What Makes an Award Winning Film Script?

Allison Ragan and Allison Shafer

STAT 696

23 April 2020

Executive Summary

Best Original Screenplay. It is one of the highest honors a film writer can be awarded by the Academy of Motion Picture Arts and Sciences, the world's preeminent movie-related organization, during the Oscars ceremony. But, what makes an original screenplay stand out amongst the other nominees and do award-winning original screenplays share common qualities and characteristics? Our study utilizes Natural Language Processing (NLP) practices to examine sentiment, and the writer's vocabulary selection shared within movie genres and between victors and nominees. Sentiment analysis, topic modeling, semantic similarity analysis, and K-means clustering are all employed to explore our hypothesis that screenplays that receive the Academy Award for Best Original Screenplay share similar emotional patterns and characteristics.

Introduction

The purpose of this study was to utilize various NLP and statistical methods, using the Python programming language, to analyze the textual data from movie scripts that were nominated for or won the Oscar for the Best Original Screenplay. Through this analysis, the goal was to determine if there is a similarity between the semantics used and the emotions portrayed throughout the textual dataset, and if any shared characteristics contributed to the chances of winning this award.

The subject dataset consisted of text from the film scripts for the movies nominated for the Best Original Screenplay Academy Award over the past five years. PDF versions of 24 of the 25 films nominated for this category during this timeframe were acquired from internet resources such as Scripts.com and SimplyScripts.com. The text was extracted from the uniquely organized PDF documents, then cleaned and normalized after careful assessment of the document's structure and cue indicators. After cleaning the text, we were able to group the data by various attributes and begin our analysis. One of the subsets of data we created was a grouping of the data by genres, which were based on a "broad genre," such as "comedy-drama" or "mystery-suspense," as opposed to a multi-descriptive genre label, such as "drama, mystery-suspense, sci-fi." We also created a data subset consisting of all of the film scripts that won the Best Original Screenplay award. We first performed sentiment analysis on the text to determine if emotional patterns played a role in defining a screenplay as a winner.

Then we conducted topic modeling on the data to see which vocabularies and themes were most associated with the winners, between genres, and among the full dataset consisting of both dialogue and visual cues. Cosine similarity was then performed to determine which films were most similar to one another and if there were patterns within the groups of winners and genres. Finally, we pursued K-means clustering to determine if there was a natural separation of scripts that won the Best Original Screenplay award and nominees that did not.

Methodology

Cleaning

Before analysis could begin, the text data had first to be extracted from the PDFs of scripts using Python's ***PDFMiner*** library. Once we had the text of the script extracted, we began cleaning the script to a standard format. We first dropped extraneous text unrelated to the script, such as the title page, the production team information, and page numbers. Next, we leveraged dialogue indicators present in most of the scripts to parse out dialogue from the rest of the script, so we were working with two corpora: a corpus of just dialogue and a corpus of the full script.

This was done so that we were able to see if the results were due to the full content of the script (including setting, visual cues, etc.) or if it was the dialogue alone that resulted in a win or not. Once the dialogue was as separated as possible, we cleaned the text to normalize it across all scripts which included expanding contractions, dropping stop words, dropping words with less than a length of three, stripping non-alphanumeric characters, lemmatizing, and tokenizing both the dialogue corpus and the full script corpus. Upon cleaning the data, we were left with 310,029 words in the dataset representing both visual cues and dialogue, and 193, 384 words representing the dialogue. We then combined the cleaned data with some supplementary data on the films--such as genre (broadened to reduce the number of unique categories), if it won Best Original Screenplay, if it won Best Picture, etc.--to enrich our analyses.

Sentiment Analysis

We performed sentiment analysis to explore the emotional aspect of our scripts and see if that contributed to a Best Original Screenplay win. The AFINN Lexicon, often referred to as the "new ANEW," is a list of English words manually labeled for valence on a scale from -5 to +5. The lexicon is included in the ***afinn*** library in Python and can be applied to text to calculate the polarity score based on that labeled scale in the lexicon. We leveraged the AFINN Lexicon due to its similarity to the ANEW word list with which we are familiar and the ease of use as a result of its creation into a Python library.

In total, sentiment analysis was performed three times: first on the corpus of just dialogue, next on the corpus of full scripts, and finally on chunks of the full script. For the third analysis, we split our script into chunks of 250 words so that we could see how sentiment changes throughout each film.

Topic Modeling

Topic modeling by way of Non-Negative Matrix Factorization (NMF) was applied to each of our primary data subsets to determine what the common themes amongst the movies were and which scripts shared related topics. We used the ***TfidfVectorizer*** from Python's ***sklearn*** module to create the TF-IDF matrix used to fit the NMF model. For each instantiation of the ***TfidfVectorizer,*** we chose hyperparameters to ensure the most meaningful words, which were not too rare or too frequent, existed in our vocabulary of terms.

To begin, we used the dialogue data, which consisted of 193, 384 words. Thresholds were carefully chosen by testing multiple ranges for the ***min_df*** and ***max_df***, which help curate a library to limit the number of times a word is utilized in the corpus or throughout the documents. Due to the size of the dataset, we knew we would have to be stricter when setting these parameters to obtain a valuable dictionary. We started the model by using unigrams and setting ***max_df*** at .8 and ***min_df*** at .1 so that we would obtain only words that were found in no more than 80% of the documents and no less than 10% of the documents. This resulted in a dictionary of 4,093 words. We tried the same ***min_df*** and ***max_df*** values for unigrams and bigrams collectively and received 6,332 words. Upon review of the vocabulary, it looked like some of the word combinations that made up the bigrams were valuable, such as "African American" and "thousand dollar," and added additional context. For the remainder of the ***TfidfVectorizer*** instantiations, we kept the n-grams parameters to include unigrams and bigrams based on this. However, we wanted to narrow down the vocabulary a little further and started by altering the ***max-df*** to 85%, so n-grams in the vocabulary were found in no more than 85% of the scripts in the corpus. This increased the number of vocabulary words. After playing around with a few more values, we decided to tweak the ***min-df*** to .2. Our vocabulary subset then consisted of words that were in no more than 85% of the scripts and no less than 20% of the movie scripts. This gave us 2,818 words in our vocabulary that appeared to be purposeful.

When conducting the topic modeling of the full script, including visuals and dialogue, we tightened the parameters up a little bit further, given that the full script lengths were often thousands of words longer than the dialogue by itself. We set the ***max-df*** at .8 and ***min-df*** at .2 so that we would only obtain words not in more than 80% of the documents and not in less than 20% of the documents. This apportioned 3,641 words to our vocabulary of words for the full scripts, and we were satisfied with this result.

We then utilized the dataset that was grouped by the broad genre to complete topic modeling by genre to compare the topics represented within each genre. First, we fit the ***TfidfVectorizer*** on the grouped dataframe using the same parameters that yielded success in past executions. This resulted in a vocabulary of over 151,000 words. Since we still wanted only to include words that were found in less than 85% of the scripts, we left our maximum parameter set at .85. We decided to choose a minimum threshold of 3, meaning any n-gram chosen would appear in at least 3 of the scripts.

This yielded in 3,146 words with substance in the vocabulary. We were satisfied with this result.

Finally, we used the subset of data that consisted only of the dialogues that won the Original Screenplay Award and fit the ***TfidfVectorizer*** on the winning scripts' data, using the same parameters that were used for the genre-related topic modeling. We received 728 unigrams and bigrams in the vocabulary. We were satisfied with this result and did not do further hyperparameter testing.

To model our topics, we chose to use NMF to reduce the dimension of the term-by-document matrix and to ensure for more interpretable results. We fit the NMF model on the term frequency-inverse document frequency (TF-IDF) matrices that were formed as a result of the ***TfidfVectorizer***.

We chose a different number of topics based on the size of the data subset. For instance, for the full dataset, we selected ten topics to model, as well as twenty-four topics. The parameter of twenty-four topics resulted in what seemed like more repetitive words amongst the topics, so we decided to drop the number of topics parameter to ten. Upon setting the topics to ten, it was a bit easier to see relation amongst the words within the topic, and the topics were more unique. When modeling the topics for the data grouped by genres, we selected five for the number of topics. We chose five for the number of since there were five genres represented in our dataset, the topics may be more clearly defined by the words selected for each when matching the number of groups. For the topic modeling for the winning scripts, we also set the number of topics to five, as five scripts were denoted as winners in the dataset.

Cosine Similarity

Cosine Similarity from the ***sklearn*** module was applied to the previously constructed TF-IDF matrices for the genres and the winning scripts subsets. This analysis allowed us to explore our hypothesis of how similar the scripts within the genres are, and which genres are most similar, and to explore whether or not winning scripts share similar textual characteristics.

K-Means Clustering

We hypothesized that winners of Best Original Screenplay would share similar characteristics that set them apart from other nominees. To test this hypothesis, we performed K-Means Clustering to group similar scripts and see if there exists a natural delineation between winners and nominees. Specifically, we clustered on two corpora of data: just dialogue as a TF-IDF vector and the full script as a TF-IDF vector.

While clustering is possible on Bag-of-Words vectors, in practice, the clustering algorithm produced clusters with all but one in one cluster and the remainder in the other. As a result, we elected to cluster just on the TF-IDF vectors as it produces clusters with more than just one member.

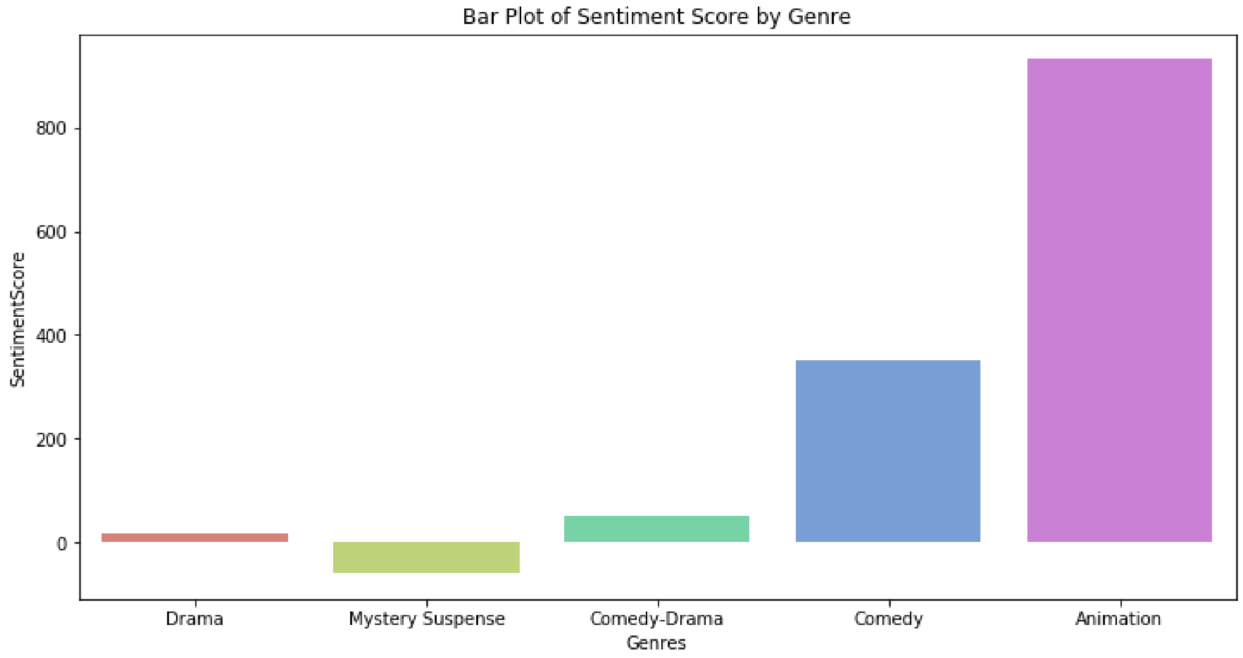
As we were performing clustering to see if there existed a natural separation of winners and nominees, we chose to use only two clusters to test this hypothesis. Additionally, due to the small sample size, clusters greater than two produced results where at least one cluster contained only one film. For these reasons, we clustered our scripts into two groups.

Results

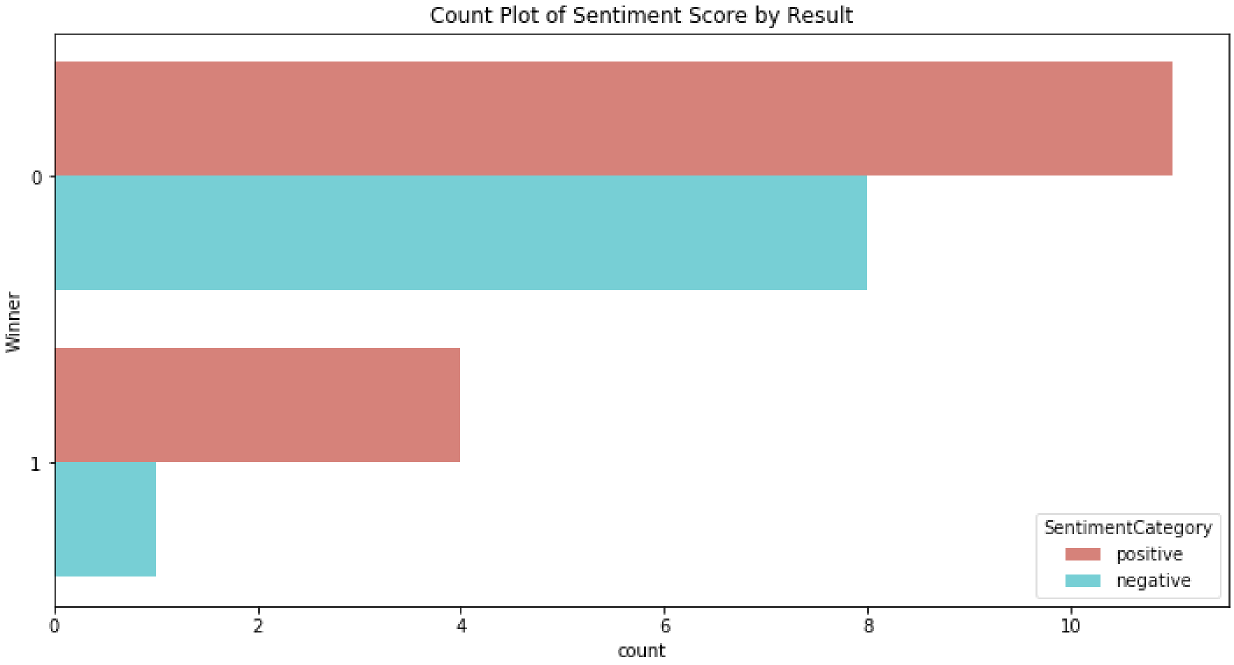
Sentiment Analysis

For every Sentiment Analysis run on our data, we compared the sentiments across genres and also between winners of Best Original Screenplay and the nominees. The results of our analyses on the corpus of just dialogue and the full script were not significantly different, so henceforth these results shall be addressed as one.

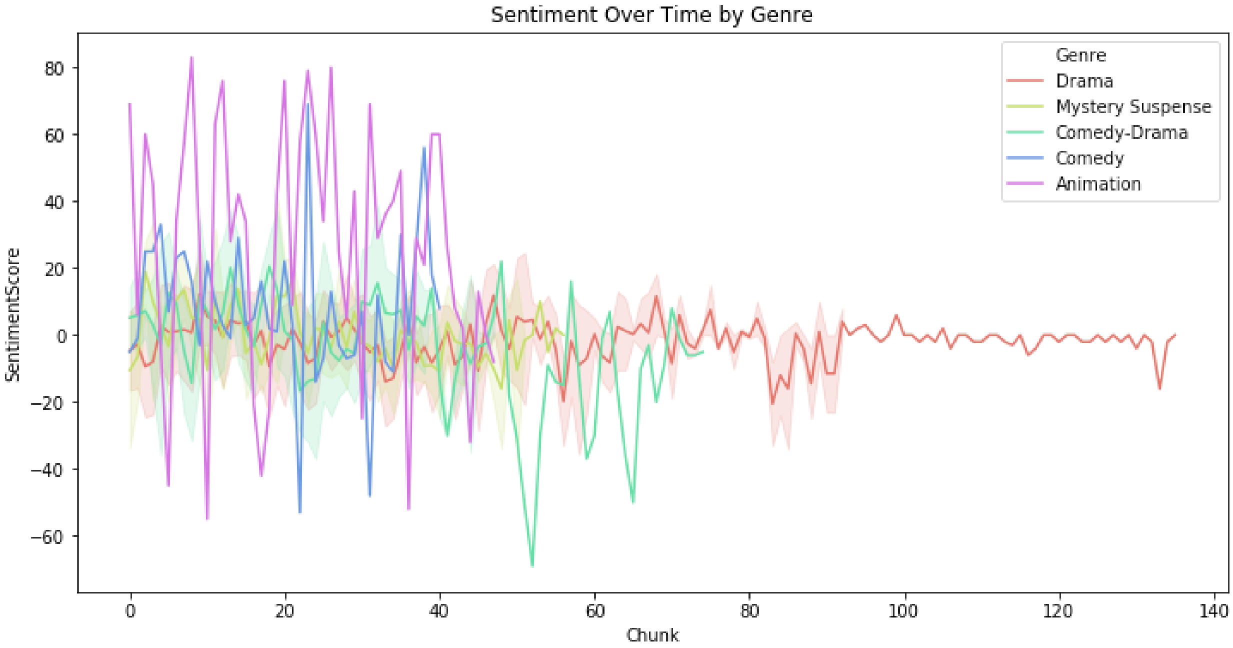
In comparing across genres, sentiment scores and categories looked how we expected them to. For instance, drama as a whole had relatively low sentiment scores, whereas a genre like comedy had much higher mean sentiment scores across all films in that genre. Animation technically had the highest sentiment compared to other genres; however, there was only one film in the animation genre--Pixar's *Inside Out*--whereas the sentiment score of all other genres is a mean score.



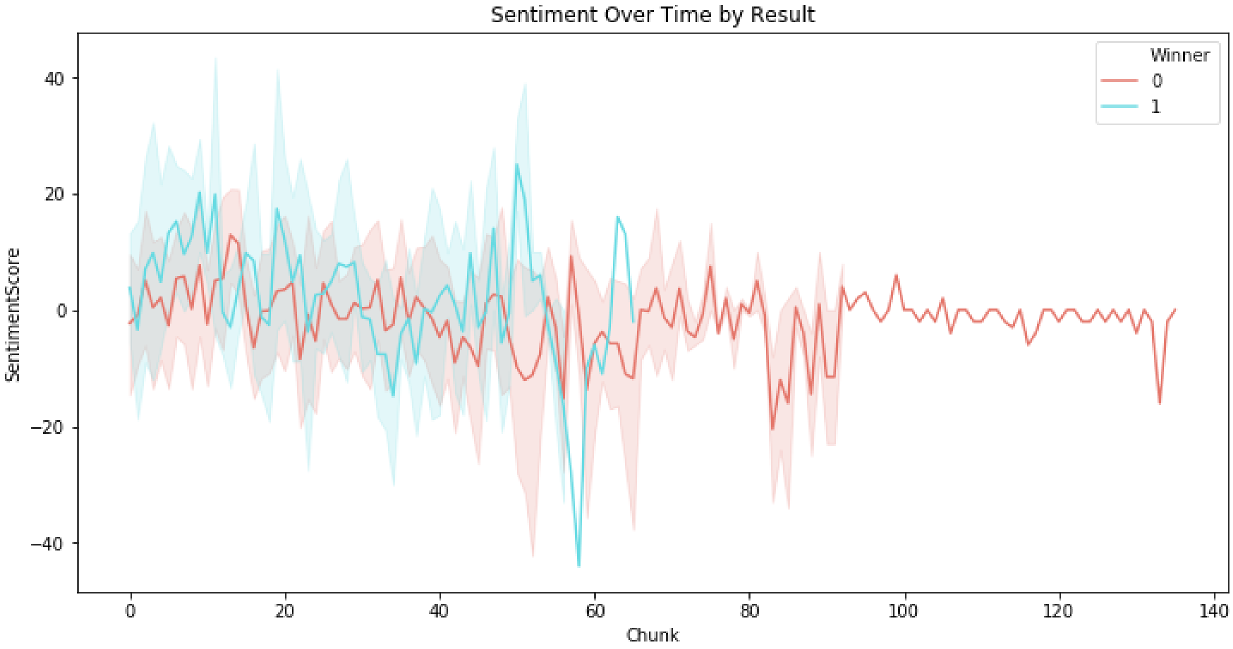
Comparing winners against nominees revealed that the majority of the winners had an overall positive sentiment, with all but one being positive. Among nominees, the same was true, although the sample size of nominees was significantly larger than winners. Future analyses should include a much larger sample size of winners to allow more opportunity to identify trends among winners in terms of sentiment.



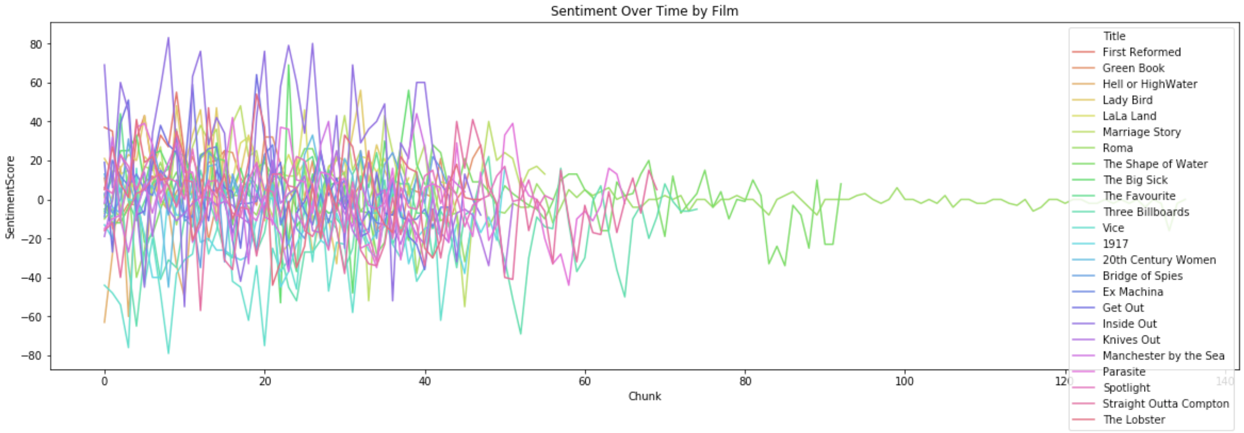
The results of sentiment over time revealed some interesting trends, complementing the results achieved by the above analyses. Genre comparison throughout the film demonstrated that sentiment score in dramas remained relatively consistent across the duration of the films. Comedy-drama had some wild fluctuations in sentiment, dropping down pretty far in the negative towards the final approximate one-third of the film. As with the bar plot above, animation encompassed only one film but demonstrated the most dramatic shifts in sentiment throughout the film. However, given that the movie was primarily focused on emotions and, in our personal experience, felt like a rollercoaster of emotions, this is not a surprising outcome.



Sentiment over time among winners and nominees of Best Original Screenplay provides some insights as to the emotional pacing differences between the two. Nominees' sentiment score remains relatively steady throughout the film, whereas winners demonstrate more emotional variation with broad shifts in sentiment from one section to the next. However, the differences in sample size between the two groups may be confounding the results on the plot.



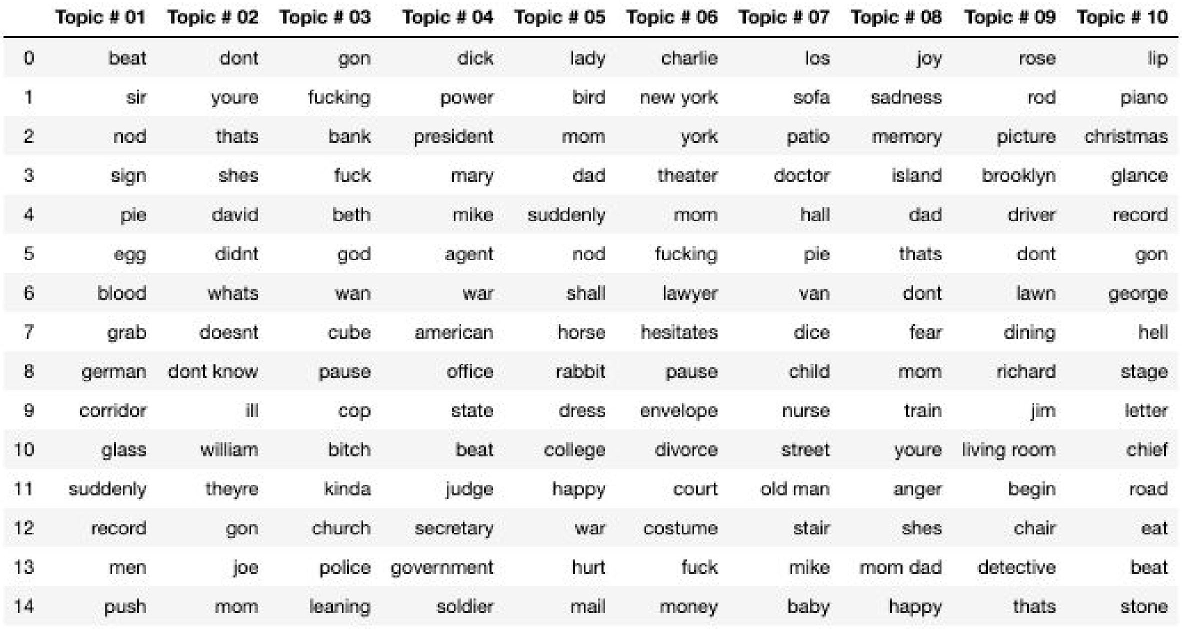
Sentiment over time for each film can be seen below. Though the plot doesn't provide any new insights, nor is it the easiest to interpret, it does demonstrate the individual shifts in sentiment over time for each film. It also highlights the incredible difference in length between *Roma* and *The Shape of Water* compared to the rest of the films.



Topic Modeling

NMF applied to the TF-IDF output from the ***TfidfVectorizer*** with the parameters selected yielded very interpretable topics for each data subset. Using the parameters described in the topic modeling section of the methodology, we were able to obtain distinguishable topics from each of the datasets. Topics were particularly well modeled and interpretable for the dialogue-specific data. It was also found that the topics modeled using the entire dataset, inclusive of the visual cues, were constructed very similarly to the topics modeled on the dialogues only.

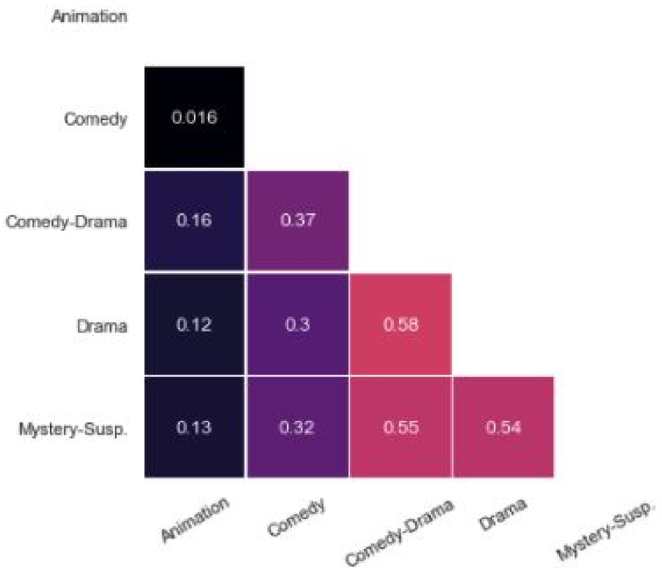
Even without much knowledge about a particular film, you can still determine to which script a topic may be relevant. This is demonstrated in the visual below. You can see that Topic 4, from the dialogue topic modeling, is probably indicative and relates to the film, "*Vice*," based on the words within the topic. Also, similar make-ups of topics can be seen across each dataset, for instance, topics modeled for the winning scripts have similar vocabularies as topics modeled for all of the data, which is also true to topics modeled solely on the dialogue data.



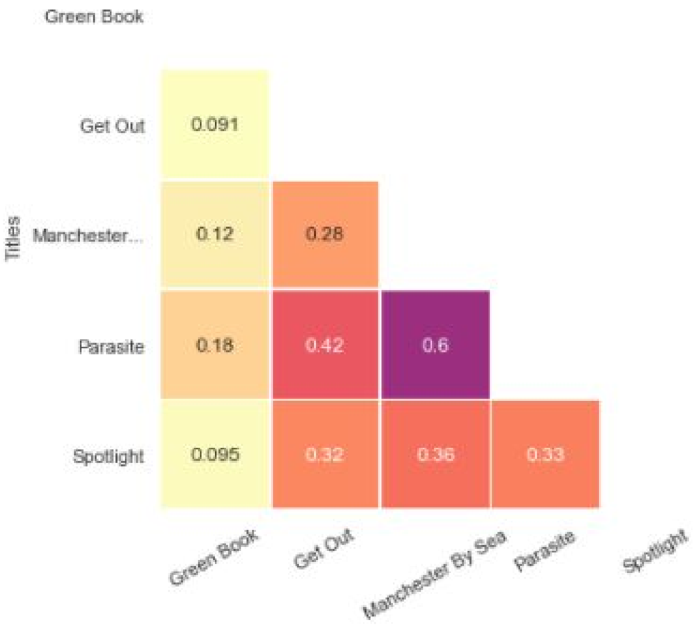
	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05	Topic # 06	Topic # 07	Topic # 08	Topic # 09	Topic # 10
0	beat	dont	gon	dick	lady	charlie	los	joy	rose	lip
1	sir	youre	fucking	power	bird	new york	sofa	sadness	rod	piano
2	nod	thats	bank	president	mom	york	patio	memory	picture	christmas
3	sign	shes	fuck	mary	dad	theater	doctor	island	brooklyn	glance
4	pie	david	beth	mike	suddenly	mom	hall	dad	driver	record
5	egg	didnt	god	agent	nod	fucking	pie	thats	dont	gon
6	blood	whats	wan	war	shall	lawyer	van	dont	lawn	george
7	grab	doesnt	cube	american	horse	hesitates	dice	fear	dining	hell
8	german	dont know	pause	office	rabbit	pause	child	mom	richard	stage
9	corridor	ill	cop	state	dress	envelope	nurse	train	jim	letter
10	glass	william	bitch	beat	college	divorce	street	youre	living room	chief
11	suddenly	theyre	kinda	judge	happy	court	old man	anger	begin	road
12	record	gon	church	secretary	war	costume	stair	shes	chair	eat
13	men	joe	police	government	hurt	fuck	mike	mom dad	detective	beat
14	push	mom	leaning	soldier	mail	money	baby	happy	thats	stone

Cosine Similarity

Based on the cosine similarity, a few genres, including drama, comedy-drama, and mystery suspense have some similarities. This is not surprising as some of these movies may be classified under multiple genres depending on the resources. It is interesting to see that there is no more similarity between the genres of comedy and comedy-drama. The figure below represents the cosine similarity between each of the genres in the dataset.



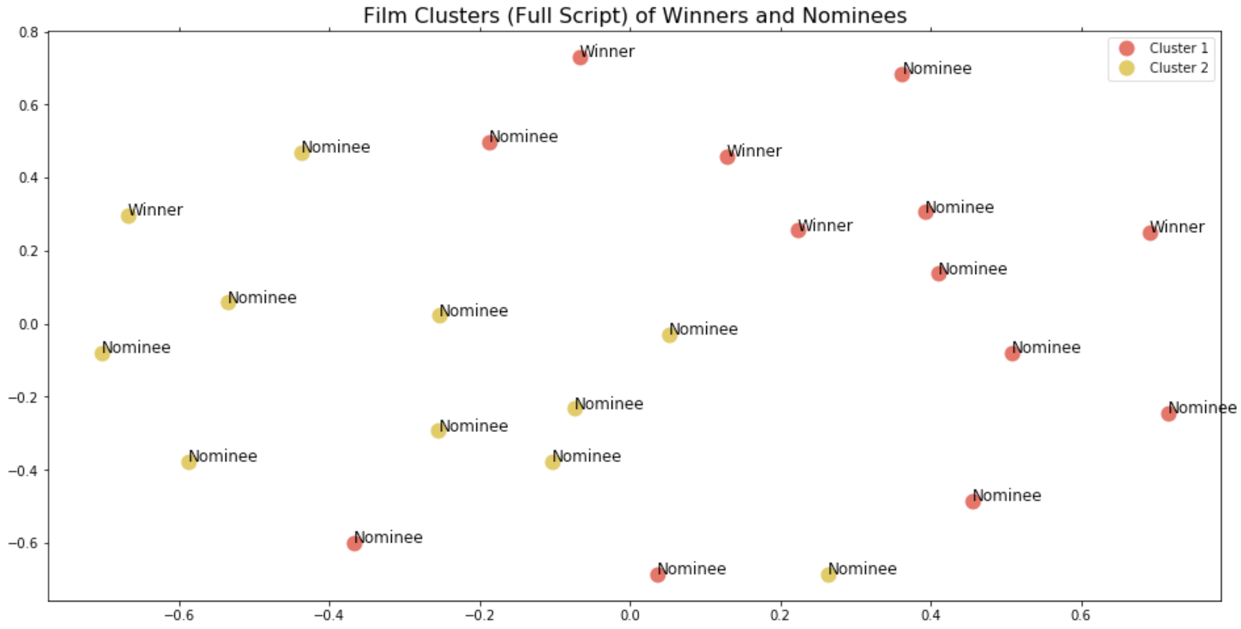
Additionally, from the cosine similarity analysis between the winning movie scripts, we can determine that the scripts indicated as winners in the dataset do not seem to share a significant amount of semantic similarity. However, Manchester by the Sea and Parasite do share the largest measure of cosine similarity seen throughout the dataset, even when incorporating the whole corpus of films. The lack of this relationship is most likely due to the small sample size of winning films.



K-Means

The results of our cluster algorithm resulted in little difference between the TF-IDF of just dialogue and the TF-IDF of the script. Henceforth, the results of these two clustering models will be addressed as one result. As several features fed into this clustering model, to demonstrate the results of our cluster, we had to project our points into a multi-dimensional space using cosine distance and visualize as seen below.

The results of this model did not, contrary to our original hypothesis, result in one group of only nominees and one group of only winners. Instead, the majority of the winners appeared in one cluster mixed in with some nominees and the remaining winner appearing in the other cluster. With higher sample sizes, especially in the winning class, there may be a more distinct separation between winners and nominees, and we plan to explore this in future research.



Future Opportunities

During our study and analysis, we identified a few areas of opportunity for additional research. First, since our subset of winners only consisted of five scripts accounting for only about 16% of the data, a larger sample size could benefit this study. Additionally, the use of Parts-of-Speech tagging could be implemented to separate the dialogue from the visual cues further. We attempted to accomplish this, but we were not satisfied with how it resulted in portions of our data. This task was also very time consuming when trying to address caveats strewn throughout each script and could be improved by grouping expected patterns of parts of speech. Finally, with additional attributes, such as the length of the film, critics reviews vs. audience reviews, or details such as the film's rating, a classification model could be implemented to help scriptwriters determine whether or not their script falls within the classification with similar award-winning films based on their composed text.