Similarity, Clustering, Topic Modeling

Scholar Works:

Fei Lan (2022) presents a hybrid text similarity measurement method that combines TF-IDF with semantic information from HowNet. While TF-IDF struggles with high-dimensional, sparse vectors and lacks semantic depth, this method reduces dimensionality by selecting the most relevant terms based on both TF-IDF values and semantic similarity. This is achieved through a term similarity weighting tree (TSWT), which enhances similarity calculations by integrating semantic relationships. This hybrid method is effective for datasets with unstructured or semi-structured text, such as text clustering tasks, where traditional TF-IDF methods may fail.

Gabriela Nathania H. et al. (2020) used TF-IDF for extractive summarization of hotel reviews, extracting key terms to create concise summaries. However, they noted that TF-IDF's reliance on word frequency misses deeper meanings, such as sentiment. Similarly, Zhang et al. (2022) used K-means clustering to segment restaurant reviews into clusters based on sentiment and features like food quality. Despite its utility, K-means struggled with mixed sentiments, making it difficult to define distinct clusters.

These findings are relevant to my project, where I’m using TF-IDF and K-means clustering for restaurant review analysis. The limitation of TF-IDF in capturing deeper semantic meaning, especially in food-related terms like "chicken" or "ramen," was evident in my trial. This confirmed that context-aware techniques would be more suitable for detecting semantic differences in reviews.

My Trial:

TF-IDF was helpful for extracting common terms across reviews, but its focus on frequency resulted in misleading insights. For example, frequent mentions of "chicken" didn’t reflect customer sentiment. Despite adjusting the n-gram range and removing stop words, the results didn’t improve significantly. When clustering with K-means, I expected sentiment-based clusters but found mixed sentiments within clusters, complicating theme analysis. Similarly, cosine similarity-based review selection showed no significant change. Whether selecting the first review in each cluster or using the average score of each cluster, the reviews selected were the same across both methods. This consistency further underscores the limitations of TF-IDF, as it suggests that TF-IDF’s reliance on frequency alone fails to capture the true semantic differences between reviews, especially in cases where sentiment or contextual nuance plays a critical role in distinguishing customer experiences.

To improve, I’ll explore alternative approaches like word embeddings (e.g., Word2Vec or BERT) for better semantic analysis and clustering accuracy.

The use of TF-IDF and K-means in restaurant reviews highlighted the challenge of capturing subtle cultural nuances in consumer sentiment. TF-IDF’s reliance on frequency alone doesn't account for deeper meanings, such as the cultural significance of dishes like "chicken" or "ramen," which vary based on personal experience and context.

Findings after LDA and Count Vectorizer:

In the LDA analysis, several topics (e.g., Topics 0, 1, 2, and 4) for Chinese cuisine had average scores below 0.01, indicating lower significance. These topics likely cover service quality, ambiance, or price. Both Chinese and Japanese cuisines showed similar topic scores for Topics 4 and 5, emphasizing aspects beyond food like service quality. Despite this, **Topic 9** in both cuisines highlighted that **food quality** and specific dishes (e.g., chicken for Chinese and sushi for Japanese) are central to consumer experiences and how dishes are named and interpreted in reviews.