Movie trend overview

Yeonie Heo (sh5874)

Project Introduction

How did the topic and content of movies shift over time?

⇒ How did the significance & attention to social problems in movies shift over time, examining the winners of Oscar (1927-2021)?

- **Social atmosphere** reflected in movies
- **Stories and narrations** receive the most **empathy**
- Hypothesis: a gradual improvement in inclusiveness and diversity (gender, race, handicapped, etc) within the *content* of the films

Data Collection

from Wikipedia & Rotten Tomatoes through Web-Scraping in python

1) **Plots** of Best Picture Award Winners (1927-2021)

oscar_plot

Year	Best Picture	Plot
1928	Wings	Jack Powell and David Armstrong are rivals in the same small American town, both vying for the attentions of pretty
1929	The Broadway Melody	Eddie Kearns (Charles King) sings "The Broadway Melody", and tells some chorus girls that he brought the Mahon
1930	All Quiet on the Western Front	Professor Kantorek gives an impassioned speech about the glory of serving in the Army and "saving the Fatherland
1931	Cimarron	The Oklahoma land rush of 1889 prompts thousands to travel to the Oklahoma Territory to grab free government la
1932	Grand Hotel	Doctor Otternschlag, a disfigured veteran of World War I and a permanent resident of the Grand Hotel in Berlin, obs
1933	Cavalcade	On the last day of 1899, Jane and Robert Marryot, an upper-class couple, return to their townhouse in a fashionabl
1934	It Happened One Night	Spoiled heiress Ellen "Ellie" Andrews has eloped with pilot and fortune-hunter King Westley against the wishes of h
1935	Mutiny on the Bounty	One night in Portsmouth, England in 1787, a press gang breaks into a local tavern and presses all of the men drinki
1936	The Great Ziegfeld	The son of a highly respected music professor, Florenz "Flo" Ziegfeld Jr. yearns to make his mark in show business
1937	The Life of Emile Zola	Set in the mid through late 19th century, the film depicts Émile Zola's early friendship with Post-Impressionist paint

2) **Critic / Audience Review**

movie_cul_data

movie	type	date	review
1980	audience	20-Mar-22	Heartbreaking and uplifting... first saw this movie in me teens and rewatching it is just as powerful and rele
1980	audience	20-Sept-21	This movie is supposed to reflect psychological disturbances of a boy who lost his brother in a small sailing And then his mother's looses her affection against her husband and younger son after death her elder son. Psychological disturbance of the boy as a result of guilt feeling is not convincing. Mother displays a character of an order freak at home, and acts like a resentful robot against her son. All acting in this movie looks amateurish. And plots are sooo boring. From the beginning to end, for every actor, one feels that casting was wrong. No one is impressive. And I remember of nothing of the soundtrack if there was one.
1980	audience	5-May-21	An emotionally heavy and insightful family drama about a family's troubles coming to terms with the death
1980	audience	15-Mar-21	Robert Redford's directorial debut promptly won him the top prizes for Best Director and Best Picture at th Timothy Hutton won the Oscar for Best Supporting Actor for his role as the son Conrad, who has just been The great strengths of Ordinary People are its outstanding cast and the realistically written script. The dialo
1980	audience	21-Feb-21	This is probably the greatest film ever made. Every single scene adds to the story. The acting is perfect. In

Scope of this project & Methods

[Plot] ~ Wikipedia

The change in content of plot over time

- **STM:** identify **expected topic proportion**
- **Textnets:** plot content connections within year_era

[Review] ~ Rotten Tomatoes

Appreciation / evaluation on movie content

- **STM with covariates:** **release_era + critic & audience** to observe film content change over time and difference of reviews by experts (technical) and the public

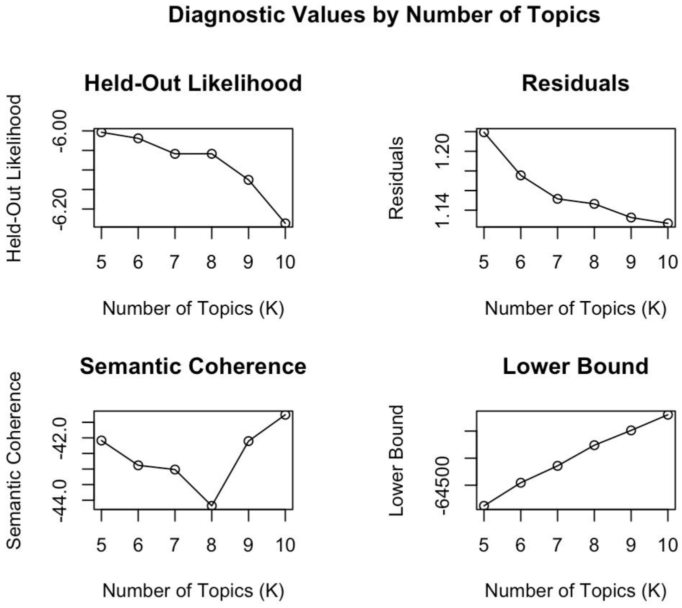
[PLOT] MOVIE TREND OVERVIEW

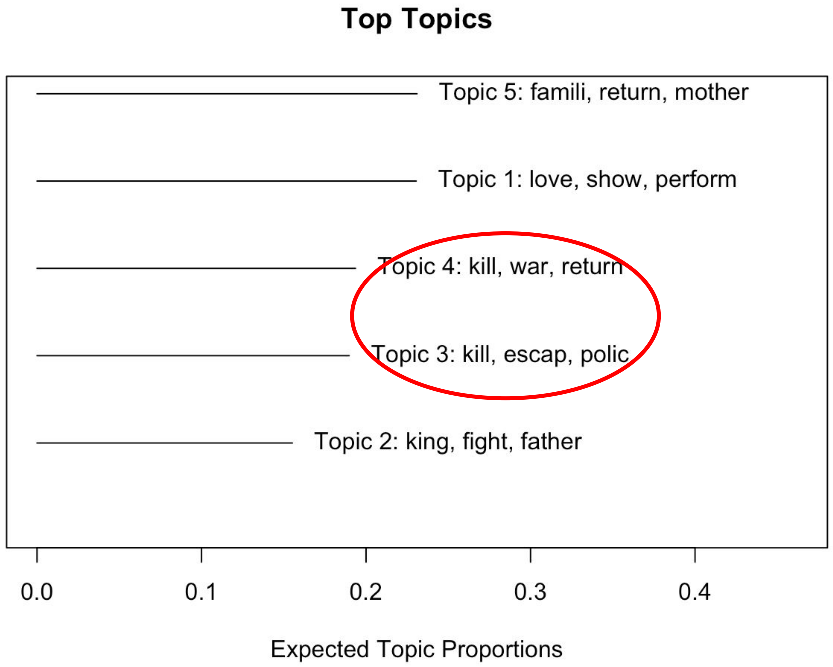
What are the movies about? How did the content change over time?

Structural Topic Modeling (STM) uncovered latent topics within a corpus of the Plot data. STM plotted each movie as distributions of topics (topic prevalence) and topics as a distribution of words (topic content).

After preprocessing (tolower, numbers/punct/stopwords/symbol removal, stemming), Plot data was categorized into 10 topics.

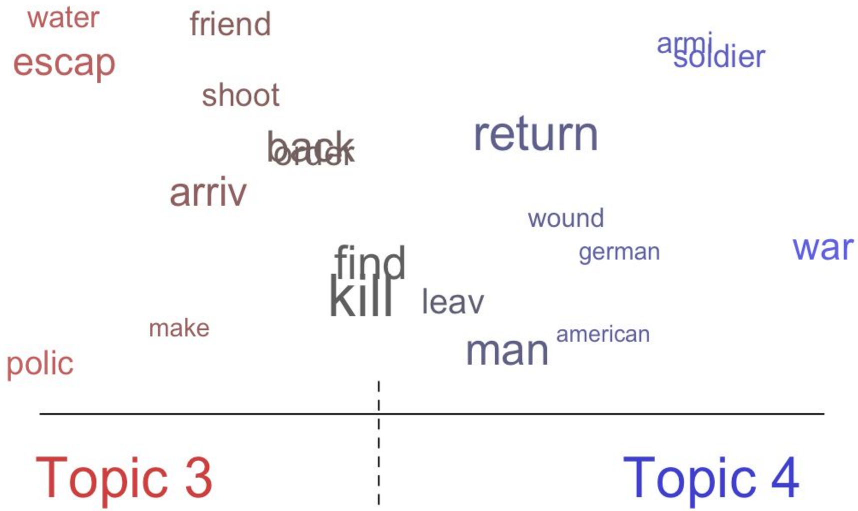
```
Topic 1 Top Words:
 	 Highest Prob: miss, invit, film, tell, home, job, find
 	 FREX: miss, invit, dinner, hire, film, job, peopl
 	 Lift: miss, invit, dinner, guest, hire, write, build
 	 Score: miss, dinner, invit, guest, apart, perform, hire
Topic 2 Top Words:
 	 Highest Prob: famili, children, return, mother, home, hous, son
 	 FREX: children, famili, mother, busi, hous, sing, parent
 	 Lift: children, famili, parent, sing, busi, neighbor, discuss
 	 Score: children, famili, sing, busi, mother, parent, relationship
Topic 3 Top Words:
 	 Highest Prob: king, die, court, leav, death, order, wife
 	 FREX: king, court, death, england, lead, die, ship
 	 Lift: king, england, court, told, desir, declar, author
 	 Score: king, court, england, speak, declar, marriag, explain
Topic 4 Top Words:
 	 Highest Prob: man, water, find, money, back, discov, car
 	 FREX: man, water, car, money, pass, hide, polic
 	 Lift: water, man, pass, hide, found, crash, observ
 	 Score: water, money, polic, phone, car, hide, shoot
Topic 5 Top Words:
 	 Highest Prob: friend, run, team, learn, stori, work, begin
 	 FREX: team, investig, run, secret, stori, game, friend
 	 Lift: team, secret, game, investig, cover, john, depart
 	 Score: team, u., investig, develop, cover, run, bar
Topic 6 Top Words:
 	 Highest Prob: father, money, return, love, meet, day, learn
 	 FREX: father, money, brother, wed, owner, learn, love
 	 Lift: wed, owner, father, brother, fellow, money, lost
 	 Score: wed, father, money, letter, lost, sister, owner
Topic 7 Top Words:
 	 Highest Prob: kill, war, soldier, armi, return, leav, find
 	 FREX: soldier, armi, enemi, war, wound, land, british
 	 Lift: enemi, victori, armi, soldier, land, camp, prison
 	 Score: enemi, armi, soldier, wound, victori, british, prison
Topic 8 Top Words:
 	 Highest Prob: show, fight, love, tell, leav, make, manag
 	 FREX: fight, show, sister, manag, spend, apart, punch
 	 Lift: punch, wrong, let, show, sister, fight, spend
 	 Score: punch, fight, show, sister, apart, perform, busi
Topic 9 Top Words:
 	 Highest Prob: arriv, escap, polic, french, discov, order, inform
 	 FREX: french, polic, escap, letter, inform, german, london
 	 Lift: french, letter, book, polic, arrang, london, situat
 	 Score: french, letter, german, polic, escap, beauti, sing
Topic 10 Top Words:
 	 Highest Prob: kill, play, tell, gun, case, murder, convinc
 	 FREX: gun, play, case, murder, convinc, claim, kill
 	 Lift: case, gun, murder, play, ident, claim, crime
 	 Score: case, gun, murder, play, shoot, crime, kill
```



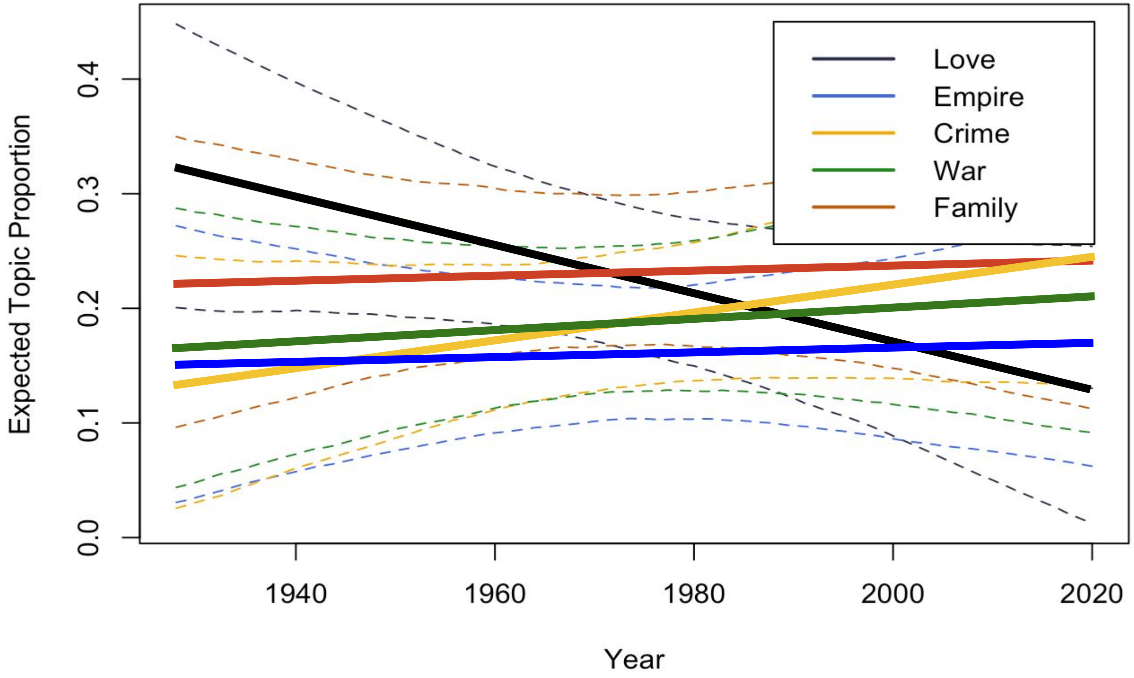


While contents of films were classified into “top topics” through STM using different word choices, some topics (like Topic 3 and 4) were harder to distinguish from the other. In that case, I manually observed associated words with high frequency for the topics as seen on the next slide.

```
Topic 1 Top Words:
 	 Highest Prob: love, show, perform, run, play, tell, miss
 	 FREX: perform, show, miss, run, play, apart, love
 	 Lift: miss, perform, show, bar, interest, star, audienc
 	 Score: miss, show, perform, apart, star, beauti, woman
Topic 2 Top Words:
 	 Highest Prob: king, fight, father, friend, die, leav, refus
 	 FREX: king, fight, court, england, father, win, lead
 	 Lift: king, england, court, fight, express, declar, speak
 	 Score: king, court, fight, father, prison, armi, win
Topic 3 Top Words:
 	 Highest Prob: kill, escap, polic, arriv, water, find, back
 	 FREX: water, polic, escap, gun, shoot, arrest, murder
 	 Lift: water, crime, gun, window, polic, escap, remov
 	 Score: water, polic, escap, shoot, gun, shot, hide
Topic 4 Top Words:
 	 Highest Prob: kill, war, return, man, soldier, armi, find
 	 FREX: soldier, german, team, war, armi, wound, enemi
 	 Lift: team, enemi, u., german, camp, releas, soldier
 	 Score: team, soldier, enemi, armi, u., wound, camp
Topic 5 Top Words:
 	 Highest Prob: famili, return, mother, mr, home, son, children
 	 FREX: famili, mr, children, mother, hous, son, sing
 	 Lift: children, mr, sing, famili, parent, mother, busi
 	 Score: children, famili, mr, mother, sing, son, busi
```



This way, I was able to clearly see the difference between Topic 3 and 4 and assign topics: “Crime” for Topic 3 and “War” for Topic 4.



Looking through the highly associated words for each topic, I assigned movie themes to each and created a visualization. While the movie industry have a stronger preference for Crime, Family, War, and Empire-related films, we see a gradual decrease in the number of Love films over time.

Next, using the Plot data again, I conducted **"textnets"** to investigate whether specific themes from specific decades are closely related to other decades. Merging movies into following 'year_era' by decades, it was interesting to find out that each content of year_era is connected to all other year_era. Looking back, I should've explored further into the **detailed correlation/weights of such connection**, instead of simply verifying the existence of such connection.

```
data_prepped_all <- PrepText(textdata = data, #this is the data
                             textvar = 'Plot', #this is the text column (the first mode)
                             groupvar = 'year_era', #this is the column of groups (the second mode)
                             node_type = 'groups',
                             # another variation = words
                             # this specifies that the nodes will be month-years, and ties between
                             # them will be based on shared words; change 'groups' to 'words' to make the terms nodes
                             # and to draw ties based on whether they share month-years
                             tokenizer = 'words', # which kind of tokenizer to use
                             pos = 'all', #which parts of speech to use (can also be, for example, just 'nouns')
                             remove_stop_words = TRUE,
                             compound_nouns = TRUE #should we save compound nouns (e.g., 'haircut', 'dry-cleaning')?
)
```

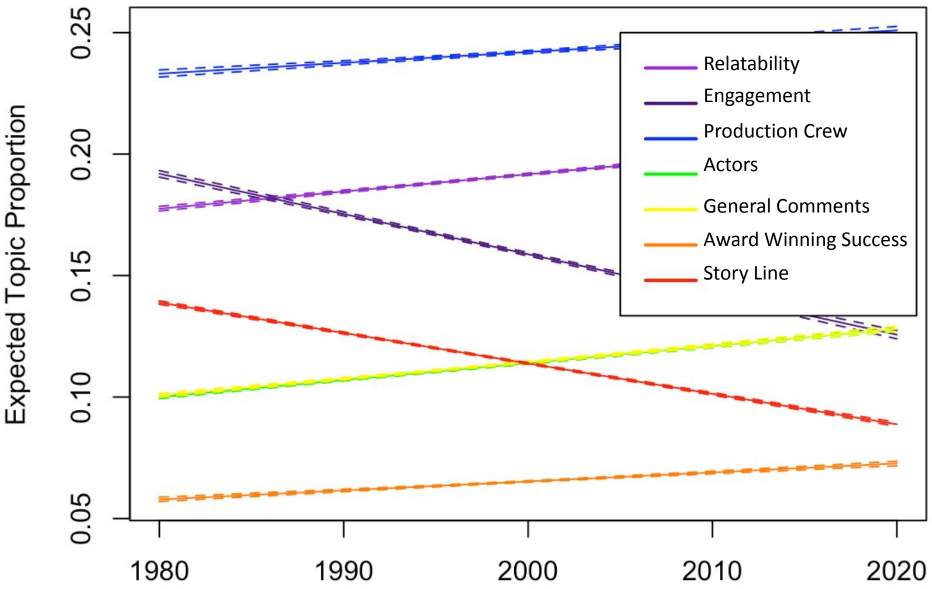
```
 [1] 1920s--1930s 1920s--1940s 1920s--1950s 1920s--1960s 1920s--1970s 1920s--1980s 1920s--1990s
 [8] 1920s--2000s 1920s--2010s 1930s--1940s 1930s--1950s 1930s--1960s 1930s--1970s 1930s--1980s
[15] 1930s--1990s 1930s--2000s 1930s--2010s 1940s--1950s 1940s--1960s 1940s--1970s 1940s--1980s
[22] 1940s--1990s 1940s--2000s 1940s--2010s 1950s--1960s 1950s--1970s 1950s--1980s 1950s--1990s
[29] 1950s--2000s 1950s--2010s 1960s--1970s 1960s--1980s 1960s--1990s 1960s--2000s 1960s--2010s
[36] 1970s--1980s 1970s--1990s 1970s--2000s 1970s--2010s 1980s--1990s 1980s--2000s 1980s--2010s
[43] 1990s--2000s 1990s--2010s 2000s--2010s
```

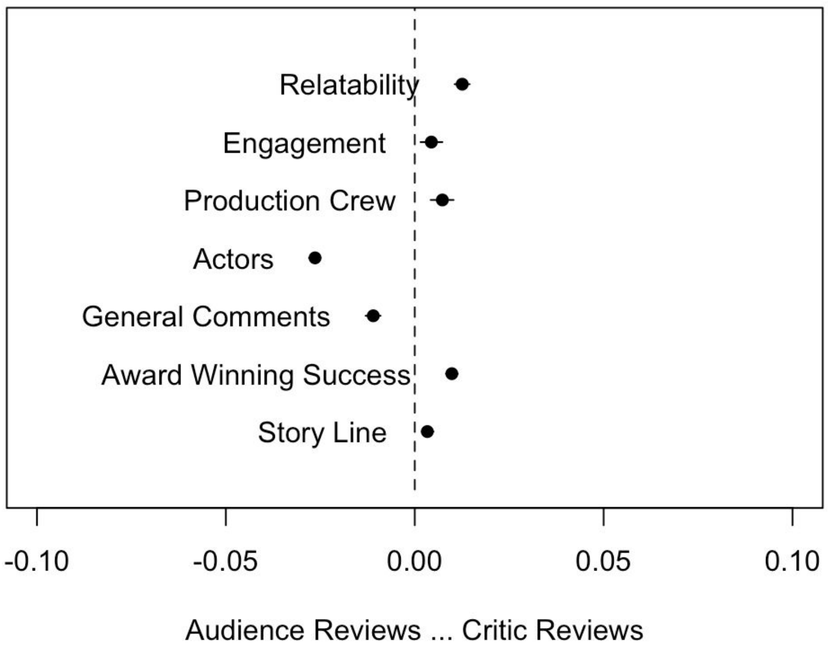
[REVIEWS] WHAT ASPECTS OF MOVIES ARE APPRECIATED THE MOST?

Covariates: Release era & Type of audience/critic

Limitations:

- Vary enormously by genres of the films
- Actors
- Producers
- Set accidents





This time, I used type of reviews(audience VS critic) as the covariate. Audiences discussed more about actors and left casual, generic comments of films, while critic tend to pay more attention to the films’ connection to reality, flow of context, production crew, success performance, and storylines of films.

Conclusion

1) **Contents of movies** across the period are **closely related to each other**, with the absence of drastic changes in specific topics.
2) **STM was not able to strictly produce distinct + clear topic categories**, causing frequent overlaps and similar results within topics.
3) The opposite from what I expected, **the reviews overall were very much casual** and recommendation-based, other than analysis and interpretation.
4) **Poor goal achievement** on “improvements in inclusiveness” & “social atmosphere” - no direct topics on such!

LIMITATIONS

1) Oscar winners, which I used as data, **may not represent the most dominant topic/content** of each year. Hence, it may not be representative of the year_era investigated.
2) Limitations in **understanding the overall flow of the context** ~ as this project classified the topics of movies/reviews by the a number of words (high prob, FREX in STM)
3) As **Plot data is a processed text**, it may be subjective to some extent, causing bias in the overall topic of the movies.

Improvements for next research:

- Usage of **original text** (ex: transcript)
- Usage of better suited analysis methods to examine the **flow of the context**