# Script Text Analyses of Movie Directed by Christopher Nolan

Haley Jiang [yufeijia@andrew.cmu.edu](mailto:yufeijia@andrew.cmu.edu) Handi Yang [handiy@andrew.cmu.edu](mailto:handiy@andrew.cmu.edu)

Abstract

When searching for movies, people usually search with movie types, so it is interesting to know if there exists some special pattern for movie scripts that can help us have a quick guess about what kind of movie it is. We investigated 11 movies directed by Christopher Nolan and divided them into four categories. Frequency analysis, keyness value analysis, collocation analysis, and PCA are used to find special patterns for these four movie types. The result shows that thriller movies tend to have the word "cutter" with high frequency and keyness value. Action movies and thriller movies both have positive relationships with rigorous instruction and negative associations with descriptive senses.

## Introduction

Christopher Nolan is one of the leading filmmakers of the 21st century. He has been nominated for five Academy Awards, five British Academy Film Awards, and six Golden Globe Awards because of his well-produced, imagination-stirring, and reality-reflecting films (2022). He started his great journey by making *Following*. With that experience, he was nominated for the Academy Award for Best Original Screenplay with his second film *Memento*. He then directed *Insomnia* to attempt studio filmmaking, and further achieved significant commercial success in Oscar nominations with *The Dark Knight Trilogy, The Prestige,* and *Inception*. *Interstellar, Dunkirk* and Tenet debuted in subsequent years. His tremendous success in film production makes it interesting to explore the characteristics of his movies. This inspires our questions:

1. What are the most frequent words and keywords for each movie genre?

2. What are the collocations of the most frequent words?

3. Are there any commonalities between different movie genres?

This study is based on the 11 movies Nolan directed and seeks to address the questions.

## Data

Our data is based on the scripts of the 11 movies that Nolan directed so far that were collected from The Internet Movie Script Database. We named the data files in the format "genre\_moviename-year", where genres were identified according to the movie introduction in Wikipedia. We then compiled and read the separate movie script files through R. We then preprocessed and tokenized the data using the preprocess\_text function in the quanteda package (Benoit et al., 2018) to separate the contractions and hyphenated words, remove punctuation, make all tokens to lowercase, replace accented characters with un-accented characters, and remove numbers. To assign metadata to our corpus, we extracted movie genre and released year from the data file name, Table 1 summarizes the number of tokens, genres, and released year of corresponding movie scripts. There are in total 4 genres. 5 movies are categorized as action; 4 movies are identified as thriller movies, and 1 movie is identified as scientific and war respectively. The movie release year ranges from 1998 to 2020.

Table

Description automatically generated

## Methods

### Frequency and keyness analyses

We used the script dataset after removing those proper nouns in frequency and keyness analysis. In each script at the beginning of each conversation, it will show the character name and for each script scenario, it will include the location of that scenario, if we do not remove proper nouns, the high-frequency words and high keyness words will all be character names.

We analyze the overall high-frequency word for the whole corpus after removing proper nouns, and also high-frequency words for each genre. We use the textstat\_frequency() function from the cmu\_textstat package which produces counts and document frequencies summaries of the features in the document-feature matrix. We use the dfm\_subset() function in the quanteda package to get the DFM of each genre and rank by frequency in descending order.

We analyze the keyness value of words in different genres. By using the textstat\_keyness() function in the cmu\_textstat package, we can calculate the log-likelihood of each word in the document-feature matrix and rank by descending order to see which word has a high keyness value in different genres. Then we use the kbl() function in the kableExtra package to generate the graph. We also calculate the keyness value of action compared with thriller movies using the keyness\_table() function in the quandeta package to investigate if there exist some special words for these two genres. Since we only have one script for war and science, we did not include these two genres in the keyness value comparison.

### Collocations

After getting the high-frequency words of different genres, we choose the word that has a high rank in both of these four genres. By using the collocates\_by\_MI() function in the quanteda package, we calculate collocational associations by Mutual Information. Because the MI value is sensitive to rare words, we make thresholds for both token frequency and MI score and left words with token frequency larger than 5 and MI score larger than 5. After that, we use the kbl() function in the kableExtra package to generate the graph.

### Principal Component Analysis

We applied the tokens\_lookup() function from the quanteda package to count and categorize features with our raw tokens based on DocuScope which was developed at CMU by David Kaufer and Suguru Ishizaki. Since DocuScope is not categorizing all of our tokens, we normalized the counts with respect to the total count from the original tokens object.

We further performed principal components analysis (PCA) on the normalized counts with the prcomp() function with variables shifted to be zero-centered and scaled to have unit variance before the analysis takes place. A Scree plot was made to visualize the amount of variance explained by each principal component. By convention, we looked into principal components that have the variance explained above 1/p, where p is the number of principal components.

To visualize the PCA biplot with variable features, we applied the fviz\_pca\_biplot() function in the factoextra package (Kassambara and Mundt, 2020) with the first principal component on the x-axis, the second principal component on the y-axis, the DocuScope feature variables plotted as arrows and different movie genres in colors.

## Result

### Frequency and Keyness analysis

Table 1 shows high-frequency words for all of these 11 movies directed by Norlan and 4 different genres: war, action movie, thriller movie, and science fiction movies. We can see that word" the" is at the top rank for all these four genres below it, proposition words like "of", "to", and pronoun words like " he", "she", and "I", all have a high frequency in these four genres. There exist some special nouns in the top 20 frequency table that differ for these four genres. However, after checking with the cast of characters, we find

that they

|  |
| --- |
| Table1: top 20 high-frequency words in the whole corpus and four different genres |

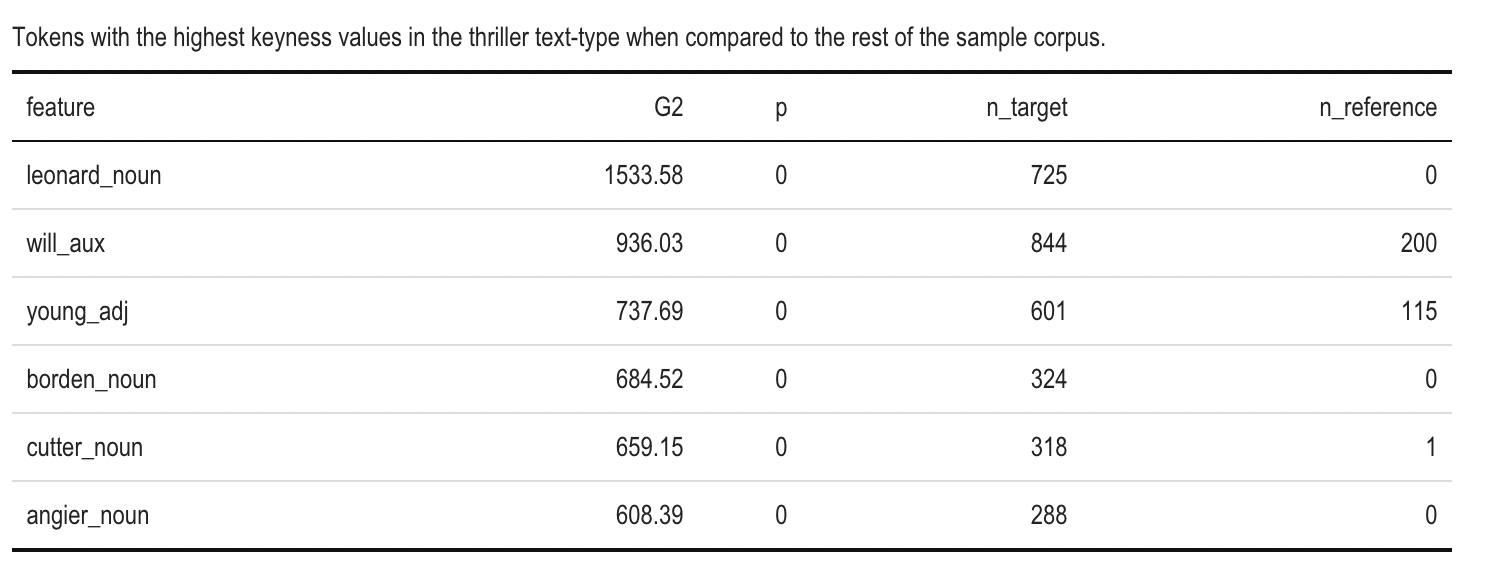
are characters’ names in each movie so they can not be counted as special properties for us to distinguish different genres. Thus, we can not gain enough information from frequency tables to see if there exists any special pattern for these movie genres.

Since we did not see any special pattern in frequency analysis, we implement keyness value analysis which calculates the log-likelihood of each word, which is G2 in table 2, in different genres and ranks it by descending to filter out words that are more likely to appear in that genre.

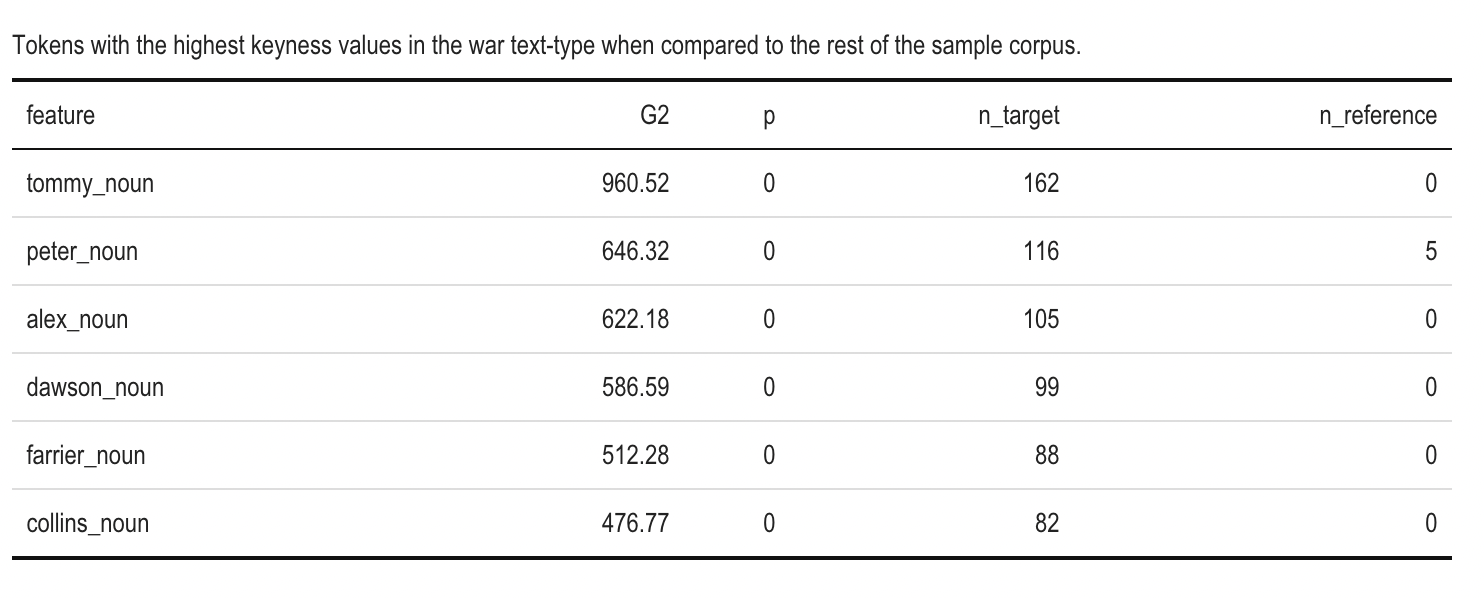
From the table2, we can see that for science fiction movies, words with high log-likelihood value are nouns such as character names. The two main characters in Interstellar are Morph and cooper.

|  |
| --- |
| Table2: word with high keyness value for science fiction movie |

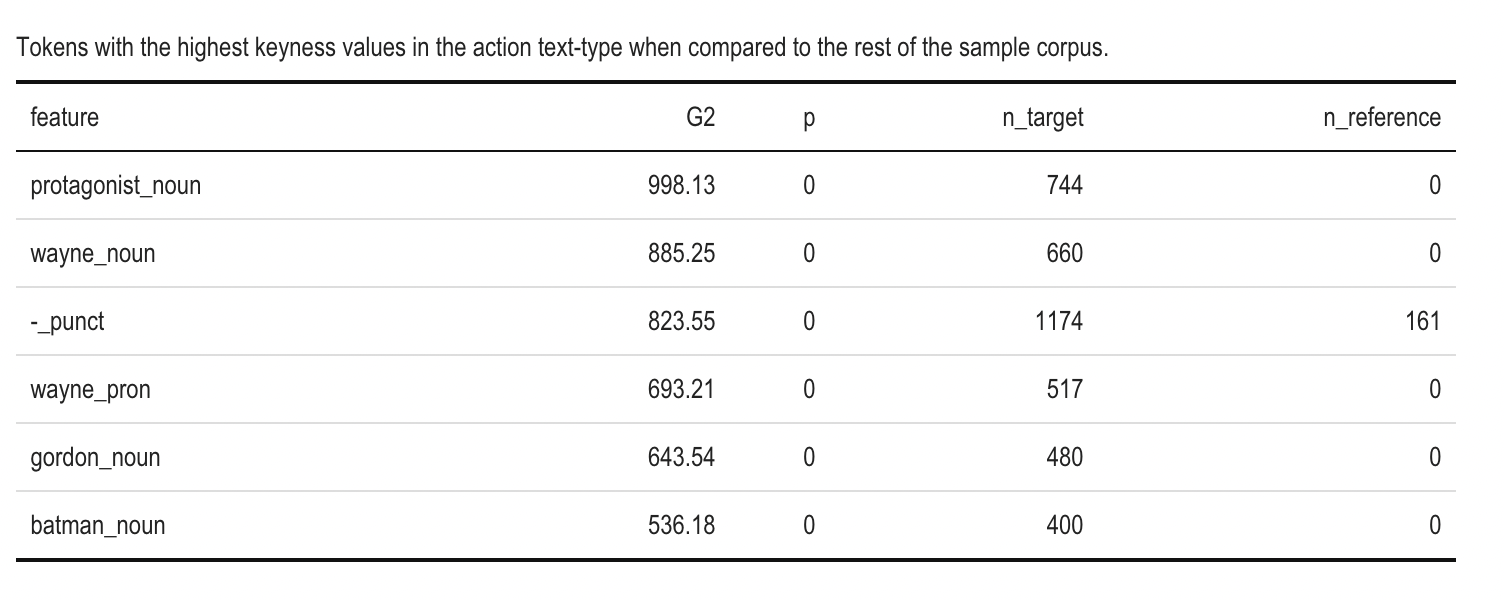
From the table3, we can see that besides those nouns that represent the main characters' names, there also exist nouns like "cutter" that are interesting to analyze. In thriller movies, there might be crime scenes that usually have "cutters" used as criminal tools. In this case, movies containing "cutter" as a high keyness value word are more likely to be thriller movies.

Table3: word with high keyness value for thriller movie

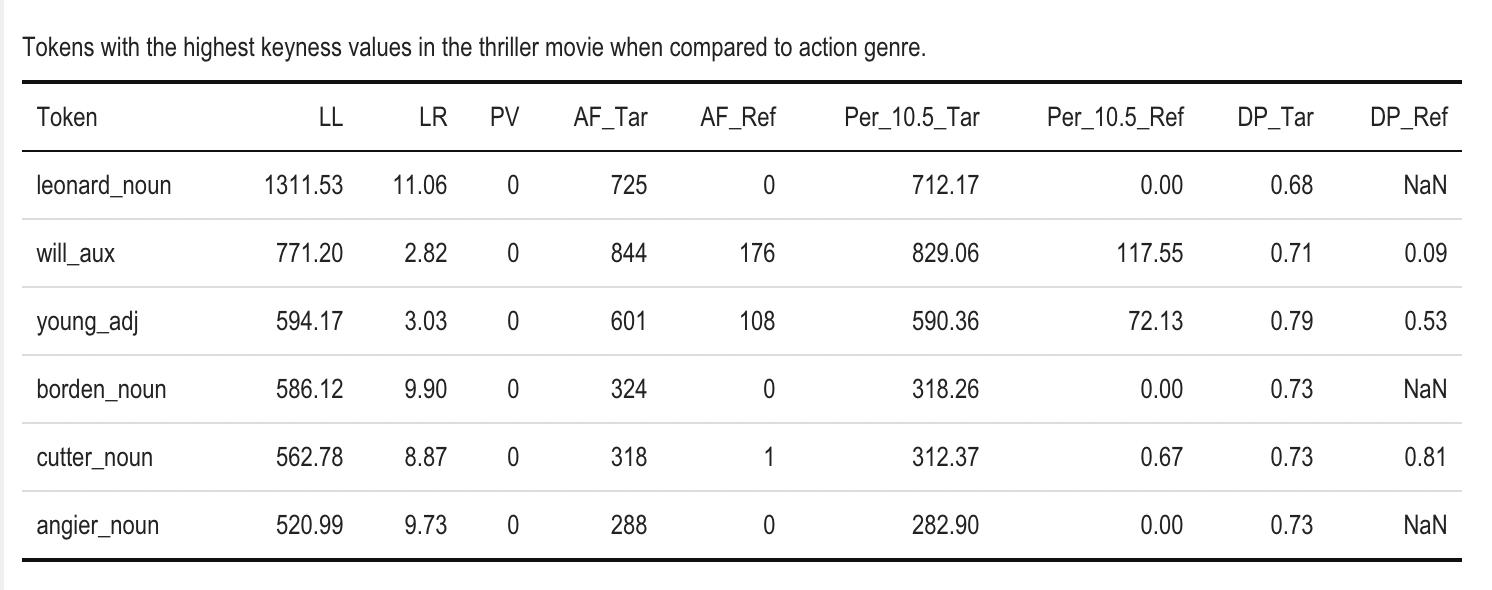
From the table4, we do not find any special pattern since words with high keyness values are all main characters' names. That might be because we only have one movie characterized as war movie that Norlan directed.

Table4: word with high keyness value for war movie

From the table5, we also did not find any special pattern, and words with high keyness values are all characters' names and most from Batman. We have a total of five movies categorized as action movies and three of them are Batman series movies. That might be the reason why these names are all from Batman movies.

Table5: word with high keyness value for action movie

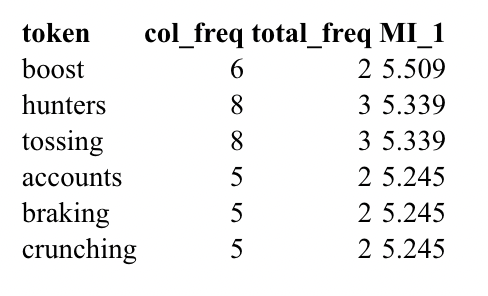
We also compare the keyness value between action movies and thriller movies since we have more examples of these two movie types. From the table6, we can see that compared with action movies, thriller movies have words like "borden", "cutter", and "young" that more often appear in thriller movies.

Table6: a word with high keyness value in thriller movie compare with action movie

### Collocations

From the frequency table shown in table1, we find that word "the" ranks at the top for all these four movie genres. Therefore, we aimed to find the collocation of the word "the" in different genres to investigate if there exist some special words that usually come with "the" for different genres.

As shown in Table 7, we filtered words with a frequency lower than 5 and MI values higher than 5. This means that these words usually appear with the word "the", and have relatively high word frequency. We can see that "hunters", and "tossing" have the highest frequency and relatively higher MI value.

Table7: word with high MI value for action movie

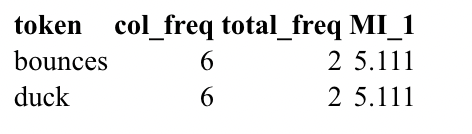
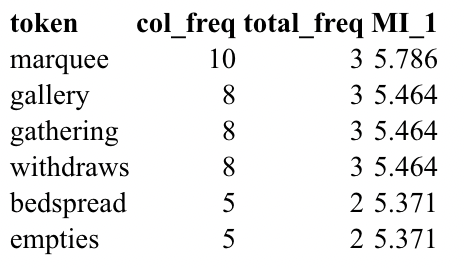
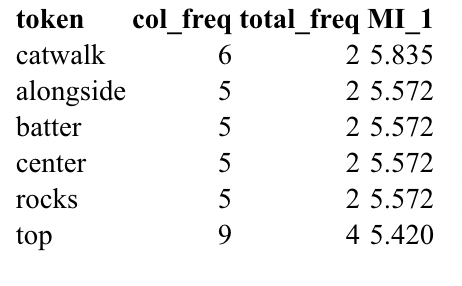
As shown in Table 8, in the war movie corpus, the word "bounces", and "duck" have MI values larger than 5 and frequency larger than 5.

Table8: word with high MI value for war movie

As shown in table9, even though the word "marquee" has the highest MI value and frequency, we will not consider it because it does not exist in actors' lines but as the inductive word in the play. The table shows that the words "gallery", "gathering", "withdraws" have high MI value and frequency in the thriller movie corpus.

Table9: word with high MI value for thriller movie

As shown in Table 10, "catwalk", "alongside", "batter", and "center" usually come with the word "the" with a high MI value and frequency value. But we only have one movie categorized as science fiction, so these words might only be unique for this movie but not represent the science fiction type.

Table10: word with high MI value for science fiction movie

### Principal Component Analysis

As shown in figure 1, the first four principal components explained variances above the average percentage among 11 principal components. The two principal components captured major features of the 11 movies, which together explained 54.8 % of the variances. We will mainly use these two principal components for our plotting and analysis.

Chart, histogram

Description automatically generated

Figure 1: variance explained by 11 principal components

According to figure 2, action movies have most of the data around the upper right corner and less data in the upper left corner. On the other hand, thriller movies have data concentrated in the lower right corner and less data in the lower left. This suggests that action and thriller movie scripts are positively associated with features of *reasoning*, *facilitate, metadiscourseinteractive, confidencehedged, informationstates*, which, in common, conveys relatively rigorous instruction that enables or directs one through specific tasks and actions. Also, action and thriller movie scripts are negatively associated with features of description which were interpreted as the language that evokes sights, sounds, smells, touches, and tastes, as well as scenes and objects.

Chart, diagram, radar chart

Description automatically generated

Figure2: PCA biplot of the first and the second principal components

## Discussion

### Recap

In this paper, we use 11 movies directed by Christopher Nolan and divide them into four categories: action movies, thriller movies, science fiction, and war movies. We have 5 movies categorized as action, 4 movies categorized as thriller movies, 1 movie as science fiction, and 1 movie as a war movie.

We first implement frequency and keyness analysis to different movie genres, the result shows that the word “the” rank at the top frequency for all these four movie genres, and we use the word “the” to do the collocation analysis and filter out words that usually appear with “the” with MI value and frequency value higher than 5. We also analyze words with high frequency for these four movie genres.

We performed PCA on a normalized count of DocuScope feature variables. By observing the relationship of movie genres and feature variables across the first two principal components, we observed that action movies and thriller movies tend to have an association with the same set of features, which implies commonalities between action and thriller movies.

### Takeaways

Combining the results above, we find that for thriller movies, the word "cutter" not only has a high frequency in the thriller genre but also has a high keyness value compared with action movies. That implies the word "cutter" might be a special word for thriller movies. If the word "cutter" ranked high frequency and has a high keyness value, it is more likely to be a thriller movie since this word usually appears in crime scenes. We also find that thriller movies tend to have words "gallery", "gathering" and "withdraws" appear closely with "the". For action movies, since Biber annotation did not delete all character names, we did not get any special pattern for high-frequency words and words with high keyness value. But we find that there exist words like "boost", "hunter", and "tossing" that might appear with the word "the" in action movies.

Our PCA suggests that action movies and thriller movies both have positive relationships with rigorous instruction and negative associations with descriptive senses. This makes sense since action movies and thriller movies involve conversations of detailed action plans instead of describing one's physical sensations.

### Limitations and Prospect

Though we applied udpipe to tag our tokens and deleted character names that were identified as proper nouns. Character names still exist in the text since some of them were identified as nouns, which still makes them significant in the keyness table.

The major problem with PCA is that there is not enough data in each movie genre. For Sci-fi and war genres, only one movie is contained in each of the genres, even for action and thriller movies that contain 5 and 4 movie scripts seem less. Also, we concluded that there are commonalities between action and sci-fi movie genres. However, it is worth noting that a movie could be assigned to multiple genres. Here we assigned movies to a genre that has more tendency towards that genre. It is also possible that the movie contains elements of sci-fi, which might be the reason they share common features. Overall, more movie scripts are required to generalize our conclusion to different movie genres. We, therefore, might consider analyzing the similarities and disparities between movie genres with a larger data set which does not limit to the movies directed by Nolan.

If we have more time, we will try to delete character names by hand to approach more accurate hypothesis testing results on meaningful keywords. For further study, we could perform factor analysis as multidimensional analysis on a generalized data set to understand the common features and differences between movie genres since correlation exists between DocuScope variables. We could also try Biber's feature variables.

## Reference

Wikimedia Foundation. (2022, December 11). *Christopher Nolan*. Wikipedia. Retrieved December 11, 2022, from https://en.wikipedia.org/wiki/Christopher\_Nolan

*The internet movie script database (imsdb)*. The Internet Movie Script Database. (n.d.). Retrieved December 11, 2022, from https://imsdb.com/

*DocuScope*. DocuScope - docuscospacy documentation. (n.d.). Retrieved December 12, 2022, from <https://docuscospacy.readthedocs.io/en/latest/docuscope.html#categories>

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). “quanteda: An R package for the quantitative analysis of textual data.” \_Journal of Open Source Software\_, \*3\*(30), 774. doi:10.21105/joss.00774 <https://doi.org/10.21105/joss.00774>, from https://quanteda.io.

Kassambara A, Mundt F (2020). \_factoextra: Extract and Visualize the Results of Multivariate Data Analyses\_. R package version 1.0.7, from https://CRAN.R-project.org/package=factoextra.