

# Improving Hotel Service Quality through Sentiment Analysis & Keywords Extraction

## Executive summary

The project focused on preprocessing a database of reviews on TripAdvisor's online site to better understand the strengths and weaknesses of a particular hotel. The analysis of the project can influence the choice of the final reservation of the customers and help the hotel to improve and enhance the service. The database consists of two main columns: text reviews and 1–5 star reviews. Sentiment Analysis and Keywords Extraction, combined with data preprocessing process methods are used to reach the goal.

## Project objectives

The study focuses on the Electronic Word of Mouth (eWOM) phenomenon, which refers to the online communication of consumer opinions and experiences through various platforms, including social media and online review sites such as TripAdvisor. The reviews posted by users on these sites have a significant impact on the public's perception of hotels, and can ultimately influence their decision to book a stay.

To address this, the proposed method involves text analysis of TripAdvisor's text-based user reviews to identify the most focused issues. The ultimate goal is to help hotels improve their services and increase customer satisfaction by addressing areas of improvement.

Positive comments indicate that users are more satisfied in these areas. For example, if “location” appears high in such reviews, it indicates that hotel location is a strong competitive advantage for them. Negative comments indicate that the hotel is in dire need of improvement in these areas, such as the presence of “parking” means that parking is difficult and users complain a lot about it.

After distinguishing between positive and negative comments, by categorizing the service situation based on emotional tendencies, the study can provide separate recommendations for each category. This approach allows hotels to address specific areas of concern and make targeted improvements to their services. Overall, the study provides a useful tool for hotels to understand their customers' needs and preferences and improve their services accordingly.

## Data description

We get the dataset that gathers hotel reviews and ratings data from the Trip Advisor website using web scraping techniques, which is a total of 20,491 reviews collected in 2016 from the research [Alam, M. H., Ryu, W.-J., Lee, S., 2016. Joint multi-grain topic senti- ment: modeling semantic aspects for online reviews. Information Sci- ences 339, 206–223.](#)

The dataset consists of two main columns: text reviews and corresponding star ratings on a scale of 1 to 5. This data provides valuable insights into customers' experiences and perceptions of hotels, allowing for further analysis on various aspects such as service quality, room cleanliness, and overall customer satisfaction.

## Methodology

We first perform data preprocessing by removing punctuation and stop words, converting text to lowercase, stemming, and lemmatizing the text, which is achieved in Text\_Normalization\_Function.ipynb file. Also, we add a column to transform ratings 1-2 to ‘negative,’ and ratings 3-5 to ‘positive.’ Then, we use VADER sentiment analysis to determine whether the reviews are positive or negative, which will assign a score from -1 to +1 to each review (see Figure 1-2). After this analysis, we separate the positive and negative reviews

using a threshold value and extract the top 15 keywords from each topic to identify the most important words in each category by LDA modeling. Also, we use the highest coherence score, lowest perplexity, and highest log-likelihood values to get the optimized value of the number of topics. Lastly, we analyze the keywords for the positive and negative reviews to identify the strengths and weaknesses of the hotel to provide business insights so that the hotel knows what they should highlight, and what they should improve.

The pros of this method are:

1. **Efficient data preprocessing:** By normalizing the text data, the model can reduce noise and extract the most important features from the reviews, improving the overall accuracy of sentiment analysis and topic modeling.
2. **Accurate:** By using VADER sentiment analysis, we can tune the parameters to get better results. Also, it is suitable for every domain knowledge including the hotel industry, so we can identify the keywords to be positive or negative more correctly.
3. **Insightful:** Analyzing the keywords and finding topics for the positive and negative reviews can provide valuable insights into the strengths and weaknesses of the hotel, helping them to make strategic decisions about what they should highlight and what they should improve.

The cons of this method are:

1. **Information Loss:** In this case, the original data rating includes a scale from 1 to 5, but we transform these data into positive and negative categories to make it easier for the VADER analysis. This process may lose some information and leads to limited analysis.
2. **Context Effects:** Sentiment analysis can be influenced by the context in which the text is used, and certain words, such as buzzwords, cannot be accurately identified. Without understanding the context, it can be challenging to interpret the results accurately.

## **Results and Discussion**

Utilizing VADER sentiment analysis, we were able to achieve a high level of accuracy with a rate of 0.895 by setting a threshold value of 0.1. In addition to this, we compared coherence scores, log-likelihood values, and perplexity values to determine the ideal number of topics 3 to 5 for positive and negative comments. Through this analysis, we identified that 3 topics were the most suitable for positive comments, while 3 topics were optimal for negative comments. This shows the comments are homogeneous, and hotels should be aware of these features.

For those positive comments, the three topics people concern about are (see Figure 3-5):

1. **Basics:** including rooms, clean, breakfast. This accounts for 54% of tokens.
2. **Medium:** including beach, resort. This accounts for 43.3% of tokens.
3. **Luxury:** including spa, club, villa, island. This accounts for 2.7% of tokens.

For those negative comments, the three topics people concern about are (see Figure 6-8):

1. **Location and Facilities:** including rooms, bathroom. This accounts for 48.8% of tokens.
2. **Experience:** including foods, resort, and beach. This accounts for 44.9% of tokens.
3. **Payment and safety:** including additional charges, stolen. This accounts for 6.3% of tokens.

Based on the results of our analysis, we can provide some useful suggestions for hotels to improve their services. Firstly, for positive comments, hotels should maintain these strengths to stay competitive such as room cleanliness, breakfast quality, and the overall room experience, and also medium-level features such as the beach and resort facilities. Lastly, hotels can consider adding luxury features such as spa, club, villa, and island experiences to further enhance customer satisfaction.

On the other hand, for negative comments, hotels should prioritize improving the facilities, particularly the quality of the rooms and bathrooms. They should also focus on enhancing the overall customer experience, such as the quality of the food, the beach, and the resort facilities. Lastly, hotels should pay attention to payment and safety issues, such as avoiding additional charges and ensuring the safety of customers' belongings.

It is also worth noting that customers in both positive and negative comments share some common concerns such as food, resort, and beach experiences. Therefore, hotels should pay special attention to these areas to ensure customer satisfaction. Overall, by addressing these key issues, hotels can improve their services and enhance customer experience, leading to positive reviews and improved public perception.

## **Conclusion**

### Recommendations:

1. Hotels should prioritize improving room and bathroom quality, food, and resort facilities based on negative reviews.
2. Hotels should continue to maintain high standards in areas identified as strengths in positive reviews, such as cleanliness and breakfast quality.
3. Given the overlap of concerns in both positive and negative reviews (food, resort, and beach experiences), these areas should be given special attention.
4. The introduction of additional luxury features has the potential to build on positive feedback and encourage more positive reviews.
5. Safety and transparency in charges should be addressed to prevent potential negative reviews.

### Shortcomings of this research:

1. **Information Loss:** The conversion of the 1-5 star rating system into a binary positive/negative category led to a loss of nuanced information. To address this, future work could consider developing a multi-class sentiment analysis model that can handle the full range of star ratings.
2. **Human Bias:** The explanations of keywords might introduce bias. Future analysis could employ more objective methods of keyword selection.
3. **Context Effects:** Sentiment analysis is context-dependent, and without understanding the context, it can be challenging to interpret the results accurately. Future work could involve more complex NLP models that can capture the context of sentences, such as transformers or recurrent neural networks.
4. **Homogeneous Comments:** The analysis showed that comments were homogeneous. Although this makes it easier to identify areas of focus, it might be beneficial to analyze more diverse reviews to capture a wider range of customer opinions.

## Appendix

VADER Scores for ACTUAL Positive Reviews  
(in decreasing order of VADER scores)

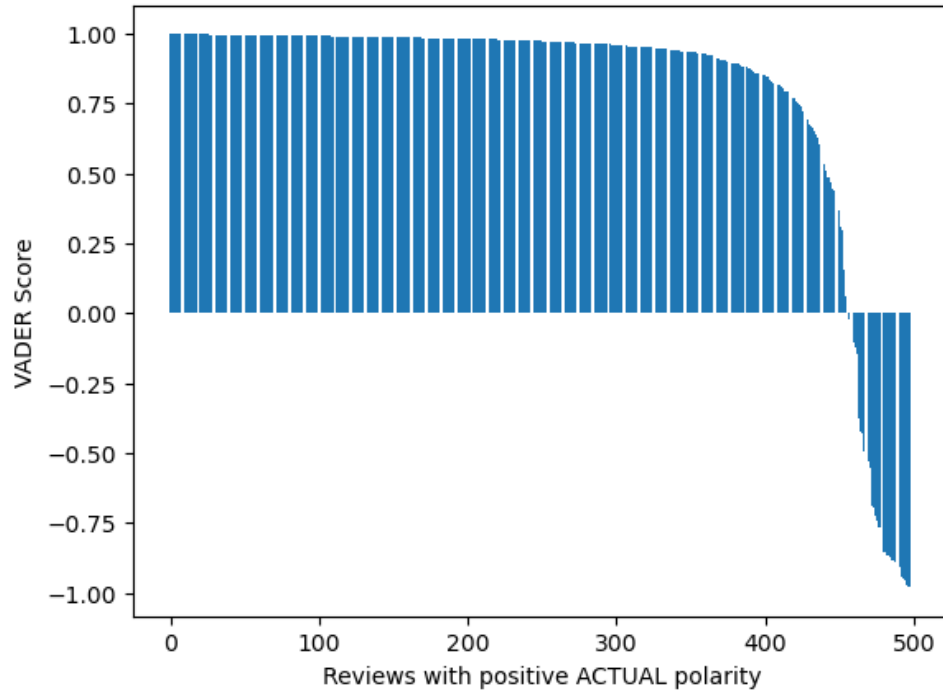


Figure 1 VADER Scores for Positive Reviews

VADER Scores for ACTUAL Negative Reviews  
(in ascending order of VADER scores)

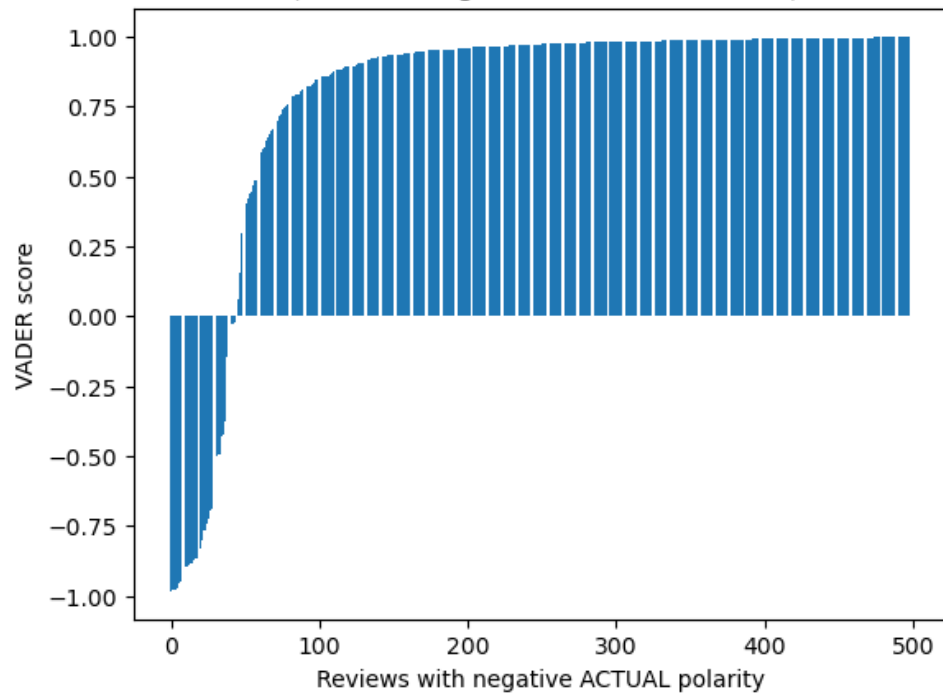


Figure 2 VADER Scores for Negative Reviews

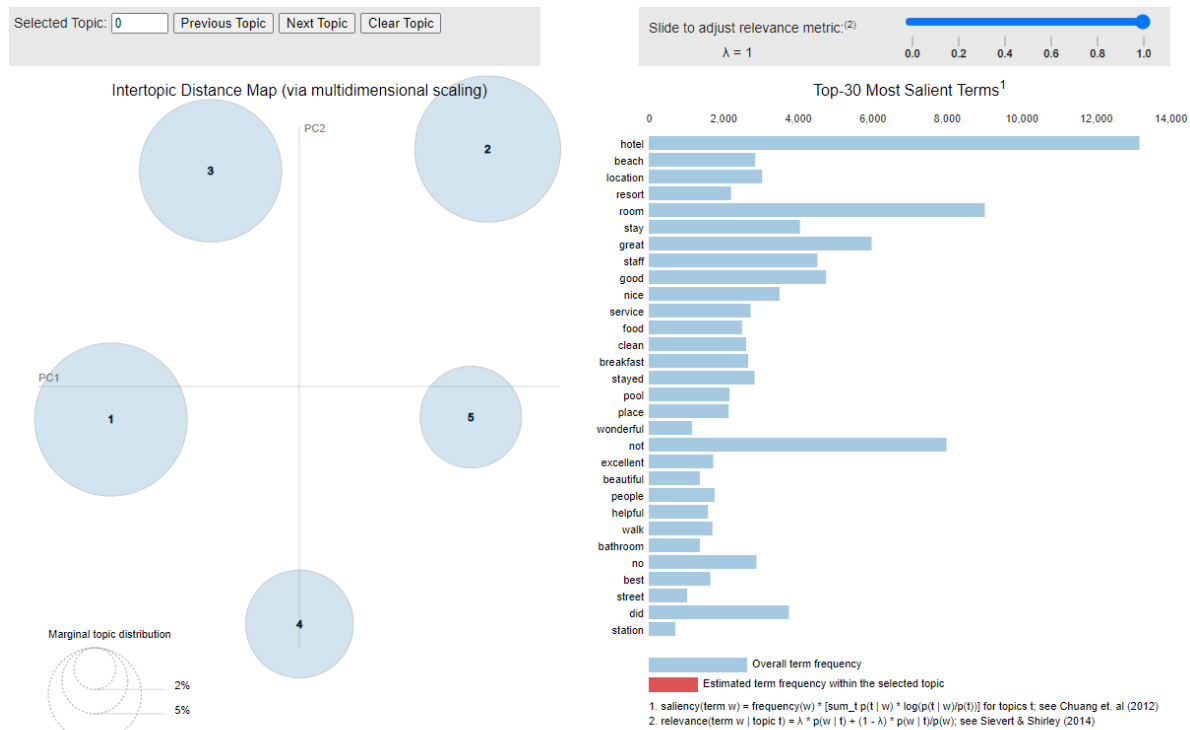


Figure 3 Topic Model Visualization for Positive Reviews with 5 Topics

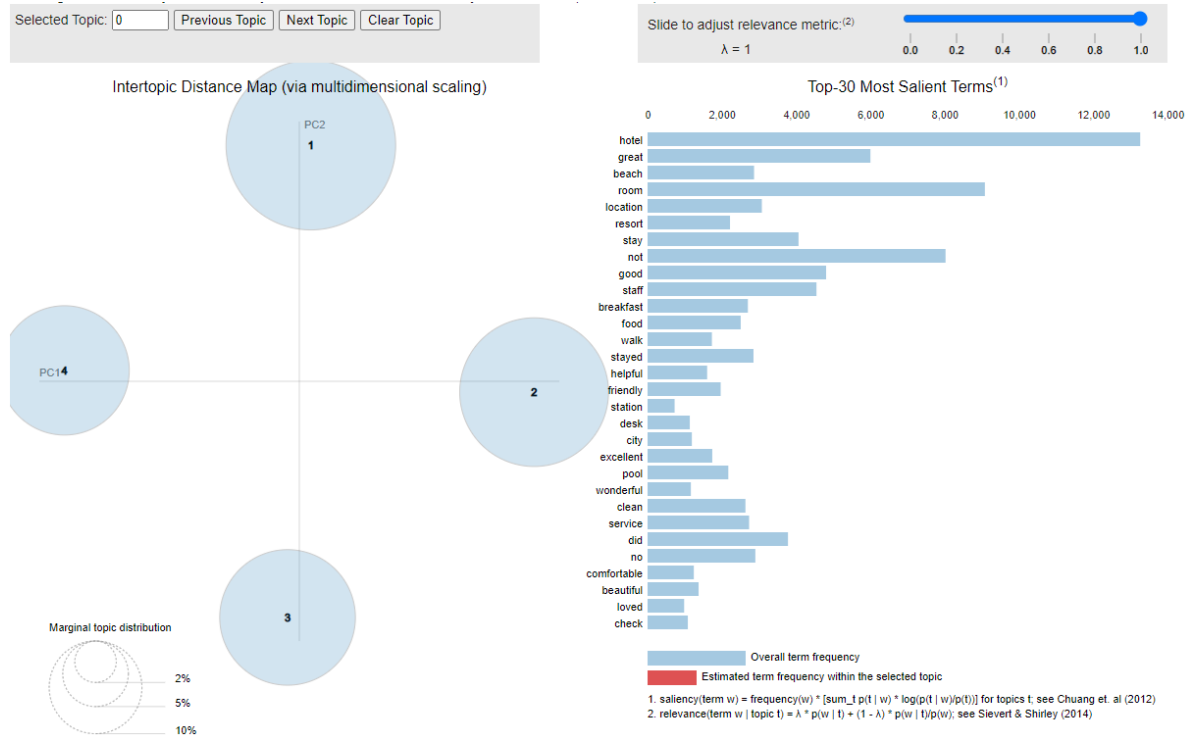
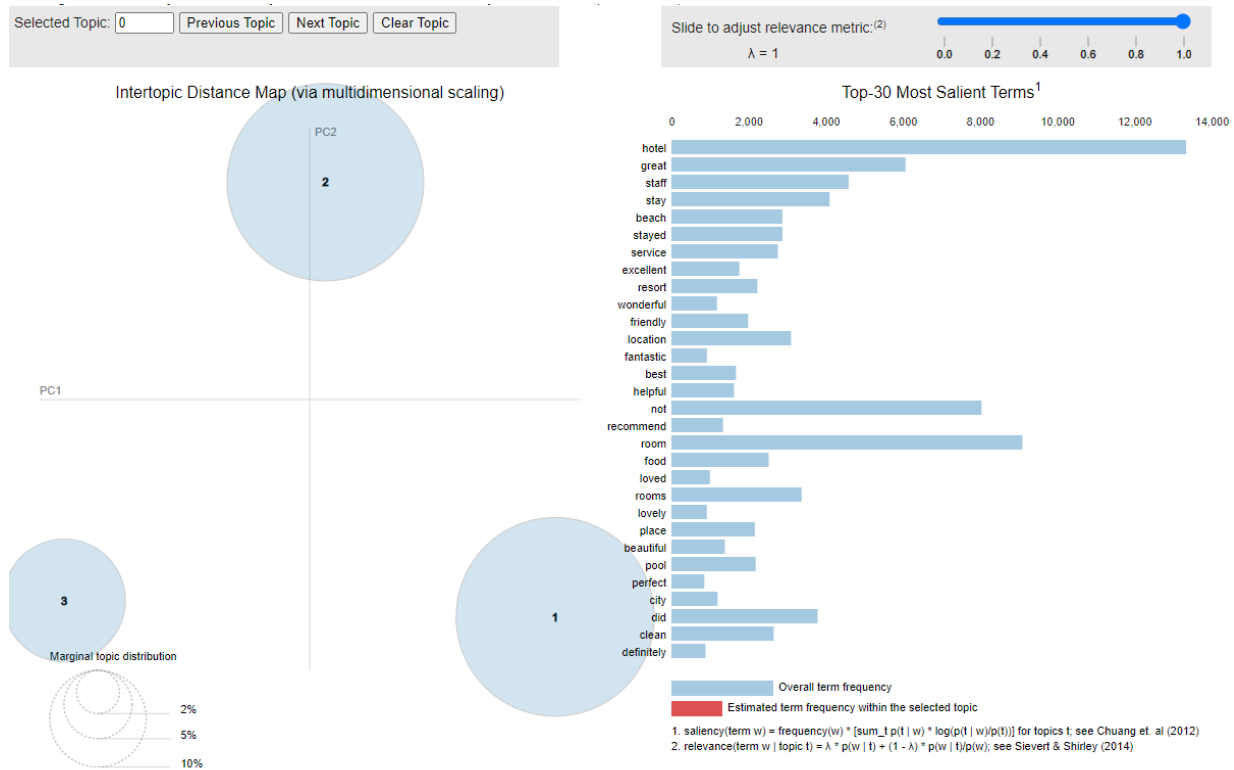
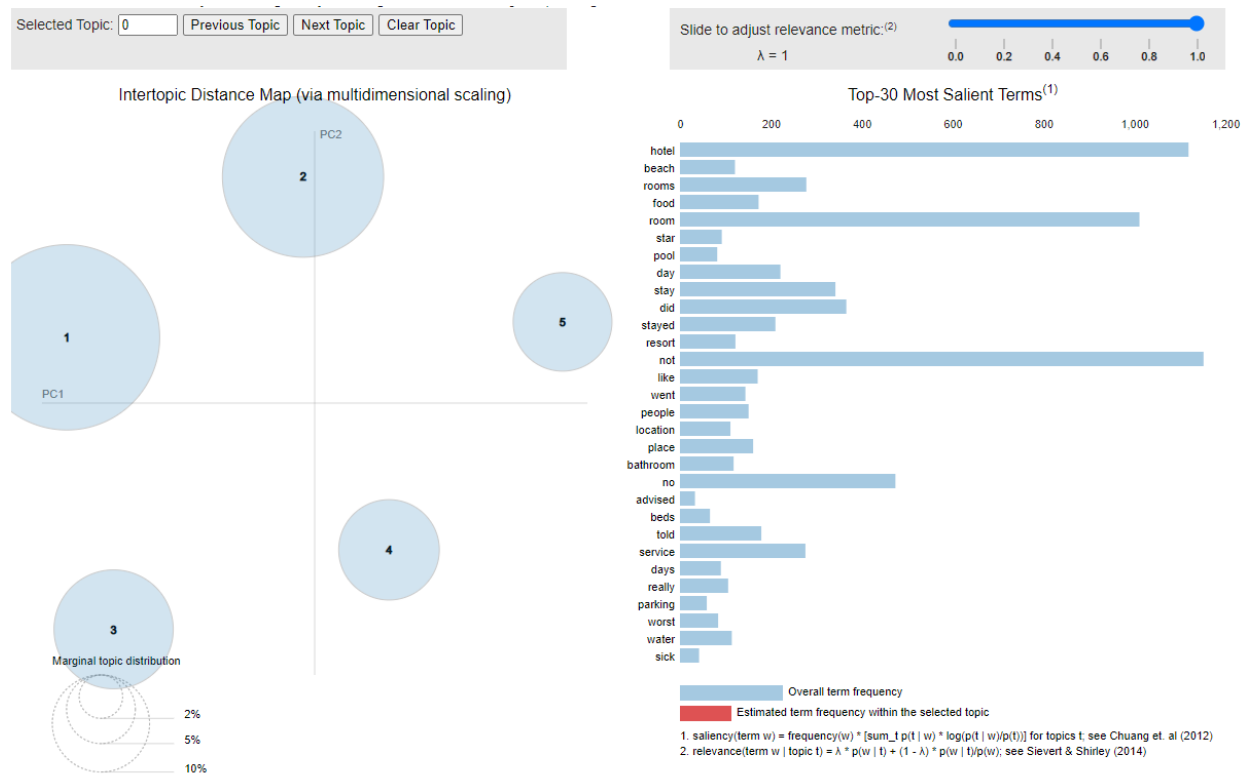


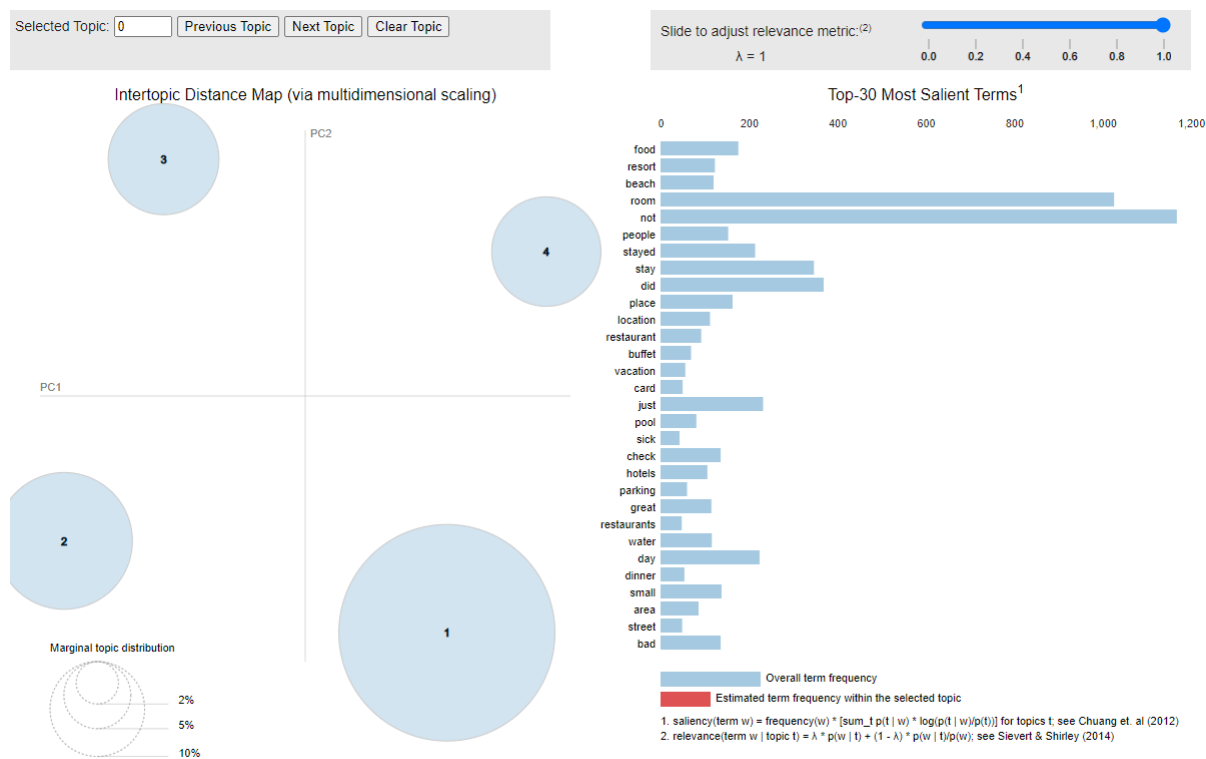
Figure 4 Topic Model Visualization for Positive Reviews with 4 Topics



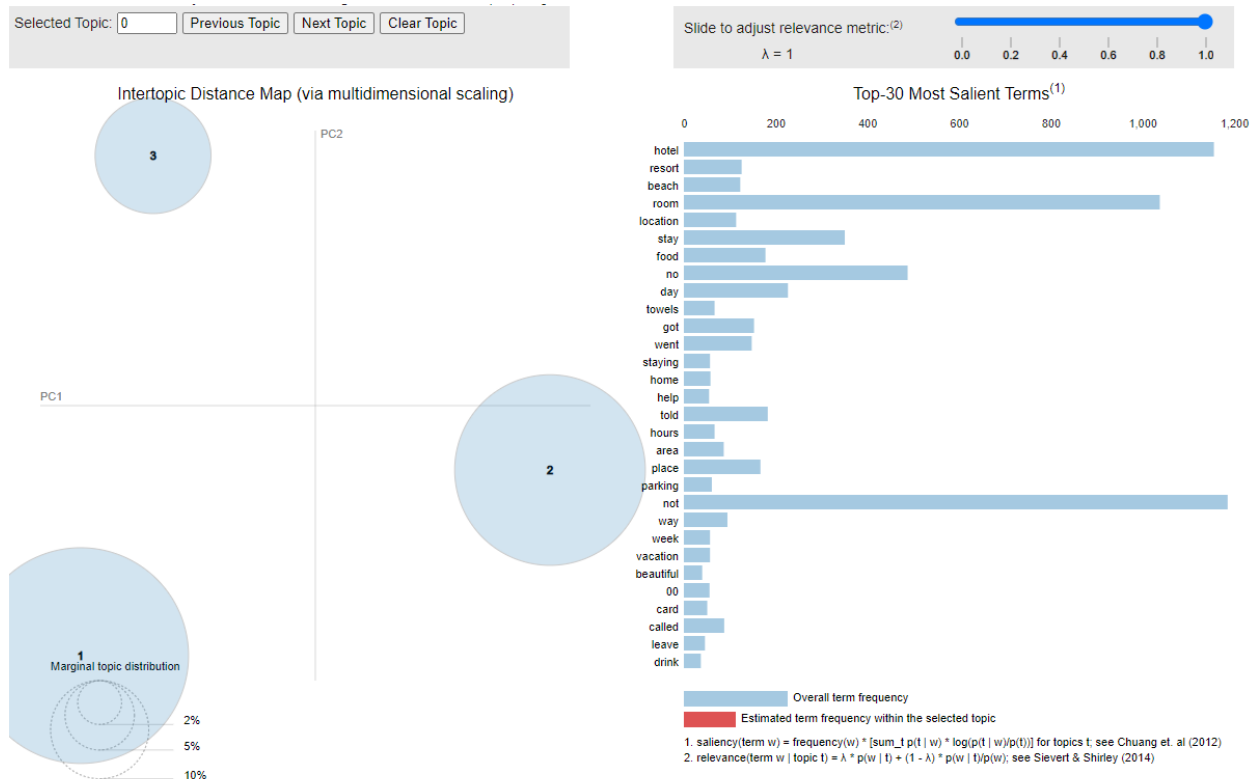
**Figure 5 Topic Model Visualization for Positive Reviews with 3 Topics**



**Figure 6 Topic Model Visualization for Negative Reviews with 5 Topics**



**Figure 7 Topic Model Visualization for Negative Reviews with 4 Topics**



**Figure 8 Topic Model Visualization for Negative Reviews with 3 Topics**

## References

Bansal, B. (2018, April 17). *Tripadvisor Hotel Review Dataset*. Zenodo. Retrieved May 1, 2023, from <https://zenodo.org/record/1219899#.ZFbsTXbMKUm>