NORTHWESTERN UNIVERSITY

SCHOOL OF PROFESSIONAL STUDIES

Assignement.2: Clustering on Movie Dataset

MS DSP 453 – Natural Language Processing

Sachin Sharma

October 27, 2024

Introduction

The primary objective of this assignment was to experiment with clustering techniques on a dataset of movie reviews. We aimed to group reviews based on sentiment and content to discover natural patterns within the data. The core task was to evaluate the alignment of automatically generated clusters with logical categories or sentiments in the reviews, identify issues in clustering, trace the causes, and explore potential improvements for better clustering fidelity.

Data

The dataset for this project comprises ten movie reviews with sentiment labels, capturing a range of perspectives on recent Batman movies. Each review is paired with a rating from 1 to 10, adding quantitative context to the overall sentiment, whether positive or negative. The language in these reviews is diverse, often informal, and includes complex expressions of sentiment, such as sarcasm, mixed opinions, and subjective observations about various aspects, including acting, cinematography, pacing, and storyline. Additional metadata, such as UserID, Post Date, Source, URL of the Review, Word Count, and Keywords, was also collected to enrich the dataset. However, for our clustering approach, we focused solely on the Review and Rating attributes. This setup allowed us to concentrate on sentiment patterns and textual characteristics, making it a fitting dataset for testing the robustness of various clustering algorithms in sentiment analysis.

Research Design and Modeling Methods

Methods Applied

1. **TF-IDF with NMF**:

- **Initial Setup**: We started with TF-IDF using 4 NMF components, resulting in clusters that loosely grouped reviews based on thematic content but failed to separate sentiment effectively. This led to mixed-positive and -negative reviews within the same cluster.
- **Refinement**: Increasing to 6 components slightly improved theme differentiation. However, sentiment remained poorly aligned, and clusters still displayed broad, overlapping topics due to vocabulary commonality across sentiments.

2. **Doc2Vec with K-Means**:
   - **Initial Doc2Vec**: Using a default `vector_size` and `window` size, we clustered reviews with K-means. This yielded clusters focused on general topics like "detective elements" but did not differentiate reviews by sentiment, as words with context-dependent meanings (e.g., "dark") were treated uniformly.
   - **Refinement**: By increasing `vector_size` and adjusting the `window` size, we aimed to capture a richer context. This improved the cohesion of thematic clusters, reflected in a higher Silhouette Score (0.56), yet still failed to achieve strong alignment with sentiment labels as indicated by a low ARI score.
3. **DBSCAN**:
   - **Application**: DBSCAN was briefly applied to explore density-based clustering. Due to the highly varied language and sparse semantic connections in movie reviews, DBSCAN produced several outliers and did not yield meaningful clusters.

Experimental Setup

To evaluate the effectiveness of these methods, we applied K-means clustering on the document vectors generated by each technique. After clustering, we assessed cluster alignment with sentiment labels (positive or negative) using metrics like **Adjusted Rand Index (ARI)** and **Silhouette Score**, gauging both alignment with true labels and internal cohesion within clusters.

Results

Document ID	Rating	Sentiment	TF-IDF with 4 NMF	TF-IDF with 6 NMF	DBSCAN	Bigram cluster	Doc2Vec	Doc2Vec Refined
1	10	Positive	0	1	0	1	1	1
2	4	Negative	0	0	-1	2	0	0
3	8	Positive	0	0	1	5	1	1
4	10	Positive	0	0	-1	4	0	0
5	3	Negative	0	0	2	0	0	0
6	1	Negative	0	0	1	5	1	1
7	10	Positive	0	1	0	1	0	0
8	3	Negative	1	0	0	3	0	0
9	8	Positive	0	0	2	0	1	1
10	10	Positive	0	1	0	1	0	0
		Silhouette Score	0.30	0.11	0.35	-	0.46	0.56
		ARI	0.10	0.07	-0.11	-	-0.07	-0.07

Insights from Results

1. Cluster Composition and Sentiment Alignment

- **TF-IDF with 4 NMF Components**: The ARI score of 0.10 indicates that 4-component NMF does not align well with sentiment. Positive and negative reviews (e.g., Document IDs 1, 5, and 8) are grouped together, revealing an inability to capture sentiment nuances.
- **TF-IDF with 6 NMF Components**: With a slightly improved ARI of 0.07, the 6-component NMF captures more variation but suffers from a low Silhouette Score of 0.11, showing weak internal coherence and difficulty in separating positive and negative tones.
- **DBSCAN**: DBSCAN's ARI of -0.11 and Silhouette Score of 0.35 reveal limitations in grouping short, sentiment-varied documents, especially those with unique terms or sparse density.

2. Bigram Clustering

- Clustering based on bigrams captured phrase-based thematic elements, but sentiment alignment was still limited. For example, Document IDs 4 and 9 were grouped due to common phrases around scenes and pacing, though their sentiments differed.

3. Doc2Vec and Refined Doc2Vec

- **Doc2Vec**: Doc2Vec's Silhouette Score of 0.46 shows better internal coherence than NMF but weak sentiment (ARI -0.07). Documents sharing thematic content, such as Document IDs 1, 3, and 6, clustered together irrespective of sentiment.
- **Refined Doc2Vec**: With a Silhouette Score of 0.56, the refined Doc2Vec captured thematic consistency more effectively, grouping documents with detailed analysis of characters or plot (e.g., Document IDs 3 and 6) but still fell short on clear sentiment.

Analysis and Interpretation

Why Results Were Not Optimal

1. **Complex Sentiment Expression**: The dataset contains nuanced language, including sarcasm and mixed sentiments. Positive reviews about visuals often coexist with negative feedback on pacing, complicating sentiment classification.
2. **Challenges with Short Texts**: The brevity of the reviews led to sparse term vectors. Techniques like DBSCAN struggled with outliers, while TF-IDF and NMF's limited term representation failed to capture the depth of sentiment effectively.
3. **Challenges with Mixed Reviews**: Many reviews contains words which are mixed. User gave good sentiment as well as bad sentiment within the same comment. Finding the overall sentiment could help to differentiate between clusters.
4. **Limitations of Basic Vector Representations**: While Doc2Vec was refined, it still struggled to separate sentiment from objective details. Both TF-IDF and Doc2Vec often produced clusters based on high-level themes rather than specific sentiments.

Suggested Improvements

1. **Enhanced Preprocessing**: Implementing robust stemming and lemmatization, along with advanced phrase extraction and synonym map, could help unify sentiment-critical terms and capture nuanced expressions.
2. **Advanced Embedding Techniques**: Using context-sensitive embeddings like BERT could improve clustering by better capturing semantic relationships, making it easier to distinguish subtle sentiment shifts.
3. **Hybrid Clustering Approach**: A two-stage clustering method—first grouping by theme and then refining by sentiment—could yield more accurate clusters.
4. **Sentiment-Aware Topic Modeling**: Incorporating a sentiment layer in topic modeling could effectively separate themes while considering sentiment, allowing for clearer differentiation between positive and negative experiences.

Conclusions

In this project, clustering methods were applied to movie review data to examine how well automated techniques capture sentiment-driven patterns. While TF-IDF and Doc2Vec both provided some clustering coherence, results indicated challenges aligning clusters with sentiment, as reflected in low ARI scores across models. Silhouette Scores improved slightly with refined Doc2Vec, suggesting some gains in internal cluster coherence. Overall, our analysis highlighted the limitations of these models in capturing nuanced sentiment without further semantic understanding, underscoring the need for advanced techniques like phrase-based extraction or supervised sentiment analysis to achieve meaningful classification.