What Makes a Movie an Academy Award Winning BEST PICTURE?

Nayan Chavan
June 2020

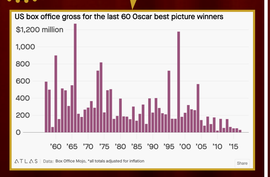
Questions for Exploratory Data Analysis:

1. Are there similarities in themes or plots in the different movies that the Academy chooses as Best Pictures?
2. Are there certain qualities in the bare script that are present in a Best Picture Winner?
3. What prominent archetypes or genders have the highest frequency in the scripts of these movies?
4. Over 50 years, do movies in the same periods have higher levels of similarities?

Did producers try to make their films stand out from the films that had been made the years prior, or did movies tend to converge to similar themes and topics within time periods?

"The **Academy Award for Best Picture** is one of the Academy Awards presented annually by the Academy of Motion Picture Arts and Sciences(AMPAS) since the awards debuted in 1929. This award goes to the producers of the film and is the only category in which every member of the Academy is eligible to submit a nomination and vote on the final ballot. Best Picture is the final award of the night and is considered the most prestigious honor of the ceremony."

Why Choose to Analyze the Academy Award for Best Picture Winners?
According to this **article**, the last blockbuster hit to win the Best Picture Oscar was in 2004 (*The Lord of the Rings: The Return of the King*). Historically, box office success does not necessarily correlate with Best Pictures. Therefore, this is a unique dataset to examine primarily because of the vastness and variability of award-winning movies. It will be exciting to possibly find any common trends, themes, or plotlines that have arisen in these films in the last 50 years.



THE OSCARS

Descriptions of Tools, Methods, and Findings:

Description of the Data Set:
• Data scraping all the movie scripts from www.imsdb.com of the Best Picture Oscar winners from the last 50 years (since 1971)
•I will be extracting the scripts from **this** database of **files** and will manually find any scripts that may be missing from www.scriptslug.com.
• **Here** is a list of the last 50 Best Picture Oscar winners

Web Scraping:
I used this script from **GitHub** that allowed me to scrape the International Movie Script Database's site of ALL its scripts. However, the original contained some bugs, so I forked it and modified it in order to download all the scripts available on imsdb.com. You can see my fork **here.** I have compiled all the 50 scripts' text files **here.** All 1158 can be found **here.**

Convert into CSV:
Using an adaptation of the "Make_csv_tutorial.ipynb" notebook, I converted all the text files of the 50 scripts I had compiled into one CSV. The first column contained the title of the file/movie and the second column contained each movie script's respective content. The CSV that was created can be found **here.** Furthermore, I created 5 more CSV over ten-year intervals over the 50 years.These 5 CSVs along with how the movies were split up can be found **here.**

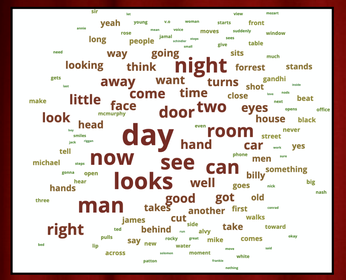
OverviewDocs:
With the CSV of my dataset I had produced, I then fed it into OverviewDocs.com to get a general idea of the content of the various scripts. I was aware that because the scripts varied in length, the word cloud might be skewed. Furthermore, the chart of entities turned out to be much more useful to my research, since it was able to filter out stopwords, the most common words in the English language, and numbers. This left me with a table of the most frequent character names that appeared in various scripts.

Entity	count	docs
forrest	1,171	1
michael	992	10
gandhi	965	2
alvy	882	1
mcmurphy	810	1
nash	763	2
jamal	757	1
patton	719	1
mozart	669	4
annie	654	5
conrad	649	1
frankie	636	3
solomon	634	4
riggan	629	1
schindler	620	1
karen	599	4
clarice	540	1

colin	512	4
salieri	512	1
strickland	510	1
bertie	503	1
shirley	489	3
hoke	484	1
munny	480	1
robby	473	1
sanborn	473	1
elisa	467	1
doyle	463	4
george	453	16
chris	451	3
maggie	442	1
lionel	428	1
lester	423	2
dunbar	419	1
giles	416	1

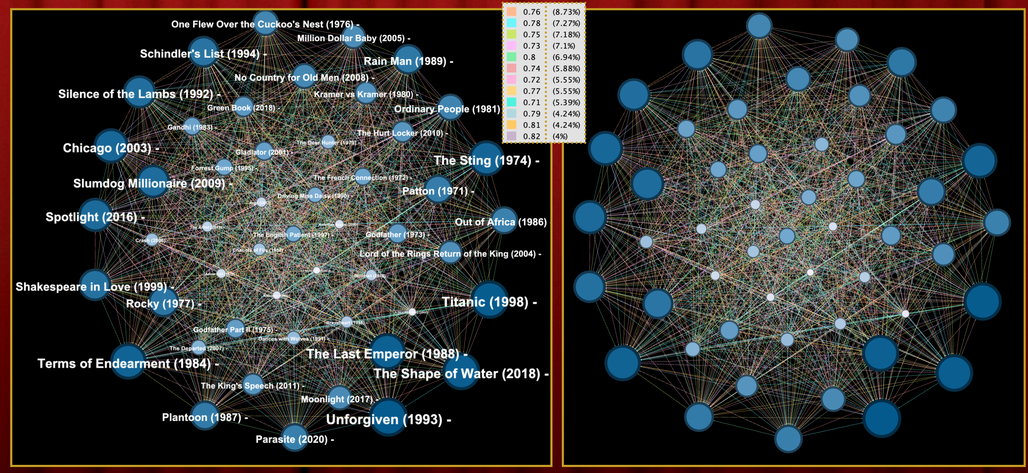
Moreover, this table indicated the places and countries that were mentioned across most scripts. These included "England" (16), "America"(27), "Africa" (8), "London" (15), "France" (14), "Brooklyn" (10), "Hollywood" (13), and "California" (15). **This indicates that many of these movies were likely to have predominantly Eurocentric views both geographically and ideologically.**

The table of entities indicated the most frequent names are male and many of these male names are common enough to appear across multiple scripts.**9 of the top 10 most frequent names are the names of male characters.** The #10 name is "Annie", spanning 5 scripts, and most likely to only be in the top 10 because of the 1977 winner "Annie Hall." Moreover, "Michael" and "George" seem to be the most common and frequented names for male characters, as they respectively span 10 and 16 scripts. Other common names include"Johnny", "William", "Charlie", "David", and "Robert", each appearing across more than 10 scripts. In fact, the phrase "a man" appeared across all 50 scripts.



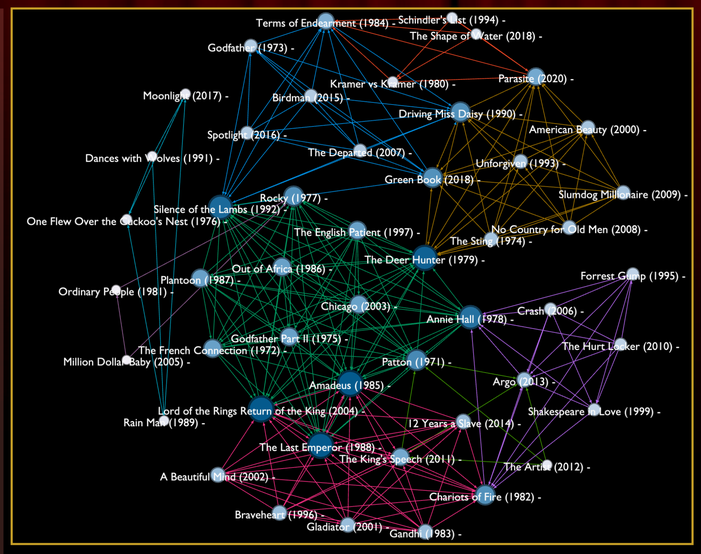
As predicted, the WordCloud did not have much use for my research. However, "man" is one of the more prominent words in the cloud (after cleaning it up). **This correlates exactly with the frequency of the type of names in the scripts because they are all predominantly traditional masculine names.**

Doc2Vec:
Doc2Vec vectors represent the theme or overall topic of a document; in this case, the documents would be each of the 50 scripts. Using the example Doc2Vec Notebook ("Doc2Vec (Kenan)"), **I made my changes and made sure to filter English stopwords along with other words that I found to be the most common in movie scripts, such as "cont", "continue", "exit", and "ext".** My notebook allowed me to create a CSV file that allowed me to set a similarity threshold of .846. The CSV produced from this threshold was relavtively small for Gephi visualiztion purposed. **Therefore, I increased the similarity threshold to .5 to produce more edges for the purpose of analysis.** For reference, if the similarity score was 1.0, that would mean that the scripts were exactly the same. The second CSV I produced was much larger with 1226 edges, and can be found **here.**



This directed graph visualization was created using Gephi and the **Force Atlas 2** layout. The first visualization includes the node labels, which are the movie titles, and the second visualization does not. There are 50 nodes since the dataset contains 50 movies. The amount of edges is based upon how many movies, set as targets and sources, exceeded or met the .5 similarity score threshold. The colors of the edges are based upon different levels of similarity ranging from .5 to .97. Unfortunately, this was the only Gephi tool that was able to provide any analysis with the Doc2Vec data. The size of nodes is dependents upon how many incoming edges the movie represented by the node had. For example, The Shape of Water (2018) and Titantic (1998) are two of the most prominent nodes in the visual. This indicates that several of the 49 other movies had several similarities with these particular films. Moreover, The Godfather (1973) and The Godfather Part II (1975) are both smaller nodes on the graph because not many other movies in the dataset had overlapping similarities with these two films. However, when looking at the dataset both The Godfather movies have a **.97 similarity** with each other! **This is most likely because they are in subsequent movies in the same movie franchise. I wasn't able to get a clustering visual to analyze this data further, but it was interesting to see that there was no obvious correlation between when the movie was released and its relevance.**

Topic Modeling Network Analysis:
Using an adaptation of the code from Kenan Jiang's TM2Net_edge_list notebook, **I produced a CSV that showed me the source, its target, the source weight, the target weight, and the topic it was associated with.** In this case, I chose 10 topics to be found. **Furthermore, I set the topic weight for the edge to be appended to the dataset to be .25.** I found this to be the best value because I was able to get a large enough spreadsheet that did not skew the visualization by making it seem like a topic was the most prevalent when it was not. The CSV produced from my notebook can be found **here.** This notebook analyzed the most prevalent topics within the 50 movie scripts, and within which movies these 10 topics showed up the most.



4.0	"GEORGE" + "nurse" + "rose" + "PISTOL" + "german" + "dressing" + "SOLDIER"	(36.99%)
9.0	"DAVID" + "airport" + "GEORGE" + "british" + "director" + "movie" + "royal"	(18.29%)
5.0	"JAMES" + "rose" + "stern" + "boat" + "ship" + "SOLDIER" + "SERGEANT"	(14.63%)
2.0	"MIKE" + "JAMES" + "dressing" + "judge" + "PISTOL" + "fucking" + "lawyer"	(11.38%)
1.0	"LIEUTENANT" + "rose" + "SOLDIER" + "master" + "boat" + "SERGEANT" + "CAPTAIN"	(8.54%)
6.0	"dances" + "bird" + "kicking" + "LIEUTENANT" + "wagon" + "DANGER" + "camp"	(4.07%)
7.0	"WILLIAM" + "SOLDIERS" + "york" + "horses" + "SOLDIER" + "stern" + "GUARDS"	(2.44%)
8.0	"master" + "british" + "christmas" + "ford" + "judge" + "queen" + "rose"	(2.44%)
10.0	"patient" + "plane" + "tent" + "sand" + "british" + "JAMES" + "RIFLE"	(1.22%)

This directed graph visualization was created using Gephi and the **Frutcherman-Reingold** layout. The colors are partitioned by the 10 topics my topic modeling notebook found to be the to be most prevalent. Furthermore, the node size is dependent upon the average of the source and target weight of the outgoing and incoming edges from each of the 50 movie nodes. Just by looking at the graph, the topic with presence across the most movies is Topic #4. This topic is described as ["GEORGE"+ "nurse" + "rose" + "PISTOL" + "german" + "dressing" + "SOLDIER"],which is a clear indication that most of the movies **have strong male leads, many ties to other Western values such as war, and reference to Eurocentric geographic locations.**

After examining the ten topic breakdown with ten words each (both of these attributes can be toggled within the topic modeling notebook), I was able to identify 3 common themes: Men, War, Colonialism/Eurocentrism. Several traditionally Angelic names appeared across many of the produced topics indicated that most of the 50 movies, if not all, had to have a very strong male figure throughout the script. The reason the "Angelic" part is important is because across all fifty scripts it is safe to assume that there was a plethora of male characters and names. However, only names like "George","James", and "William" make it to the topics. Some male names appear more than once. Moreover, several terms in the realm of "pistol", "soldier", and "rifle" appear across most of the topics. I made the assumption that these terms likely refer to war or violence. This leads me to believe that wartime movies, action movies, and/or movies with prevalent violence are weapons are common within the fifty Best Picture database. Lastly, the terms indicating movie themes centered around colonialism and Eurocentrism are strung through the topics, as well. Although not as obvious as the two previous themes, but it is still noticeable. Words like "german", "british", and "queen" are seen more than once. This was also hinted at through my OverviewDocs analysis, but this further confirms many of these movies will have a strong Western ideology within the script's dialogue, location, and plot. These terms of interest have also been indicated with varying fonts in the percentage breakdown above.

What Makes a Movie an Academy Award Winning BEST PICTURE?

Descriptions of Tools, Methods, and Findings (continued):

Nayan Chavan
June 2020

Questions for Exploratory Data Analysis:

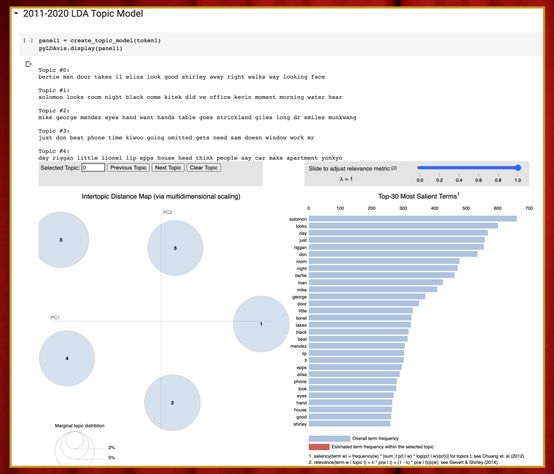
1. Are there similarities in themes or plots in the different movies that the Academy chooses as Best Pictures?
2. Are there certain qualities in the bare script that are present in a Best Picture Winner?
3. What prominent archetypes or genders have the highest frequency in the scripts of these movies?
4. Over 50 years, do movies in the same periods have higher levels of similarities?

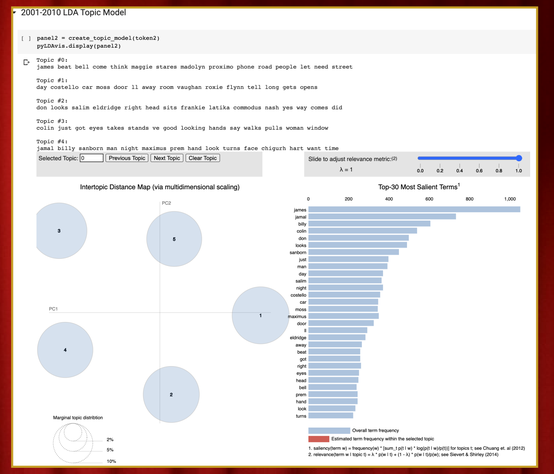
Did producers try to make their films stand out from the films that had been made the years prior, or did movies tend to converge to similar themes and topics within time periods?

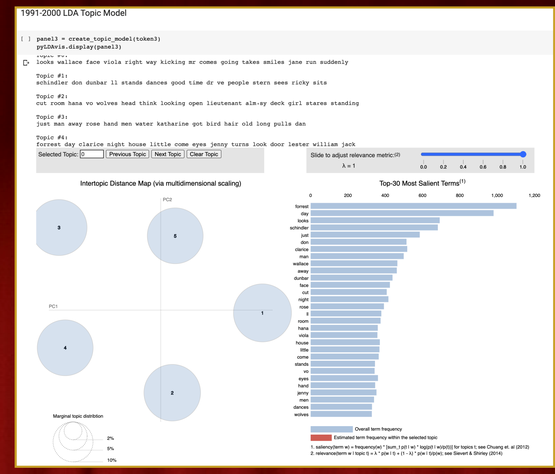
"The **Academy Award for Best Picture** is one of the Academy Awards presented annually by the Academy of Motion Picture Arts and Sciences(AMPAS) since the awards debuted in 1929. This award goes to the producers of the film and is the only category in which every member of the Academy is eligible to submit a nomination and vote on the final ballot. Best Picture is the final award of the night and is considered the most prestigious honor of the ceremony."

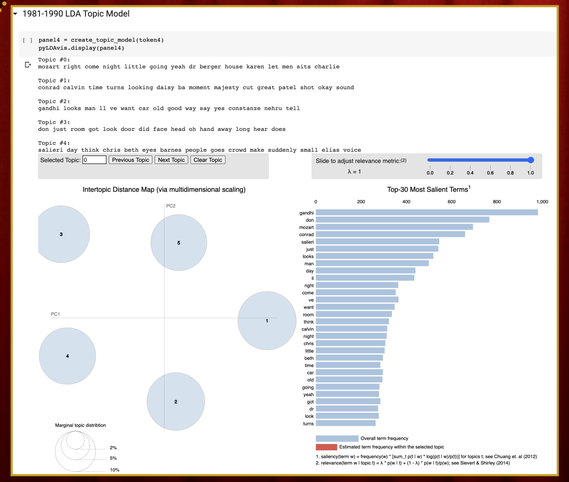
Why Choose to Analyze the Academy Award for Best Picture Winners? According to this **article**, the last blockbuster hit to win the Best Picture Oscar was in 2004 (*The Lord of the Rings: The Return of the King*). Historically, box office success does not necessarily correlate with Best Pictures. Therefore, this is a unique dataset to examine primarily because of the vastness and variability of award-winning movies. It will be exciting to possibly find any common trends, themes, or plotlines that have arisen in these films in the last 50 years.

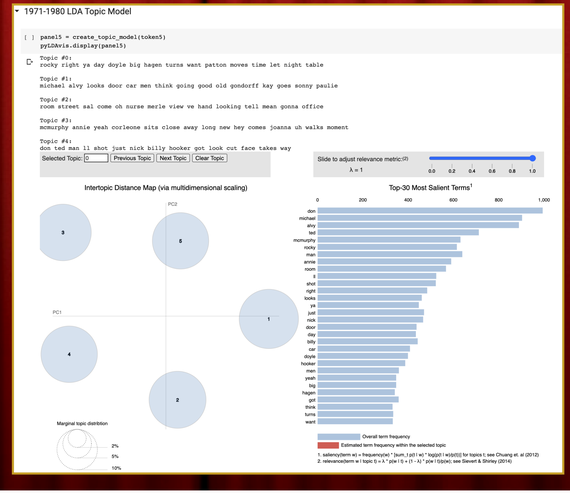
Latent Dirichlet Allocation: With an adaptation of the BasicTopicModel.ipynb, I will use the pyLDAvis's LDA visualization function to further investigate the shared topics amongst the movies. LDA will be able to provide me with a rough idea of what the various topics are that caused the interconnectedness we saw the Topic Model network visualization. **It will be worthwhile to use the LDA model because it uses a bag-of-words (meaning the order of the words do not matter) representation of the documents.** LDA topic modeling will be able to match a particular set of words to a topic. I could then compare the models produced from Doc2Vec versus LDA to find any blindspots with my analysis or any discrepancies in my data.











The five LDA Topic Models are of movies over five ten-year intervals to analyze any correlation with the time period in which a movie was released and the topics. A closer view of these data models can be found **here.** There are three key aspects of the LDA visualization to note: saliency, relevance, and bubble size. **Saliency measures how much a particular term can tell you about the topic. Relevance is the "weighted average" of the probability of a word in a particular topic normalized by the probability of the topic. The size of each bubble indicates how important a topic is concerning the dataset.** In the notebook, this model is interaction and users can hover over each topic bubble to see the most salient terms, along with the top 30 most salient terms overall. Furthermore, bubble clustering means similar topics or themes within the movie dataset. **Due to the bag of words approach, the LDA model was not as effective as I would have hoped because it clumped each set of ten movies within each time-period together to produce five topics, with no regard to sentences or context.** Because there are multiple possibly contextually unrelated movie script documents being meshed together, it is difficult to see if each script had any topics that overlapped with its neighboring Best Picture winner's scripts. **I believe I would have been able to do this if the LDA model analyzed each of the ten movies within each time-period as ten individual bags of words** (totaling to 50 bags of words), rather than one bag of words per time-period. Based on my previous three methods, I believe that there should be clustering amongst the bubbles, at least in the 1971-2020 model (which can be found **here**).

Right off the bat, the most salient term for each of the five time-periods appears to be a traditionally masculine name. Indicating and further affirming that most of these movies were almost guaranteed to have prominent male characters throughout the storyline. After looking at the 5 topics produced from the various time-periods, I was able to identify some trends. The 1971-1980 era closely fits the trends that have already been found, many prevalent male names and characters along with several American terms. 1981-1990 shifts into including more European such as "Mozart" and "Salieri", and "Gandhi" appears because of the Mahatma Gandhi biopic "Gandhi" winning in 1983. Perhaps this era is when movies moved out of the American setting and instead to other countries and cultures. This seems likely because of movies like Amadeus (1985), set in Vienna Austria, Gandhi (1983), set in India, and Out of Africa (1986), set in various parts of Europe and Africa. The following three decades relatively fit the original trends that had already been discovered, but we see the presence of traditional female names increase between 1991-2010. **This is particularly interesting because, in the latest 2011-2020 decade, there is only one obvious female character name that shows up in the topic breakdown.** In conclusion, there may have not been a consistent trend within each decade of movies. However, as I stated earlier, this could be because of the LDA methodology and how it approaches each group of documents. Furthermore, I believe there was a pattern emerging from the Eurocentric default between the years of 1981 and 1990. **Moreover, my LDA analysis did not seem to make anything else too obvious about each set of movies other than many movies tend to converge to male-dominant scripts with Western themes and ideas.**

ALL OF THE VISUALIZATIONS, DATASETS, AND NOTEBOOKS REFERENCED IN THIS POSTER CAN BE FOUND HERE! THE NOTEBOOK I HAVE INCLUDED ALSO LINKS TO THE VARIOUS NOTEBOOKS FROM WHERE I ADOPTED SOURCE CODE. MY WORKS CITED PAGE CAN BE FOUND HERE!

WHAT IS THE SECRET FORMULA TO THE BEST PICTURE?

Interpretation of Results:

- The OverviewDocs analysis was able to give me an idea of what my dataset would indicate moving into my Doc2Vec, Topic Modeling, and LDA methodology. OverviewDocs provided a rather surface-level answer to my EDA question regarding the most common genders that show up across Best Pictures. The entities feature of OverviewDocs proved to be the most useful because it counted the frequency of words, names, and locations across the 50 movie scripts and counted how many scripts these terms appeared. **The top half of the list was overwhelmingly names. The names that were the most common and prominent spanning the 50 scripts were traditional Angelican masculine names, indicating that the leading characters or narrators of many of these movies are likely to be Caucasian and male. When examining geographic locations, many of the most prevalent locations were either American or European, further perpetuating a focus on a Eurocentric lifestyle.**
- Moving on to the Doc2Vec analysis, CSV that was produced from this part of the notebook included targets, sources, and similarities (where the similarity threshold was set to .5). The CSV had approximately 1226 edges spanning across the node representation of the 50 movies. Unfortunately, Gephi was unable to provide me with any information past the Force Atlas 2 visualization. I am not entirely sure why, but it might be because the size of the dataset was too small to find any clustering or trends. However, the node size was dependent upon the weighted in-degree, in other words, this is a count of how many movies had similarity vectors targetting the node. It was interesting to see that most of the movies are visible on the directed graph which means that many of these movies have intertwining themes and plots. **"Unforgiven" (1993), "Titanic" (1998), and "The Shape of Water" (2018) were the top three movies with the most incoming edges. "12 Years a Slave" (2014), "A Beautiful Mind" (2002), and "Amadeus" (1985) were the three movies with the least number of incoming edges.**
- The CSV created from the Topic Modeling section of my notebook and the subsequent Gephi visual were able to provide the most information about these particular films. **I already provided an analysis of my findings earlier in this poster, but this analysis proved that many of these movies were likely to be action or wartime movies. This is quite plausible because historically, male leads are the most common in action movies or movies with any weapons and/or violence. Of the ten topics, none of them were outliers in the sense that all of the topics included words that alluded to at least one of the three themes we have seen to be the most prevalent in the fifty movie scripts: men, Western ideology, and violence.** The most prominent character seems to be male in almost all, if not all, of the scripts.
- The LDA Topic model was unable to provide me with the information I had hope for, as I stated earlier. I cannot necessarily answer whether or not producers tried to make their movies "stand out" amongst the Best Pictures within their time eras, but **it seems that most of the decades fell into the Eurocentric mold. One of the decades, 1981-1990,** seemed to disassociate from this default but it is unclear whether this was by chance or implicitly intentional. Unfortunately, besides the one outlier, it does not seem that I can make an educated conclusion whether or not there was a correlation with the decade in which a movie won Best Picture and the plot lines or themes of the particular movie.
- **Ultimately, I was able to answer most of my questions for EDA besides the prevalence of topics in regard to the year released. There was some oversight on my part because LDA was not the best suited for the way the dataset was split up, and I should have instead looked at each movie individually in order to find overlapping topics within movies and decades.** However, this was out of scope because I did not have enough time to conduct a more specified and individualized analysis. On the other hand, Overview Docs, Doc2Vec, and the Topic Modeling notebook were able to provide me with very interesting conclusions. These conclusions have implications that lead me to believe that the Academy is most likely to choose movies with Caucasian men in power possibly paired with a less prominent female counterpart, filled with action scenes with a high chance of violence, and include references to the Western world whether that be geographically or ideologically. There are obviously winning outliers to this trend which is important to recognize and can be seen amongst the smaller nodes in the Doc2Vec model. **Most recently, the 2020 Best Picture winner, Korean film "Parasite" directed by Bong Joon-ho is a very obvious and strong outlier.**