LAB for week 3: PCA and K-mean Clustering

Investigate the Method's Current Usage:
Recent advancements in recommendation systems have demonstrated the effectiveness of combining Principal Component Analysis (PCA) with K-means clustering to improve accuracy and computational efficiency. For example, in "A New Approach for Movie Recommender System using K-means Clustering and PCA" (2022), researchers used PCA to reduce the dimensionality of user rating data, capturing key patterns in user preferences. K-means clustering was then applied to group users with similar tastes, enhancing recommendation accuracy and reducing processing time. This approach proved effective in managing high-dimensional data and addressing sparsity issues.

Similarly, in the e-commerce domain, Thakur and Mandal (2022) employed PCA and K-means to handle sparse user-product interactions. PCA was used to capture essential features of user behavior, while K-means grouped users with similar purchase patterns. This combined method yielded a notable improvement in recommendation accuracy by lowering RMSE, illustrating the benefits of PCA and clustering in personalized recommendations.

Moreover, the integration of multimodal data—specifically text and images—has become essential in recommendation systems, especially for visually driven domains like e-commerce. A recent survey by Zhu et al. (2023) emphasized that combining textual and visual data allows systems to capture complementary information, enhancing the recommendation quality. By leveraging both text (e.g., product descriptions) and images, multimodal systems can provide a richer, more accurate understanding of user preferences.

In our system, we employ PCA and K-means clustering on text and image data, enabling us to capture core patterns and group similar items effectively. This approach, validated by recent studies, might allow for a robust recommendation system that meets the diverse needs of users while efficiently handling high-dimensional and sparse data.

Test the Method on Your Data:

Throughout my project, I explored several approaches to develop an effective recommendation system, beginning with text-based features and later transitioning to image-based analysis. Initially, I experimented with TF-IDF and KNN, adjusting the numFeatures parameter for TF-IDF from 500 to 2000 and testing various PCA dimensions from 50 to 150. However, the results were disappointing: the explained variance per principal component never exceeded 0.0006. This low variance is typical in sparse text data, where each word only appears in a limited subset of documents. In my case, the product titles were often short and filled with character names from anime or movies, which limited the overlap or context between items. When running KNN, I observed that the Silhouette score only improved when the number of neighbors matched the number of PCA dimensions, suggesting that the variance was spread thinly across many components. These findings led me to conclude that text data alone would not provide enough predictive value.

Pivoting from text, I moved to image-based analysis, implementing OpenCV’s ORB descriptor to extract image features. This approach initially seemed promising, as visual information could add context not captured by text. I processed each image, flattened ORB descriptors into vectors, and stored them as features. However, transferring and maintaining these high-dimensional vectors in a scalable format presented logistical challenges; some vector dimensions were lost in the process, complicating downstream analysis. I resolved this issue by storing the vectors in a Parquet format, which ensured data integrity for use with PySpark.

With the image vectors intact, I implemented PCA to reduce dimensionality, making the dataset more manageable for computationally intensive tasks like K-means clustering. While PCA effectively reduced feature dimensions, the clustering phase still proved costly in terms of computation. Moreover, tuning the k parameter for optimal clustering was impractical due to the dataset’s size. The evaluation metrics did not indicate a significant improvement, leading me to reconsider this approach.

In conclusion, both text and image-based feature extraction methods presented limitations in terms of computational efficiency and predictive power. Given these insights, I plan to shift my focus to collaborative filtering, which may offer a more efficient and effective path for building a recommendation system tailored to my data.

Reflect on Social/Cultural Implications of your findings:

Reflecting on my analysis using PCA and K-means clustering, I've found that these methods may not be the best fit for my dataset due to the nature of my data and the computational demands of these techniques. My project aims to develop a recommendation system, initially exploring text-based and image-based features to enhance the recommendation quality. However, both PCA and K-means encountered limitations with the structure and scale of my data. PCA effectively reduced dimensionality but yielded low explained variance for each principal component in the text data, suggesting that the sparse, fragmented nature of short product titles (often consisting of anime or movie character names) doesn’t lend itself well to this method.

K-means clustering further presented challenges; the large size of my dataset made tuning the k parameter computationally infeasible, and the evaluation metrics showed limited improvement in clustering quality. This implies that alternative approaches, such as collaborative filtering, may be better suited for this context.

From a social perspective, my findings raise questions about the limitations of algorithmic systems when applied to culturally specific or niche datasets. The methods I used here rely on identifying underlying patterns within high-dimensional data, but they seem less effective with sparse, specialized content like anime and movie-related product titles. This underscores the importance of tailoring recommendation systems to reflect the nuances of cultural and social interests, as one-size-fits-all methods may not fully capture the diversity of user preferences within specific fandoms. Using standard clustering and dimensionality reduction techniques on culturally unique datasets may yield superficial patterns that don’t align with users’ true interests.