Table of Contents

INTRODUCTION

In the illustrious realm of cinema, the title of "Academy Award for Best Picture" not only signifies critical acclaim but also serves as a hallmark of box office success. These films, carefully selected by the Academy of Motion Picture Arts and Sciences, have transcended the silver screen to become cultural landmarks. This project delves into the heart of these cinematic masterpieces.

Through advanced text mining and topic modeling techniques, this project aims to unravel the underlying themes and patterns woven into the fabric of these acclaimed narratives. This endeavor is not merely an exploration of text but a journey into the narratives that have stood the test of time. By applying topic modeling to the scripts of Best Picture Winners, this project aims to distill the essence of these films, identifying recurring themes and uncovering the hidden threads that bind them together.

For aspiring filmmakers, this project serves as a treasure trove of insights. Beyond the glitz and glamour of awards, it unveils the blueprint for storytelling success – a fusion of narrative brilliance that captivates both critics and the box office.

BACKGROUND RESEARCH & SCOPE

The project is centered around the exploration of Best Picture-winning films spanning the years 1970 till present. However, the endeavor faces a limitation as film scripts, being confidential properties, are not accessible for movies released post-2015. Consequently, the analysis is confined to the available scripts, covering 34 out of 46 Best Picture winners from 1970 to 2015. A comprehensive list of Oscar-winning films, indicating their availability for public access is made available in the "Academy Award for Best Picture (1970 to 2015)" section. Films coded in bold green are those that are available for downloads, whereas those in Red indicate their non-availability.

As a part of background research, the project delved into the financial dimensions of these cinematic achievements. Preliminary findings reveal intriguing insights into budget, gross revenue, and calculated profits. The profit margins vary significantly, ranging from 84% for "The Last Emperor" to an astounding 20355% for "Rocky." It's crucial to note that these figures are estimates, sourced from internet searches, and lack validation from specific credible sources. This initial exploration serves as a foundational step, emphasizing the importance of understanding the commercial success of Best Picture-winning films. A full list of this profit figures can be viewed in "Box office – Academy Award Winning Films (1970 to 2015)."

To facilitate the data scraping process, HTML links to the available scripts have been compiled into a CSV file. This file, containing columns for film names and corresponding URLs, is made accessible for reference here : 'Movies_List.csv'

FAIR USE DISCLAIMER

Academic and Research Considerations

This project involves the topic modeling analysis of movie scripts, which were sourced from publicly accessible websites, namely IMSDB.com and Dailyscript.com. It is essential to emphasize that the usage of these scripts is strictly intended for academic and research purposes only. The inclusion of copyrighted content within this project is undertaken with the belief that it falls under the principles of "fair use" as defined by copyright law. The analysis and insights derived from these scripts aim to contribute to scholarly discussions surrounding text mining, natural language processing, and thematic exploration in the context of award-winning cinema.

We recognize and respect the intellectual property rights of the scriptwriters, filmmakers, and relevant copyright holders. Any unintended infringement of copyright is sincerely regretted, and appropriate attributions have been made where applicable.

DATA SCRAPING

The project initiated with the acquisition of data through web scraping. The target was a curated list of movie names and corresponding URLs stored in a CSV file. The R programming language, specifically leveraging the `rvest` and `XML` libraries, was employed for this purpose. The chosen approach involved a systematic loop through each entry in the CSV file. For each movie, the associated URL was used to download content from the web. A local folder structure is specified for storing the downloaded content. This content, primarily in HTML format, underwent a cleaning process to extract relevant text information. This iterative cleaning process comprised parsing the HTML content, removing extraneous elements such as JavaScript code, and consolidating the text into a coherent format. The resulting cleaned text was then saved into individual files for further analysis.

For a list of all the packages, functions and other objects used for the project, please refer to "List of Packages Used in R Code" section in the appendix.

A Folder that contains the files used for the analysis is made available in the following link: "Topic_Modeling_Data."

DATA LOADING

This phase involved the extraction and preparation of the textual content necessary for the analysis. This stage encompassed the transformation of raw text files into a structured and manageable format. The process commenced by ensuring the availability of essential R libraries. These included `tidytext`, `topicmodels`, and `tm`, each playing a pivotal role in text processing and analysis. The data source consisted of a directory containing text files, each representing a movie. The `dir` function was employed to extract the list of relevant files based on a specified pattern. A set of custom functions, crucial for efficient text processing, were sourced and integrated into the workflow. These functions facilitated tasks like joining lines of a document into a single string. A loop iteratively processed each text file. For every file, the lines of the document were concatenated into a single string. This string, representing the text content of the document, was then structured into a data frame. The resulting data frames were consolidated to form a comprehensive data set.

A corpus was created from the processed text data. This corpus served as the foundation for the text analysis tasks.

TEXT PRE-PROCESSING

The raw text data acquired during the scraping phase underwent a series of transformations to prepare it for effective analysis. This pre-processing stage involved several key steps:

- **Lowercasing:** All text was converted to lowercase, ensuring uniformity, and eliminating potential discrepancies arising from variations in letter casing.
- **Apostrophe Removal:** Apostrophes ('s) were systematically removed from the text, simplifying word structures, and facilitating more accurate analysis.
- **Punctuation Removal:** Extraneous punctuation was stripped from the text to ensure that only meaningful words contributed to the analysis.
- **Numeric Removal:** Any numerical values present in the text were removed, as they were deemed non-essential for the analysis.

STOP WORDS REMOVAL

The stop word removal process involved the exclusion of common and domain-specific terms that typically do not contribute significantly to the semantic content of the text. Four different stop lists, in addition to the default English stopwords, were employed for a comprehensive elimination process:

- **stoplist.csv:** This stoplist, sourced from the book "Text Analytics with R" by Matthew L. Jockers and Rosamond Thalken, contributed to the removal of general stopwords relevant to text analytics.
- **characternames.csv**: To enhance the quality of the corpus, a meticulous process was implemented to eliminate character names. This step, while time-intensive, significantly improved the subsequent analysis by removing noise associated with character references. A unique stoplist was created from character names extracted from the "credits.csv" file in Kaggle's "the Movies Dataset." This involved a detailed text cleaning and pre-processing phase on the credits file. The subsequent application of Latent Dirichlet Allocation (LDA) modeling resulted in a vocabulary list, which was saved as "characternames.csv."
- **scriptkeywords.csv**: This stoplist comprised approximately 200 common terms used by scriptwriters during screenplay drafting. It aimed to eliminate frequently occurring script-related keywords that might introduce noise into the analysis.
- **customstoplist.csv**: Initially not part of the text pre-processing phase, this stoplist evolved from the observation of irrelevant words and additional character names in the initial topic model. Around 100 words were manually identified, compiled into a CSV file ("customstoplist.csv"), and subsequently utilized as a stoplist to enhance the precision of the final model.

The systematic application of these stoplists collectively refined the corpus by removing non-informative and contextually irrelevant terms, thereby preparing the text for subsequent analysis and modeling. The resulting text was then transformed into a structured data frame, setting the stage for subsequent stages of analysis and topic modeling.

TOPIC MODELING

The topic modeling phase of the project aimed to uncover latent themes within the movie script dataset. This involved converting the raw text data into a structured format, exploring the optimal number of topics, and creating a meaningful representation of these topics.

CREATING 'Document-Term' MATRIX

To facilitate topic modeling, a Document-Term Matrix (DTM) was constructed from the movie script text. This process involved tokenization, symbol removal, and eliminating repetitive words. The DTM served as the foundation for the modeling.

FINDING 'Optimal Number of Topics'

The identification of an optimal number of topics is a crucial step in topic modeling. This process involves determining the number of themes or subjects that best represent the underlying structure within a collection of documents.

To accomplish this, the project employed the `ldatuning` package in R. This package facilitates the exploration of different topic numbers and their associated coherence scores, allowing for the selection of an ideal number of topics.

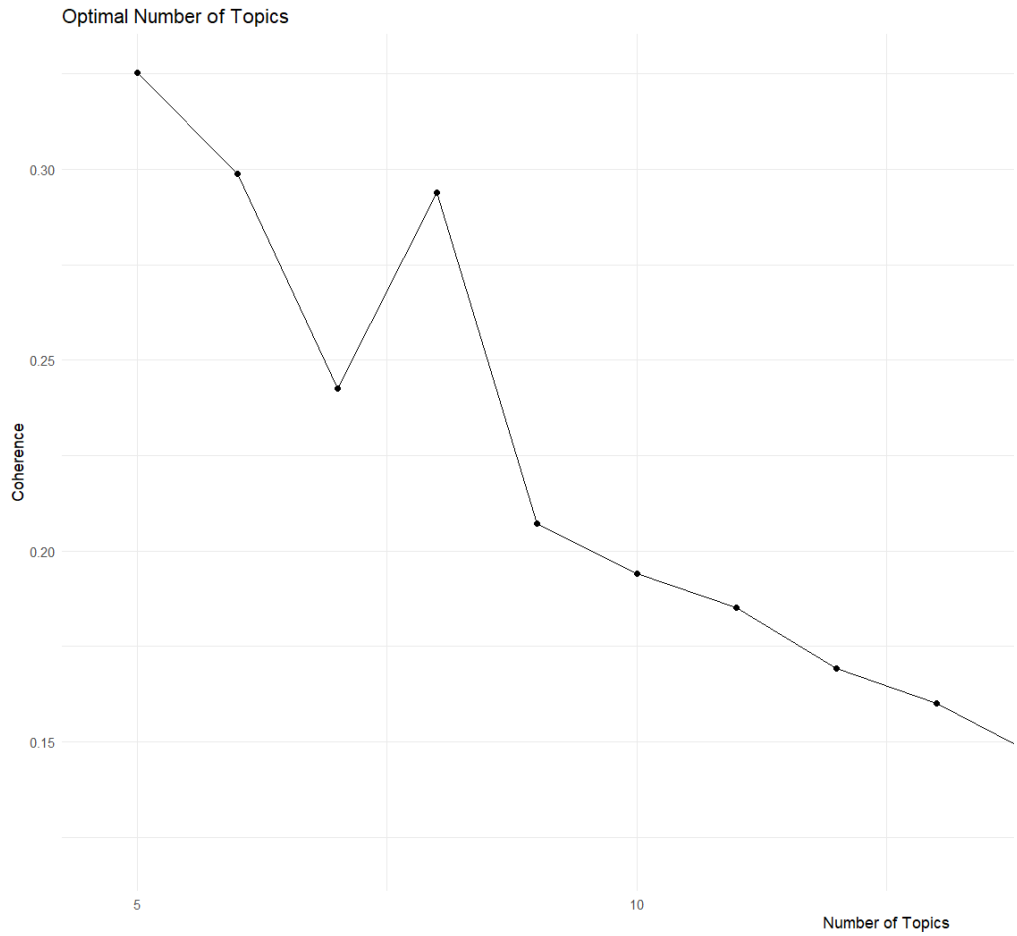
The Document-Term Matrix (DTM) obtained from the text data was converted to a matrix format. Then, the coherence score was evaluated for a given number of topics.

Coherence Score:

Coherence score plays a pivotal role in fine-tuning the topic modeling process. It not only aids in identifying the optimal number of topics but also ensures that the topics generated are coherent, meaningful, and aligned with the inherent structure of the textual data. It provides a quantitative measure of how interpretable and meaningful the identified topics are within a given corpus of text.

- A higher coherence score implies that the words within a topic are more semantically connected and provide a clearer, more interpretable theme.
- By evaluating coherence across a range of topic numbers, one can identify the point at which adding more topics no longer contributes to a more interpretable representation of the data.
- Topics with higher coherence scores are considered more meaningful, relevant, and representative of the underlying structure within the text corpus.
- The coherence score, assessed through the CaoJuan2009 method, was utilized to determine the optimal number of topics for the project.

The `FindTopicsNumber` function from the `ldatuning` package was applied to explore a range of topic numbers (from 5 to 20) using the Gibbs sampling method. The results were visualized using the `ggplot2` library, creating a line plot that displayed the coherence scores across different topic numbers.



From the above plot, we can see that the coherence score is highest (0.325) when there are 5 topics. Though it improved when the number of topics rose to 8 (0.28) , it was lower than the initial score we observed. So, the final model assumed 5 underlying themes within the movie scripts.

LDA MODEL

To distill meaningful insights from the movie scripts, Latent Dirichlet Allocation (LDA) was employed. LDA is a powerful technique for uncovering hidden thematic structures within a collection of texts. The LDA model was constructed using the `LDA` function, a part of the `topicmodels` package in R. This step involved converting the Document-Term Matrix (`dtm`), representing the frequency of terms across documents, into a coherent model. Several control

parameters were fine-tuned for model stability and reproducibility. The `seed` parameter was set to 999 to ensure consistent results across different runs.

The LDA model, in its native format, can be challenging to interpret. Hence, the model output was transformed into a tidy data frame using the `tidytext` package. This data frame, denoted as `lda_df`, facilitates a more straightforward exploration of the topics and associated terms.

DISPLAYING KEYWORDS FOR EACH TOPIC

Once the topic modeling process is complete, it's crucial to understand and interpret the top terms associated with each identified topic. The key steps taken to extract and present these terms are outlined below.

Utilizing the `dplyr` and `knitr` libraries in R, the top terms for each topic were systematically retrieved. A specified number of terms (in this case, 30) were selected to provide a concise yet comprehensive overview of each topic. The extracted top terms for each topic were organized into a structured data frame. This facilitated a clear and organized representation of the significant terms associated with each identified topic.

To enhance accessibility and enable further exploration, the results were saved in an Excel workbook and are reproduced below.

	Keywords
Topic 1	man, fuck, turns, hear, opens, gonna, woman, give, starts, stares, fucking, enters, pulls, puts, young, find, stops, voice, moment, things, shit, nods, police, slowly, listen, suddenly, job, feel, driver, follow
Topic 2	man, turns, gonna, starts, shit, voice, moves, suddenly, moment, pulls, hear, woman, stops, young, walkie, position, opens, girl, watches, give, reaches, perimeter, nods, makes, slowly, ahead, passes, yards, fuck, shoulder
Topic 3	man, turns, woman, begins, pulls, moment, slaves, girl, stares, moves, young, gonna, puts, suddenly, holds, lights, give, nods, opens, stops, police, grabs, nigger, hear, gotta, makes, driver, guard, watches, find
Topic 4	man, turns, moves, police, nods, moment, young, glances, slowly, hear, pulls, things, suddenly, voice, woman, stops, officer, stares, girl, holds, makes, give, starts, opens, quietly, finally, find, pause, featuring, reaches
Topic 5	man, turns, suddenly, give, hear, moves, moment, voice, stares, young, majesty, stops, watches, pulls, slowly, woman, starts, opens, begins, holds, nods, girl, find, guard, enters, kisses, appears, puts, finally, makes

The interpretation of these topics is discussed below.

INTERPRETATION OF TOPICS

TOPIC	INTERPRETATION	POSSIBLE GENRE	THEME (s)
1	This topic seems to revolve around intense and confrontational scenes. The presence of words like "fuck," "stares," and "police" suggests a potentially dramatic and high-stakes scenario. The actions of "enters," "pulls," and "puts" contribute to the sense of urgency and tension. The inclusion of "driver" and "follow" hints at a possible pursuit or escape, adding an element of suspense.	Drama/Thriller	Intense Confrontations / Pursuits
2	This topic seems to capture dynamic and sudden events. The use of words like 'suddenly' and 'moment' implies quick actions and changes in the narrative, creating a sense of unpredictability. The presence of words like "walkie," "position," and "perimeter" hints at strategic or covert activities.	War/Thriller	Strategic Maneuvers and Unpredictability
3	This topic delves into emotional moments and struggles, possibly depicting interpersonal relationships or challenges. Words like "holds," "gives," and "stares" suggest emotional depth. The inclusion of "slaves" and "girl" hints at themes of oppression or societal struggles. The presence of "police" and "guard" suggests conflict and obstacles.	Historical/Social Drama	Emotional Struggles and Social Issues
4	This topic revolves around police-related scenarios and moments of suspicion. Keywords like "police," "officer," and "suspicion" suggest law enforcement themes. Actions such as "moves," "glances," and "pauses" indicate a careful and observant atmosphere. The use of "finally" and "reaches" suggests a resolution or culmination.	Crime/Thriller	Police Encounters and Tense Moments

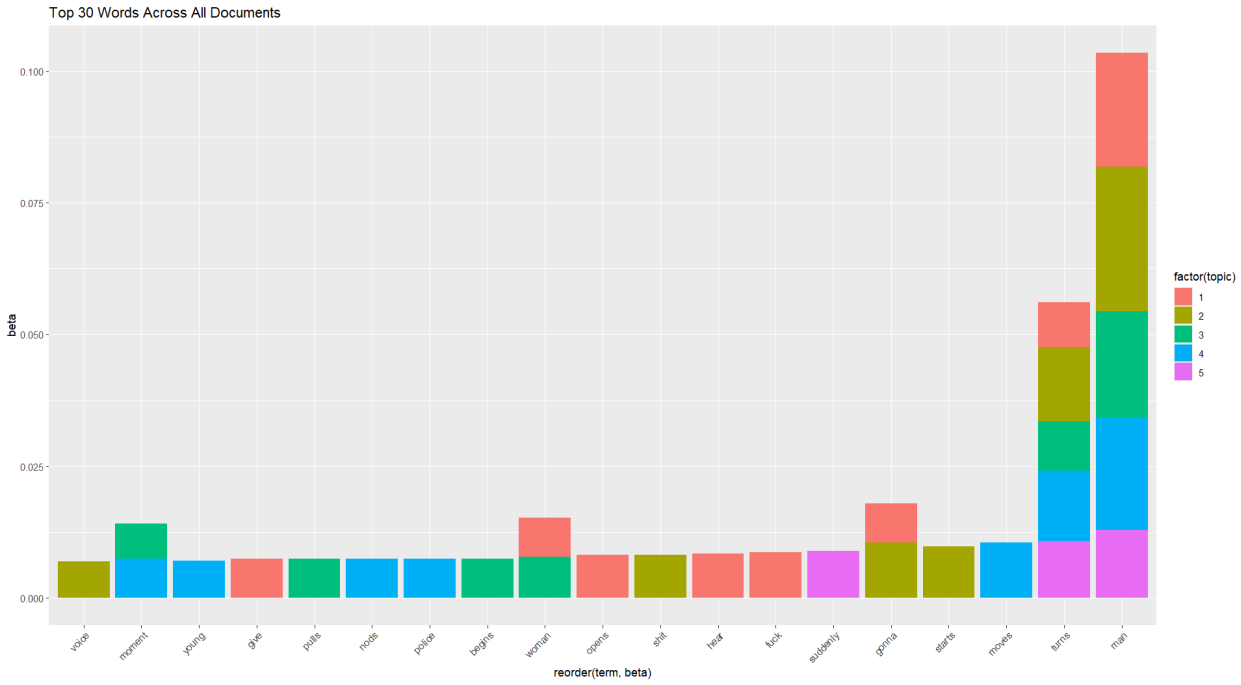
TOPIC	INTERPRETATION	POSSIBLE GENRE	THEME (s)
5	This topic captures tender and emotional moments. Keywords like "give," "stares," and "kisses" evoke emotions and intimacy. Inclusion of words like “majesty,” suggests a royal or grand element.	Romance/Fantasy	Romantic Encounters and Grandeur

INSIGHTS FROM TOPIC MODELING

GROUPED BAR PLOT : Dominant Words in Overall Corpus

The project delved into a comprehensive analysis of the textual data through the application of a grouped bar plot, a graphical representation illustrating the frequency distribution of the top 30 words across all documents. The process was initiated by extracting the relevant data from the LDA model results. Focusing on the top 30 words overall, the analysis aimed to uncover prevalent terms that transcended individual topics, providing a holistic understanding of the corpus.

Each bar represents a distinct term, while the height of the bars indicated their prevalence. The utilization of color-coded bars facilitated the differentiation of words across topics.



Key Insights from the Grouped Bar Plot Analysis

The examination of the grouped bar plot has unearthed compelling insights that shed light on recurring themes and pivotal elements within the corpus. These findings are crucial in deciphering the underlying narrative and thematic patterns inherent in movie scripts.

- **Universal Themes**: The terms 'Man' and 'turns' emerge as singularly omnipresent, threading through all five topics. In the context of cinematic storytelling, 'Man' could symbolize the pervasive influence of the male perspective or gaze, while 'turns' may signify intense actions that convey profound emotion. These recurring words intricately weave a common thread across diverse topics, suggesting their overarching significance in cinematic discourse.
- **Multifaceted Significance:** 'Woman,' 'moment,' and 'gonna' stand out as words bridging more than one topic, each contributing to the thematic tapestry in distinct ways. 'Woman' likely encapsulates themes of femininity, portraying the multifaceted roles and perspectives of female characters across different narratives. 'Moment' may signify pivotal scenes within movies, capturing instances of heightened emotional impact, potentially serving as narrative focal points. The informal contraction of 'going to,' 'gonna' introduces an element of anticipation, suggesting future events that may unfold. This linguistic choice can introduce suspense and intrigue within the storyline.

These discerning insights are further underscored by the accompanying visualizations of word clouds for each topic.

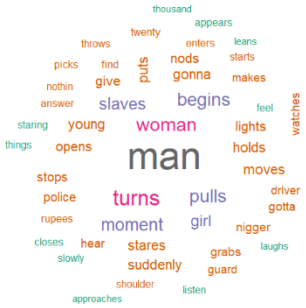
WORD CLOUDS : Dominant Words in Each Topic

Visualizing Word Clouds involved extracting terms associated with the chosen topic, emphasizing the top 50 terms based on their relevance (beta values). This visualization provides an insightful representation of the most prominent terms associated with a particular topic.

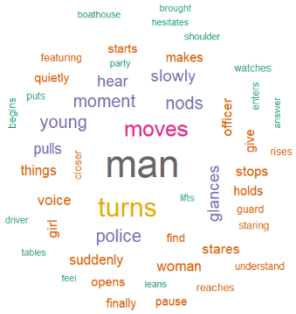
Word Cloud for Topic 1

Word Cloud for Topic 2

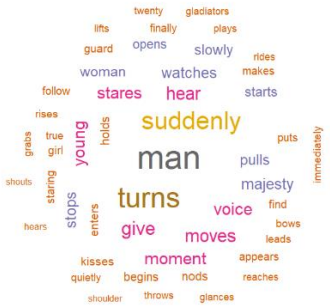
Word Cloud for Topic 3



Word Cloud for Topic 4



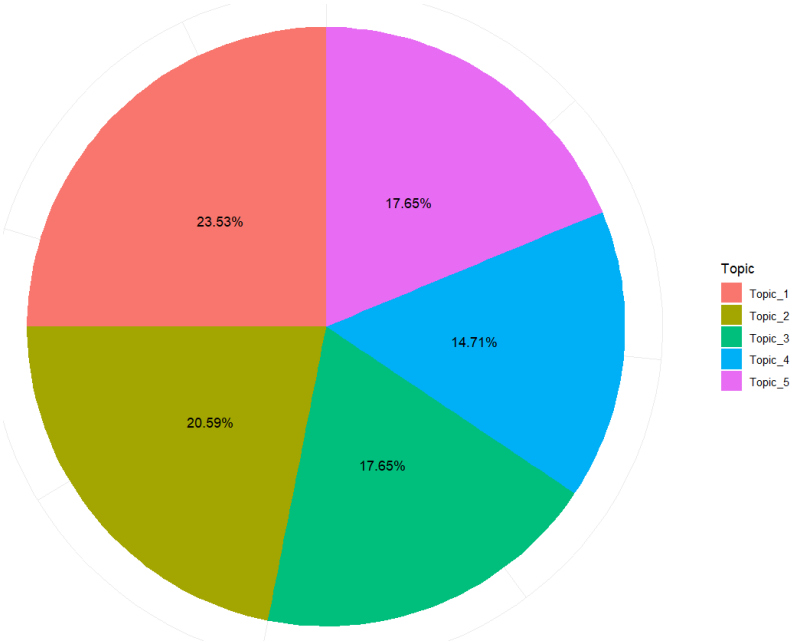
Word Cloud for Topic 5



The above Word Clouds offer a quick overview of the most significant terms within the selected topic. The size of each term is proportional to its importance, allowing for an immediate identification of the terms that strongly characterize the topic. From the word clouds, we can see that the most common words in all the topics are ‘man’, ‘turns.' This suggests the thematic significance of these two words in each topic, which was already discussed by visualizing the grouped bar plot.

PIE CHART : Breakdown of Topic Compositions

To gain a comprehensive understanding of the corpus, a pie chart was generated to visualize the distribution of topics across the entire dataset. The pie chart vividly illustrates the percentage composition of each topic. Each segment corresponds to a specific topic, with the size of the segment proportional to the prevalence of that topic in the entire dataset.



Insights

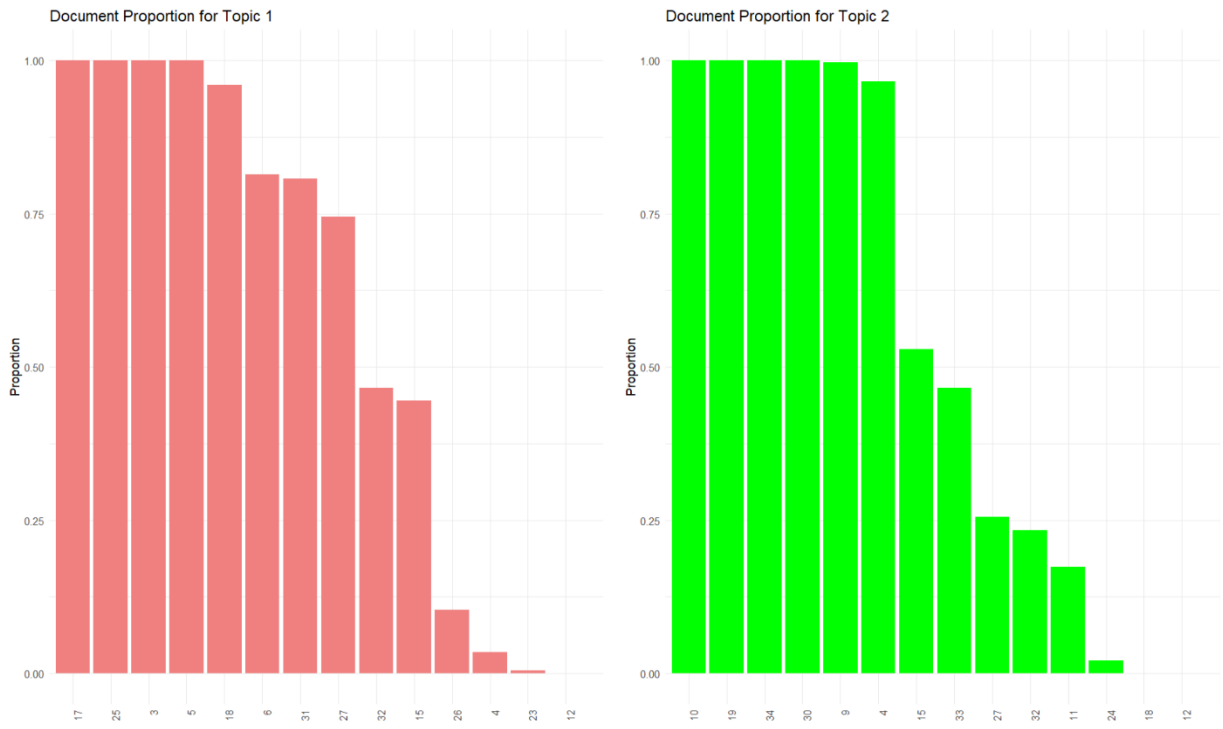
- A significant portion of Oscar-winning movies revolves around the theme of "Intense Confrontations/Pursuits," as indicated by the distribution of Topic 1 in the corpus.
- Additionally, approximately 21% of the content in the best picture winners is dedicated to exploring "Strategic Maneuvers and Suspenseful Moments," encapsulated in Topic 2.

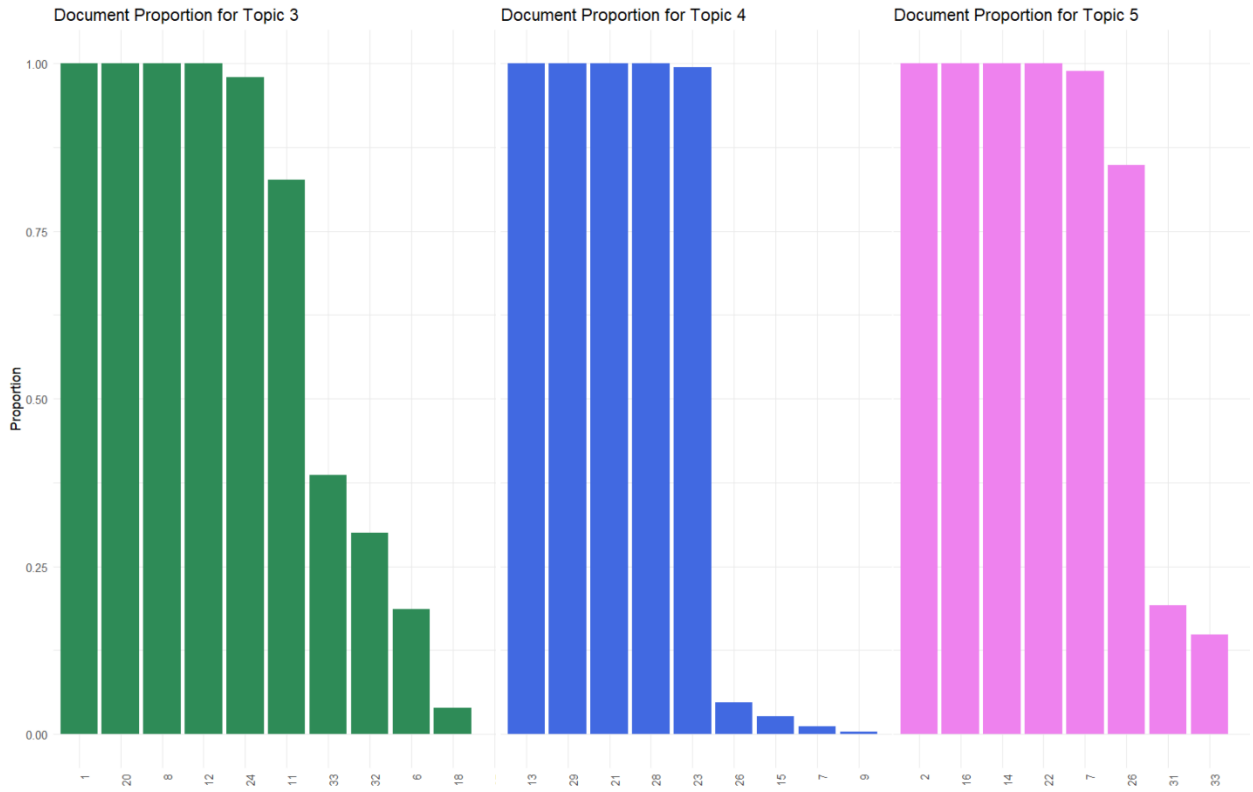
BAR PLOTS : Distribution of Each Topic Across Documents

The exploration of document-topic proportions provides valuable insights into the distribution of topics across the corpus. The process begins with extracting document-topic proportions from the model, where each document is associated with a distribution of topics. To facilitate a clear understanding, bar plots are employed to represent the document proportions for each specified topic. The bar plots for each topic are presented where:

- The x-axis represents the 'document number,' which corresponds to individual movies in our dataset. To facilitate better interpretation, a table containing the movies names and their document ids has been provided in the "Document – Movies Identification" in the appendix section.
- Meanwhile, the y-axis gauges the relevance of a particular topic in a given movie.

The bar charts provide a nuanced understanding of how each film engages with specific topics and to what extent. By examining the bars, we can discern not only which films delve into a particular topic but also the degree to which each movie is intertwined with that specific thematic element.





From the above plots, we can see that most of the movies predominantly deal with a certain topic. To validate the model's efficacy, documents that exhibit a considerable emphasis on specific topics are meticulously curated in the below table. Then, the insights obtained from the thematic interpretations made earlier, are incorporated. Later, the corresponding movies are identified from their document ID references in the "Document – Movies Identification" in the appendix section.

Docs	Possible Genre	Underlying Theme (s)	Movies Covered
17 **25** **3** **5** **18**	Drama/Thriller	Intense Confrontations / Pursuits	No Country for Old Men The Departed American Beauty Argo Ordinary People
10 **19** **34** **30** **9**	War/Thriller	Strategic Maneuvers and Unpredictability	Dances With Wolves Platoon Unforgiven The Hurt Locker Cuckoo's Nest

Docs	Possible Genre	Underlying Theme (s)	Movies Covered
1 **20** **8** **12**	Historical/Social Drama	Emotional Struggles and Social Issues	12 Years a Slave Rocky Crash Forrest Gump
13 **28** **29** **21**	Crime/Thriller	Police Encounters and Tense Moments	Gandhi The Godfather The Godfather 2 Schindler's List
2 **16** **14** **22**	Romance/Fantasy	Romantic Encounters and Grandeur	Amadeus Lord of the Rings 3 Gladiator Shakespeare in Love

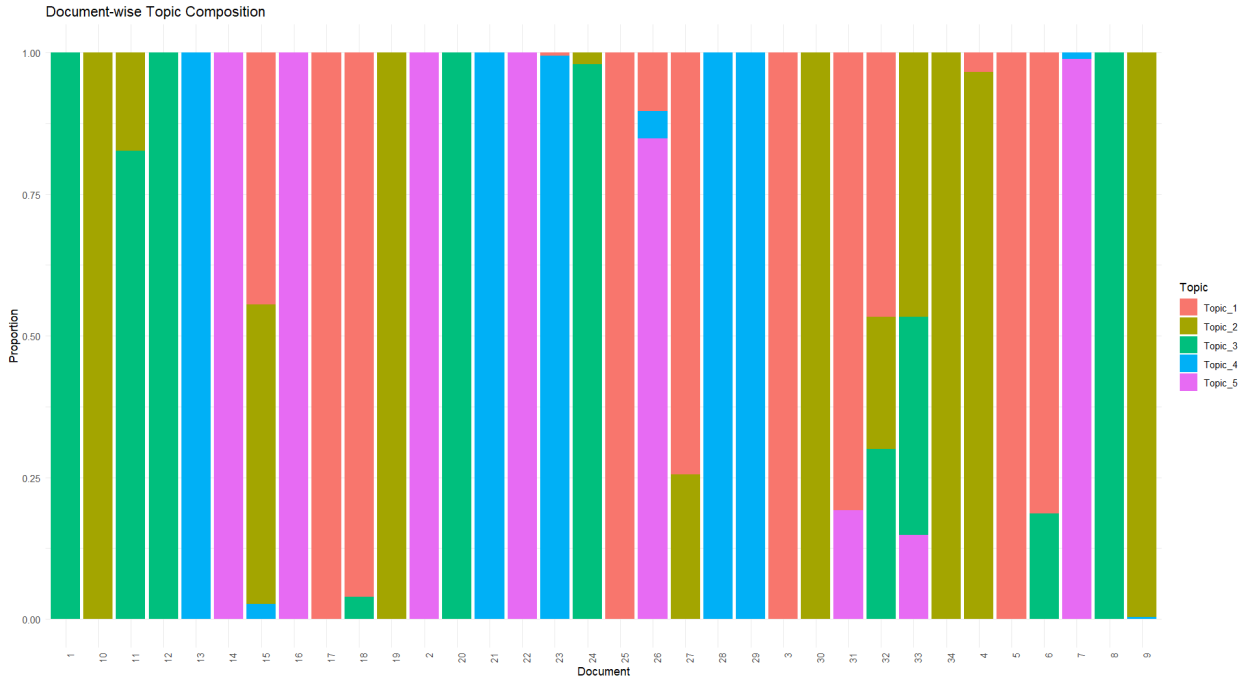
Despite the dataset being small, the LDA model proved to have great interpretability as the results of the model were spectacular. We can see from the above table that the results demonstrate a high level of precision, aligning well with the actual content and genres of the movies.

STACKED COLUMN CHART: Document-wise Topic Composition

The document-wise stacked bar plot provides a visual representation of the distribution of topics across different documents. This analysis is crucial for understanding the thematic composition of each document in the corpus.

The below stacked bar plot is nothing, but a consolidation of the individual plots discussed earlier Each bar represents a document, and its segments correspond to the proportions of different topics within that document. The x-axis represents individual documents, while the y-axis indicates the proportion of each topic. However, in addition to the assessment of the prevalent topics in each document, it aids in identifying the diversity of topics covered in the corpus and understanding the relative significance of each topic within individual documents.

In this section, an emphasis is made only on those movies that had dealt with more than one topic. As more than one topic is involved in certain movies, a fresh interpretation is made based on the combination of keywords of the combined topics. The table that follows the stacked bar plot identifies the documents and the set of topics that the document deals with. The interpretation of the blended topics and the possible genre is incorporated in the subsequent columns. Later, the movie names are identified from the "Document – Movies Identification" table.



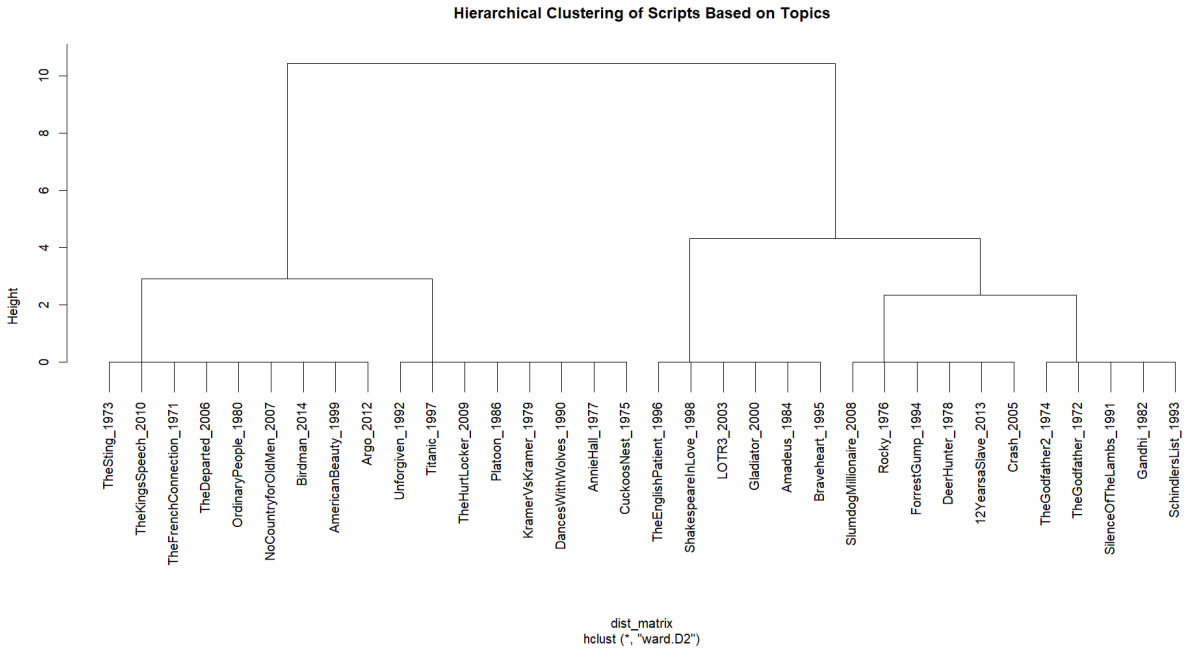
Doc	Topics	Theme (s)	Genres	Movies
33	2, 3 & 5	Tense and Emotional in a Grandeur setting	Historical Thriller Romance	Titanic
27	1 & 2	Police Investigation and Tension	Drama Thriller	The French Connection
32	1, 2 & 3	Tense Confrontations, Law Enforcement, and Power Struggles	Crime Drama	The Sting

The model's results are promising even for the movies that encompass more than one topics, as reflected by how closely the underlying themes in the combination of these topics align with the actual story of these movies.

DENDROGRAM : Hierarchical Clustering

Hierarchical clustering is implemented to organize scripts into clusters based on their similarity in topic distribution. The document-topic matrix, derived from the LDA model, is used as the foundation for this analysis. To facilitate comprehension, the dendrogram is cut at a specified height, resulting in a predetermined number of clusters. For this analysis, five clusters are selected to align with the number of topics.

The cluster assignment provides valuable insights into the thematic similarities between scripts, paving the way for deeper analyses and targeted exploration of specific thematic clusters. It should be noted that the insights from the dendrogram presented below aligns with the observations made from the stacked bar plot. That is, we can say which Oscar winning movies from the corpus have dealt with similar themes.



SUMMARY

Key Insights

- **Universal Themes:** Our exploration revealed recurrent themes cutting across diverse genres. The words 'Man' and 'turns' emerged as universal threads, suggesting their overarching significance in cinematic discourse. These terms intricately connected various topics, symbolizing the pervasive influence of the male perspective and intense actions conveying profound emotion.
- **Multifaceted Significance:** Words like 'Woman,' 'moment,' and 'gonna' bridged multiple topics, adding depth to the thematic tapestry. 'Woman' captured the multifaceted roles of female characters, 'moment' signified pivotal scenes, and 'gonna' introduced anticipation, injecting suspense into storylines.

Topic Exploration:

1. **Intense Confrontations/Pursuits (Drama/Thriller):*** Revolving around dramatic and high-stakes scenarios, this topic depicted urgent and tense situations.
2. **Strategic Maneuvers and Suspenseful Moments (War/Thriller):** Capturing dynamic and sudden events, this topic hinted at strategic or covert activities.
3. **Emotional Struggles and Social Issues (Historical/Social Drama):** Delving into emotional moments and struggles, this topic explored interpersonal relationships and societal challenges.
4. **Police Encounters and Tense Moments (Crime/Thriller):** Focused on police-related scenarios, this topic conveyed moments of suspicion and careful observation.
5. **Romantic Encounters and Grandeur (Romance/Fantasy):** Centered on tender and emotional moments, this topic captured romantic encounters with a touch of grandeur.

Visualizations:

1. **Grouped Bar Plot:** Highlighted the frequency distribution of top 30 words across all documents, emphasizing universal and multifaceted terms.
2. **Word Clouds:** Reinforced the thematic significance of 'Man' and 'turns' in each topic, providing a visual representation of dominant words.
3. **Pie Chart and Bar Plots:** Illustrated the distribution and prevalence of topics, offering insights into the thematic composition of Oscar-winning films.

4. **Stacked Column Chart:** Examined document-wise topic composition, emphasizing movies dealing with more than one topic, showcasing model precision.
5. **Dendrogram:** Hierarchical clustering organized scripts into thematic clusters, aligning with observations from the stacked bar plot.

CONCLUSION

This project, delving into the heart of Academy Award-winning films through advanced text mining and topic modeling, offers invaluable insights for aspiring screenwriters, filmmakers, and enthusiasts alike. As the curtains draw on this exploration, three key facets emerge, underlining the significance and future potential of this endeavor.

1. **A Treasure Trove for Aspiring Creators:** This project unfolds a treasure trove of insights for budding screenwriters and filmmakers. It unveils the secret sauce behind Academy Award-winning scripts, providing a blueprint for narrative success. Aspiring enthusiasts can glean from the distilled themes, learning to captivate both critics and audiences, steering their own stories toward acclaim.
2. **Strategic Guidance for Studios:** Studios aiming for both critical acclaim and box office triumph can leverage this project's findings. Understanding the recurring themes in Best Picture winners offers a strategic advantage. By aligning narratives with these insights, studios can increase the odds of crafting films that resonate deeply, ensuring both critical acclaim and commercial success.
3. **A Call for Broader Horizons:** Expanding the dataset to include more films is the key to unlocking deeper insights. A broader sample would offer a more nuanced understanding of evolving and recurring cinematic trends, allowing creators and studios to stay at the forefront of storytelling. The more films included, the richer the tapestry of insights, guiding the industry toward even greater heights.

APPENDIX

Academy Award for Best Picture (1970 to 2015)

Year	FILM	Genre
1970	Patton	Drama, War
1971	The French Connection	Action, Crime, Drama
1972	The Godfather	Crime, Drama
1973	The Sting	Comedy, Crime, Drama
1974	The Godfather Part II	Crime, Drama
1975	One Flew Over the Cuckoo's Nest	Drama
1976	Rocky	Drama
1977	Annie Hall	Comedy, Romance
1978	The Deer Hunter	Drama, War
1979	Kramer vs. Kramer	Drama
1980	Ordinary People	Thriller, Drama
1981	Chariots of Fire	Drama, Sport
1982	Gandhi	Biography, Drama
1983	Terms of Endearment	Comedy, Drama
1984	Amadeus	Biography, Drama
1985	Out of Africa	Biography, Drama, Romance
1986	Platoon	Drama, War
1987	The Last Emperor	Biography, Drama, History
1988	Rain Man	Drama
1989	Driving Miss Daisy	Comedy, Drama
1990	Dances with Wolves	Adventure, Drama, Western
1991	The Silence of the Lambs	Crime, Drama, Thriller
1992	Unforgiven	Drama, Western
1993	Schindler's List	Biography, Drama, History
1994	Forrest Gump	Drama, Romance
1995	Braveheart	Biography, Drama, History, War
1996	The English Patient	Drama, Romance, War
1997	Titanic	Drama, Romance
1998	Shakespeare in Love	Comedy, Drama, History, Romance
1999	American Beauty	Drama
2000	Gladiator	Action, Adventure, Drama
2001	A Beautiful Mind	Biography, Drama
2002	Chicago	Comedy, Crime, Musical
2003	The Lord of the Rings: The Return of the King	Action, Adventure, Drama, Fantasy
2004	Million Dollar Baby	Drama, Sport
2005	Crash	Crime, Drama, Thriller
2006	The Departed	Crime, Drama, Thriller

Year	FILM	Genre
2007	No Country for Old Men	Crime, Drama, Thriller, Western
2008	Slumdog Millionaire	Drama
2009	The Hurt Locker	Drama, Thriller, War
2010	The King's Speech	Biography, Drama, History
2011	The Artist	Comedy, Drama, Romance
2012	Argo	Biography, Drama, Thriller
2013	12 Years a Slave	Biography, Drama, History
2014	Birdman	Comedy, Drama
2015	Spotlight	Biography, Crime, Drama

Box office – Academy Award Winning Films (1970 to 2015)

Year	Best Picture Winner	Budget (in millions)	Gross (in millions)	Est. Profit	% of Profit
1976	Rocky	1.1	225	223.9	20355%
1972	The Godfather	6.5	286	279.5	4300%
1973	The Sting	5.5	159.6	154.1	2802%
2010	The King's Speech	$15.00	$414.20	399.2	2661%
1971	The French Connection	1.9	51.7	49.8	2621%
2008	Slumdog Millionaire	$15.00	$377.90	362.9	2419%
1975	One Flew Over the Cuckoo's Nest	4.4	109	104.6	2377%
1999	American Beauty	$15.00	$356.30	341.3	2275%
1986	Platoon	$6.00	$138.50	132.5	2208%
1989	Driving Miss Daisy	$7.50	$145.80	138.3	1844%
1990	Dances with Wolves	$22.00	$424.20	402.2	1828%
2005	Crash	$6.50	$98.40	91.9	1414%
1993	Schindler's List	$22.00	$322.00	300	1364%
1991	The Silence of the Lambs	$19.00	$272.70	253.7	1335%
1988	Rain Man	$25.00	$354.80	329.8	1319%
1983	Terms of Endearment	$8.00	$108.40	100.4	1255%
1979	Kramer vs. Kramer	8	106.3	98.3	1229%
1994	Forrest Gump	$55.00	$678.20	623.2	1133%
2003	LOTR: The Return of the King	$94.00	$1,120.30	1026.3	1092%
1998	Shakespeare in Love	$25.00	$289.30	264.3	1057%
1992	Unforgiven	$14.40	$159.20	144.8	1006%
1997	Titanic	$200.00	$2,187.50	1987.5	994%
1981	Chariots of Fire	$5.50	$59.00	53.5	973%
1977	Annie Hall	4	38.3	34.3	858%
2013	12 Years a Slave	$20.00	$187.70	167.7	839%
2011	The Artist	$15.00	$133.40	118.4	789%
1980	Ordinary People	$6.20	$54.80	48.6	784%
1996	The English Patient	$31.00	$231.90	200.9	648%
1985	Out of Africa	$31.00	$227.50	196.5	634%
2004	Million Dollar Baby	$30.00	$216.80	186.8	623%
2007	No Country for Old Men	$25.00	$171.60	146.6	586%
2002	Chicago	$45.00	$306.80	261.8	582%
1982	Gandhi	$22.00	$127.80	105.8	481%
2014	Birdman	$18.00	$103.20	85.2	473%
2001	A Beautiful Mind	$58.00	$313.50	255.5	441%

Year	Best Picture Winner	Budget (in millions)	Gross (in millions)	Est. Profit	% of Profit
2012	Argo	$44.50	$232.30	187.8	422%
2015	Spotlight	$20.00	$98.30	78.3	392%
2000	Gladiator	$103.00	$460.50	357.5	347%
1974	The Godfather Part II	13	57.3	44.3	341%
1970	Patton	12.6	45	32.4	257%
2009	The Hurt Locker	$15.00	$49.20	34.2	228%
1978	The Deer Hunter	15	48.9	33.9	226%
2006	The Departed	$90.00	$289.80	199.8	222%
1995	Braveheart	$72.00	$210.40	138.4	192%
1984	Amadeus	$18.00	$51.90	33.9	188%
1987	The Last Emperor	$23.80	$43.90	20.1	84%

Document – Movies Identification

Document	Movie Name
1	12YearsaSlave_2013
2	Amadeus_1984
3	AmericanBeauty_1999
4	AnnieHall_1977
5	Argo_2012
6	Birdman_2014
7	Braveheart_1995
8	Crash_2005
9	CuckoosNest_1975
10	DancesWithWolves_1990
11	DeerHunter_1978
12	ForrestGump_1994
13	Gandhi_1982
14	Gladiator_2000
15	KramerVsKramer_1979
16	LOTR3_2003
17	NoCountryforOldMen_2007
18	OrdinaryPeople_1980
19	Platoon_1986
20	Rocky_1976
21	SchindlersList_1993
22	ShakespeareInLove_1998
23	SilenceOfTheLambs_1991
24	SlumdogMillionaire_2008
25	TheDeparted_2006
26	TheEnglishPatient_1996
27	TheFrenchConnection_1971
28	TheGodfather_1972
29	TheGodfather2_1974
30	TheHurtLocker_2009
31	TheKingsSpeech_2010
32	TheSting_1973
33	Titanic_1997
34	Unforgiven_1992

List of Packages Used in R Code

S.No	PARTICULARS	TYPE	USAGE IN THE PROJECT
1	rvest	Package	In the project, the **rvest** package is used to scrape data from web pages. Specifically, it is employed to download the content for each URL from a CSV file containing movie names and URLs.
2	xml	Package	The **XML** library is used for parsing HTML content. Specifically, the **htmlParse** function from the **XML** library is employed to parse the HTML content downloaded from movie URLs.
3	tidytext	Package	**tidytext** is used for loading the data and in text processing tasks, such as creating a document-term matrix (**dtm**) from the text data.
4	topicmodels	Package	The **topicmodels** package is used for creating an LDA model, identifying optimal topics, and for analyzing the distribution of topics across documents.
5	Tm	Package	**tm** package is used for text preprocessing. Specifically, it's employed to create a Corpus, which is a fundamental structure in the text mining process.
6	ldatuning	Package	The **ldatuning** package is used for tuning parameters of the LDA model, for finding the optimal number of topics.
7	ggplot2	Package	**ggplot2** is employed to generate various plots, such as bar plots, pie chart, and coherence plot, to visualize and interpret the results.
8	rbind	Function	**rbind** is used to combine data frames during the processing phase for organizing the data.
9	read.csv	Function	**read.csv** is employed to read the CSV file containing movie names and URLs, which is part of the data scraping process. It is also used for reading the stopword files during text pre-processing.
10	gsub	Function	**gsub** is used to replace specific patterns in text data, such as removing JavaScript code from HTML content during the data cleaning process.
11	Corpus	Object	For this project, the term "corpus" is used to create a Corpus object from the text data present in the pre-processed text data from the movie text files. This object is the foundation for the topic modeling steps in the project.
12	tm_map	Function	**tm_map** is used for various text pre-processing tasks, such as converting text to lowercase, removing stopwords, and removing punctuation, as part of preparing the data for topic modeling.

S.No	PARTICULARS	TYPE	USAGE IN THE PROJECT
13	split	Function	The **split** function is used to split the character names into chunks. In this project, it is specifically used to split the 'charactername' stoplist into smaller chunks for more efficient removal of stopwords during text pre-processing.
14	sapply	Function	The **sapply** function is used to convert the corpus into a data frame, where each document is represented as a row with its corresponding text.
15	as.matrix	Function	The **as.matrix** function is applied to the DTM created from the corpus to prepare the data for finding the optimal number of topics using ldatuning.
16	LDA	Function	In this project, the **LDA** function from the **topicmodels** package is used to implement the Latent Dirichlet Allocation modeling technique, which helped in discovering topics present in the corpus.
17	dplyr	Package	The **dplyr** package is employed to perform operations like filtering, grouping, and arranging data, particularly in the context of analyzing and visualizing the top terms for each topic.
18	knitr	Package	The **knitr** package is used for dynamic report generation in R. In this project, it is used in creating a reproducible document that includes the topics and keywords.
19	readxl	Package	The **readxl** package is used to read the results of topic modeling from the Excel file into a data frame for further analysis or visualization.
20	wordcloud	Package	the **wordcloud** package is used to generate word clouds for the topics in the model.
21	function	Keyword	In the project, a keyword "function" is used to define custom functions – for Word clouds, bar plots and Pie chart within the project.
22	as.data.frame	Function	**as.data.frame** is is employed to convert the document-topic matrix obtained from the LDA model into a data frame, making it easier to work with and visualize.
23	dist	Function	The **dist** function is utilized to calculate the Euclidean distance matrix for hierarchical clustering.
24	hclust	Function	The **hclust** is applied to the Euclidean distance matrix obtained from the **dist** function, resulting in a hierarchical clustering dendrogram.