Assignment 2: Assess Clustering and Classification Outputs

Vivian Xia

Northwestern University

MSDS453: Natural Language Processing

Syamala Srinivasan

February 6, 2022

Introduction

Using the numerical representation of the documents, the goal is group similar documents into clusters. This objective can be accomplished through clustering, sentiment analysis modeling, and topic modeling. Clustering is an unsupervised learning method that clusters similar documents together based on similar features, based off the vectorization of the documents. Sentiment analysis modeling can be created with supervised learning methods such as support vector machines, naïve bayes, logistic regression, random forest classifier to classify the documents into positive or negative sentiment. These models use the numerical representation of the documents as the inputs to train the classification model. Topic modeling will use latent semantic analysis, LSA, as well as latent dirichlet allocation, LDA, models to uncover latent variables to represent the topics of the clusters. These topics and its associated terms will be used to identify similar documents.

The corpus contains 250 movie review documents, 25 movies with 5 positive and 5 negative reviews each. By using this corpus, the objective of k-means clustering is to find the optimal number of clusters that best groups similar documents together with little overlap. This includes clustering to see if this method could cluster based on genre and movie. The sentiment analysis modeling uses the inputs with its positive or negative sentiment labels to see how well the model can classify the documents correctly. For topic modeling, the goal is to see if the LSA and LDA models can find the latent variables or topics that can classify similar documents into clusters including four genres and two sentiments. Each method for grouping similar documents will be evaluated to see which models perform the best.

Data

The corpus includes ten reviews, five positive and five negatives, for twenty-five movies. The twenty-five movies include six action, horror, science fiction and seven comedy movies. There is a total of 250 documents in the corpus.

The library os was used to connect to the Google Drive folder to load the class corpus into the notebook. The library matplotlib is used to visualize the results in clustering, classification, and topic modeling. Seaborn was used to create the confusion matrix used to evaluate the performance of the classification models. The library pandas and numpy were used for formatting purposes.

The re and NLTK libraries were imported for preprocessing the text. The NLTK contains a package to tokenize the text to output tokens for the corpus vocabulary. It also contains a list of stop words which are all removed from the text to filter out the noise. The method of lemmatization from the NLTK library is then applied to the text to convert the tokens with the same roots to their real English base term. This technique reduces the number of tokens and combines the same root term into one. The library re is used for removing punctuation, non-alphabetic tokens, tokens with four or less characters, the term "movie," as well as normalizing the tokens to lowercase.

The sklearn library includes packages to split the data into training, validation, and test sets as well as to perform tf-idf vectorization on the documents. This library also contains the algorithms to perform k-means clustering, support vector machines, random forest classifier, logistic regression, and naïve bayes. It includes packages to evaluate those models' performance such as silhouette score, accuracy score, AUC score, and confusion matrix. The genism library contains the packages for Word2Vec and Doc2Vec models as well as packages to implement Latent Semantic Analysis and Latent Dirichlet Allocation. With the results from LSA and LDA, the similarities between the documents is compared with the similarities package in genism.

After data wrangling the text, the text is vectorized using term frequency-inverse document frequency, tf-idf, to create a dimensional space containing 250 points or documents where the dimensions equal the number of features or tokens. This vectorization converts the text to a numerical form using the frequency of the word in a document to represent the text. The vectors can then be used as inputs into the models. For sentiment analysis, the text is processed into Doc2Vec matrices as well as tf-idf matrices. To implement LSA and LDA, the documents are converted into document term matrices.

Research Design and Modeling Methods

Part 1: Clustering

Clustering is a type of unsupervised learning that is used to group similar documents together. K-means clustering takes the tf-idf scores of the documents to group them by similarity. The “k” in k-means represents the number of clusters. This clustering algorithm first randomly chooses the k centroids where k is the number of clusters. The distance between each point and the centroids are calculated to group the documents that are close to those respective centroids together. Once all documents are assigned a cluster, the average of these points are used as the new centroid. The distances are calculated, and points are put into a cluster again. This iteration continues until the average stops changing, indicating that the points in each cluster are the most similar to each other distance-wise (Srinivasan, 2022).

The elbow method is used to determine the optimal k. This method calculates the sum of squared distances between the centroid and the points. Within the cluster, a smaller sum of squared distance value is better, denoting thar there is less dissimilarity within the cluster. The visualization of the elbow method shows where the sum of squared distance falls, indicating the optimal k (Banerji, 2021).

Silhouette score is one performance metric for clustering. It is also used for determining the best k value (Srinivasan, 2022). It is calculated by getting the difference between the mean nearest-cluster distance of the sample and its mean intra-cluster distance then divided by the max mean distance to get a score in the range of -1 to 1, 1 being the best. As the score gets closer to 0, the score indicates that there are overlapping clusters. If the score becomes negative, there are samples that were assigned to the wrong cluster because there is one cluster is more similar ("sklearn.metrics.silhouette_score," n.d.). A silhouette score plot is visualized to see which number of clusters will yield the best silhouette score.

Part 2: Sentiment Analysis

The documents are processed into tf-idf scores as well as word-embedding model Doc2Vec to produce vectors to be inputted into the classification models. The Doc2Vec represents the documents as dense vectors using the average of the features/tokens from the word embeddings. Similar documents will have similar embedding vectors as well as meanings (Srinivasan, 2022).

Supervised learning methods can be used in the classification of these documents. Through classification, the semantics of the documents can be captured. Support vector machines (SVM) is one type of classifier that uses a kernel trick to map the data to a higher dimensional space as well as a maximum margin that maximizes the margin of the decision boundary to better discriminate between the two classes. Logistic regression is another way to classify the documents into one of the two classes, positive or negative. It uses the vectorization of the documents as the input and processes those inputs using the logistic function to map the predicted probabilities for each class. Naïve Bayes is a conditional probability model that assigns a probability for each to the document using numeric features. Random Forest Classifier builds multiple decision trees while using the Gini and entropy performance metrics to discriminate the classes as each branch is formed. With the multiple trees, the prediction is made by considering the class prediction of the majority of the trees (Srinivasan, 2022). These four types of classifiers are trained on two-thirds of the data then evaluated using the remaining one-third of the data.

The accuracy and ROC AUC curve and score are used to assess the performance of the classifiers. The ROC curve shows the tradeoff between true positive rate and false positive rate of the classifications. The ideal AUC score is 1, representing perfect accuracy. A score of 0.5 represents random classification. A confusion matrix for the chosen classifier is visualized with the classification rates to see how well the model does in discriminating the sentiments (Srinivasan, 2022).

Part 3: Topic Modeling

Topic modeling groups similar documents together based on topic. Each document is made up of a variety of topics, each of which is made up of a set of words. These topics are latent variables that are hidden but shape the meaning of the documents. Topic modeling can be performed through Latent Semantic Analysis, LSA, and Latent Dirichlet Allocation, LDA.

To create the LSA and LDA models, the documents have to be converted into document term matrix with m rows of documents and n columns of words with its frequency of each term in the document making up the matrix. This matrix is sparse and noisy, so in LSA, dimensionality reduction using single value decomposition, SVD, is performed on the matrix to break it up into singular values/vectors. These vectors with weights for each term in the topic is identified for each document. These weights are then used to evaluate the similarity between the documents. With the LSA output weights, a document similarity matrix can be built. LDA is a generative probabilistic method that uses a tdm matrix with frequencies or tf-idf to output probabilities for each topic for the documents. This method calculates the probability distributions for a mixture of topics in each document (Xu, 2018).

The LSA and LDA models will be built using 2, 4, and 10 topics with 10 words used in each topic. The similarity in the terms used to form the topics are usually skewed, so only the largest scored ten terms will be observed. The LSA and LDA models are plotted into a document similarity matrix to assess the similarity of the documents to one another based on the latent variables. The coherence can be used to measure word similarity within a topic. It measures the similarity of documents in a single topic; therefore, a higher coherence score is better (Srinivasan, 2022).

Results

Part 1: Clustering

To evaluate which k value to choose for k-means clustering, the elbow method is plotted. However, there is no evident elbow, value of k that falls, so this plot does not help in choosing an optimal k value. The silhouette score plot, Figure 1, was also visualized. The graph flattens a little from 10 to 12, 20 to 22, and 24 to 26. Overall, the peak from the plot is when k equals 24, but 20 follows closely behind in its score. The k-means algorithm will be implemented with 24 clusters as well as 20. To explore further, the algorithm will also be executed when k is 4 to see if the documents of the movies in each of the genres of comedy, action, horror, and sci-fi movies would be clustered together. The tf-idf matrix for each document is plotted with its corresponding cluster class when k is 20, 24, and 4. From the documents in each cluster and the silhouette score, the optimal k is 24.

Part 2: Sentiment Analysis

Using the tf-idf vectorization of the documents, support vector machines, logistic regression, naïve bayes, and random forest classifiers were used to classify the sentiment of each of the documents. Figure 4 shows the performance metrics of each model. The random forest classifier had the best test accuracy among the four but is still low and does not do a good job classifying the sentiments. In Figure 3, the ROC curves are visualized. The random forest classifier had the best ROC AUC score. A confusion matrix is visualized in Figure 5 to assess the random forest classifier on the classification of the test data.

The predictions are also made using document embedding vectors with 200 dimensions. The test accuracy and AUC score of the models, SVM, logistic regression, and random forest, are shown in Figure 6. The random forest classifier model had the best test accuracy and AUC score among the three models that were built.

Part 3: Topic Modeling

With the LSA models of 2, 4, and 10 topics and 10 words, the coherence scores are 0.52, 0.57, and 0.37, respectively. The model with the highest coherence score is the 4 topic and 10 words, so the 4 topic 10 words model should be used. The coherence scores of the LDA models of 2, 4, and 10 topics with 10 words were 0.27, 0.25, and 0.25, respectively. The LDA model with 2 topics and 10 words had the best coherence score and should be used instead of the other two. The LSA model with 4 topics and 10 words had the best coherence score among all the LSA and LDA models.

Analysis and Interpretation

Part 1: Clustering

As seen from the silhouette score plot, the peak and highest silhouette score is when k is 24. The optimal k at 24 makes sense considering that there are 24 unique movies in the corpus.

The k-means clusters when k is 20 and 24 are plotted, as well as 4 to observe if the resulting clusters represent the 4 genres.

The documents and their corresponding tf-idf scores are grouped into 20 clusters. Cluster 0, 2, 4 through 9, 11, 14 through 17, and 19 all have only one movie's corresponding documents in their cluster, so these clusters make sense. Cluster 1 has all the *Hereditary*, *Poltergeist* movie documents but also has one *The Matrix Resurrection* and one *Arrival* document. The *Hereditary* and *Poltergeist* documents together do make sense considering they are both horror movies. However, *The Matrix Resurrection* and *Arrival* are both science fiction films. *Arrival* has a supernatural element in its plot, like the two horror films, so *Arrival* makes some sense in this cluster. Cluster 3 has four of *The Matrix Resurrection* documents, all the *Casino Royale,* and all the *Mission Impossible Fallout* documents. When collecting the documents, *The Matrix Resurrection* was labeled as a science fiction film. However, *The Matrix Resurrection* also has action and adventure, so can be labeled as the genre action as well. *The Matrix Resurrection* in the cluster with the two other action films *Casino Royale* and *Mission Impossible Fallout* makes sense. Cluster 10 has all *The Ring* documents with an additional *The Matrix Resurrection* document. This does not make much sense considering *The Matrix Resurrection* is a science fiction and action movie. There is no horror or thriller element to it. Similarly, Cluster 12 has all *Groundhog Day* documents but also one *The Matrix Resurrection* document. *Groundhog Day* is a comedy, so *The Matrix Resurrection* does not make sense in this cluster. Cluster 13 has three *The Matrix Resurrection* documents and all the *Guardians of the Galaxy* documents. This cluster makes sense since *The Matrix Resurrection* and *Guardians of the Galaxy* are both science fiction and action movies. Cluster 18 has *Interstellar* and *Arrival* documents, which makes sense because they are both science fiction films. By creating 20 clusters, observations can made on which movies can be categorized in more than one genre. As seen, for example, *The Matrix Resurrection* is both science fiction and action and makes sense to be categorized in either or both. The tf-idf plot, Figure 2, for this cluster shows that most of the documents clustered together are close to each other. There are some documents in some assigned clusters that are more scattered from the cluster.

The documents are then grouped into 24 clusters. All clusters except for Cluster 2, 7, 8, and 9 grouped only one movies' documents in their clusters. Cluster 2 had all *The Matrix Resurrection* and *Pirates of the Caribbean: The Curse of the Black Pearl* documents. Both movies have a lot of action and adventure, so despite not splitting these two sets of documents, they do make sense together. Cluster 7 and 23 both contained *Red Notice* documents but did not group them altogether into one cluster. This may be due to some *Red Notice* documents being more similar than others in sentiment or topic. Cluster 7 also had one document from *The Matrix Resurrection*. *The Matrix Resurrection* and *Red Notice* are both action genres, and this particular document from *The Matrix Resurrection* may be more focused on that action elements rather than science fiction. Cluster 8 includes the all the documents from *Interstellar* and most of *Arrival,* which makes sense as they are both science fiction movies. Cluster 9 includes two *Pirates of the Caribbean: The Curse of the Black Pearl* documents, two *Arrival* documents, and one *Pacific Rim*. This cluster does not make very much sense because *Arrival* does not have any action or adventure or fantasy genre that are in *Pirates of the Caribbean: The Curse of the Black Pearl* and *Pacific Rim. Pirates of the Caribbean: The Curse of the Black Pearl* does not have any science fiction element that is in *Arrival* or *Pacific Rim.* These documents may be outliers of their own respective set of movie documents which is perhaps the reason they were grouped together. The tf-idf plot looks very similar to the plot from the previous k-means clustering with 20 clusters. There is still overlapping colors between each cluster.

For the k-means cluster with 4 clusters, Cluster 0 had only and all the *Red Notice* documents except for one of its documents. Cluster 2 had all the *Spiderman 3, Sudden Impact,* and *Guardians of the Galaxy* documents. These documents make sense together. Those three movies are all action films. Cluster 3 had all the *Cruella* and *The Conjuring 3* documents. This cluster does not make a lot of sense because *Cruella* and *The Conjuring 3* do not share the same genres. *Cruella* is not a horror film, and *The Conjuring 3* is not a comedy drama. Cluster 1 had the other *Red Notice* document as well as the documents of the remaining 18 movies. This cluster was a mixed bag of comedy, action, science-fiction, and horror films, which did not make sense altogether. The plot shows that all the clusters except for Cluster 0 have points that are further away from the cluster and overlap into the other clusters. Because there are more clusters that make sense than the others, the optimal k will be 24 to form 24 clusters. As seen by the silhouette scores, 24 clusters also yield the largest silhouette score, indicating that there are fewer overlapping clusters.

Part 2: Sentiment Analysis

Tf-idf Vectorization

Of the four classification models, the random forest had the best accuracy as shown in Figure 4. The accuracy score is not very high, so this model does not do a good job classifying the positive and negative sentiments. The test accuracy is less than randomly guessing each document's sentiment. The AUC score for random forest is also greater than that of the other models but only by 0.02. The AUC score for random discrimination is 0.50, which is denoted by the dashed red line. The random forest's score of 0.52 is not much better than randomly classifying the documents. The other three models' ROC curves are aligned with the red dashed line that represents random discrimination, so those models are no better than randomly classifying the documents.

The confusion matrix shows that the random forest classifier did a good job classifying the positive sentiments correctly but did a very bad job of classifying the negative sentiments. The negative sentiments were only correctly classified 13% of the testing data. Overall, this model does not do a good job discriminating between the sentiments.

Document Embedding Vectorization

The random forest classifier had the best test accuracy and AUC score among the three models. The test accuracy is 0.45, which is a better test accuracy than that of the random forest classifier accuracy using tf-idf vectors. However, this accuracy score is very bad in general. Randomly guessing among the two classes would give an accuracy of 50%, but this model and the model using tf-idf both do worse than that. The ROC curve, Figure 7, also supports that the model is very bad and is pretty much random discrimination. The SVM ROC curve shows that it actually does worse than random discrimination with a score of 0.48. The logistic regression ROC curve does about the same as random discrimination and the random forest barely does any better.

The confusion matrix, Figure 8, for the random forest classifier shows that this model does a little better than the tf-idf model in terms of classifying the negative sentiment documents correctly at 19%. This model does a little worse at classifying the positive sentiments than its counterpart. Overall, the models using tf-idf and Doc2Vec do a poor job discriminating the sentiment. However, the random forest classifier model with Doc2Vec does the better job among the models experimented with.

Part 3: Topic Modeling

Latent Semantic Analysis: 2 topics 10 words

The output of the ten highest weighted terms of the LSA model with 2 topics and 10 has general terms such as “story,” “thing,” “would,” and “character” in its first cluster. The second topic cluster had more specific words such as “cruella,” “baroness,” “disney” and “villain.” These ten highest scored terms in each cluster show that the two topics do not correspond to the sentiment of the documents, especially the broad terms in the first cluster. There are no distinct positive or negative sentiment words in either cluster. The first cluster seems to be more generalized about all movies while the second cluster has terms that seem to refer to the movie *Cruella* and other Disney movies with women, so perhaps, *Frozen 2.*

The plot, Figure 9, does not show two chunks of yellow. Instead, there are two small chunks and one very large chunk. The first top chunk seems to include the documents for *Frozen 2* and *Spiderman 3,* indicating that these two movies are similar to each other. There is another chunk containing documents from just the movie *Cruella*, so there was no topic similarity found between that movie and the others. The large chunk contains the other 220 movie documents. This also shows that there is an uneven number of documents in one cluster than the other, so this model does not find any similarity via positive or negative sentiment within the documents.

Latent Semantic Analysis: 4 topics 10 words

The next model uses 4 topics and 10 words to find similarity between the documents. The terms for Cluster 0 are similar to Cluster 0 of the preceding model. It has mainly very broad terms such as “there,” “character,” “story,” “thing”, etc. This cluster seems to represent the topic of general family movies. Cluster 1 is similar to the last model’s Cluster 1 as well with the same terms as “cruella,” “villain,” and “fashion.” The term “cruella” in Cluster 1 has a significantly larger weight in this cluster compared to the other words. Even though “cruella” is in Cluster 0 as well, the term in that cluster has a much smaller weight. Cluster 1 terms seem to point towards *Cruella* and other Disney movies with women.

Cluster 2 and 3 have more specificity for the terms in their groupings than Cluster 0. Cluster 2 and 3 consists of more character and actor names. Cluster 2 has the actor names “johnson,” “reynolds,” “gadot,” “thurber” in its cluster which seems to refer to *Red Notice.* It, however, also has the terms “horror” and “action” but are weighted less than the other terms. The topic seems to be action movies, specifically *Red Notice,* with the mentioned terms of actors.

Cluster 3 includes terms such as “harry”, “eastwood,” “spiderman,” “interstellar,” “space.” These terms refer to the movies *Sudden Impact*, *Spiderman 3,* and *Interstellar*. The inclusion of the word “space” may include *Guardians of the Galaxy* and *The Martian* into this topic cluster as well. This topic seems to be primarily action movies that includes space action as well. Based on the words and the clustering, the use of 4 topics did not result in genre groupings. The ten highest weighted terms in each cluster overlap as well as the first cluster has very broad terms compared to the other clusters’ terms.

The plot, Figure 10, of the document similarity shows approximately 9 clusters. There is a lot of overlap and colors are distributed everywhere, so there are no mutually exclusive groupings using 4 topics. The largest chunk with the least overlap consists of *Mission Impossible: Fallout, Hereditary, The Conjuring 3*, *Us, The Ring, Lamb, Poltergeist, The Martian, Pacific Rim, Guardians of the Galaxy, Interstellar,* and *Arrival.* This grouping represents action, horror, and science fiction genres, so there is still some overlap in genre groupings as well as an uneven number of movies within each of the four topics.

Latent Semantic Analysis: 10 topics 10 words

This model uses 10 topics and 10 words. Its first two clusters are identical to the first two clusters of the prior models. Similarly, the third and fourth cluster have the same terms as that of the third and fourth cluster of the model with 4 topics. Cluster 4 has terms such as “interstellar,” “space,” “horror,” “devil,” “mission” which seems to refer to the topic of science fiction and horror genres. In Cluster 5, the terms “spiderman,” “peter,” “eastwood,” “sudden,” “impact,” and “parker” give the context that the topic is action movies since Eastwood directed *Red Action*, an action movie, and the action movies *Spiderman 3* and *Sudden Impact* are also mentioned in the terms. Cluster 6 has terms such as “family,” “conjuring,” “devil,” “encanto,” “racer” so there are horror, action, and comedy genres within this cluster. The topic for this cluster would be more general then. Cluster 7 has “speed,” “racer,” “fallout,” “spiderman,” “action” as terms which refers to many of the action movies in the corpus, so the topic may be action movies. Cluster 8 refers more of the horror movies such as *The Ring* and also the action movie *Speed Racer*. The topic for this cluster may be action thrillers. Cluster 9 has terms that refer to the movies *Speed Racer, Mission Impossible: Fallout, The Conjuring 3*. Because the terms in Cluster 8 and 9 are similar, this cluster may be more focused on action rather than the thriller angle of the movies.

This plot, Figure 11, shows much more overlapping colors within its chunks than the prior plot, so there does not seem to be many mutually exclusive similarities between the documents. There are approximately 18 chunks in this plot. Most of the chunks some that the similarity is between its own documents.

LSA Model Coherence

A larger coherence score is better because it represents more similar of documents within a topic. Therefore, the 4 topic and 10 word model is the better model of the three. The plots for each of the models also support that the 4 topic and 10 word model is better. The plot for 2 topics was very skewed toward one topic as seen by the very large chunk in its plot. The 10 topic plot had a lot of overlapping colors including in the inside of the chunks along the diagonal, so there were no clear topics from that clustering. The plot with 4 topics shows more yellow chunks with less overlapping colors than the 10 topic plot.

Latent Dirichlet Allocation: 2 topics 10 words

The LDA model uses 2 topics and 10 words. The first cluster, like the LSA counterpart, has very general terms such as “character,” “story,” “first,” and “thing.” There are only three meaningful words “world,” “family,” and “little,” but are given the least weight in that cluster. The second cluster also has terms that are very broad such as “scene,” “action,” and “would.” However, it does have two meaningful terms “family” and “action.” The terms that make up these two topics do not seem to correspond to two sentiment clusters. The first cluster may seem like it could refer to negative sentiment with the word “little.” However, the other words do not imply any other negative sentiment. The second cluster conveys a more genre-related topic of family action movies. The plot, Figure 12, shows a lot of overlapping colors and no clear chunks. The diagonal cannot even be seen clearly. The yellow is distributed all throughout the graph rather than in along the diagonal in chunks.

Latent Dirichlet Allocation: 4 topics 10 words

This model uses 4 topics and 10 words. The first cluster contains broad words such as “character,” “there,” “story.” The grouping of these terms in this cluster represents a general scope of maybe family movies. Cluster 1 also has some general terms such as “thing” and “character,” but also has “action” and “family.” This topic could be more family-oriented action movies. Cluster 2 has a lot of noise in its terms as well but does also contain “horror,” so the topic of the cluster may be horror movies. Cluster 3 also has very generalized terms that do not have much meaning. It also has “action” and “family” in its cluster. The topic may be family action movies. However, most of the terms in the clusters are very general and do not contribute to any specific features in the movies. These 4 topic clusters do not differ much in their terms. There are a few terms that imply genre, but they seem to overlap and also are not specific enough. Many science fiction movies are also action movies, so these clusters could split the science fiction movies. The clustering on the plot, Figure 13, is not present. There is a lot of overlap in colors and no distinctive chunking among it, so there is no mutually exclusive clustering and similarities between documents.

Latent Dirichlet Allocation: 10 topics 10 words

The LDA model uses 10 topics and 10 words. The first cluster has some broad terms but not as many as the first cluster in the prior model. There are terms such as “never” and “still” that could be meaningful towards negative sentiment as well as “cruella” that refers to *Cruella* and “harry” which could refer to *Spiderman 3, Sudden Impact, Casino Royale, The Martian*. Those movies all have action in common, so the topic could be negative sentiment on action movies.

Cluster 1 has all broad terms except for “family,” “little,” and “peter.” The “little” could refer to negative sentiment, but there are no other words to support that sentiment except for that term. The terms “family” and “peter” could refer to *Spiderman 3* which is both action and a family film, so the topic for this cluster could be documents that are for family action where the review is slightly negative. Cluster 2 has the term “cruella,” “family,” and “original.” The term “cruella” and “family” could refer to the Disney film *Cruella* which is family oriented. *Cruella* is a comedy but also has some action. The term “original” has a positive connotation, so the topic could be positive reviews for family action films. Similarly, Cluster 3 has many of the same terms as Cluster 2 but not “original,” so its topic could be similar to Cluster 2 as family action films. Cluster 4 has many general terms but also has “family,” “johnson,” and “cruella.” The term “johnson” refers to the film *Red Notice* and “cruella” refers to *Cruella*, which both are action films. This cluster also seems to have the same topic as the Cluster 2 and 3 of family action.

Cluster 5 has all general terms except for “harry,” which could refer to *Spiderman 3, Sudden Impact, Casino Royale,* or *The Martian*. So this cluster’s topic could be action. Cluster 6 has “never” and “action” in its list that conveys negative sentiment and action films, respectively. The other terms in this cluster are meaningless. Cluster 7 has “action,” “horror,” and “little” in its list. The “little” could refer to negative sentiment and the other two terms could refer to genres, so this cluster topic could be action horror documents that have a negative sentiment. Cluster 8 has “family,” “horror,” and “speed” which could refer to the overlapping of genres of family, action, and horror. Cluster 9 has only one meaningful word of “action” in its list, conveying that the topic may be action. Overall, there is a lot of noise in these clusters as the same meaningless words keep appearing as important terms in each cluster. This makes it hard to see if there is a more specific and unique topic among the clusters. The plot, Figure 14, shows that there are less areas that are yellow compared to the last model and its plot, which means that the clusters are better at discriminating the documents from one another. However, there is still overlapping colors, but the diagonal is more visible in this plot.

LDA Model Coherence

The coherence for each LDA model is compared to see which model did the best in clustering similar documents under its topics the best. The 10 topic model had a better score than the 4 topic, which was also evident by their document similarity matrices. However, the model with 2 topics and 10 words had the overall best coherence score, which means there was more similarity found between the documents within each of the two topics than with each of 4 or 10 topics. The LSA models, overall, had the better scores and plots than the LDA models. The LSA model with 4 topics and 10 words was the best model out of all the LDA and LSA models and was able to find the most similar documents in a single topic.

Conclusion

There are several methods to cluster and classify similar documents including clustering, sentiment analysis modeling, and topic modeling. The k-means clustering did the best job with grouping similar documents together. In particular, the k-means clustering with 24 clusters provided the clusters that made the most sense and had the highest silhouette score, which indicates there were fewer overlapping clusters.

The sentiment analysis models using support vector machines, naïve bayes, logistic regression, and random forest classifier all did not do a good job in discriminating between the positive and negative sentiment of the documents. Their accuracy scores for both versions, tf-idf and Doc2Vec, were less 0.50, indicating that a random guess between the two sentiments would have yielded better results than the models. Of the models in both versions, the random forest classifier model performed the best.

The LSA model in general had better coherence scores than LDA. The LSA model with 4 topics did the best in finding the latent similarities between documents. The results from all three methods to group similar documents reinforce the importance of data wrangling. As seen from the topic modeling, the terms with its defined weights are what identifies the topic. If the terms are meaningless and too broad, the topic modeling will not yield very good topics and weighted terms to identify clusters of similar documents. From the topic modeling, it was evident that the data needed to be cleaned more to take out common and meaningless terms such as “character” and “story” to reduce the noise in topic modeling as well as the other clustering and classification methods. Clustering and classification also did not group the documents perfectly within, for clustering, its four genres or 24 movies and, for classification, its two sentiments.

These methods show that the cleaned and vectorized text is the base of building these models to be able to group similar documents together. The importance of data wrangling is enforced in this assignment when assessing the results of the clustering, classification, and topic modeling. From experimenting with all these different methods, a better understanding is gained behind the concept of each method and how to analyze its performance as well as consider further steps to improve on the model performance.

References

Banerji, A. (2021, May 18). *K-Mean: Getting The Optimal Number Of Clusters*. Analytics Vidhya. Retrieved 2022, from https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

scikit-learn developers. (n.d.). *sklearn.metrics.silhouette_score*. scikit-learn. Retrieved 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

Srinivasan, S. (2022). *MSDS 453 Natural Language Processing: Week 5 & 6 – Classification and Clustering, Knowledge Graphs* [Zoom Cloud Recordings].

Xu, J. (2018, May 25). *Topic Modeling with LSA, PLSA, LDA & lda2Vec*. Medium. Retrieved 2022, from https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05

Appendix

Figure 1. Silhouette score plot for k-means clustering.

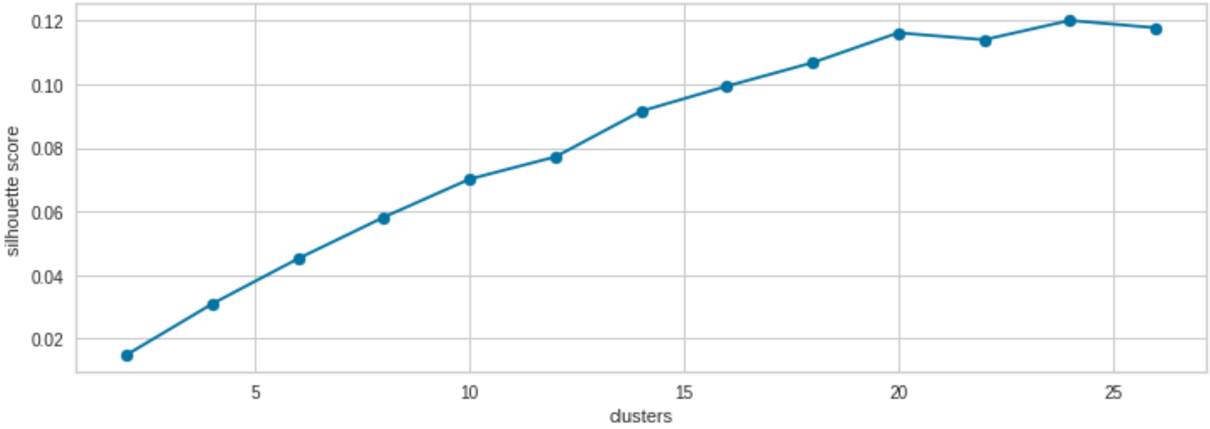


Figure 2. K-means clustering, k = 20, tf-idf plot.

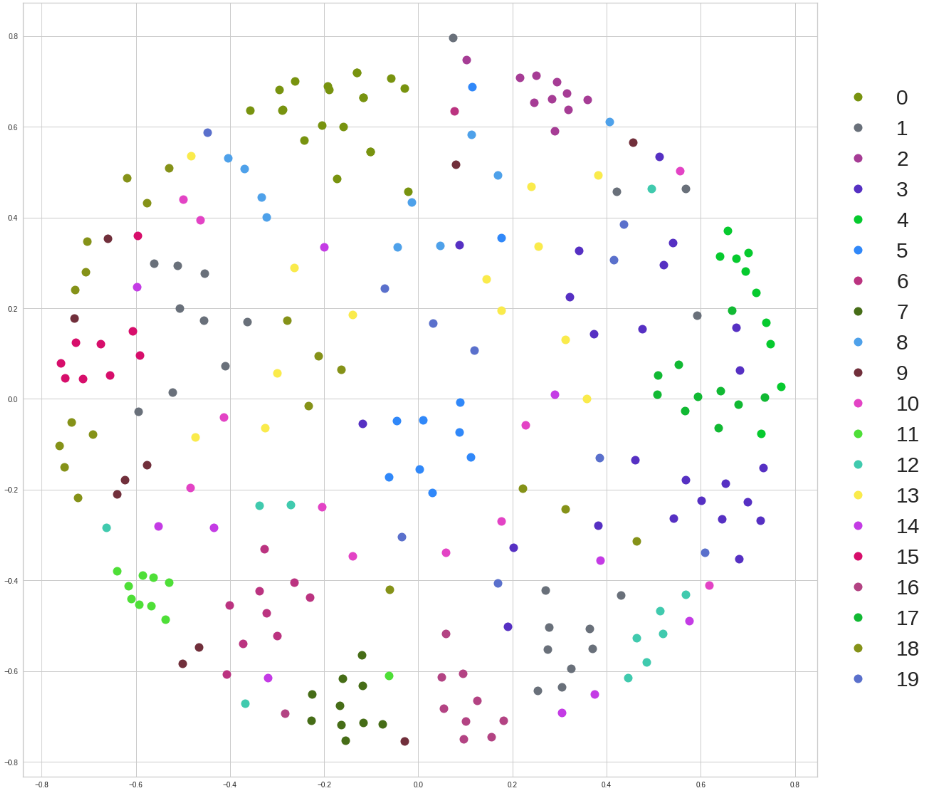


Figure 3. Predicting sentiment analysis with tf-idf ROC curves.

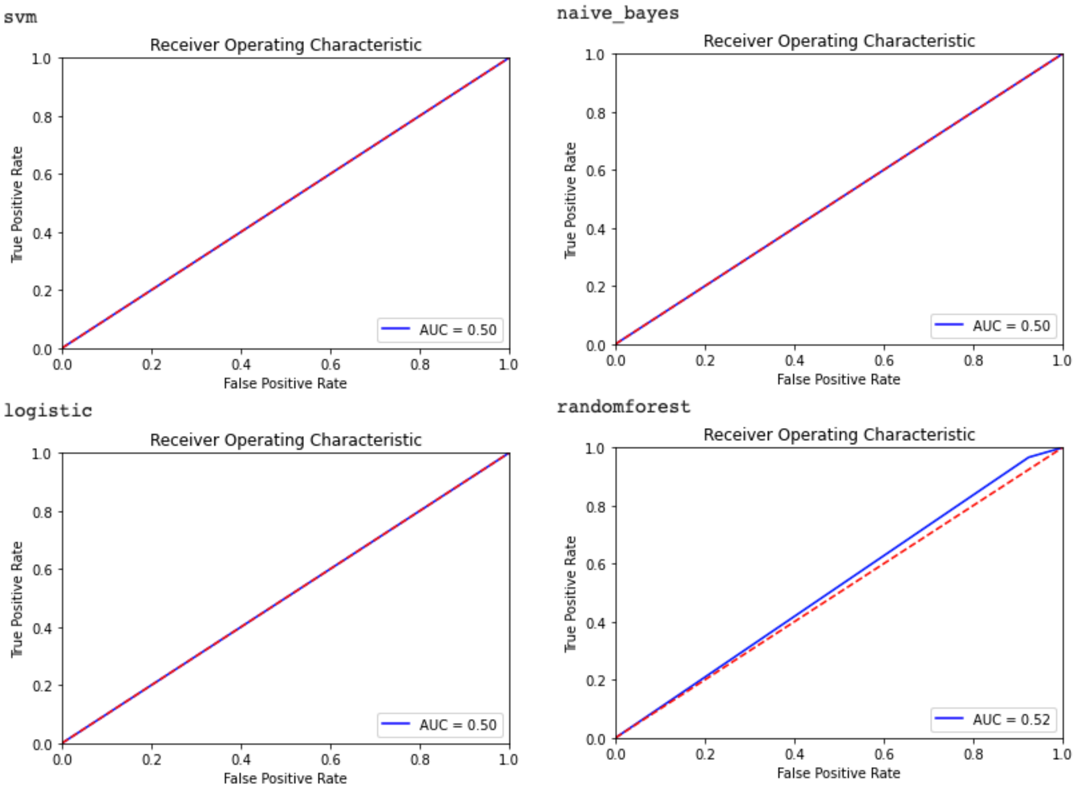
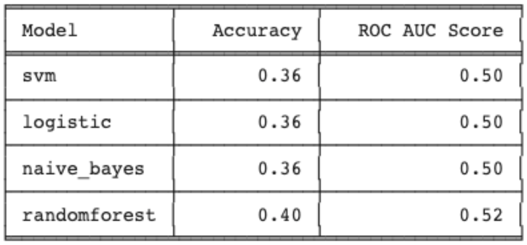


Figure 4. Accuracy of sentiment classifier models with tf-idf.



Model	Accuracy	ROC AUC Score
svm	0.36	0.50
logistic	0.36	0.50
naive_bayes	0.36	0.50
randomforest	0.40	0.52

Figure 5. Random forest classifier model with tf-idf to predict sentiment confusion matrix.

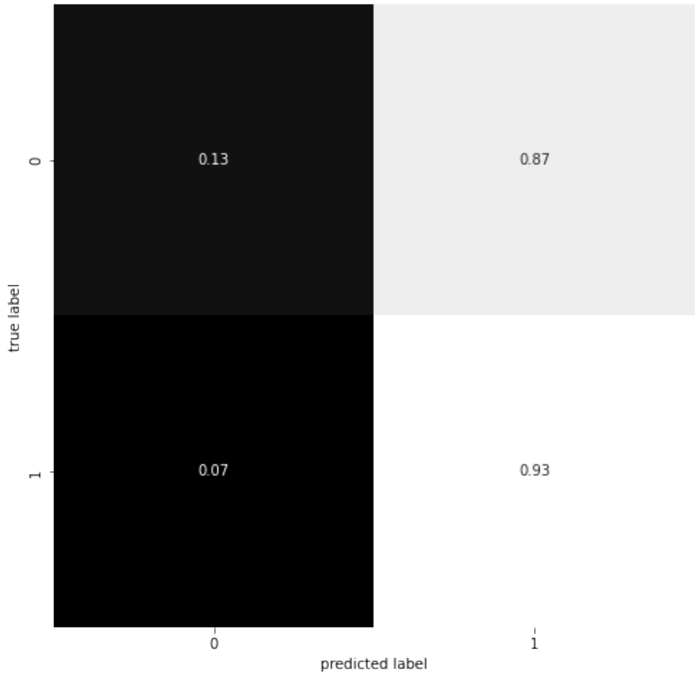


Figure 6. Predicting sentiment analysis with Doc2Vec ROC curves.

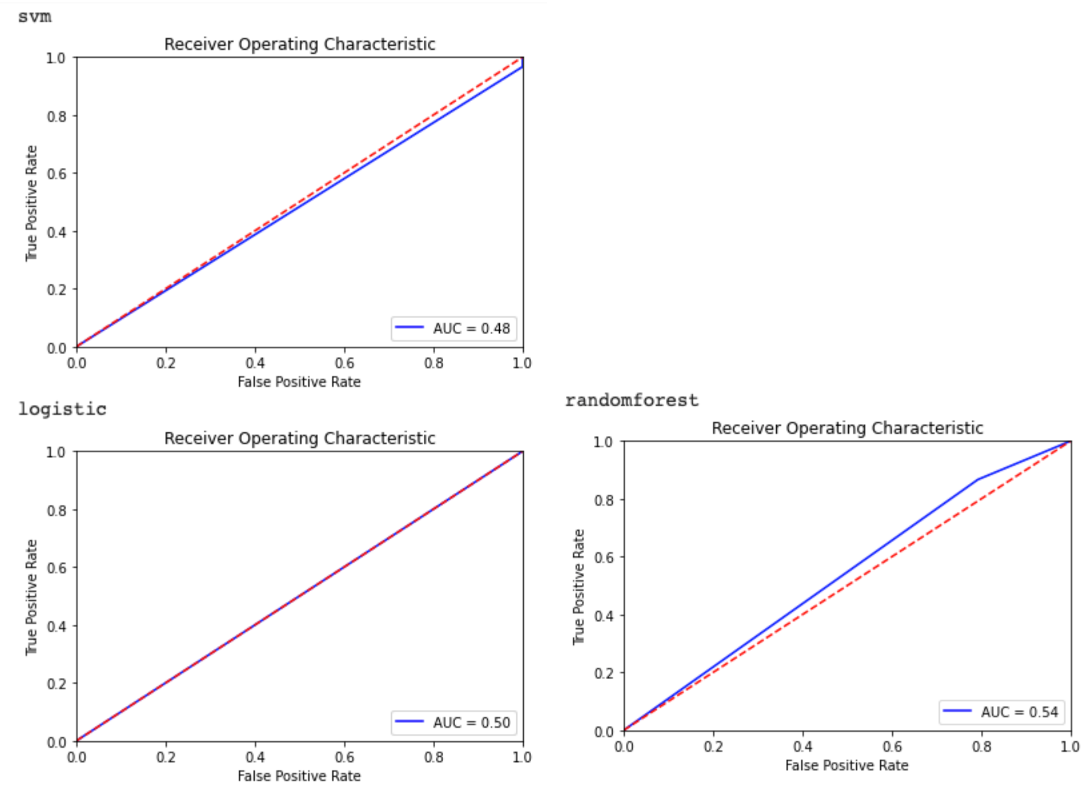
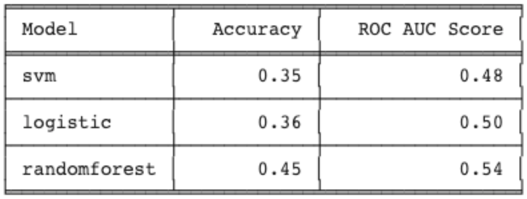


Figure 7. Accuracy of sentiment classifier models with Doc2Vec.



Model	Accuracy	ROC AUC Score
svm	0.35	0.48
logistic	0.36	0.50
randomforest	0.45	0.54

Figure 8. Random forest classifier model with Doc2Vec to predict sentiment confusion matrix.

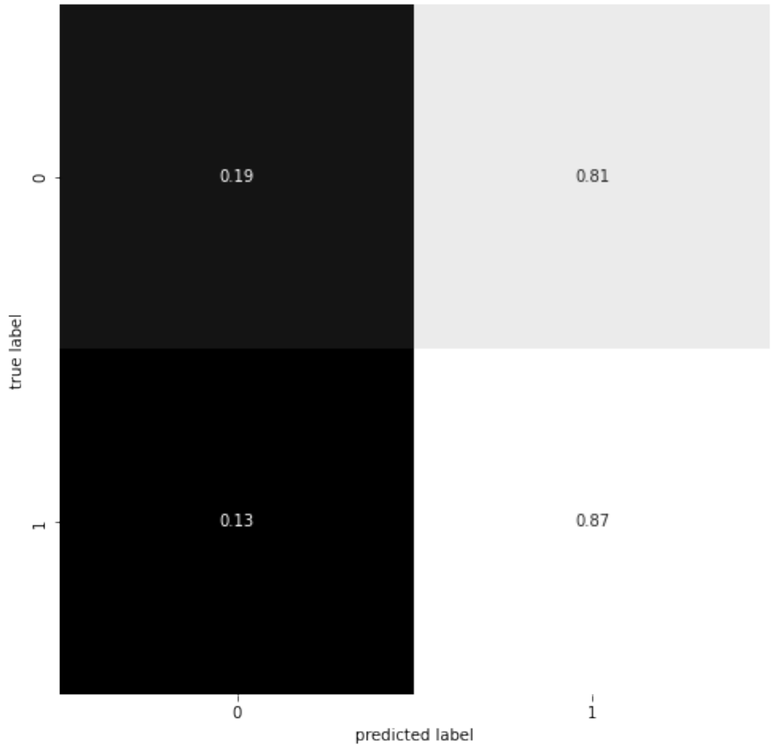


Figure 9. LSA 2 topics and 10 words plot.

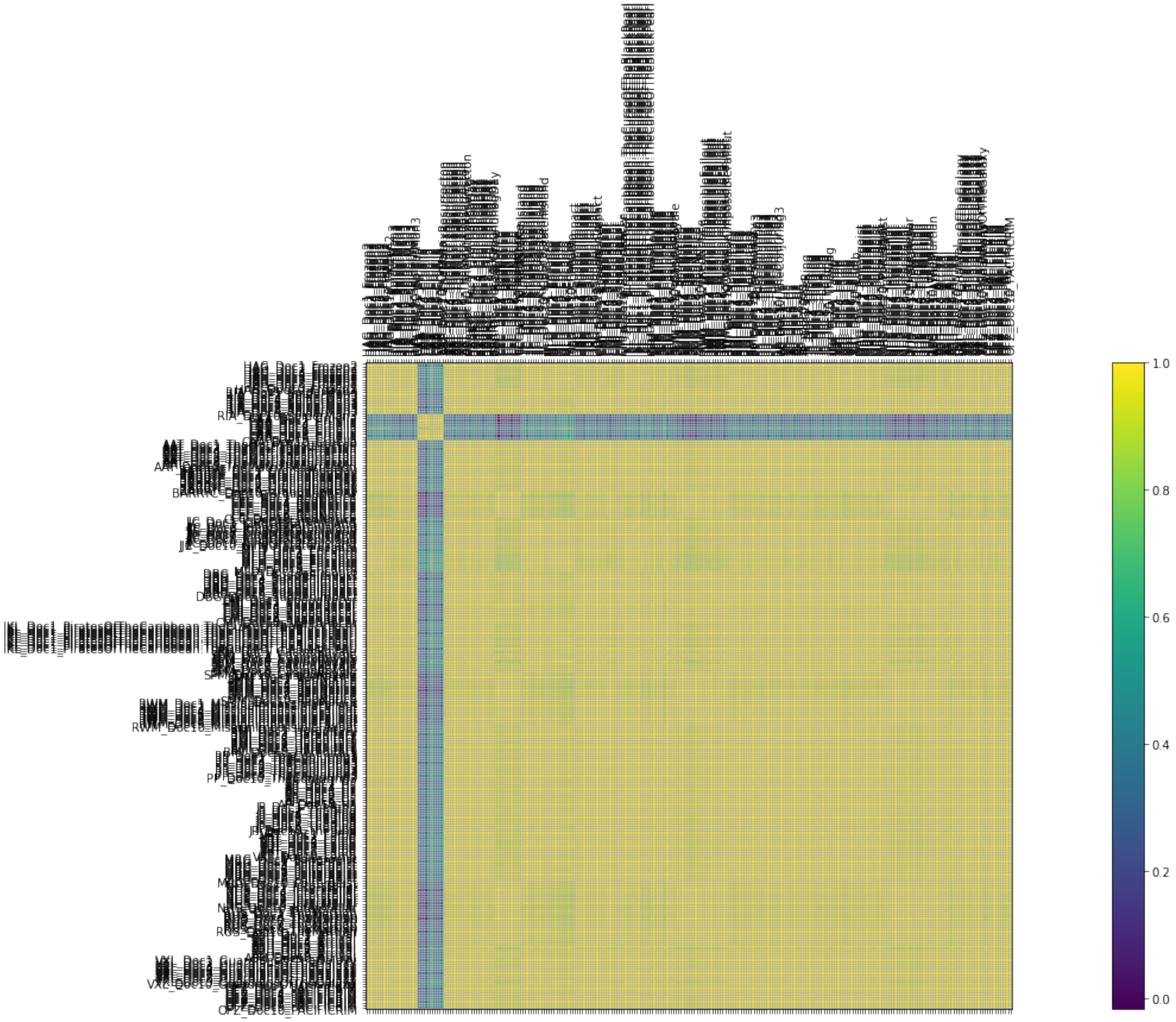


Figure 10. LSA 4 topics and 10 words plot.

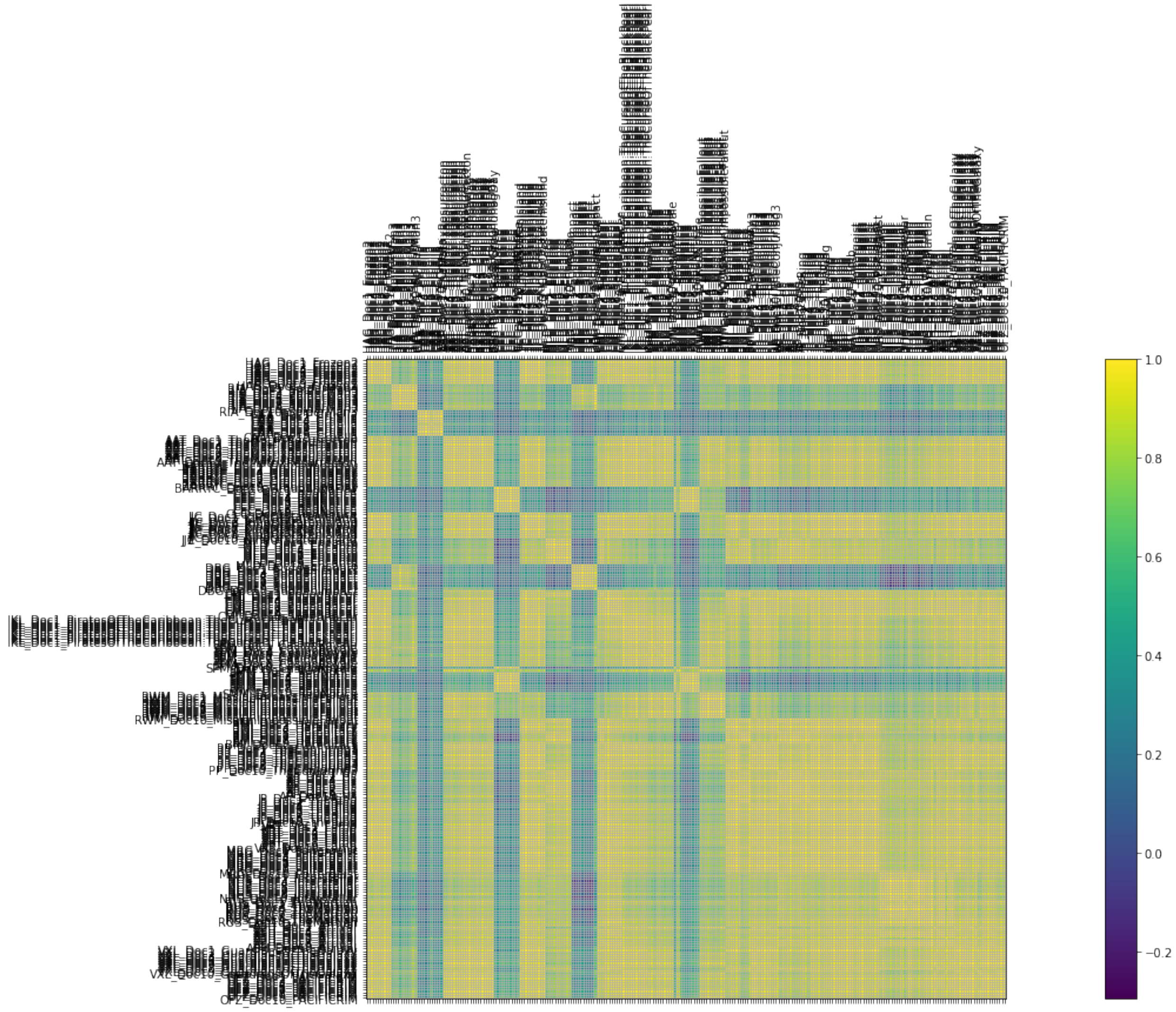


Figure 11. LSA 10 topics and 10 words plot.

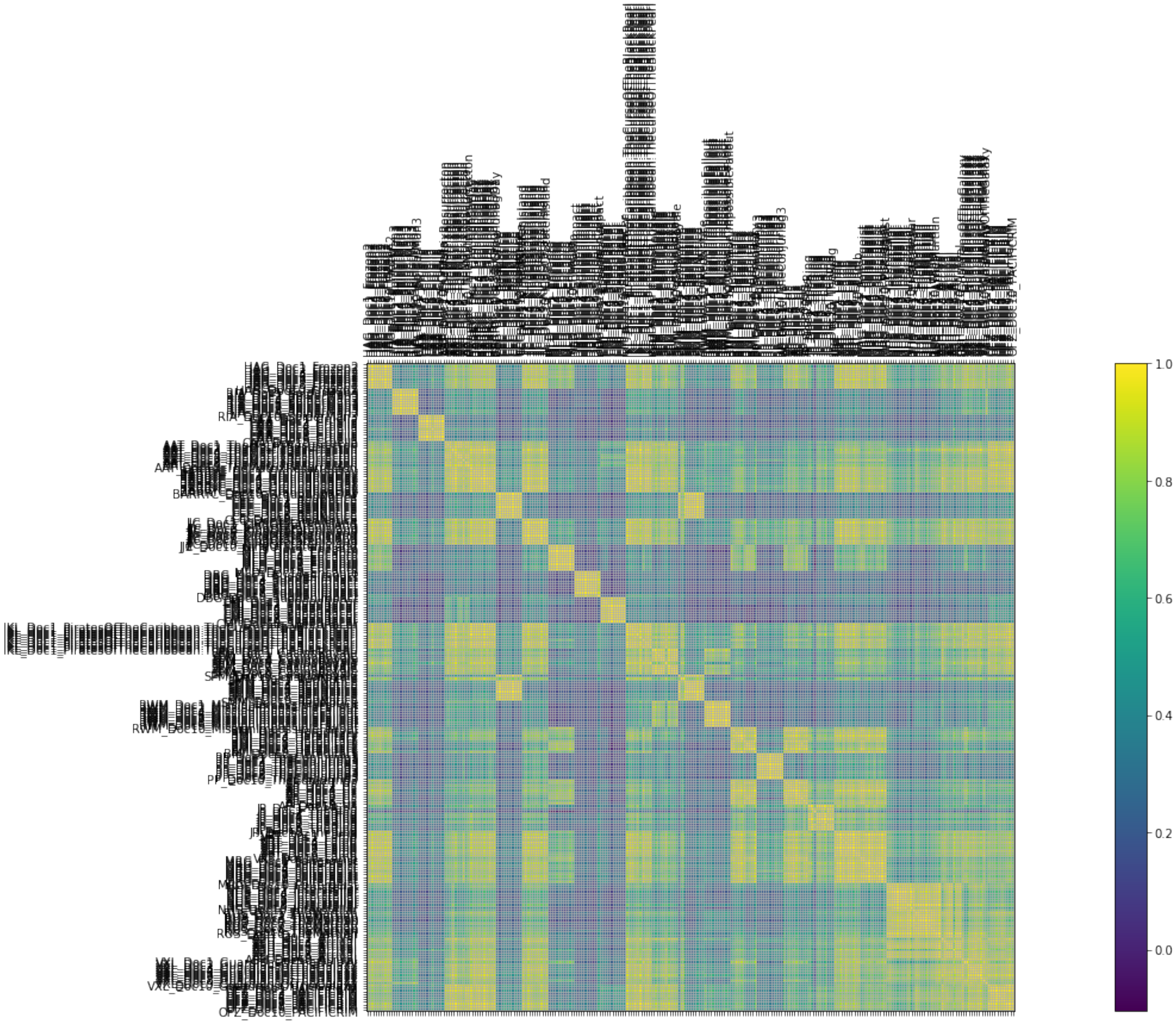


Figure 12. LDA 2 topics and 10 words plot.

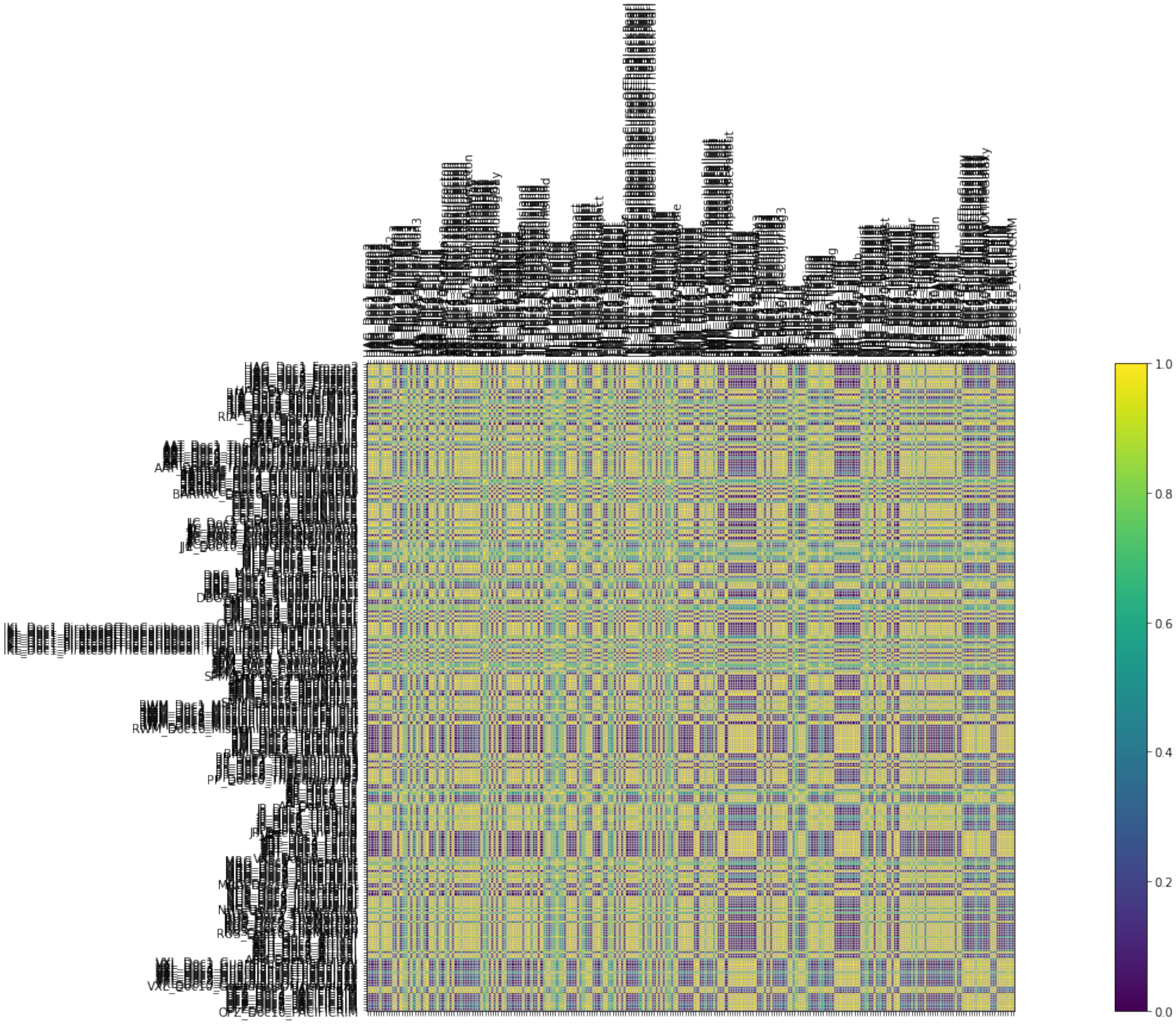


Figure 13. LDA 4 topics and 10 words plot.

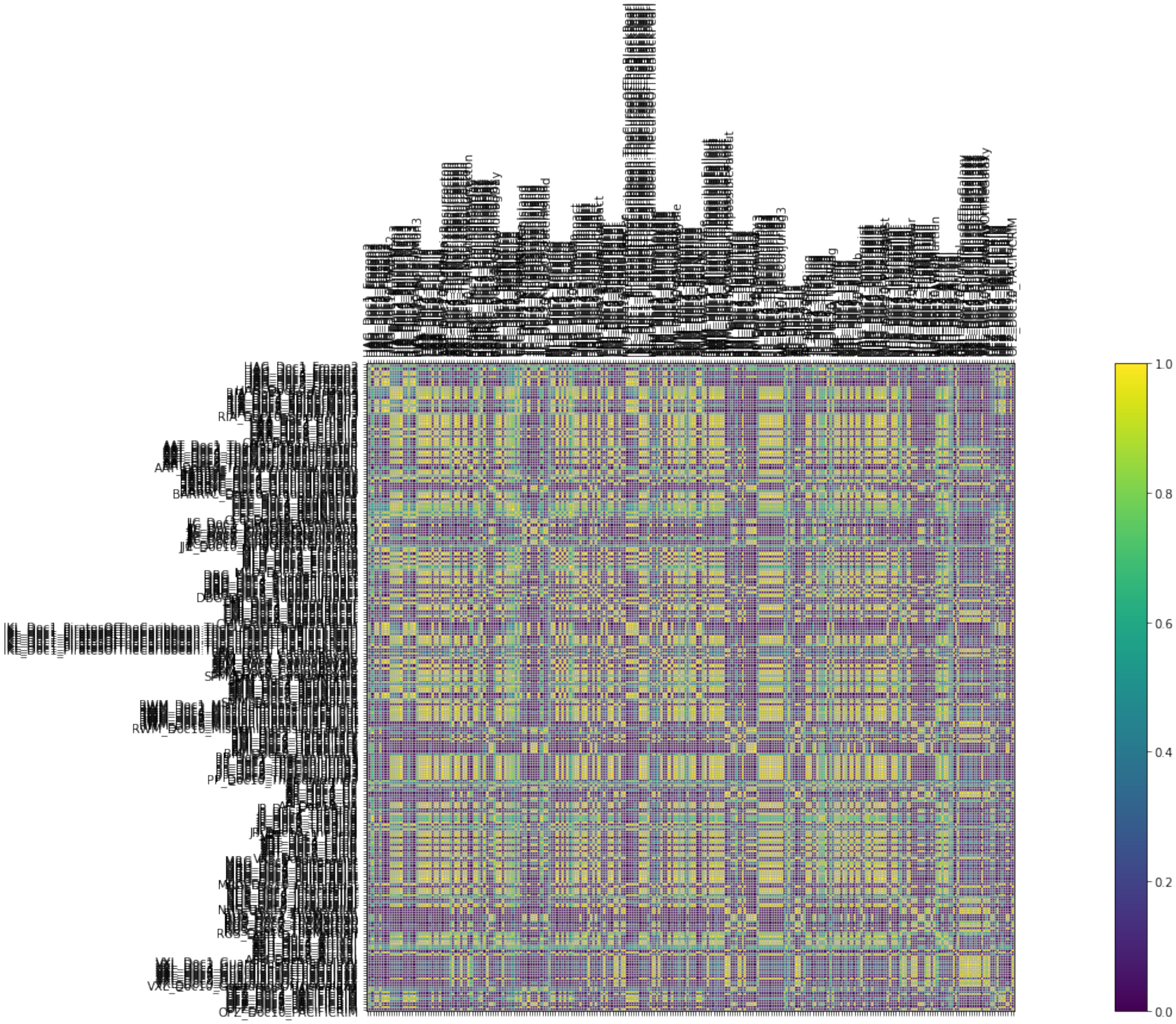
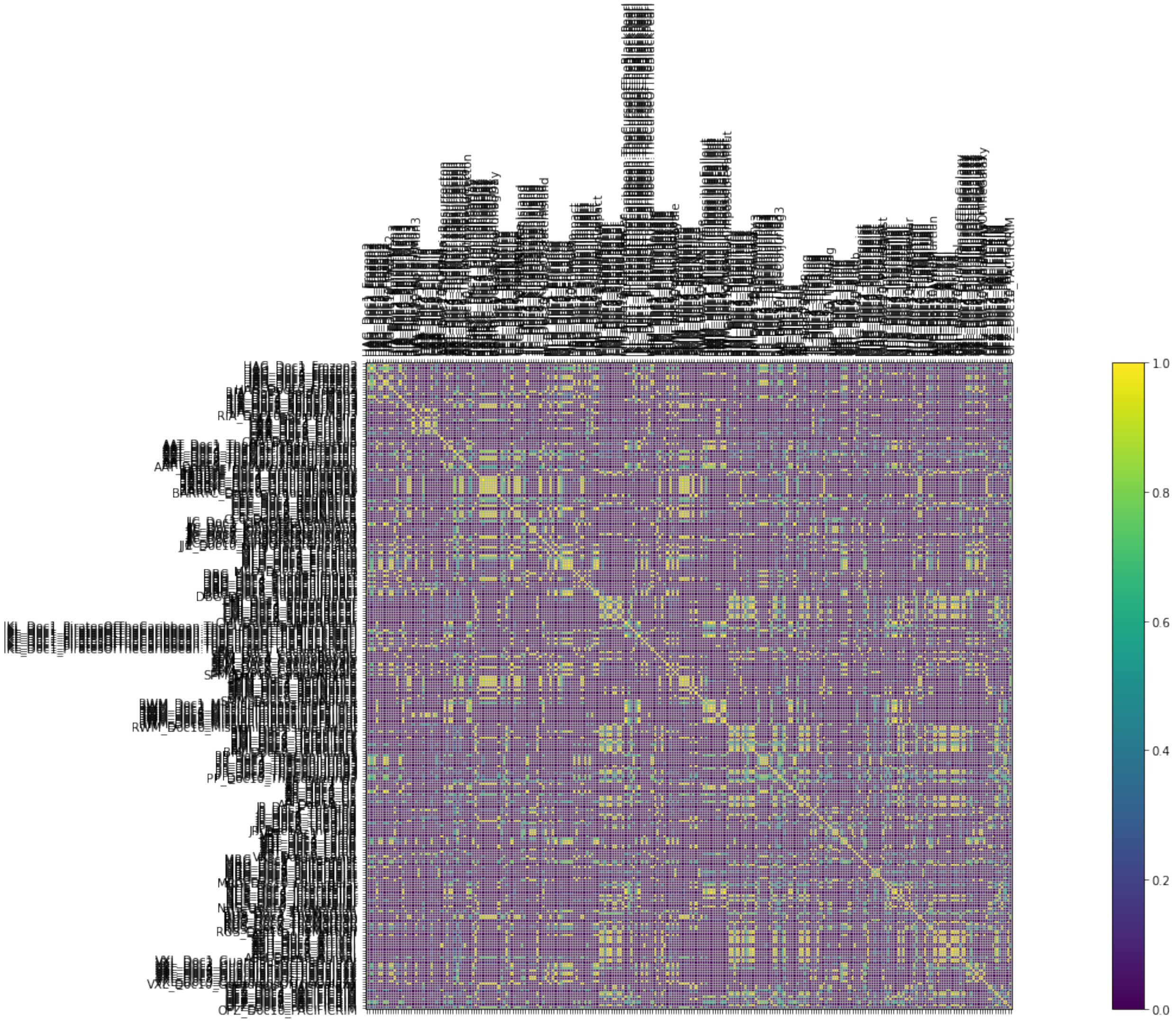


Figure 14. LDA 10 topics and 10 words plot.


HAC_Doc1_Frozen2
OFZ_Doc10_PACIFICRIM