**Assignment 2.  
 Clustering, Classification and Topic Modeling**

MSDS 453: Natural Language Processing

Husein Adenwala

Northwestern University

1. Introduction & Problem Statement

The purpose of this assignment is to perform unsupervised clustering analysis, sentiment analyses and Topic modeling on the 249 movie reviews. The unsupervised clustering analysis is done to decipher structure and patterns in the data and to see if the assigned groups in the data reflect the clustering pattern. I will also evaluate the ability of various classification algorithms to classify positive and negative sentiment/reviews in the data. I will also perform Topic Modeling using LDA and LSA techniques to understand the word patterns in the documents and to see if they correlate within the movie labels.

In this research, I will experiment with K means DBSCAN and agglomerative hierarchical clustering techniques on the TF-IDF vectors and Doc2Vec embedding vectors. I will also experiment with SVM, Naïve Bayes, Random Forest and Logistic Regression classification algorithms for sentiment analysis and finally, I will evaluate LSA and LDA topic modeling methods to understand and compare patterns of topics in the data.

## Data preparation, exploration, visualization

The data consists of 249 movie reviews of 25 movies that was stored and uploaded in a csv file. The tables and plots below show the movies’ names, the number of movie reviews and movie genres.

All 25 movies have 10 reviews except the movie Martian and Red Notice, which have 9 and 20 reviews respectively. The number of positive and negative reviews are equal for all movies except Martian. Also, it is important to note that the count for each movie genre is unbalanced; the Drama genre has only 9 reviews.

Table

Description automatically generatedTable

Description automatically generated Table

Description automatically generated with medium confidence

## Data Processing

All movie reviews were combined to create the Corpus. This corpus was tokenized and normalized by removing punctuation, stop word removal, stemming, lemmatization, lower cases and removing non-alphabet words.

The Corpus was tokenized using TF-IDF and Doc2Vec and Word2Vec embeddings.

# 3. Research Design and Modeling Method(s)

### Step 1 Clustering

I performed K means, DBSCAN, and Agglomerative Clustering using the TFIDF and Doc2Vec word embedding and evaluated the clustering methods using the silhouette score, R2,homogeneity score and by visualizing the elbow plot/method.

### Step 2 Sentiment Analysis

I performed sentiment analysis by implementing SVM, Naïve Bayes, Random Forest and Logistic Regression models on TFIDF and Doc2Vec embeddings. The data for this was split between test and train in a 3:2 ratio. I also experimented with feature reduction/selection techniques using scikit learn select k best function which improved the accuracy. I used accuracy, F1 score, confusion matrix and an ROC curve to evaluate the models.

### Step 3 Topic Modeling

I used LSA and LDA topic modeling methods on TFIDF and Word2Vec embeddings. I experimented with the number of topics and numbers of word parameters to get optimal coherence and perplexity score for the model.

# Results

### Clustering :

Using TFIDF, I found the optimal silhouette score of 0.14 did correspond to approximately 25 clusters. The plot below shows K-means clustering for the 25 groups which approximately correspond to the movie titles.

Chart, scatter chart

Description automatically generated Table

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Using Doc2vec, I found the optimal silhouette score of 0.62 did correspond to 2 clusters. The plot below shows K-means clustering for the 2 groups. However, those groups don’t correspond to the positive or negative reviews; it appears to be random as it clusters the positive and negative reviews accurately approximately 50% of the time.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Application

Description automatically generated with low confidence Chart

Description automatically generated

Silhouette Score plot Elbow plot

Chart, line chart

Description automatically generated Chart

Description automatically generated

### Sentiment Analysis:

I conducted 7 experiments for sentiment analysis with different classification algorithms and word embedding and dimension reduction techniques. The results are as follows:

Calendar

Description automatically generated

In addition to the experiments above, I used the NLTK TextBlob pretrained sentiment model, which provides the polarity and subjective score of text. Polarity is between -1 and +1 where greater than 0 is positive sentiment and less than 0 is negative. According to this model, 90% of the reviews were positive, which obviously does not correlate with the human coding, in which the split between negative and positive reviews was 50/50.

**228** reviews have polarity greater than 0.

Chart, scatter chart

Description automatically generated

### Topic Modeling:

For LDA 22, topics have the highest coherence of 0.28; however, based on the similarity matrix, as shown below, the topics for the most part don’t correlate with the movie titles.

**LDA**

Chart, line chart

Description automatically generated A picture containing graphical user interface

Description automatically generated

The LSA model has a coherence score that is approximately the same for all the topics (0.26) that range from 6 to 50. The similarity matrix shows that the topics correlate with the movie title.

Chart

Description automatically generated with low confidence

# Analysis and Interpretation

#### Key Findings:

##### Clustering.

As seen in the results above, clustering on TFIDF provides 25 clusters, which represent the clusters by movie names. However, there is some overlap of movie titles between a few clusters. I found K means clustering to provide the best clustering. Agglomerative hierarchical clustering also provided good results, but DBSCAN was unable to provide distinct clusters (using KNN to find epsilon). See appendix A for Agglomerative and DBSCAN plots.

Sentiment Analysis.

The important thing to note is that the movie reviews are usually based on rating scales and a low review does not necessarily mean there is negative sentiment in the review. The pretrained NLTK TextBlog model identified 90% of reviews as positive.

Models in method 3 and method 5 as shown above provided the highest test accuracy. These models used Random Forest or logistic regression classifiers for feature selection which reduced the noise in the data. However, that is a supervised learning method which would not generalize well on text data. Also, Doc2Vec embedding performed poorly in the classification model.

I also experimented by removing nouns using NTLK pos tagging package but it adversely affected the model’s performance.

I believe that Method 6, 7 and 2 would generalize better. In Method 6, I used binary bag of word embedding (term document matric has 0 and 1 only) and in Method 7, I used PCA to reduce dimension on the TFIDF embedding. In Method 2, I used TFIDF metric on the clean data.

All models overfit the training data and had high variance; this is partially because we have a small dataset.

The best model without feature selection was the **logistics regression classifier** in **Method 6** that used binary bag of word vectorization because of its simplicity and highest accuracy among the models that did not perform feature selection using a supervised classifier.

The data is balanced with an equal number of negative and positive reviews and the test train split was stratified to maintain the balance; therefore, test accuracy was a good metric and was the same to the f1 score in Method 6.

The AUC score of 64% suggests that the model can predict better than a random guess (50%).

Table

Description automatically generated Text

Description automatically generated Chart

Description automatically generated with medium confidenceChart, histogram

Description automatically generated Chart, line chart

Description automatically generated

Topic Modeling

LSA is analogous to principal component analysis applied to text data. LDA is probabilistic and uses Dirichlet prior.

LSA gives a correlated movie similarity matrix and LDA does not. This does not mean the LDA model is not selecting the correct topics. The reviews written for the movies often refer to other movies and topics to give analogies and there can be a lot of variation between reviews for each movie.

LSA and LDA topic modeling techniques had a low coherence score of 26-28%. This suggests that the words selected for each topic as a group are not very meaningful for human interpretation.

# Conclusion

The clustering result showed that the word embedding can discriminate between movie titles, but not between the positive and negative reviews. This is because a movie with a low rating does not necessarily mean the reviews contain negative sentiment. Most of the movies in the corpus were successful based on Box office revenue and the low rating for these movies do not always translate to negative reviews and this can be seen the NTLK TextBlob polarity score.

The classification algorithm did not perform well unless the features were reduced by selecting only the important features by using a supervised classification model. However, this is not suitable dimension reduction method for text data and the model will not generalize on unseen data. The key improvement for the model comes from data cleaning and from the vectorizing/word embedding technique. The best model i.e., the logistic regression classifier performance improvement came from using a simple binary bag of word model (0 and 1 only).

Removing noise from the data by removing punctuation, stop words, non-alphabet words, lemmatization and converting all the text to lower case also improved the model performance but based on the results of unsupervised clustering and topic modeling, vector space cannot distinguish between the two labels of positive and negative reviews, which suggests an inherent limitation in the dataset.

Finally, the LSA and LDA models don’t provide much insight into the data. The topics selected by LSA are correlated within each movie title and not among movies, where the topics selected by LDA show very little correlation within or among the movies. This shows the LDA topic selection has more variance and does not group the topics for the movie review to its labeled group of movie title or genre or positive/negative review type.

# Appendix:

Appendix A

**Clustering**

Agglomerative

Chart

Description automatically generated

**Chart, scatter chart

Description automatically generated** **Chart, scatter chart

Description automatically generated**

Chart, scatter chart

Description automatically generated

DBSCAN

Chart

Description automatically generated

Appendix B

Model with highest accuracy and f1 score

Chart

Description automatically generated

Graphical user interface

Description automatically generated

T