PCA, Word Embeddings, Topic Modeling

Scholar:

PCA has become a fundamental method for dimensionality reduction, especially in high-dimensional datasets such as text data. Sparse PCA, in particular, has been explored for its ability to improve interpretability by selecting the most significant variables while maintaining the dimensionality reduction properties of traditional PCA. Zou et al. (2006) highlighted that Sparse PCA is ideal for high-dimensional, sparse data, like the one found in text mining, where many features are not informative. They emphasized how Sparse PCA allows for better model interpretability by focusing on key components, which is crucial in domains like social media or consumer review analysis, where many features (words) contribute little to the variance. Similarly, Bingham and Mannila (2001) introduced random projections as an efficient alternative to PCA for dimensionality reduction, particularly in large datasets. Their approach minimizes computational costs while preserving important pairwise distances, making it applicable for text data, which often has large and sparse representations. Moreover, Aggarwal and Zhai (2012) emphasized the combination of clustering with dimensionality reduction methods like PCA to improve the classification of textual data. They illustrated how combining K-means clustering with dimensionality reduction enhances the interpretability of topics and improves clustering performance, especially when dealing with high-dimensional text.

These findings are closely related to my project, where I use PCA to analyze restaurant reviews for Chinese and Japanese cuisines. I initially tried applying PCA to bag-of-words representations of the review data, following established approaches. However, the results were unsatisfactory due to the sparse nature of the bag-of-words model. Sparse data often leads to suboptimal performance in PCA, as PCA struggles to effectively handle the sparse matrix without losing interpretive granularity. The sparsity in the bag-of-words representation meant that PCA could not identify the underlying structures or meaningful patterns in the reviews effectively. To address this issue, I shifted to Word2Vec embeddings, which provide dense vector representations of words and help capture semantic meaning. The embeddings improved PCA's performance, as the dense vectors provided richer, more informative features. However, despite this improvement, interpreting the principal components (PC1 and PC2) was still challenging. In particular, components were dominated by terms like "incident," "serious," and "customers," which, although relevant to the reviews, were not the central themes related to food quality or dining experience.

Trail:

During my experiment, I faced several difficulties with applying PCA on the Word2Vec embeddings. Initially, I struggled to interpret the components, as they often included terms that did not seem directly relevant to food-related themes, such as "incident" and "serious." I also had to consider how to handle the word embeddings' sparseness and its effect on PCA's ability to retain useful variance across the components. These issues raised several questions: How can I optimize PCA to better capture the variance in consumer reviews? Should I experiment with alternative methods, like BERT embeddings, to capture deeper contextual meanings in the reviews? Would using a different dimensionality reduction method be more suitable for this dataset? Also, the results were not always as interpretable as expected, and additional preprocessing or using more context-sensitive models like word embeddings could potentially improve the quality of the analysis.

Meaning:

In terms of social and cultural implications, these findings suggest that cultural differences play a significant role in shaping consumer perceptions of food, especially when analyzing restaurant reviews. In both datasets, terms related to food and specific dishes (like "buns" for both cuisines) appear prominently. However, the weight of terms related to service and price seems higher in Chinese reviews, suggesting that service quality and cost might be more significant factors for consumers when evaluating Chinese restaurants compared to Japanese ones. Both cuisines have negative connotations with terms like "incident," "serious," and "eaten," but the prominence of terms like "$100+" and "maintained," and the repeated appearance of "again!!!" in Chinese reviews could imply that pricing and service quality have a more direct impact on negative sentiments in Chinese reviews. While my initial approach using PCA struggled with capturing the finer cultural distinctions, this method still provided a starting point for exploring the broader societal narratives that underlie consumer experiences. For the next step, I am considering extending this analysis by integrating word embeddings or even sentiment analysis techniques to get a more nuanced understanding of the relationships between words, topics, and negative reviews.