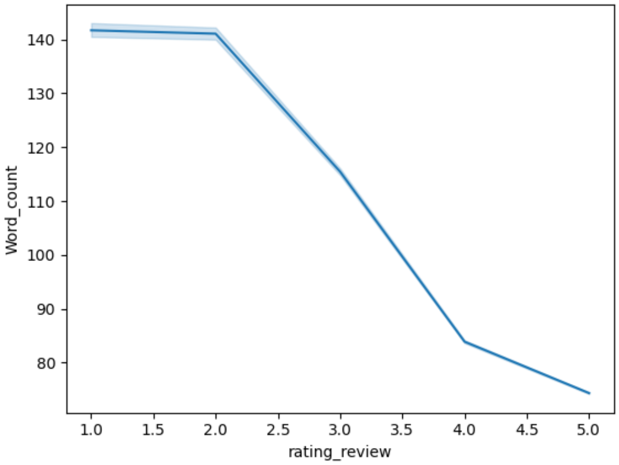
6. Results and Analysis:

- The reviewers were enthusiastic about leaving positive reviews mostly.

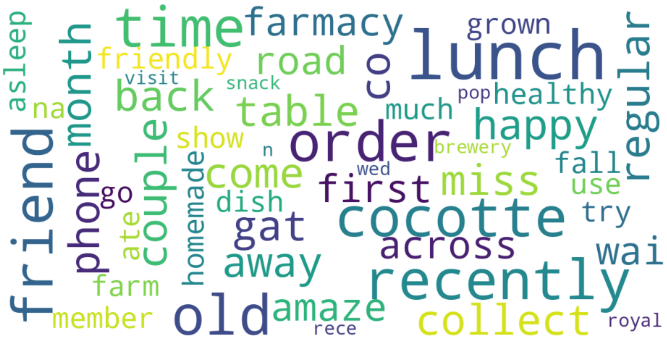




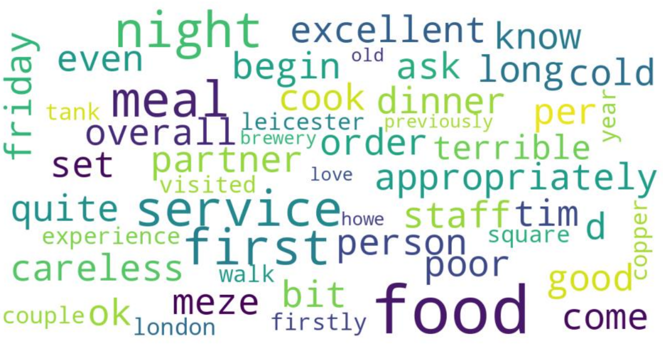
- But, while leaving negative reviews, people use more words. Afterwards, the wordcount in the reviews was found to be a significant attribute to predict negative sentiments.



- The word cloud on the Positive and Negative Comments were as following:

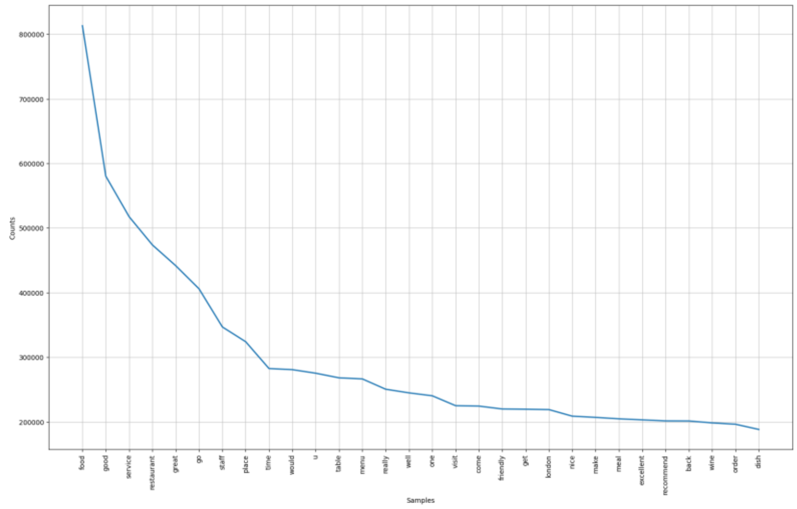


Positive



Negative

From the word clouds it was visible that there were several words that had very high frequency but could impact the sentiment analysis. To avoid further bias, those high frequency words were included in the stopwords. ('food', 'place', 'restaurant', 'go', and 'london')



- The identified 8 topics were as following:

Topic 0: perfect, fish, chip, authentic, chocolate, french, recently, extensive, die, nicely, shop, risotto, cod, modern, ice_cream, proper, create, candle, brasserie, dim_sum, favorite, theme, play, market, prepared, divine, grand, fit, buzzy, warn, pickle, carafe, limited, tempt, non, fully, tartare, pleasure, difference, struggle

Topic 1: table, book, bar, seat, birthday, next, people, sit, walk, treat, show, tea, open, cake, soon, customer, reservation, party, door, window, queue, pm, available, right, call, brilliant, plan, accommodate, manager, outside, greet, guest, become, class, front, owner, advance, close, move, celebrate

Topic 2: menu, wine, course, main, delicious, amaze, special, starter, dessert, choose, selection, set, list, glass, include, evening, relax, bottle, wife, present, superb, fabulous, follow, experience, recommendation, pie, outstanding, desert, surprise, start, remember, din, occasion, watch, care, item, ambiance, presentation, interesting, sommeli

Topic 3: especially, warm, coffee, breakfast, prepare, hot, venue, polite, chef, space, server, truly, pre_theatre, gorgeous, request, like, roast, cold, pricey, highly, turkish, pop, vegan, short, champagne, delight, kitchen, arrival, egg, allow, pre, casual, noisy, english, dress, mushroom, romantic, indeed, loud, toast

Topic 4: good, service, great, staff, really, visit, friendly, excellent, nice, well, lovely, recommend, atmosphere, definitely, price, lunch, enjoy, dinner, love, find, look, little, quality, always, attentive, feel, cocktail, choice, tasty, wonderful, busy, experience, bit, night, helpful, try, make, area, quite, welcome

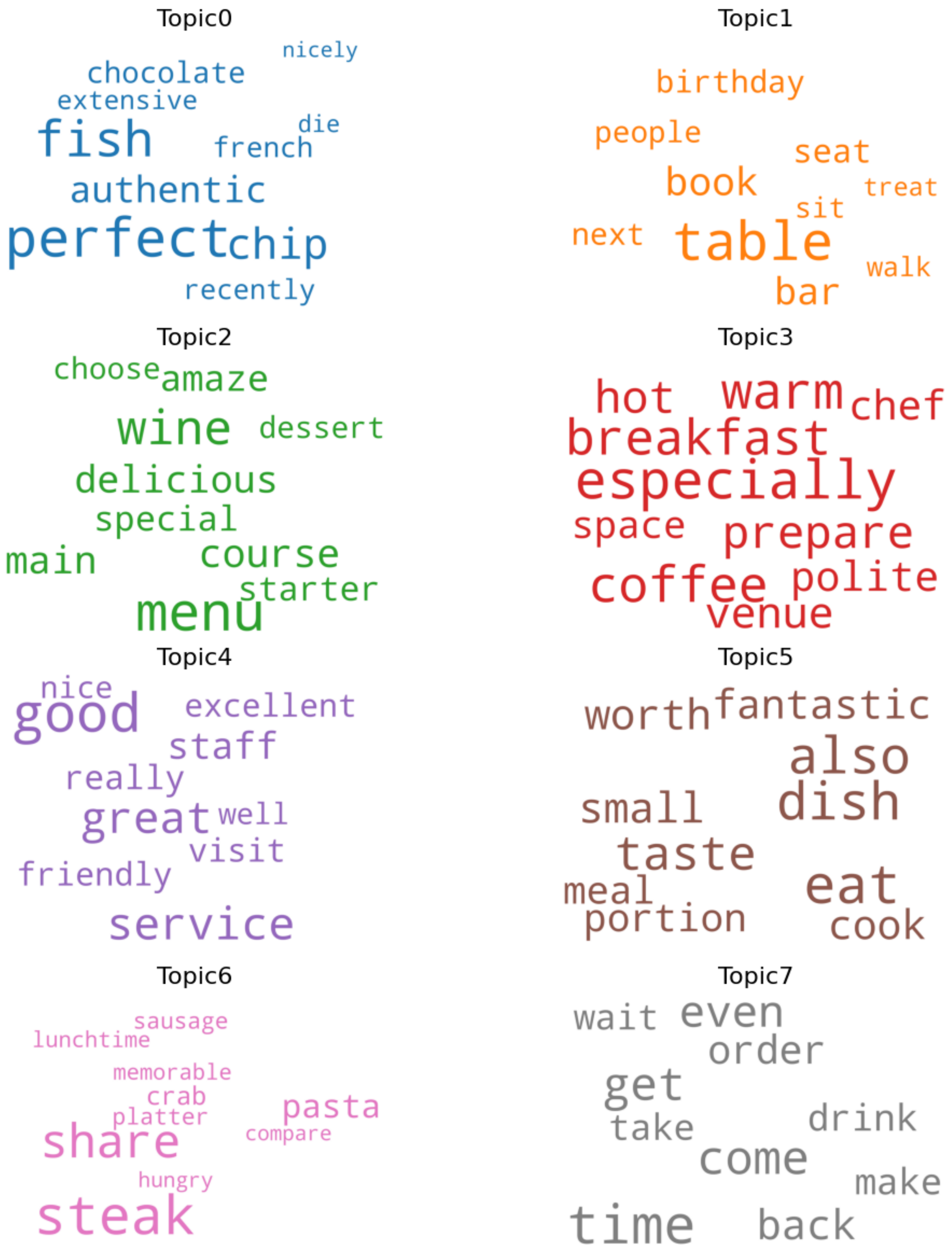
Topic 5: eat, dish, also, taste, worth, small, cook, fantastic, portion, meal, fresh, serve, italian, side, family, absolutely, burger, flavour, top, amazing, meat, ever, salad, pizza, large, sauce, full, try, plate, chicken, use, think, bread, option, favourite, beef, home, cheap, enough, cheese

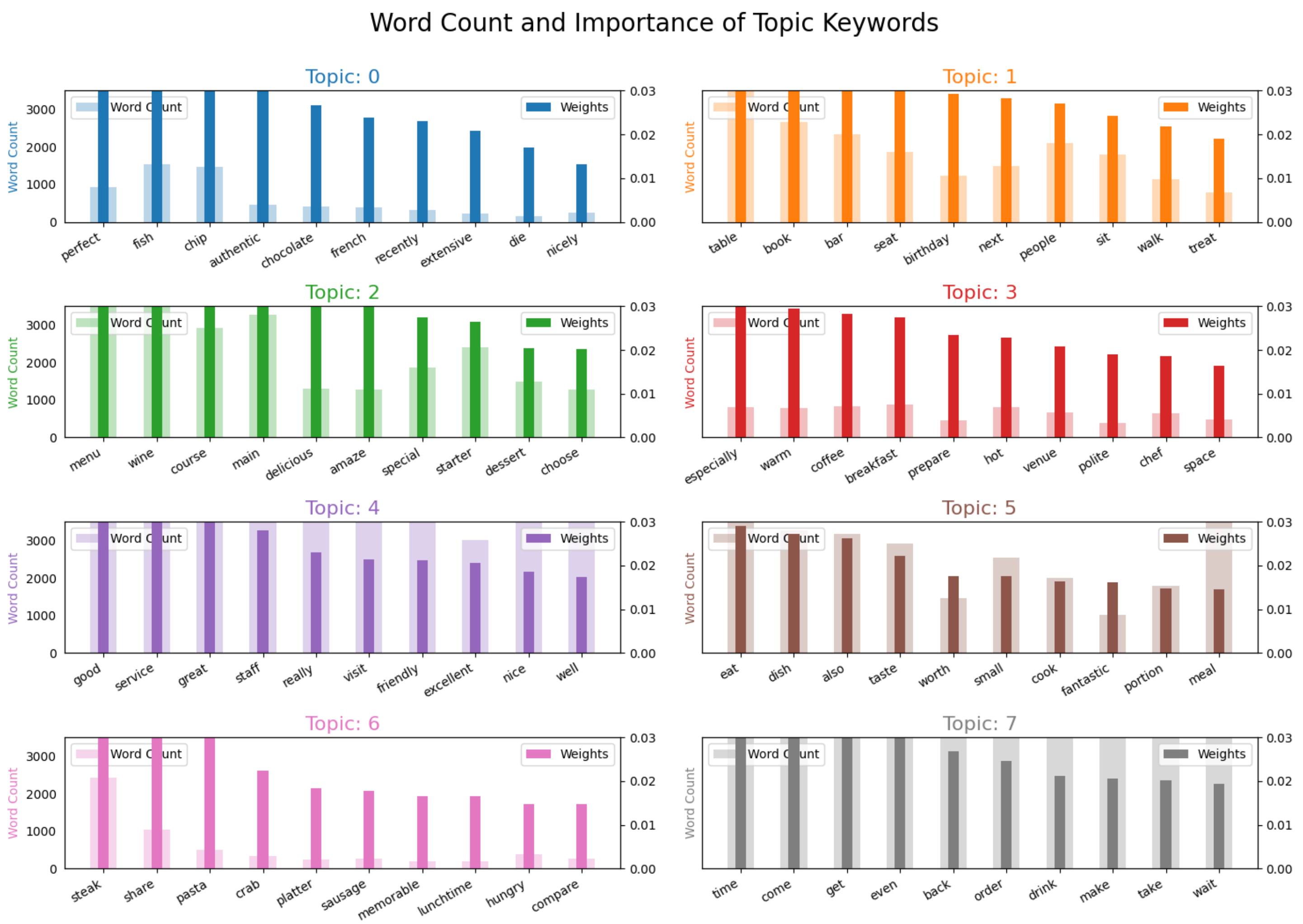
Topic 6: steak, share, pasta, crab, platter, sausage, memorable, lunchtime, hungry, compare, cafe, equally, opinion, green, medium_rare, informal, morning, concept, attention_detail, creamy, term, feed, relatively, vibrant, shrimp, gaucho, wooden, tend, appetizer, hit, knowledgable, situate, german, involve, pure, cat, willing, cooked_perfection, rock, counter

Topic 7: time, come, get, even, back, order, drink, make, take, wait, meal, want, say, first, waiter, give, friend, see, ask, thank, never, last, long, arrive, serve, still,

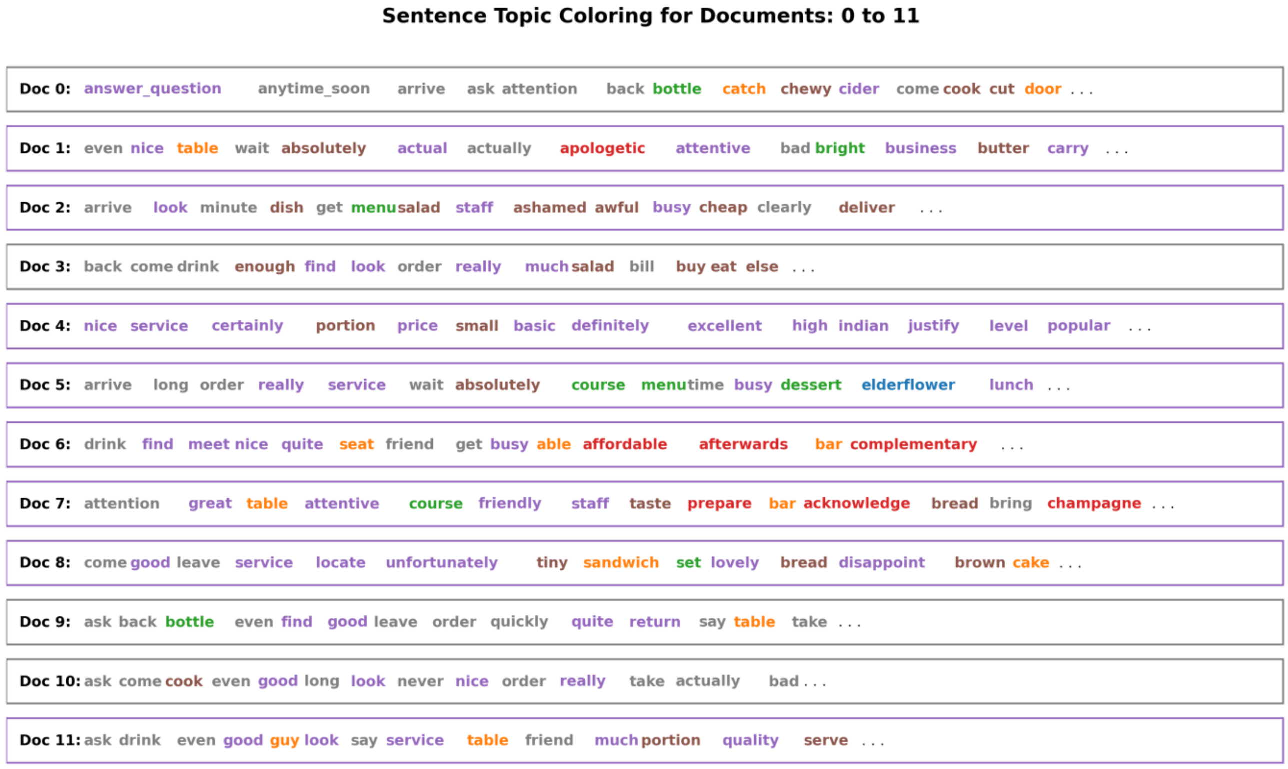
```
leave, know, happy, early, hour, decide, experience, minute, need, finish, waitress, bill,
think, thing
```

Following is the wordcloud of the identified topics:





Word Count and Importance of Topic Keywords



Sentence Topic Coloring for Documents: 0 to 11

Doc 0: answer_question anytime_soon arrive ask attention back bottle catch chewy cider come cook cut door . . .

Doc 1: even nice table wait absolutely actual actually apologetic attentive bad bright business butter carry . . .

Doc 2: arrive look minute dish get menu salad staff ashamed awful busy cheap clearly deliver . . .

Doc 3: back come drink enough find look order really much salad bill buy eat else . . .

Doc 4: nice service certainly portion price small basic definitely excellent high indian justify level popular . . .

Doc 5: arrive long order really service wait absolutely course menu time busy dessert elderflower lunch . . .

Doc 6: drink find meet nice quite seat friend get busy able affordable afterwards bar complementary . . .

Doc 7: attention great table attentive course friendly staff taste prepare bar acknowledge bread bring champagne . . .

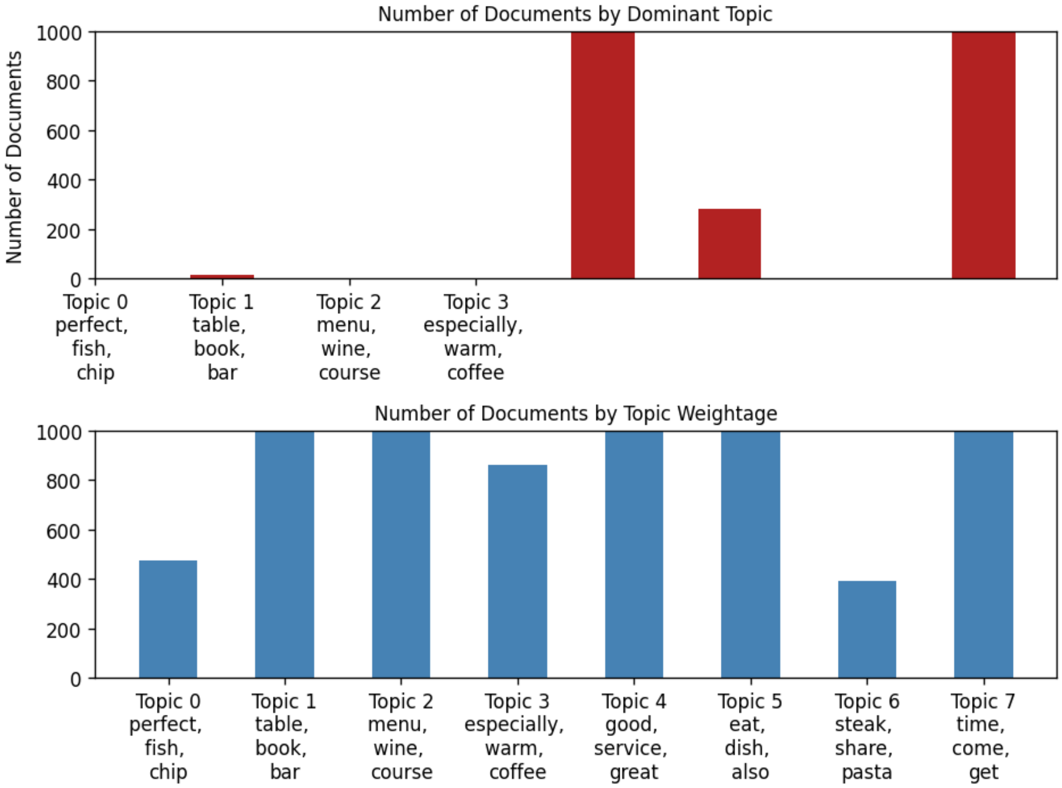
Doc 8: come good leave service locate unfortunately tiny sandwich set lovely bread disappoint brown cake . . .

Doc 9: ask back bottle even find good leave order quickly quite return say table take . . .

Doc 10: ask come cook even good long look never nice order really take actually bad . . .

Doc 11: ask drink even good guy look say service table friend much portion quality serve . . .

- Among these 8 Topics, we analyzed to find out the Dominant topics., which were Topic 4 and Topic 7 –

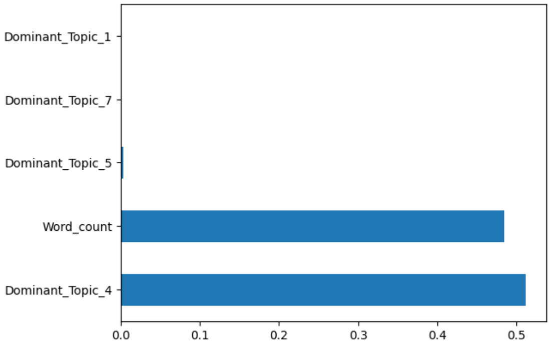


```
Topic 4: good, service, great, staff, really, visit, friendly, excellent, nice, well,
lovely, recommend, atmosphere, definitely, price, lunch, enjoy, dinner, love, find, look,
little, quality, always, attentive, feel, cocktail, choice, tasty, wonderful, busy,
experience, bit, night, helpful, try, make, area, quite, welcome
```

```
Topic 7: time, come, get, even, back, order, drink, make, take, wait, meal, want, say,
first, waiter, give, friend, see, ask, thank, never, last, long, arrive, serve, still,
leave, know, happy, early, hour, decide, experience, minute, need, finish, waitress, bill,
think, thing
```

From both Topics, it is evident that these keywords are associated with Service. (Topic 4- Quality of the Service and Topic 7 – Time related to the Service)

- To determine the key features that contribute to sentiment prediction, our analysis reveals that 'Topic 4' and 'Word Count in the Reviews' are the most influential factors.



7. Conclusion

Summarize the outcomes of the project, its contributions, and potential areas for future work.

This project successfully identifies the factors that influence restaurant review sentiments, focusing on a comprehensive analysis of London restaurant reviews from TripAdvisor. Through detailed preprocessing, topic modeling, and sentiment classification, the analysis highlights that topics related to service quality (Topic 4) and review length are crucial in predicting review sentiment. Logistic Regression, after tuning, provided the most balanced accuracy and interpretability, making it an effective model for sentiment prediction.

The results emphasize that restaurant service quality and the depth of customer feedback play a significant role in review sentiment, offering actionable insights for restaurant owners and managers. Future work could explore integrating other features such as reviewer demographics and time of visit to enhance predictive power further, as well as extend the analysis to other cities for comparative insights across different regions.

There are scopes to for-
The further analysis has prospect for:

- There are opportunities for further analysis, including:
- Developing a processing pipeline.
- Implementing a neural network model for topic modeling.
- Applying cross-validation on sample selection and topic identification.
- Automating topic labeling based on high-frequency keywords.
- Expanding the analysis across datasets from all six cities to develop a large language model (LLM) for sentiment prediction.

8. References

- Dataset: The details on the dataset can be found here at Kaggle.
- Analysis: The Github Repo of this Analysis is here.
- Tools and Software: Python 3.9 (Most Commonly libraries listed below):
    - Numpy,
    - Pandas,
    - Matplotlib,
    - Seaborn,
    - NLTK,
    - Scikitlearn
    - Genism
    - Spacy.
- Hekpful Resources:
    - Machine Learning Plus: Topic Modeling in Python
    - Eugenia Anello's NLP Tutorial Series on Medium
    - Nanonets: Topic Modeling with LSA, PSLA, LDA, and LDA2Vec
    - Kaggle: A TripAdvisor Dataset for NLP Tasks
    - Zenodo Dataset Record