Andrew Huo
HUM 217
Professor Haverals
05/02/2024

Tracking the Evolution and Success of Film Genres from 1980 to 2024 using GuidedLDA Topic Modeling

INTRODUCTION

When we are asked what films we like, we automatically come up with already-established categories such as “romantic comedies,” “coming-of-age,” or “science fiction/fantasy.” The other person immediately understands the concept and can think of definitive examples and the tropes of each category. Humans are extremely good at establishing order and consistency in defining abstract boundaries, especially in art, literature, and, in the case of this research proposal, film. These categories have evolved over time, largely influenced by the audience's changing preferences. But at a certain point, they begin to follow shared formulas that we call genres. Classifying genres becomes complicated since they are constantly evolving, driven by a symbiotic relationship between the audience (box-office results) and the creators (producer, studios, filmmakers). Andrew Tudor, a professor at the University of York specializing in genre theory, suggests that “Genre is what we collectively believe it to be” (Matthews and Glitre 1512). Genres have vague parameters and are, by definition, unquantifiable. Early genre theory literature proposes that genres enter life cycles from primitive, the establishment of newly formed genre tropes and conventions, to classical, in which these conventions and formulas are mutually expected between filmmakers and audience, to revisionist, subversion and challenging of the genre norms, to finally parodic, where the genre conventions are ridiculed (Matthews and Glitre 1512). Genre lifecycles also coincide with popularity cycles, where particular genres fall out of public interest and resurge, such as the neo-spaghetti Westerns of the 70s or the current fatigue of the superhero genre. Many films fall on a genre lifecycle, popularity cycle, and combine multiple genre tropes. For instance, *Star Wars: Episode IV - A New Hope* (1977) is part western, science fiction, fantasy, family-friendly, and fairy-tale. Many factors go into defining and understanding the film genre and its place in the film canon. But what if there was a way to track the evolution of film genres over recent decades and visualize the data?

This research proposal sets out to quantify the evolution of a constantly evolving concept. What film genres were the most successful (highest-grossing) in a given year or decade? What combination of genre tropes or conventions was most successful? Lastly, based on previous patterns, can we predict the direction in which genres are heading in their lifecycle and popularity cycle in the near future? If so, production studios will be able to have a better understanding of what consumers are looking for, putting more money into predictably successful projects and receiving a higher probability of box-office earnings. On a broader scale, this project could provide better communication between audiences and filmmakers, serving their ever-evolving symbiotic relationship and allowing the audience to become more aware of general film trends. Incorporating the synopses of the 100 highest-grossing films per year (from 1980 to 2024) from the Box Office Mojo by IMDb Pro into a GuidedLDA topic modeling algorithm will categorize each film by a percentage of the most prevalent genre labels. We can then compare the film's breakdown of genre tropes to its monetary box-office success and, therefore, track the evolution of genre popularity over time.

DATASET

The International Movie Database (IMDb), a subsidiary of Amazon, provides public data on the worldwide box office of highest-grossing films on its separate webpage, Box Office Mojo. Ranging from 1977 to 2024, each year displays a ranking of the highest-grossing films. However, the number of films per year varies in the first 10 years, most likely due to a lower production rate in the late 70s and early 80s compared to more recent decades or a lack of information gathered during those early periods. 1977 has a list of only 37 films, slowly increasing to 118 films in 1980 and finally a recurring number of 200 films from 1986 onwards. To create consistency, I decided to use the top 100 films from 1980 to 2024 as my dataset. Each listed film is given a domestic release date (*Barbie*: Jul 21, 2023), opening box office (*Barbie*: $162,022,044), and box office gross (*Barbie*: $636,238,421), and the same three pieces of information for the European, Middle East, African, Latin American, Asia Pacific, and Chinese markets. Moreover, the domestic distributor (*Barbie*: Warner Bros.), MPAA rating (*Barbie*: PG-13), running time (*Barbie*: 1 hr 54 min), and IMDb's genre labels (*Barbie*: Adventure Comedy Fantasy) are also provided and will be used for specific visual comparisons and a basis

for the guided topic modeling algorithm (“Barbie”). Each plot synopsis will be extracted from the specific film page from IMDb’s website. For example, *Barbie* (2023)’s plot synopsis is as follows:

> The film begins with The Narrator (Helen Mirren) explaining the societal impact of the Barbie doll on history, accompanied by a clip of the original 1959 Barbie towering over a desert landscape, while little girls playing with baby dolls begin to destroy them.
>
> Hidden from the real world is Barbieland, where the Barbies and Kens, alongside other dolls like Allan (Michael Cera) and Midge (Emerald Fennell), live. The Barbies preside over Barbieland in a matriarchal system and work high-position jobs while the Kens spend time as futile subordinates living in the Barbies' shadow. Beach Ken (Ryan Gosling) has feelings for Stereotypical Barbie (Margot Robbie) and constantly vies for her attention, but she doesn't recognize.
>
> During a dance party at her house, (Stereotypical) Barbie suddenly starts questioning her mortality. The next day, Barbie suffers an existential crisis, experiencing a series of mishaps including her perfect skin with blemishes and her arched feet going flat. The other Barbies suggest she visit Weird Barbie (Kate McKinnon), who informs her the human girl who is playing with her is unhappy. To fix her crisis, Barbie must travel to the real world, find the girl and help her.
>
> On her way to the real world, Barbie finds (Beach) Ken stowed away in her car. He convinces her to let him join and the two travel to Los Angeles, where they accidentally get arrested several times. Barbie learns to cry upon taking in how flawed the real world is, before complimenting an old woman on her beauty. Ken wonders off and discovers the patriarchy, feeling accepted for the first time. He excitedly travels back to Barbieland.

At a local school, Barbie finds her human girl, Sasha (Ariana Greenblatt), and tries to help her. However, Sasha and her friends condemn Barbie for glorifying bimbo culture and unhealthy life goals, causing Barbie to run off in tears. Meanwhile, the Mattel CEO (Will Ferrell) discovers Barbie's existence and orders her deportation to Barbieland, sending his men after her.

Barbie arrives at Mattel Headquarters and meets the CEO and his male subordinates. They try to send her back via a life-sized doll box, but Barbie deduces their intention and escapes in a pursuit. She is helped by Sasha and her mother Gloria (America Ferrera), who is revealed to be the incitement for Barbie's worries. Gloria started playing with Barbie during a midlife crisis, relinquishing her concerns over to Barbie. The trio, unaware they are being followed by the CEO and his colleagues, travel back to Barbieland.

Arriving in Barbieland, the three find that Ken has led his fellow Kens in overthrowing the system, enslaving the Barbies as compliant girlfriends. They also plan to enshrine their new patriarchy in the Barbieland constitution the next day. Barbie tries to persuade Ken to change it back, but he refuses as he finally feels worthy for the first time. Barbie sinks into a depression before Gloria gives a speech on being a woman, inspiring Barbie to save Barbieland.

Barbie, Gloria, Sasha, Weird Barbie, Allan (who is against the Kens' new rule) and other discontinued Barbies and Kens hatch a plan. Using Gloria's speech, they free the Barbies of their subjugation before turning the Kens against each other to distract them from changing the constitution. As the Kens fight on the beach, the Barbies restore Barbieland's matriarchy back into the constitution.

A distraught Ken expresses disappointment in being nothing other than an accessory to Barbie, to which she encourages him to be his own person. The two apologize to each other for their mistakes. President Barbie (Issa Rae) makes a friendly deal with the Mattel CEO before agreeing to equality for the Kens and

> other discontinued toys. Barbie, still unsure of who she is, meets with the spirit of Ruth Handler (Rhea Perlman), Mattel's co-founder and her creator. Ruth states that Barbie doesn't have a specific purpose, as her evolution will always exceed her roots. She shows Barbie visions of motherhood, encouraging her to choose her own path.
>
> Barbie decides to live in the real world as a human, going by the name Barbara Handler. Gloria and Sasha drive her to an appointment, where Barbie proudly declares she is there to see her gynecologist. (“Barbie (2023) - Plot”)

Film synopses provide a detailed description of the plot and are the most objective written representation of a film (Wang 1). Originally, I set out to run a topic modeling algorithm on each screenplay as that would have the most detailed corpus of words to extract different genre tropes. However, finding each screenplay's final drafts or shooting versions is a logistical nightmare for multiple reasons and is not always the best representation of a finalized film. First, multiple screenplay databases (IMSDb, The Script Lab, Simply Scripts) do not have the rights to all the films in the dataset. Second, screenplays available online may not be the final draft; scripts are even changed at the last minute to accommodate budget constraints, actor and creative conflicts, and sudden issues. Lastly, a screenplay is not always the best representation of a film’s end result because the director can change dialogue, scenes, and other aspects to serve their vision. I decided to focus the corpus of the dataset on plot synopses as they are readily available to the public and are a clear, unbiased interpretation of the film itself and not an element of a film’s production process. Although summaries and synopses serve “as a proxy for the work itself, and [ignore] many other aesthetic and multimodal aspects” (Matthews and Glitre 1521) and thus have limitations, they are the most credible and easily accessible written records of a final film.

IMDb GENRE DEFINITIONS

I will use IMDb’s definitions of genres in its “Help Center” as a starting point for the guided topic modeling algorithm. The webpage states that a genre is a “category of artistic composition, characterized by similarities in form, style, or subject matter for a piece of content” (“IMDB | Help”), and specifically a film genre is a “motion picture category based on the narrative

elements relating to the main driving force behind the story arc" ("IMDb | Help"). They provide a simplified equation to help visualize the different elements that make up a film genre:

Story (Action) + Plot + Character + Setting = Genre

Moreover, IMDb labels each genre as either an "objective genre," which represents facts and set rules, or a "subjective genre," which may be "influenced by viewer opinions, interpretations, points of view, emotions, and judgment" ("IMDb | Help"). Finally, they list 28 genres: action, adult, adventure, biography, comedy, crime, documentary, drama, family, fantasy, film noir, game show, history, horror, musical, music, mystery, news, reality-TV, romance, sci-fi, short, sport, talk-show, thriller, war, and western ("IMDb | Help"). Each genre is given a description prefaced as a "guideline" and a label of "subjective" or "objective" (for the purpose of this research proposal, I will disregard the subjectivity versus objectivity labeling). For instance, an action film:

> Should contain numerous scenes where action is spectacular and usually destructive. Often includes non-stop motion, high energy physical stunts, chases, battles, destructive crises (floods, explosions, natural disasters, fires, etc.) **Note**: if a movie contains just one actions scene (even if prolonged, i.e. airplane-accident) it does not qualify. **Subjective**. ("IMDb | Help")

Each genre description will be a helpful resource for providing seed words to train the GuidedLDA algorithm. For "Action" films, words such as "destructive," "physical," "chase," "battle," and "explosion" will help converge topics into already defined genres, making it easier to focus the algorithm on extracting a combination of different genres.

METHOD

I decided to use topic modeling to extract each genre's probability distribution of terms from a film synopsis. The simplest and most common topic model is latent Dirichlet allocation (LDA) (Blei). Still, since I already have an idea of the topic parameters for each film, taken from the 28 genres listed on IMDb and the multiple labels attached to each film's page, I will use

GuidedLDA. Also known as SeededLDA, GuidedLDA can converge topics into a specific direction by introducing some seed words for every topic ("GuidedLDA"). The program uses an open-source Python library on GitHub (Singh) and essentially adds a "seed" word element to the n-topics decided beforehand ("Topic Modelling"). There is an added parameter: "how much extra boost should be given to a term" (Singh) called seed_confidence. It ranges from 0 to 1, with 0% or 100% bias towards a topic. IMDb-assigned genres will have a seed_confidence of 1 (100%), and the rest of the 28 genre parameters, a total of 28 topics, will be based on how common each genre label is assigned to the 4,400 films in the dataset as a percentage of all assigned labels. Let's say there are 13,200 overall labels, and the "Action" genre label occurs 2,700 times (arbitrary number), then the percentage of action films in the dataset will be ((2,700/13,200)x100) = 19.7%. If the specific film does not have "Action" as one of its IMDb-assigned labels, then the seed_confidence for the Action topic will be 0.197. This is just to provide a gauge to boost the topics in a certain direction, making the guided topics specific for each film synopsis.

Before running the GuidedLDA algorithm, I will prepare each synopsis with stop word filtering. First, I will remove words with 10 occurrences or fewer. Second, I will remove conjunctions, pronouns, articles, and other consistent grammatical filler words. Lastly, film synopses often contain character names and actor names in parentheses, which will also be filtered out. Now, I will run through the method using the example of *Barbie*.

IMDb assigns *Barbie* the genre labels of "Adventure," "Comedy," and "Fantasy." Thus, these three genres will be the primary topics with a seed_confidence parameter of 1, with seed words chosen from the IMDb Genres webpage definitions combined with other specific words associated with the genre.

1. (Seed adventure): journey, hunt, quest, explore, conquest (seed_confidence = 1)
2. (Seed comedy): comedic, humor, satire, parody, slapstick (seed_confidence = 1)
3. (Seed fantasy): magical, mystical, supernatural, mythology (seed_confidence = 1)

The rest of the 25 genre topics will receive similar seed words but a lower parameter of seed confidence.

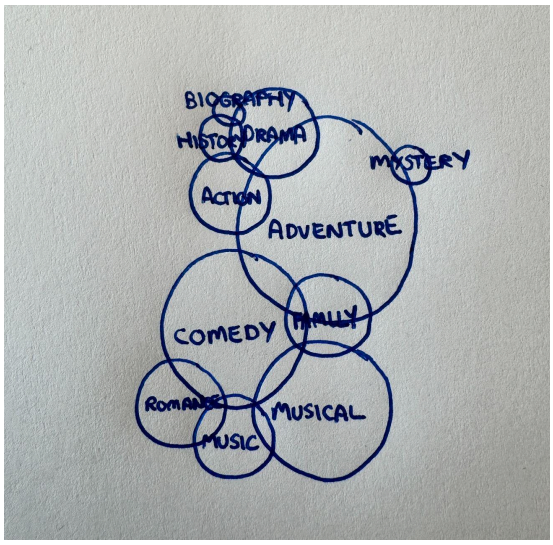
4. (Seed musical): song, dance, musical numbers, perform (seed_confidence = n-musical labels / n-all genre labels)

   And so forth.

Every 13,200 synopses will receive the same treatment, which will build data on the prevalence of genres used in each film. Then, each film will be envisioned as a topic-in-a-box visualization and an interactive line graph with detailed filters.

DATA VISUALIZATION

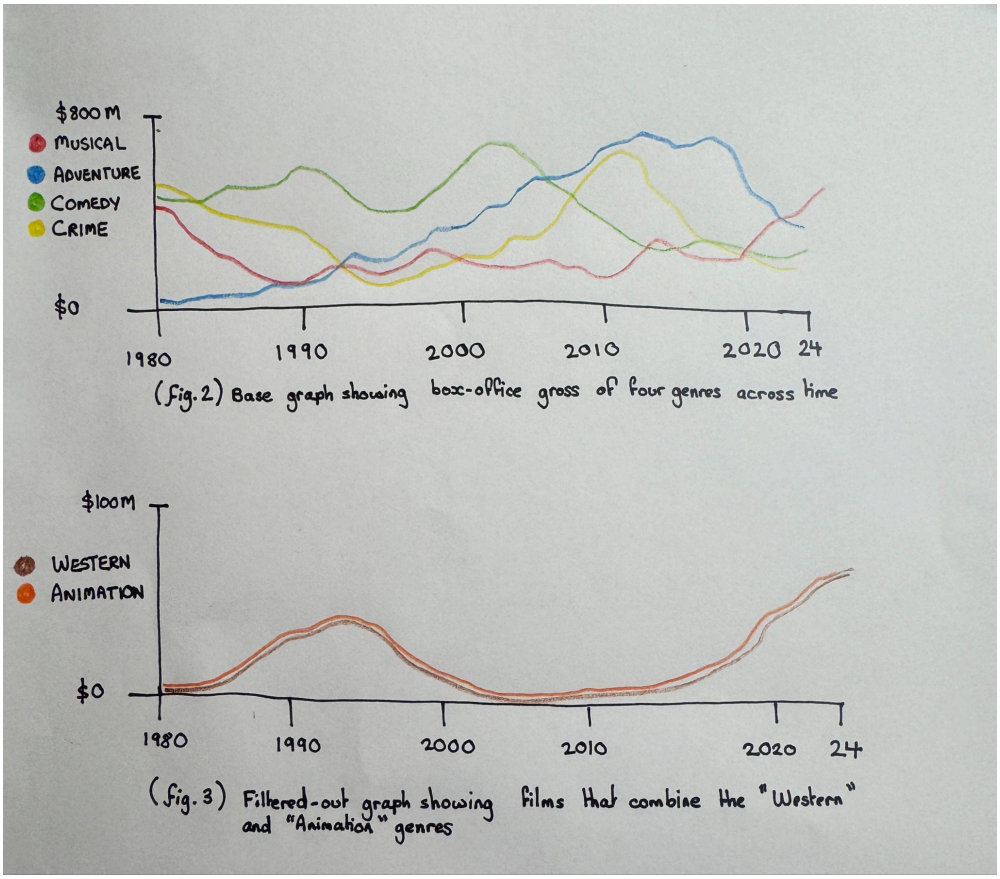
An interactive webpage wil be created to present all findings. First, the topics of each film synopsis will be presented as a topic-in-a-box visualization, which behaves similarly to a word cloud but uses local co-occurrence to choose how topics are connected (Boyd-Graber et al. 40). For instance, Barbie could have a visualized genre makeup of:



(*Fig. 1)* Topic-in-a-box visualization of *Barbie*'s genre makeup

Next, multiple interactive line graphs will be created to represent the success and popularity of genres across time (1980 to 2024). To start off with, the main graph will track the combined

box-office gross (adjusted for inflation) of each of the 28 genres (filtered per month, season, or year from specific release dates) from 1980 to 2024. Each genre will include all films with the topic equal to or more than 25% of the film's overall genre makeup. The graph will also provide zoomed-in information on specific films, such as opening gross, domestic distributor, running time, and MPAA rating, as seen on IMDb's film pages (and can track these elements against box-office gross). The graph will be able to be remodeled for different global markets, and most importantly, the success and popularity of any combination of genres such as (action and sci-fi), (musical, comedy, and adventure), or (western and animation).



DISCUSSION & LIMITATIONS

Hopefully, the visualization of the success of different genres intersecting and their combinations will unearth trends and patterns that will illustrate each genre's lifecycles and popularity cycles. It will be interesting to see which combination of genres was more successful than others at different times. However, the most important purpose of this data visualization would be to predict the direction in which film genres are heading and which combination of genres are

becoming more notorious. Even with the ambiguity and difficulty of finding final screenplay drafts, it would be interesting to see future research focusing on screenplays or the film itself as image and video recognition and machine learning software continue to improve.

Film synopses are a good starting point. However, the have multiple limitations. First, as stated before, they are a human-created representation of a film, not the film itself. The individual writing the synopses can have biases on which descriptions and scenes are more important than others; their subjective opinion of the film can seep through, and they might even have subjective definitions of film genres. Second, the author for each IMDb synopsis is not stated; therefore, it is hard to discern if they are writing from an industry or scholarly perspective of a fan perspective (Matthews and Glitre 1523). Third, the date of the synopsis is unknown. Thus, it is impossible to tell when it was written in relation to the film's release date. Older film synopses could have been written from a retrospective perspective (Matthews and Glitre 1524). It is very difficult to note the changing interpretation and meaning of language and words over the decades.

Overall, this research proposal would be an interesting project for an industry professional, a cinephile, a film scholar, or an average filmgoer. The patterns and evolution of what film genres people are interested in provide a larger lens through which to see the evolution of modern society. Film is one of, if not the most accessible and prevalent form of entertainment. It has multiple impacts on society, from celebrity culture to general pop culture, and therefore, the evolution of its success formula is a very relevant type of data analysis.

Works Cited

"2022 Worldwide Box Office." *Box Office Mojo*, www.boxofficemojo.com/year/world/?ref_=bo_nb_in_tab.

"Barbie." *Box Office Mojo*, www.boxofficemojo.com/releasegroup/gr629756421/?ref_=bo_ydw_table_1.

"Barbie (2023) - Plot - IMDb." *Www.imdb.com*, www.imdb.com/title/tt1517268/plotsummary/?ref_=tt_stry_pl#synopsis.

Blei, David M. "» Topic Modeling and Digital Humanities Journal of Digital Humanities." *Journal of Digital Humanities*, journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/.

Boyd-Graber, Jordan, et al. "Applications of Topic Models." *Foundations and Trends® in Information Retrieval*, vol. 11, no. 2-3, 2017, pp. 143–296, https://doi.org/10.1561/1500000030. Accessed 28 May 2020.

"IMDb | Help." *Help.imdb.com*, help.imdb.com/article/contribution/titles/genres/GZDRMS6R742JRGAG#.

Matthews, Paul, and Kathrina Glitre. "Genre Analysis of Movies Using a Topic Model of Plot Summaries." *Journal of the Association for Information Science and Technology*, vol. 72, no. 12, May 2021, pp. 1511–27, https://doi.org/10.1002/asi.24525. Accessed 16 Mar. 2022.

Singh, Vikash. "GuidedLDA: Guided Topic Modeling with Latent Dirichlet Allocation." *GitHub*, 29 Apr. 2023, github.com/vi3k6i5/GuidedLDA.

---. "How We Changed Unsupervised LDA to Semi-Supervised GuidedLDA." *FreeCodeCamp.org*, 16 Oct. 2017, www.freecodecamp.org/news/how-we-changed-unsupervised-lda-to-semi-supervised-guidedlda-e36a95f3a164. Accessed 4 May 2024.

"Topic Modelling, but with Known Topics?" *Stack Overflow*, 28 May 2013, stackoverflow.com/questions/16782114/topic-modelling-but-with-known-topics. Accessed 4 May 2024.

Wang, Jingcheng. "Using Machine Learning to Identify Movie Genres through Online Movie Synopses." *2020 2nd International Conference on Information Technology and Computer*

Application (ITCA), Dec. 2020, https://doi.org/10.1109/itca52113.2020.00008. Accessed 10 Mar. 2022.