Topic Modeling and Visualization Report

Overview

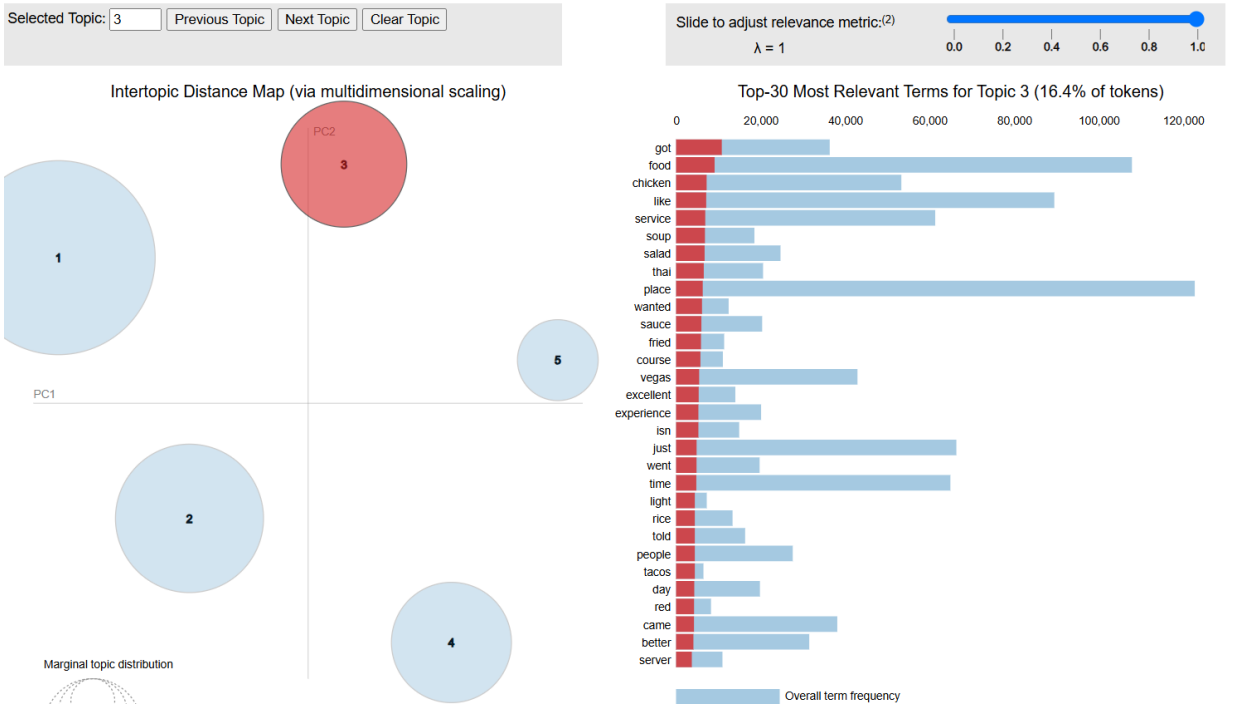
In this project, I applied Latent Dirichlet Allocation (LDA) to extract and analyze topics from Yelp reviews. The dataset, sourced from the Yelp Academic Dataset, contains reviews that provide valuable insights into user experiences. I focused on three subsets of reviews: all reviews, positive reviews, and negative reviews. To manage computational constraints and fit the memory requirements, I analyzed 60% of the data for each subset, sampled randomly.Methodology

used tools:

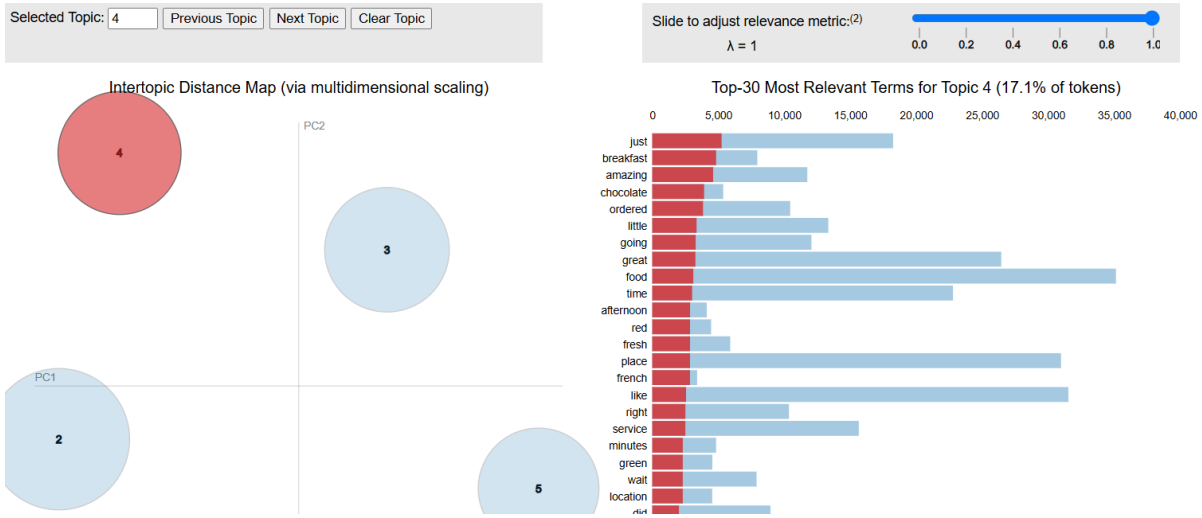
- **Scikit-learn** for LDA modeling and preprocessing
- **CountVectorizer** for creating the document-term matrix
- **pyLDAvis** for interactive topic visualization

I set the number of topics to 5 for all experiments and limited the vocabulary by excluding extremely common or rare words.
Reviews were filtered based on their star ratings to form subsets. Text data was tokenized and transformed into a document-term matrix using CountVectorizer. Stop words were removed, and terms appearing in fewer than 2 reviews or more than 95% of the reviews were excluded.
LDA was applied to identify underlying topics in each subset. I analyzed the distributions of words within topics and their relevance.
I used pyLDAvis to generate topic visualizations. This interactive tool allowed us to explore how topics were distributed and interpret key terms for each.

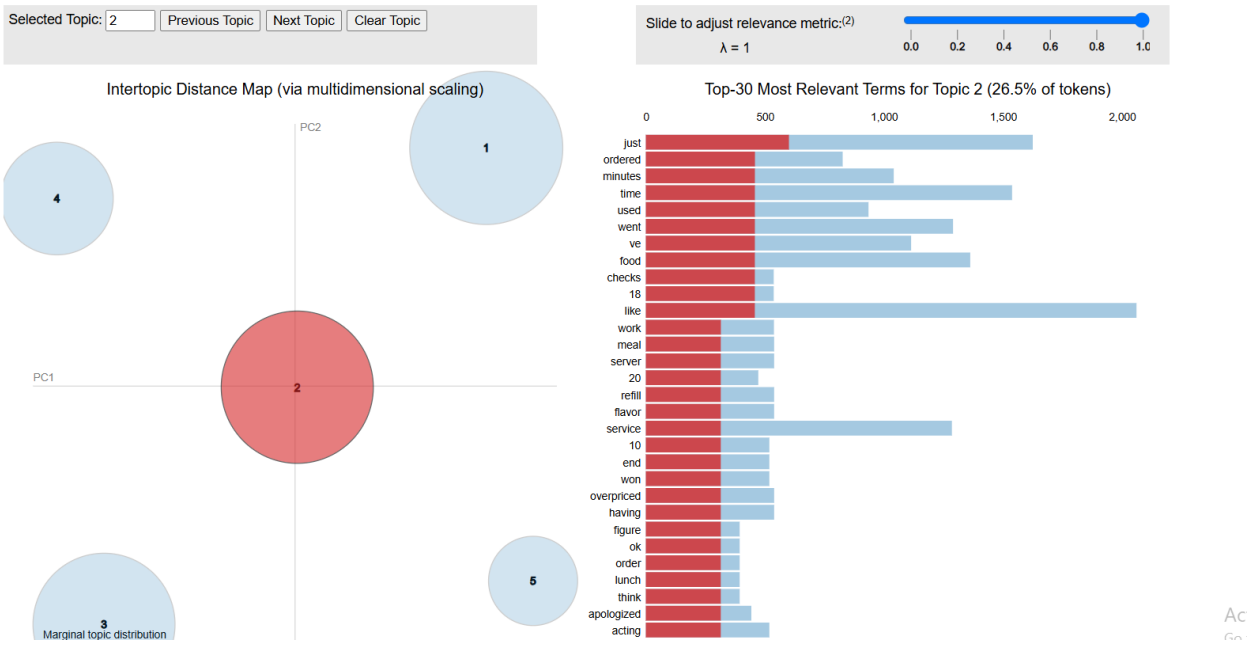
1. **All Reviews**:



2. **Positive Reviews**:



3. **Negative Reviews**:



The results make intuitive sense and provide actionable insights. For example, positive reviews tend to focus on experiences and quality, while negative reviews emphasize dissatisfaction and specific issues. The visualization effectively highlights the overlap and distinctions between topics in different subsets.

This analysis demonstrates the power of topic modeling for understanding customer feedback. Future work could expand by incorporating sentiment analysis or evaluating how topics evolve over time.