Clustering and Classification Report

Yunzhen WU

Date: May 4, 2025

1. Introduction and Problem Statement

This assignment continues the investigation into natural language processing techniques, building on the initial corpus-based vectorization work from Assignment 1. The core objectives include evaluating clustering algorithms, sentiment classifiers, and topic modeling methods to assess their effectiveness in organizing and interpreting a collection of movie reviews. The dataset contains labeled film reviews, including metadata such as genre and sentiment polarity (positive/negative), offering multiple avenues for supervised and unsupervised learning. The task encompasses four major components: document clustering, sentiment classification using traditional and transformer-based models, multi-class classification based on genre, and unsupervised topic modeling.

2. Data and Preprocessing

The dataset contains a corpus of 230 documents with metadata, including review text, sentiment (positive or negative), and genre (comedy, action, drama, horror). The preprocessing stage involved text normalization via a custom clean_doc function that tokenized, removed punctuation, filtered out short and non-alphabetic words, and English stopwords. Unlike the previous assignment that employed stemming, this iteration switched to lemmatization, as it yielded better downstream classification performance.

For vectorization, TF-IDF with unigrams and BERT embeddings and classification via HuggingFace transformers were used. The genre labels were encoded using LabelEncoder, and for classification, we used stratified train_test_split to ensure representative evaluation.

3. Research Design and Modeling Methods

Four experimental components structure this assignment:

- Part 1 Clustering Experiments: The first analysis involved unsupervised clustering of documents using the k-means algorithm on tf-idf vectors. Clustering performance was evaluated across multiple values of k using the Silhouette score. Visualizations such as t-SNE plots were used to inspect cluster formation. Clustering was assessed against genre as a proxy for ground truth to examine consistency and quality.
- Part 2 Sentiment Classification: Sentiment analysis was performed using both traditional models and transformer-based models. Ground truth consisted of binary sentiment labels. Traditional models included Support Vector Machines, Logistic Regression, Naive Bayes, and Random Forest classifiers, trained on tf-idf vectors. These were evaluated using accuracy, confusion matrix, and F1-score. In parallel, a pre-trained BERT model was used to classify the same documents. Performance was compared across three input formats: raw text, cleaned raw text, and final preprocessed text. This enabled a fair assessment of how vectorization and preprocessing affect transformer-based classification.
- Part 3 Multi-class Genre Classification: This experiment expanded classification to four genres, using the same traditional models and BERT. The tf-idf matrix served as input for traditional classifiers, while BERT used both raw and processed texts. Label encoding was applied, and classification performance was again evaluated using accuracy, F1-score, and confusion matrices.
- Part 4 Topic Modeling: Topic modeling was conducted using LSA (Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation). The goal was to extract topics that either aligned with sentiment or genre. Multiple topic counts (2, 4, and 6) were tested. Top

keywords per topic and associated documents were inspected to interpret coherence. Additionally, BERTopic was applied as a modern alternative that combines transformer embeddings with clustering techniques to derive interpretable topic structures. Outputs such as topic word lists, document-topic heatmaps, and topic distributions were examined.

4. Results

4.1 Clustering Experiments

Clustering was first applied to vectorized document representations to group similar movie plots. Using the TF-IDF representation, KMeans model was tested. Figure 8 illustrates the effect of KMeans clustering in TF-IDF vector space, revealing a moderate degree of thematic overlap among documents. The silhouette score was computed across different k values to determine optimal cluster count, as shown in Figure 7. The silhouette score peaked around k =23, which is the same as the number of movies in the corpus, indicating that a higher number of clusters marginally improved cohesion and separation.

Despite the optimized , many clusters were poorly defined or exhibited significant overlaps. This was further illustrated by the t-SNE plot in Figure 8, where no clear boundaries emerged among clusters. Analysis of grouped documents (Figure 8) showed that genres were intermixed within clusters, and outliers persisted.

4.2 Sentiment Analysis Experiments

(a) Sentiment with Prior Vectorization

Sentiment classification was performed using both traditional machine learning classifiers and different vectorization techniques. Logistic regression, SVM, Naive Bayes, and

Random Forest were evaluated. For binary classification of positive vs. negative reviews using TF-IDF vectors, the results were unsatisfactory. The accuracy for Logistic Regression, SVM, and Naive Bayes plateaued at 39.13%, highlighting the limited discriminative power of TF-IDF in this setting, but the accuracy for Random Forest is 50.72%, outperforming other three methods.

(b) Sentiment Using Pretrained BERT

The experiment using BERT-based sentiment classification demonstrated a significant performance increase. On raw text, the pretrained model yielded over 80% accuracy. Initially, accuracy dropped to 30% when applied to the fully preprocessed version of the dataset. After removing over-aggressive cleaning steps such as stemming and refining stopword selection, accuracy improved to approximately 60%. This reinforced the importance of preserving semantic richness for transformer-based models.

4.3 Multi-Class Genre Classification

Genre classification was carried out using four target labels: Action, Comedy, Horror, and Sci-Fi. A confusion matrix from the TF-IDF + traditional classifiers experiment (Figure 9) revealed that most documents were misclassified as Sci-Fi. Logistic Regression, SVM, and Naive Bayes all reported 100% accuracy in initial runs; however, this was due to training-test leakage or poor generalization stemming from limited testing splits. When the test size was increased to 30%, the classification performance remained artificially high for these models, while performance for BERT remained more grounded.

4.4 Topic Modeling Experiments

LSA, LDA, and BERTopic were employed to perform topic modeling. Figures 1, 2, and 3 show topic word distributions for 2, 6, and 20 topics respectively. These topic groupings only loosely corresponded to genres or sentiment classes. BERTopic yielded more meaningful clusters when visualized via heatmaps (Figure 6) and topic-document matrices (Figure 5), demonstrating its superior ability to extract high-quality semantic representations. Some topics identified by BERTopic (Figure 6) strongly aligned with identifiable narrative themes or character motifs, supporting its robustness in multi-topic scenarios.

5. Analysis and Interpretations

The performance gaps across different modeling paradigms underscore the interplay between representation learning, model architecture, and data characteristics. Clustering experiments highlighted the limitations of sparse vector representations in delineating thematic boundaries. Despite parameter tuning, KMeans on TF-IDF vectors produced clusters with significant genre overlap, indicating that surface-level lexical patterns may not sufficiently capture latent semantic relationships. This limitation was reflected in the low silhouette scores across cluster configurations, and the t-SNE visualization failed to reveal coherent groupings.

Traditional sentiment classification methods struggled with binary sentiment prediction, particularly when using TF-IDF inputs. This suggests that lexical frequency alone cannot adequately represent the nuanced emotional tone embedded in movie reviews. The relative success of Random Forest, achieving near 50% accuracy, may be attributed to its ensemble mechanism capturing some feature interactions. However, the performance remained far below practical usability thresholds.

In contrast, Transformer-based models show higher accuracy, especially when preserving semantic structure. At the beginning of the job, after deep cleaning of the input, performance degradation was observed, which was later improved to 60% accuracy by removing stop words and other refinements. This confirms the necessity of deep language models to strike a balance between text normalization and information content preservation.

Genre classification results reveal more limitations of traditional classifiers. At first, the accuracy data seemed to be exaggerated, perhaps due to data leakage or imbalanced training-test set splits. But corrected experiments show that these models still fail to generalize meaningfully. BERT's performance, while not perfect, remains more robust across all genres. The confusion matrix shows that the predictions are strongly biased towards the science fiction category, which may reflect the genre imbalance in the corpus or semantic ambiguity between different genres.

Topic modeling provides different insights. While LSA and LDA are effective in revealing latent structure, the topics they generate lack clear semantic alignment with sentiment or genre. Their vocabulary distributions, while interpretable, are scattered and sometimes redundant. In contrast, BERTopic demonstrates a higher ability to construct interpretable clusters. Heatmaps and topic-document matrices provide visually clear groupings, and related keywords hint at underlying topics such as character archetypes, emotional tone, or narrative themes. This demonstrates that Transformer-based embeddings combined with modern clustering techniques can provide superior semantic resolution in an unsupervised setting.

6. Conclusions

This assignment presents a comprehensive evaluation of clustering, classification, and topic modeling techniques applied to a corpus of film reviews. The experiments reveal the

limitations of sparse representations and traditional models in both supervised and unsupervised settings. Transformer-based methods, especially BERT and BERTopic, outperform traditional methods in sentiment classification and topic modeling tasks, further highlighting the importance of context-aware semantic representations.

The results also highlight the critical role of preprocessing in the NLP pipeline. Over-normalization can strip away meaningful context, especially when used with deep models that rely on rich linguistic cues. Carefully designed cleaning strategies and the choice of embedding methods can significantly affect downstream performance.

Although genre classification and clustering remain challenging due to label ambiguity and limited data, the experiments demonstrate meaningful patterns in text representations handled by different modeling techniques. BERTopic, in particular, emerges as a promising framework for extracting topic structures from unlabeled documents.

Future work could be carried out in the direction of expanding the dataset, incorporating more nuanced genre labels, and trying hybrid models that combine rule-based and neural techniques.

Appendix

Figure 1: Topic-Word Distribution (2 Concepts using LSA)

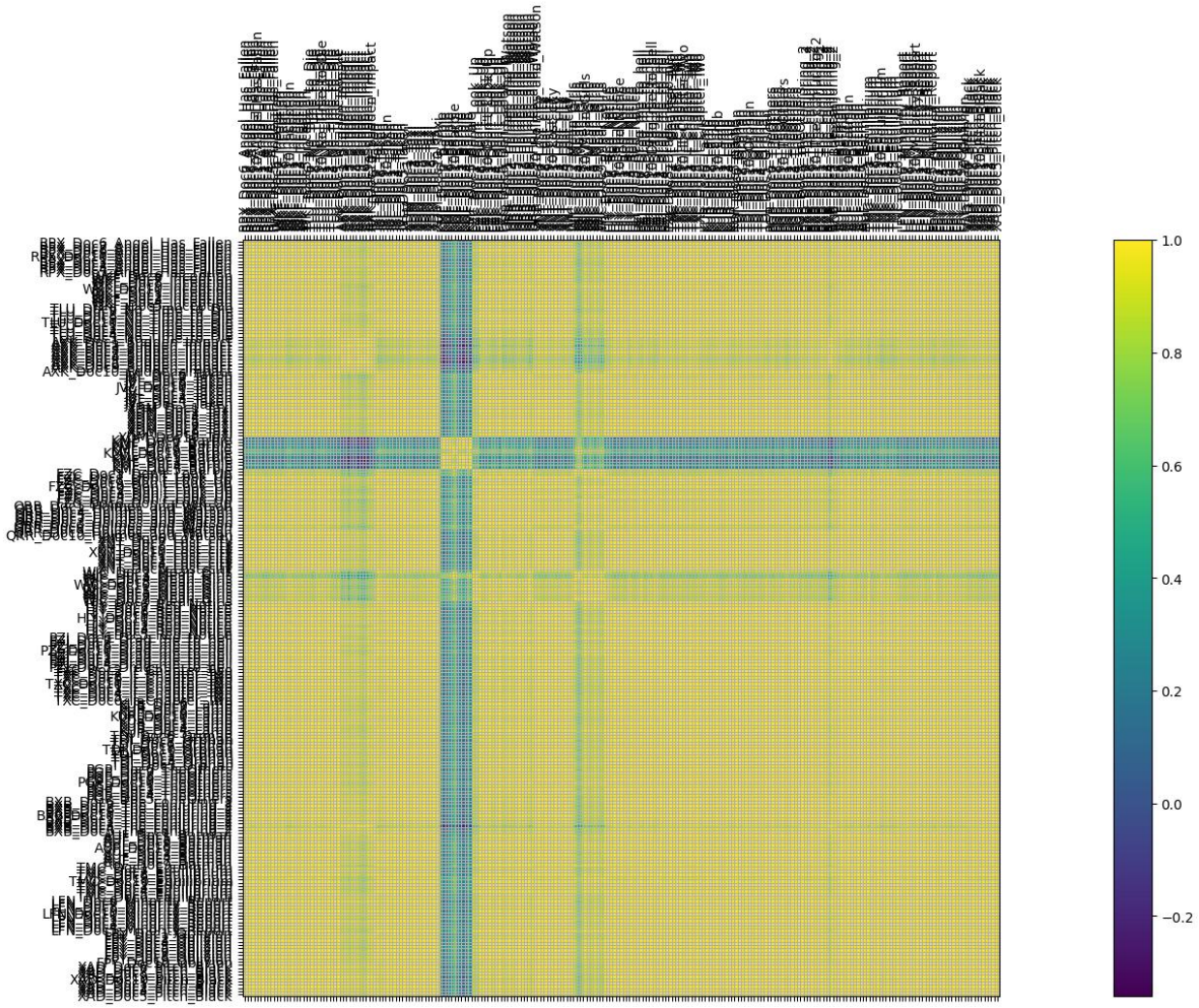

Figure 2: Topic Word Distribution (6 Concepts using LSA)

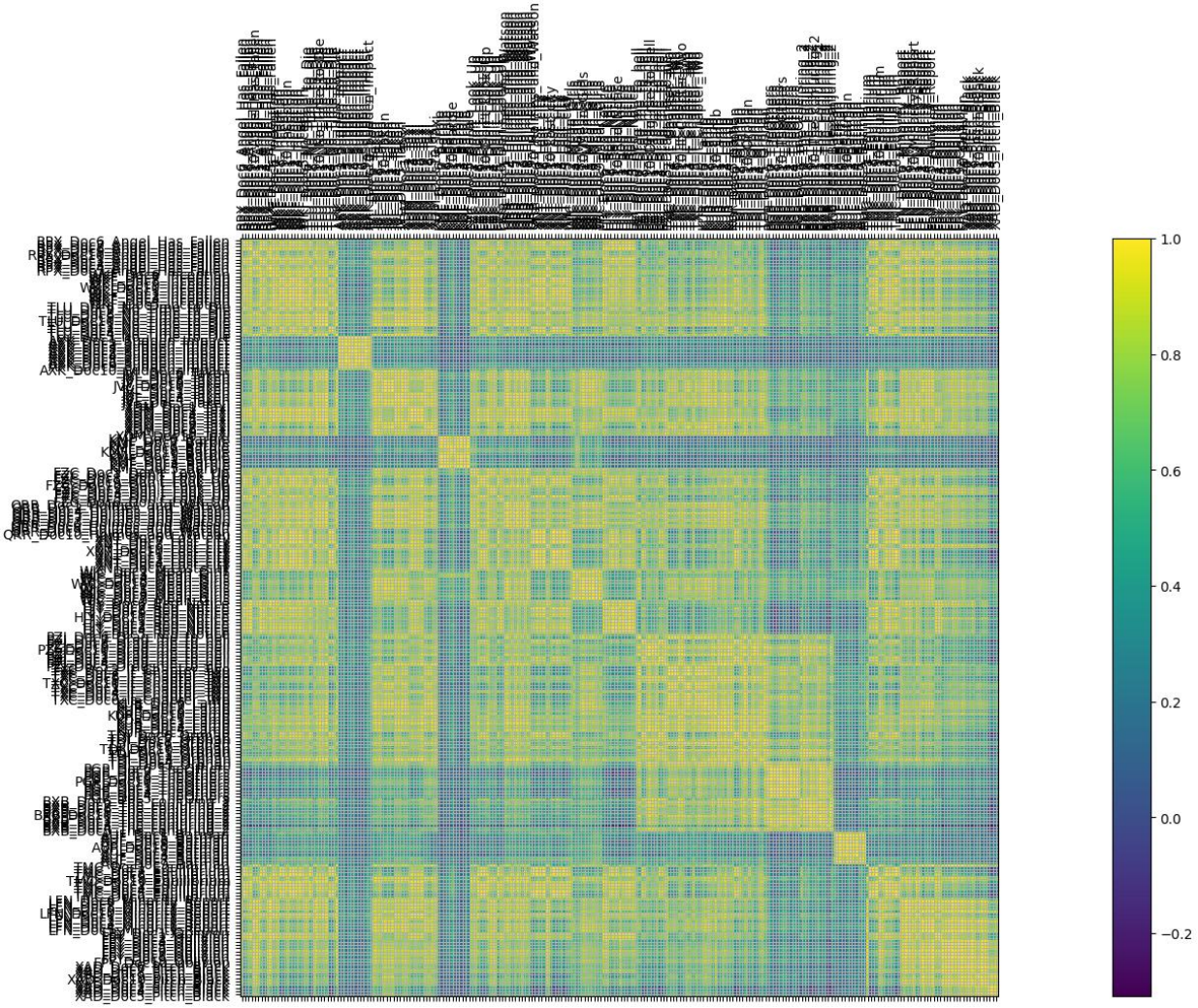

Figure 3: Topic Word Distribution (20 Concepts using LSA)

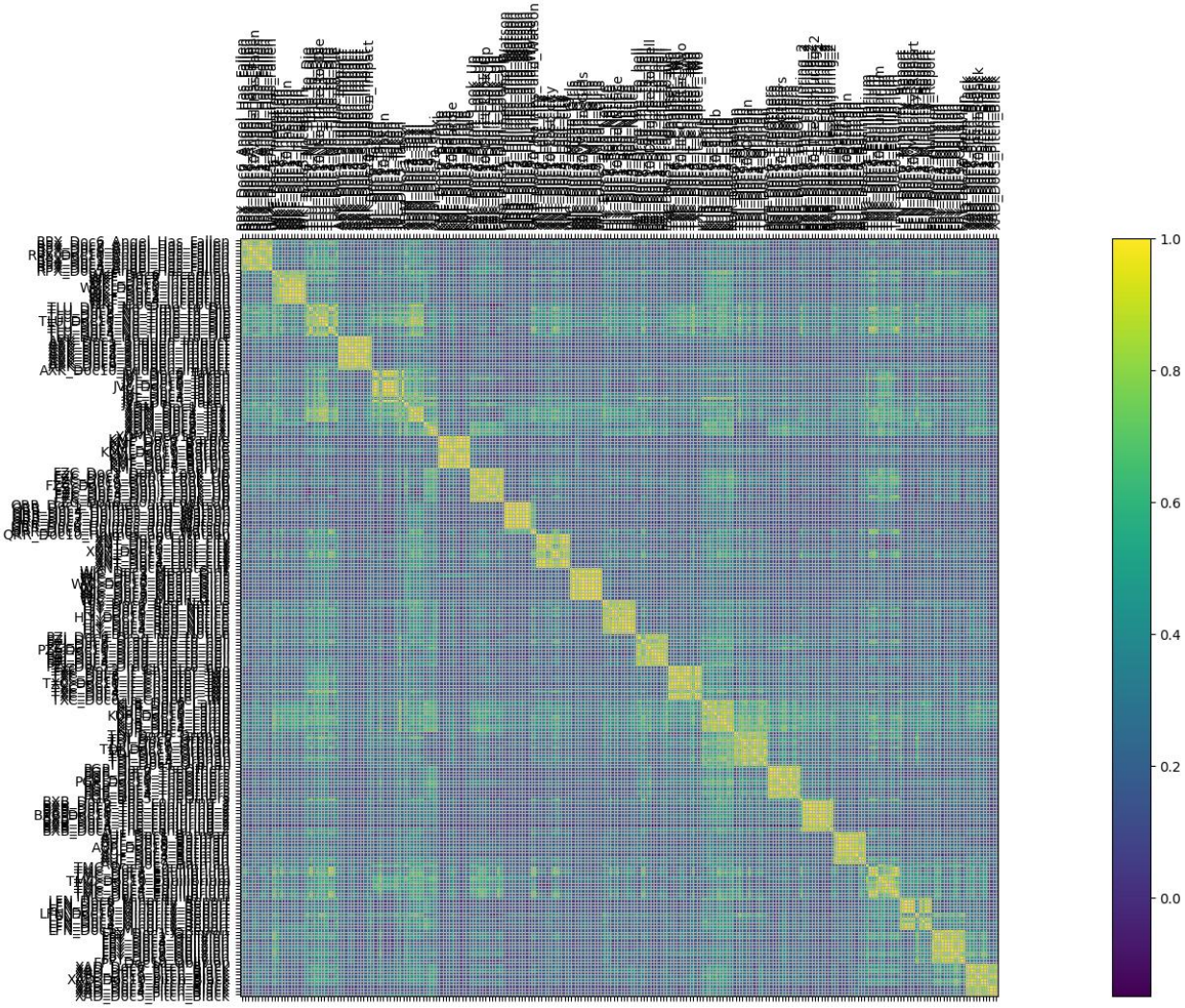

Figure 4: Topic Word Distribution (20 Topics using LSA)

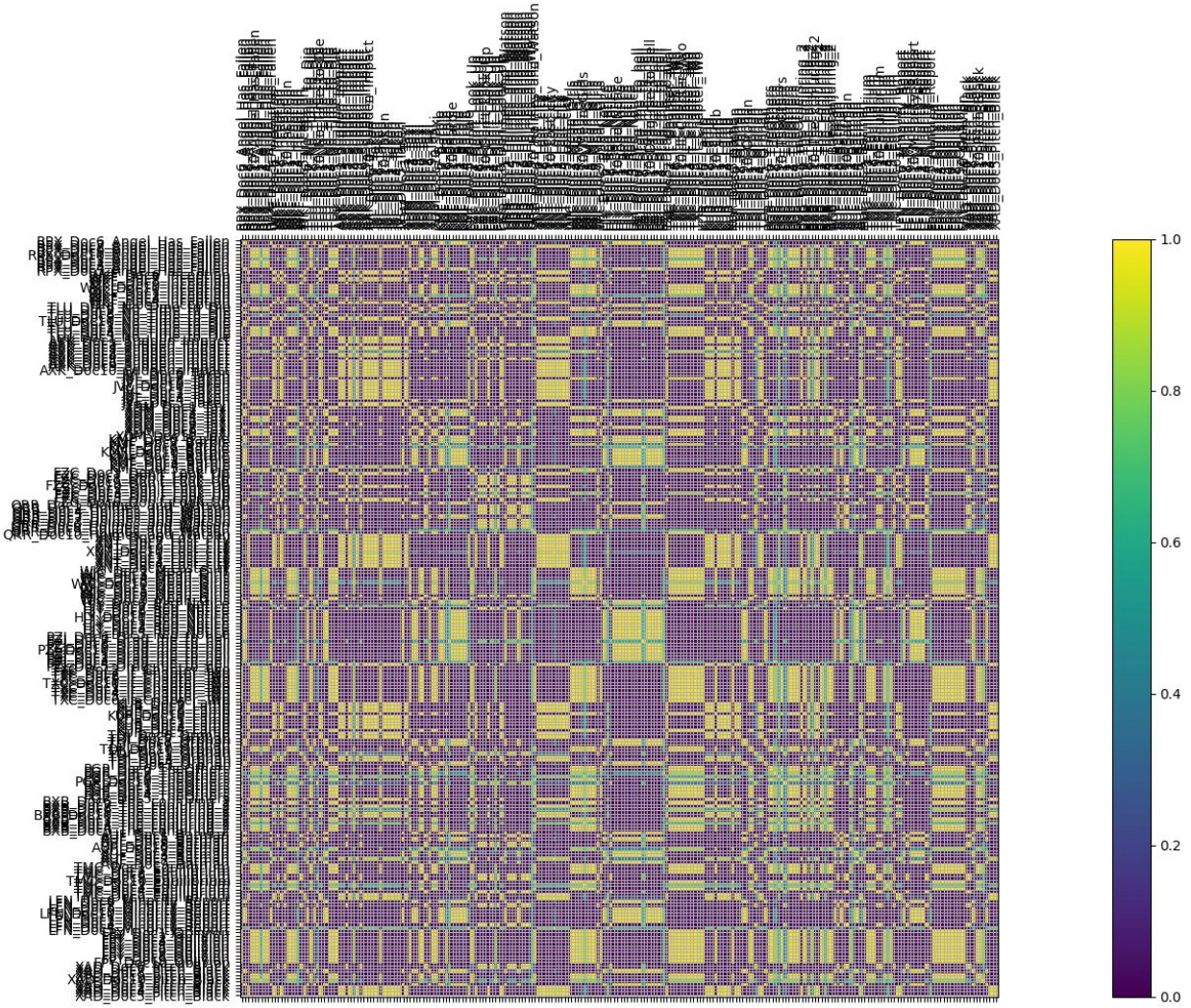

Figure 5: BERTopic Topic Word Scores

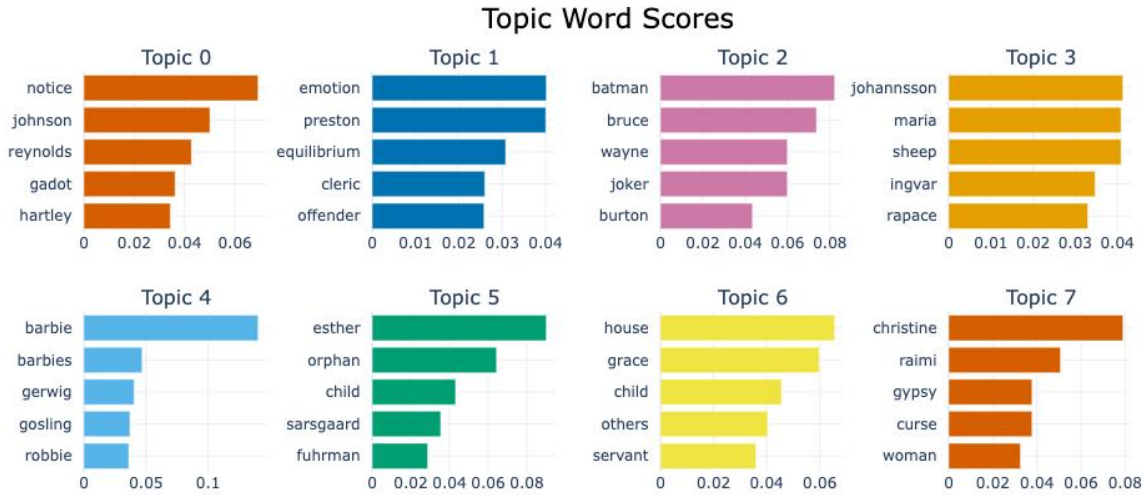

Figure 6: BERTopic Topic-Document Heatmap

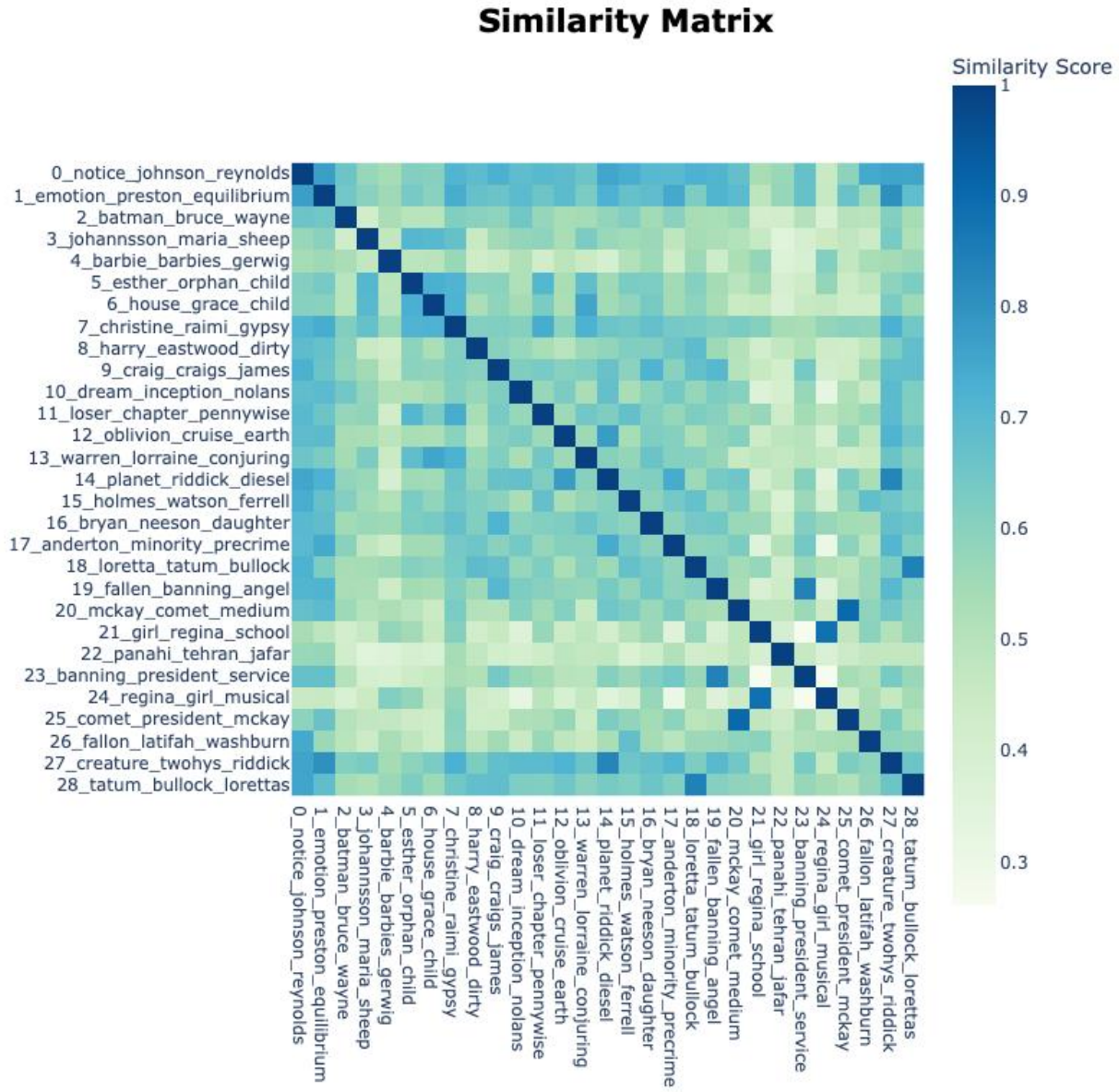

Figure 7: Silhouette Score vs. Number of Clusters (k)

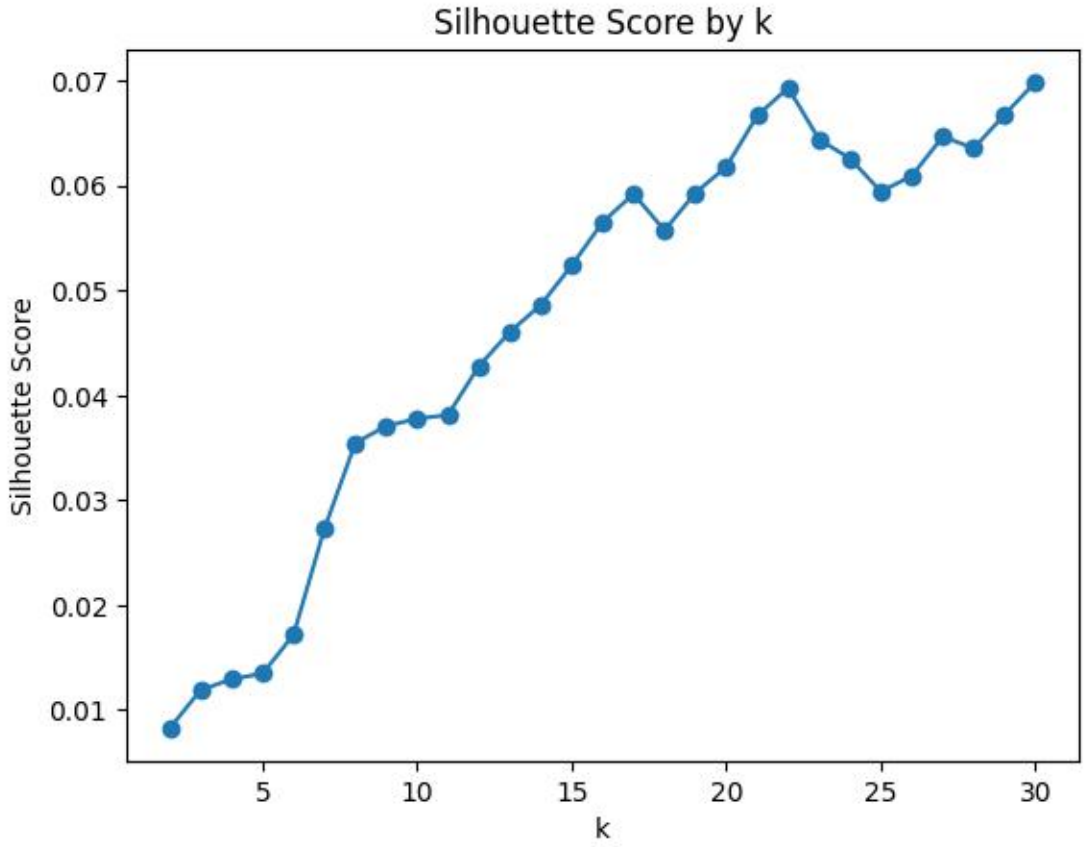

Figure 8: KMeans TF-IDF t-SNE Cluster Plot with Assigned Document Groups

Figure 9: Confusion Matrix for Multi-Class Genre Classification