Movie Reviews Under Review

Part 2: Clustering and Classification

Aubrey Molitor

MSDS 453: Natural Language Processing

Introduction

Continuing the process to compare and classify movies based on their reviews, it is necessary to perform three tasks: clustering, classification, and topic modeling. Each of these methods aims to group documents in a way that identifies similarities between them, forming the basis for personalized recommendations that have become so characteristic of services around the globe, from Netflix to Spotify. Performing accurate classification is key to understanding the meaning of documents. For each goal - clustering, classification, and topic modeling - alternative approaches will be evaluated to determine which is most apt at differentiating between documents and identifying uniting themes.

The Data

The data set for this project is the class corpus data set, compiled of ten reviews for each student in the class. The "trained data" was a 80% split of the corpus data frame, excluding the Genre of Movie (target variable). These texts were processed with a clean method to normalize, tokenize, lemmatize, stem documents and remove stopwords, then recombine the tokens to create a string per document. If train and test data were not needed, the data was processed as mentioned above directly from the corpus data frame, without dropping the target variable.

Research Design and Methods

Part 1: Clustering

For data clustering, the texts were run through a KMeans model trained on the training set of documents after processing as described above. Three models were created, one for each of 3, 4, and 5 clusters. After training these on the training data set, the models were used to predict clusters for both the training and test data sets. The silhouette score and cluster placements were saved for each of these trials, and the cluster distribution based on genre was printed. Additionally, the document titles in the testing model (3, 4, or 5 clusters) with the highest silhouette score were printed and used as the basis for refinement of the model.

Part 2: Classification

For the classification portion, the training and testing data will be used and transformed based on a TF-IDF vector fit to the training data. Then, four tests will be performed: naive bayes, SVM, and pre-trained BERT classification with both binary and multi-models. These models will be trained on the training data, and then will predict classifications for the testing data based on

movie genre. However, the BERT binary classification will be based on the sentiment of the documents rather than movie genre.

For the BERT classification models, the classification accuracy will be determined for the raw text, the cleaned raw text (with smart double/single quotes and long dashes, and non-ASCII characters removed), and the final text (cleaned as described in the data section). Then, accuracy measures and confusion matrices will be made for each to determine the best classification model. After determining the best model, measures will be considered on how to adjust the model for improvement.

Part 3: Topic Modeling

Topic modeling will include latent semantic analysis and latent dirichlet allocation. The text, processed in the clean method as described in the data section, will be modeled in LSA and LDA in tokenized format. Heatmaps will be made for each of these models to analyze the topic similarity, and the matrix similarities will be analyzed as well. Each model will be created with five topics each consisting of 10 words.

The Results

Clustering

The 3, 4, and 5 cluster runs of the training data had a silhouette score of .515, .533, and .454, respectively. The testing data's cluster runs had silhouette scores of .355, .489, and .523, respectively. Since the testing cluster with five documents had the highest silhouette score, the composition of its clusters were analyzed to adjust the model. The clusters had reasonably grouped distinctions, with a cluster for some of the more thoughtful, mind-mending action movies like Inception grouped with similarly intense Sci-Fi movies like Equilibrium. However, one issue with the clusters was that they were very uneven. For example, the third cluster had about 14 movies, whereas the first had 5 and the fourth had only 3. Additionally, though the cluster number was 5, there were only 4 groups made. To fix this, the embedding size for the Doc2Vec input vectors to run the KMeans tests was increased from 200 to 300. It seems the higher number of characteristic features in the Doc2Vec vector was able to better differentiate the documents, reducing the document numbers for clusters 1, 3, and 4 to 7, 11, and 5, respectively.

Classification

The Naive Bayes test had an accuracy score of .78, with 7 false positives in the confusion matrix. The SVM test had an accuracy score of 1.0 with a false positive rate of 0. The binary

classification model for sentiment classification had an accuracy of .5625 for each of the raw, raw cleaned, and final processed text. The multi-class classification BERT model had an accuracy of .1875 for the raw and raw cleaned text, and an accuracy of .375 for the full processed text. Based on the evidence, the BERT models were less apt at performing multi-class classification relative to the Native Bayes and SVM. Given that the Naive Bayes and SVM tests were only performed for movie genre, there is not a comparison to the BERT model for binary classification. However, between the naive bayes and SVM models, the SVM test performed the best in predicting genre, with a rate of 100% accuracy. This would be the chosen model for production and implementation based on its performance.

Given that the next best model was Naive Bayes, this model was chosen for improvement attempts. Instead of using maximum features of 5000 for the TF-IDF vector, the limit was set to 4000. This did slightly increase model performance by increasing the Naive Bayes accuracy score to .84 and reducing the false positive number to 5 documents.

Topic Modeling

The LSA model's five topics had the following words with its highest coefficients: movie, toxic, holmes, squad, and movie. The LDA model's first five words were: movie, movie, first, movie, movie. Additionally, manual observation of the LSA models' 10 words resulted in some ability to recognize a topic. For example, the "toxic" cluster centered around The Toxic Avenger characters and themes, as well as some action and superhero callouts. The "Squad" cluster was a more action-forward cluster, with Suicide Squad appearing as the main theme, with several character or actor-specific terms. LDA, in contrast, did not have as much differentiation; each cluster was a mix of loosely related specifics like "Taken" and very general film terms like "story" and "character." It is clear that based on this process, the LSA was more successful in forming some sort of topic, though it was still not a clear segmentation of the data.

In an attempt to improve these models, the number of topics was increased from 5 to 10. Though the first five topics for LSA were relatively unchanged, the added topics were fairly clustered, differentiated to the point that, for some, specific movies could even be picked out as defining the topic. More importantly, for those with shared movies, their similarities were clear, such as the shared superhero theme of The Toxic Avenger and Suicide Squad. For LDA, the topic specificity did improve, as there seemed to be less disparate themes within each topic. An example is the first topic, which had words centering around the Liam Neeson action movie.

However, each topic still did have several more general words like “scene,” “though,” and “since.” A recommended improvement for this LDA model would be to remove words that are commonly fillers, like “though” and “since,” or words with a very low TF-IDF score.

Conclusions

For each of clustering, classification, and topic modeling, there were models that stood out as the best to use and others that had specific improvement areas. The clustering approach was best completed by the KMeans models with a higher number of clusters, though general performance on testing data was still unsatisfactory. Changing the embedding dimensions of the Doc2Vec vector did slightly improve the performance of this model when the cluster count was held constant.

For classification, SVM stood out as the best multi-class classification model. Naive Bayes was not unsatisfactory, as the BERT binary and multi-class classification models were, but Naive Bayes was still clearly worse than SVM. However, decreasing the value for maximum features on the TF-IDF vector did help to improve the performance of the Naive Bayes model, bringing it to a usable accuracy.

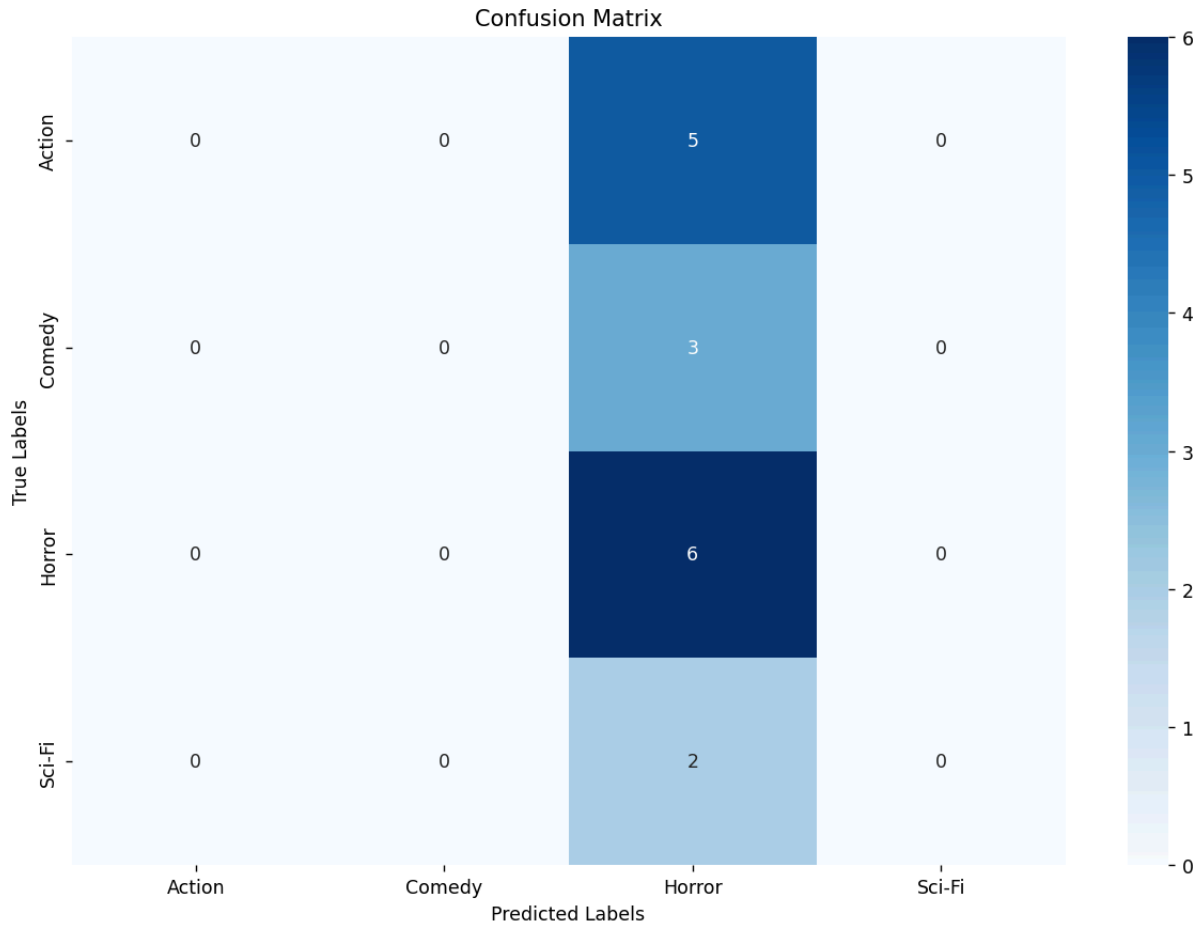
Finally, the LSA approach to topic modeling stood out relative to LDA as having more defining topics and more unique words in its chosen ten. Increasing the number of topics increased performance for both models, though there could be a performance drop-off with too many topics.

Appendix

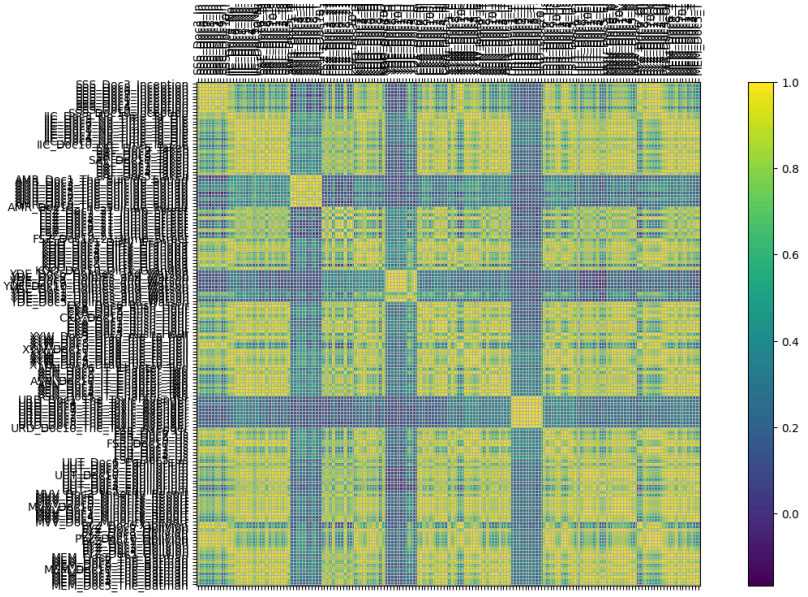
1. Classification Confusion Matrix Naive Bayes, post-adjustment

```
[[8 0 0 0]
[0 8 0 0]
[0 0 7 0]
[0 0 5 4]]
```

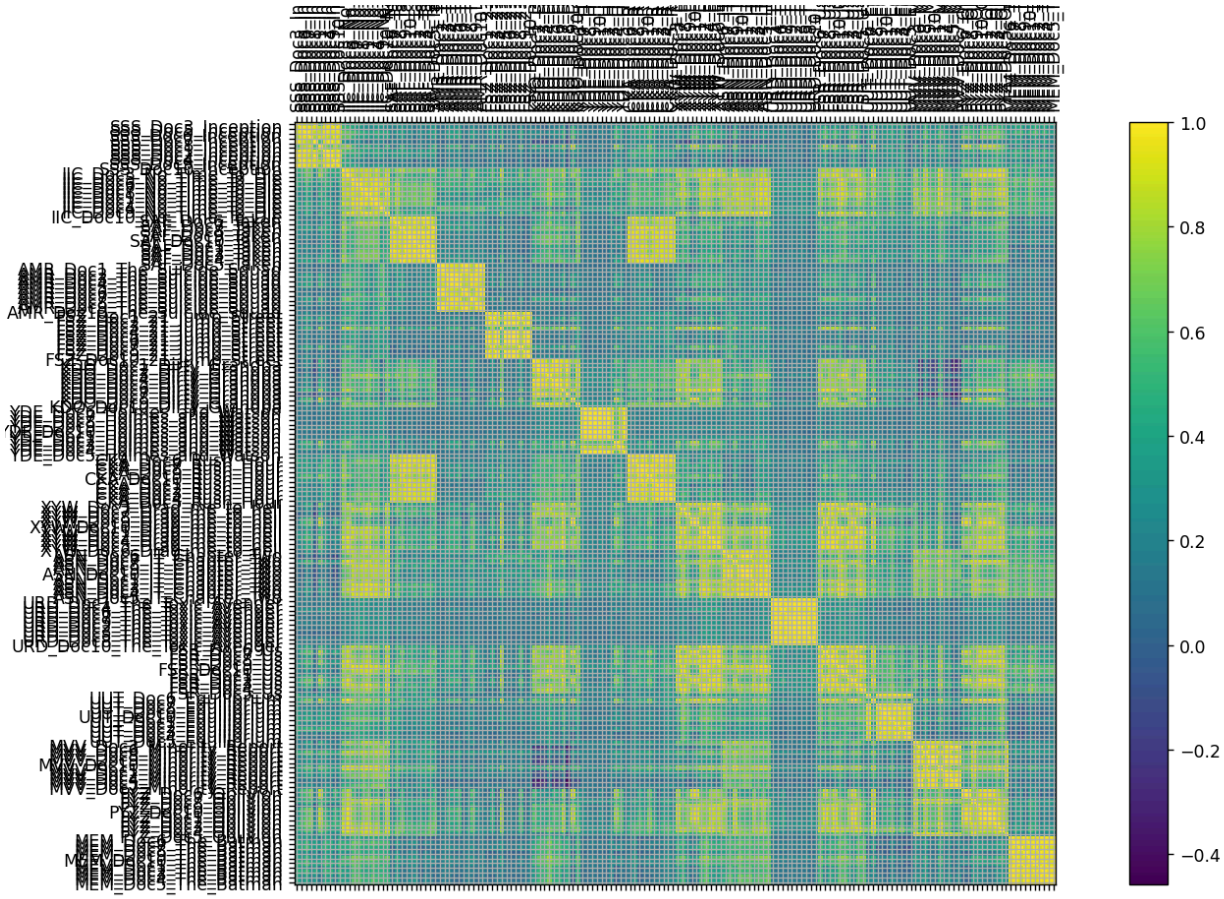
2. BERT Multi-class Classification Confusion Matrix



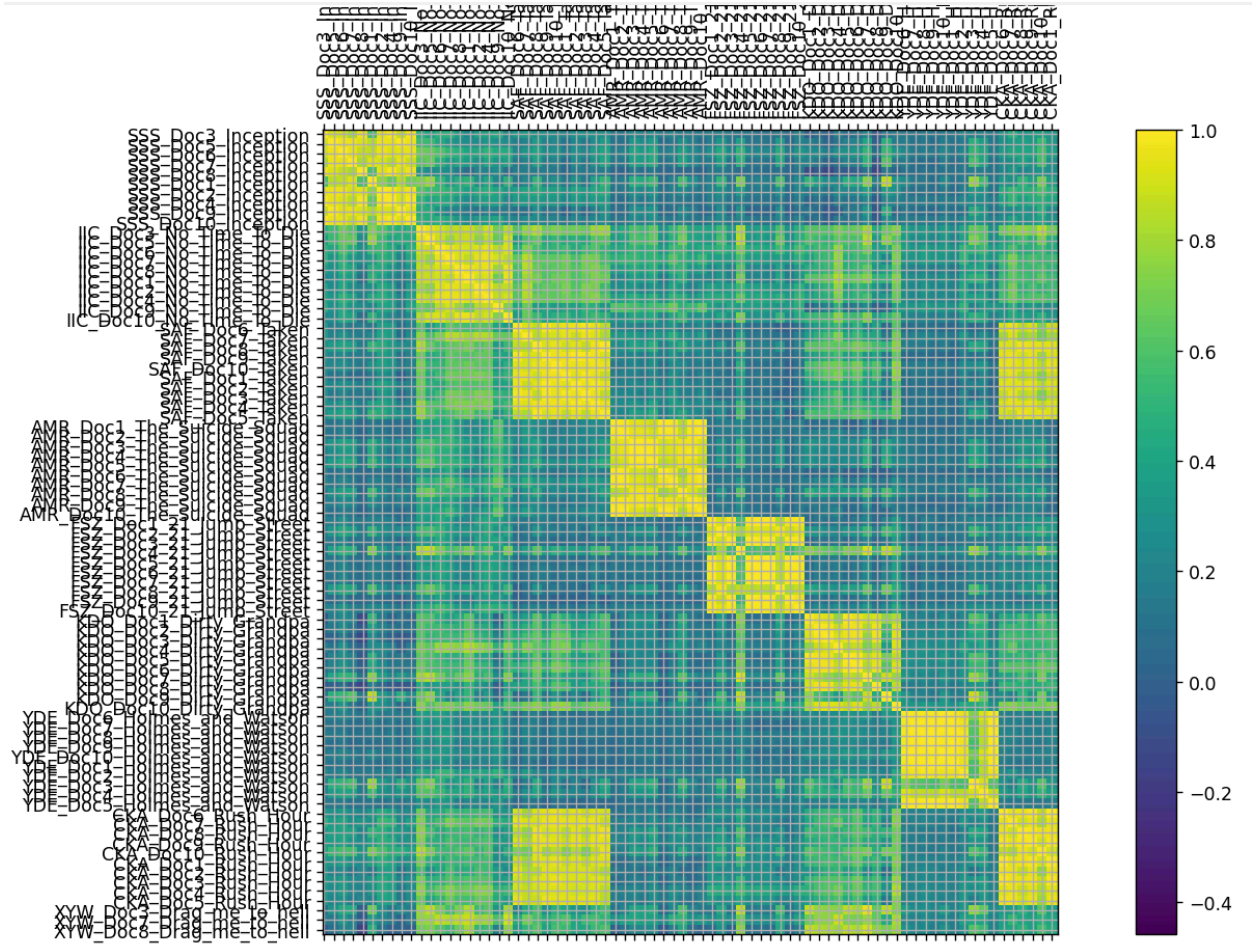
3. LSA Model, 5 topics of 10 words each

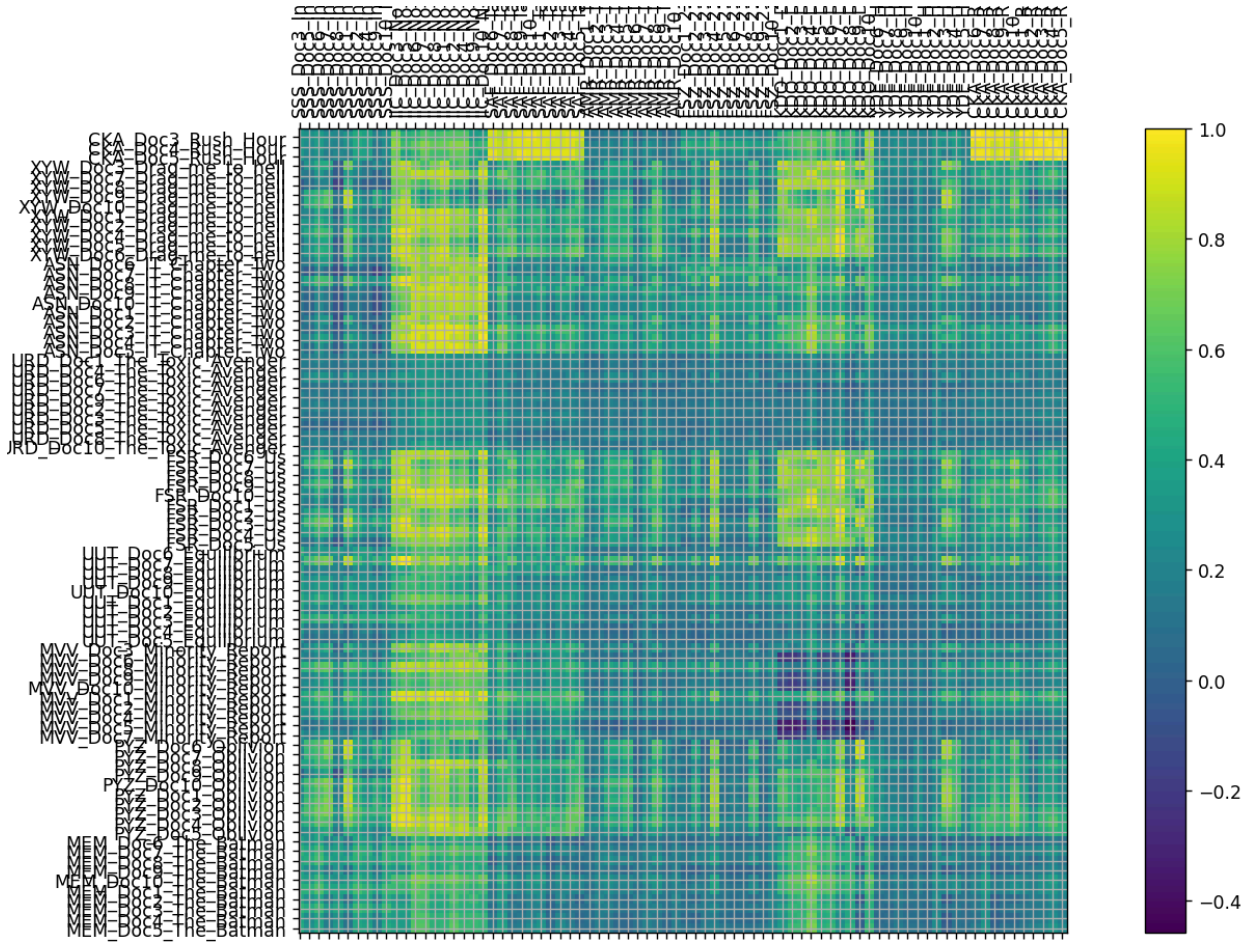


4. LSA Model, 10 topics of 10 words

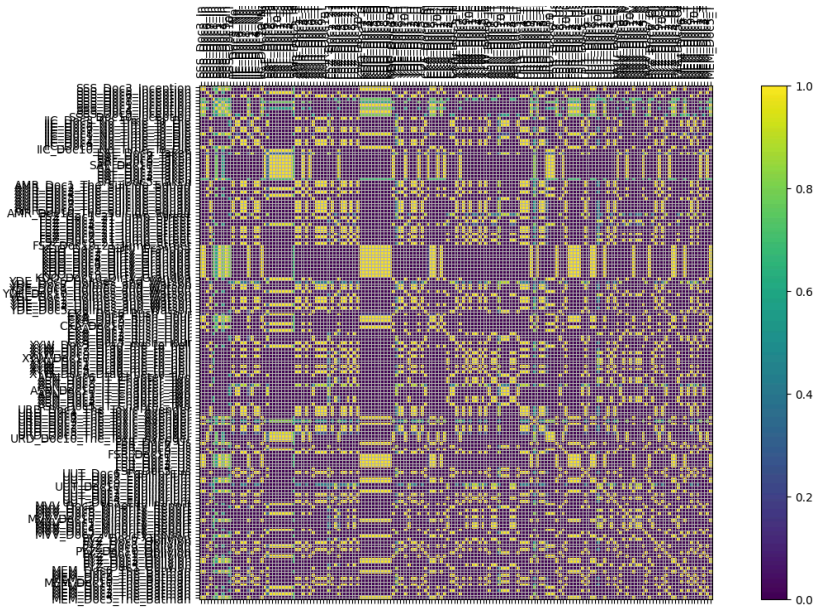


Close up for header legibility:

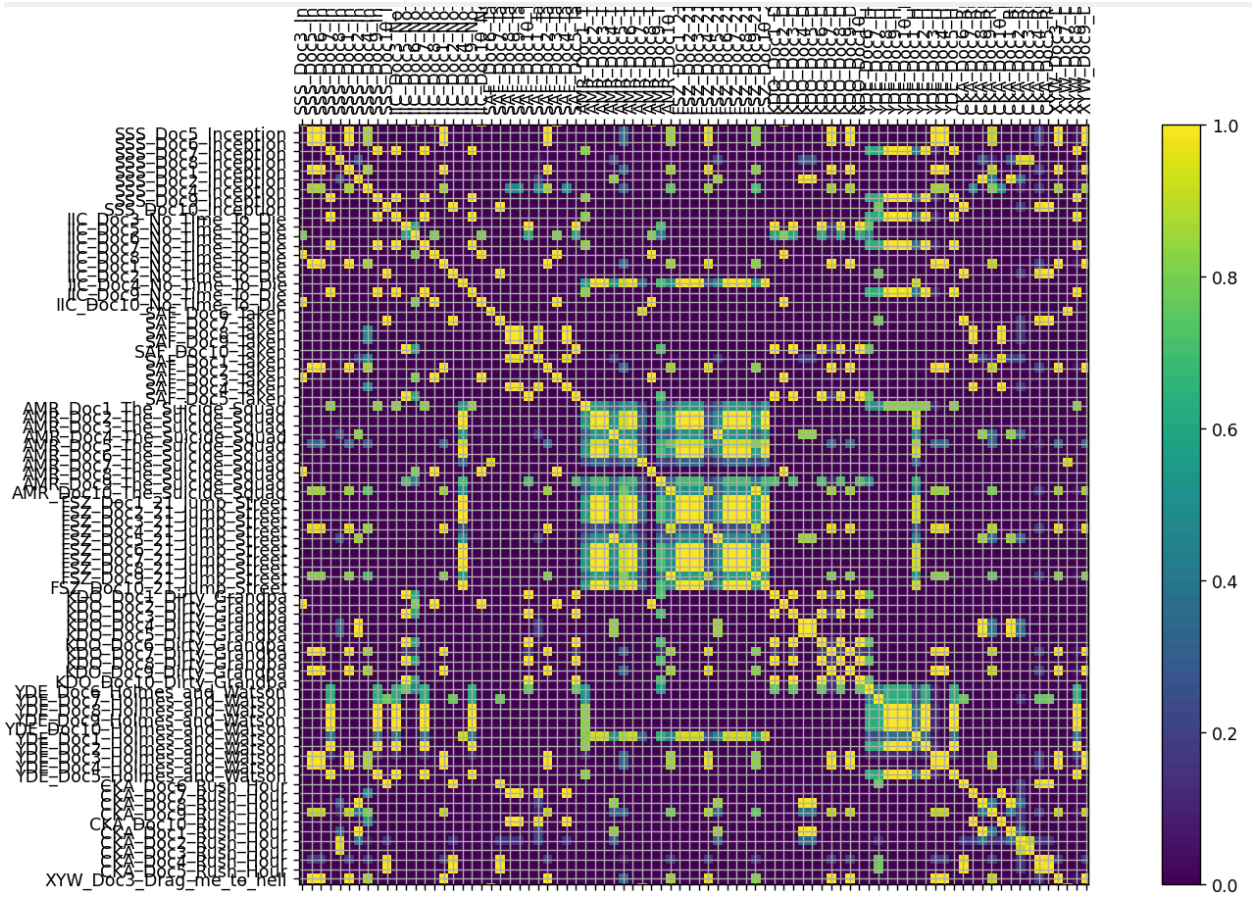


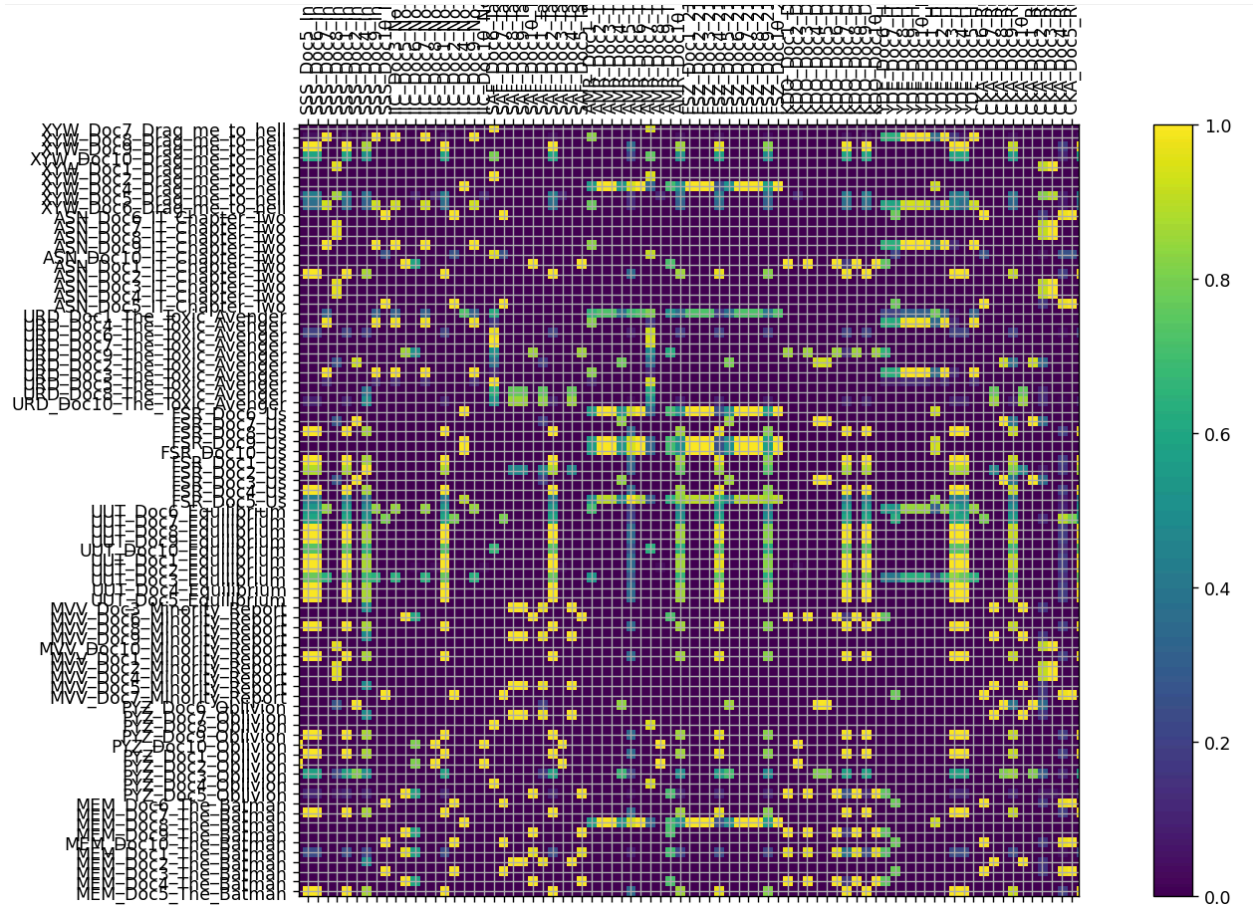


5. LDA Model, 5 topics of 10 words each



Close-up for header legibility:


SSS_Doc5_Inception
SSS_Doc6_Inception
SSS_Doc7_Inception
SSS_Doc8_Inception
SSS_Doc9_Inception
SSS_Doc4_Inception
SSS_Doc3_Inception
SSS_Doc10_Inception
IIC_Doc2_No_Time_To_Die
IIC_Doc3_No_Time_To_Die
IIC_Doc7_No_Time_To_Die
IIC_Doc6_No_Time_To_Die
IIC_Doc1_No_Time_To_Die
IIC_Doc5_No_Time_To_Die
IIC_Doc4_No_Time_To_Die
IIC_Doc9_No_Time_To_Die
IIC_Doc10_No_Time_To_Die
SAF_Doc6_Taken
SAF_Doc7_Taken
SAF_Doc8_Taken
SAF_Doc9_Taken
SAF_Doc10_Taken
SAF_Doc1_Taken
SAF_Doc2_Taken
SAF_Doc5_Taken
SAF_Doc4_Taken
SAF_Doc3_Taken
AMR_Doc1_The_Suicide_Squad
AMR_Doc2_The_Suicide_Squad
AMR_Doc3_The_Suicide_Squad
AMR_Doc4_The_Suicide_Squad
AMR_Doc5_The_Suicide_Squad
AMR_Doc6_The_Suicide_Squad
AMR_Doc7_The_Suicide_Squad
AMR_Doc8_The_Suicide_Squad
AMR_Doc9_The_Suicide_Squad
AMR_Doc10_The_Suicide_Squad
FSZ_Doc1_21_Jump_Street
FSZ_Doc2_21_Jump_Street
FSZ_Doc4_21_Jump_Street
FSZ_Doc5_21_Jump_Street
FSZ_Doc6_21_Jump_Street
FSZ_Doc7_21_Jump_Street
FSZ_Doc8_21_Jump_Street
FSZ_Doc9_21_Jump_Street
FSZ_Doc10_21_Jump_Street
KDO_Doc3_Dirty_Grandpa
KDO_Doc2_Dirty_Grandpa
KDO_Doc4_Dirty_Grandpa
KDO_Doc5_Dirty_Grandpa
KDO_Doc6_Dirty_Grandpa
KDO_Doc7_Dirty_Grandpa
KDO_Doc8_Dirty_Grandpa
KDO_Doc9_Dirty_Grandpa
KDO_Doc10_Dirty_Grandpa
YDE_Doc6_Holmes_and_Watson
YDE_Doc7_Holmes_and_Watson
YDE_Doc8_Holmes_and_Watson
YDE_Doc9_Holmes_and_Watson
YDE_Doc10_Holmes_and_Watson
YDE_Doc1_Holmes_and_Watson
YDE_Doc2_Holmes_and_Watson
YDE_Doc3_Holmes_and_Watson
YDE_Doc4_Holmes_and_Watson
YDE_Doc5_Holmes_and_Watson
CKA_Doc6_Rush_Hour
CKA_Doc7_Rush_Hour
CKA_Doc8_Rush_Hour
CKA_Doc9_Rush_Hour
CKA_Doc10_Rush_Hour
CKA_Doc1_Rush_Hour
CKA_Doc2_Rush_Hour
CKA_Doc3_Rush_Hour
CKA_Doc4_Rush_Hour
CKA_Doc5_Rush_Hour
XYW_Doc3_Drag_Me_To_Hell
1.0
0.8
0.6
0.4
0.2
0.0


XYW_Doc7_Drag_me_to_he
XYW_Doc8_Drag_me_to_he
XYW_Doc9_Drag_me_to_he
XYW_Doc10_Drag_me_to_he
XYW_Doc1_Drag_me_to_he
XYW_Doc2_Drag_me_to_he
XYW_Doc3_Drag_me_to_he
XYW_Doc4_Drag_me_to_he
XYW_Doc5_Drag_me_to_he
XYW_Doc6_Drag_me_to_he
ASN_Doc7_It_Chapter_Two
ASN_Doc8_It_Chapter_Two
ASN_Doc9_It_Chapter_Two
ASN_Doc10_It_Chapter_Two
ASN_Doc1_It_Chapter_Two
ASN_Doc2_It_Chapter_Two
ASN_Doc3_It_Chapter_Two
ASN_Doc4_It_Chapter_Two
ASN_Doc5_It_Chapter_Two
URD_Doc4_The_Toxic_Avenger
URD_Doc6_The_Toxic_Avenger
URD_Doc7_The_Toxic_Avenger
URD_Doc9_The_Toxic_Avenger
URD_Doc3_The_Toxic_Avenger
URD_Doc8_The_Toxic_Avenger
URD_Doc1_The_Toxic_Avenger
URD_Doc2_The_Toxic_Avenger
URD_Doc10_The_Toxic_Avenger
FSR_Doc9_Lust
FSR_Doc7_Lust
FSR_Doc8_Lust
FSR_Doc10_Lust
FSR_Doc1_Lust
FSR_Doc2_Lust
FSR_Doc3_Lust
FSR_Doc4_Lust
FSR_Doc5_Lust
UUT_Doc6_Equilibrium
UUT_Doc7_Equilibrium
UUT_Doc8_Equilibrium
UUT_Doc9_Equilibrium
UUT_Doc10_Equilibrium
UUT_Doc1_Equilibrium
UUT_Doc2_Equilibrium
UUT_Doc3_Equilibrium
UUT_Doc4_Equilibrium
UUT_Doc5_Equilibrium
MVV_Doc7_Majority_Report
MVV_Doc8_Majority_Report
MVV_Doc9_Majority_Report
MVV_Doc10_Majority_Report
MVV_Doc1_Majority_Report
MVV_Doc2_Majority_Report
MVV_Doc3_Majority_Report
MVV_Doc4_Majority_Report
MVV_Doc5_Majority_Report
PVZ_Doc6_Oblivion
PVZ_Doc7_Oblivion
PVZ_Doc8_Oblivion
PVZ_Doc9_Oblivion
PVZ_Doc10_Oblivion
PVZ_Doc1_Oblivion
PVZ_Doc2_Oblivion
PVZ_Doc3_Oblivion
PVZ_Doc4_Oblivion
PVZ_Doc5_Oblivion
MEM_Doc6_The_Batman
MEM_Doc7_The_Batman
MEM_Doc8_The_Batman
MEM_Doc9_The_Batman
MEM_Doc10_The_Batman
MEM_Doc1_The_Batman
MEM_Doc2_The_Batman
MEM_Doc3_The_Batman
MEM_Doc4_The_Batman
MEM_Doc5_The_Batman
1.0
0.8
0.6
0.4
0.2
0.0