Amazon X-ray mapped with Screenplays and Subtitles

Yeonie Heo*
Computer Science, NYUAD
sh5874@nyu.edu

Safal Shrestha*
Computer Science, NYUAD
ss13750@nyu.edu

Advised by: Minsu Park

ABSTRACT

This study introduces a novel dataset designed to revolutionize cultural studies research through a detailed analysis of movies. Traditional research in this field has been limited by a narrow focus on a select range of cultural products and heavily reliant on textual analysis of screenplays. To address these limitations, our dataset integrates scene-by-scene breakdowns derived from Amazon X-ray with subtitles and complete screenplays sourced from diverse providers. This integration offers a rich, multidimensional view of movie content, enhancing the consistency, availability, and accuracy of data compared to previous methods. By doing so, it allows for a more nuanced exploration of cultural themes across a broader spectrum of films, potentially altering the discourse on the influence of cultural artifacts on societal norms and behaviors. This dataset not only broadens the scope of research possibilities but also deepens the analytical potential, making it a valuable resource for advancing the study of movies as cultural artifacts and opening up new avenues for research.

KEYWORDS

Data, Character Representation, Web Scraping, Computational Social Science

Reference Format:
Yeonie Heo* and Safal Shrestha*. 2024. Amazon X-ray mapped with Screenplays and Subtitles. In *NYUAD Capstone Report Reports, Spring 2024, Abu Dhabi, UAE.* 7 pages.

*These two authors contributed equally to this work.

This report is submitted to NYUAD's capstone repository in fulfillment of NYUAD's Computer Science major graduation requirements.

جامعة نيويورك أبوظبي
NYU ABU DHABI

Capstone Report, Spring 2024, Abu Dhabi, UAE

1 BACKGROUND AND SUMMARY

Previous research across various cultural domains—such as literature, art, and music[1–5]—has been instrumental in uncovering societal norms and cultural intricacies through the ages. This research, however, has often focused intensively on a limited range of cultural products, resulting in a disproportionate emphasis on certain areas. For example, literary studies tend to delve into stylistic elements, historical contexts, and narrative structures, whereas music research primarily examines lyrical content, performance styles, genres, and their societal impacts as evidenced through critiques. This concentrated focus has inadvertently marginalized the broader spectrum of cultural expressions, particularly in film studies.

Research on movies as cultural artifacts, in contrast, remains notably restricted. Most existing studies primarily engage in textual analyses of films, targeting specific demographic groups such as women[6] or individuals with disabilities[7]. These analyses usually prioritize sentiment analysis[8, 9] and frequently overlook wider societal implications, with notable exceptions addressing topics like violence[10] and smoking in films[11]. This narrow approach underscores a significant gap in harnessing the full potential of movies to reflect and influence societal change.

To address these gaps, our goal is to significantly broaden the scope of societal and cultural analysis through movies. By introducing a comprehensive movie dataset, we aim to facilitate research that encompasses a wider range of thematic and stylistic inquiries and encourages the exploration of underrepresented areas within movie analysis.

Screenplays (movie scripts) and subtitles are invaluable resources for movie data analysis, offering rich details such as scene markings, character names, and explicit dialogues. While subtitles, specifically subtitles for the d/Deaf and hard of hearing (SDH), primarily provide dialogues and setting descriptions, like sound descriptions and speaker name, screenplays include dialogues, detailed settings, and technical information essential for shooting. However, their utility in research is hampered by issues of consistency, availability, and accuracy. These documents often exhibit format variations, character mismatches, and revisions. Authentic documents

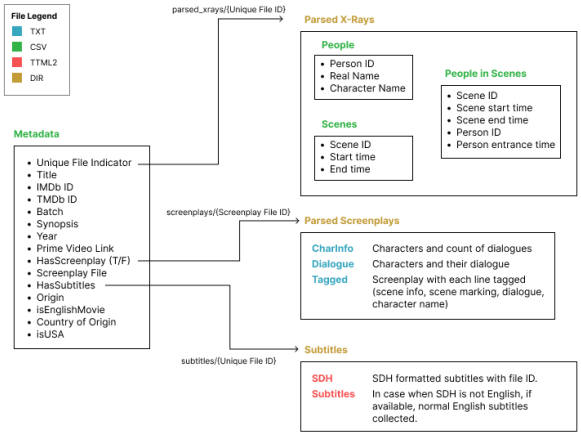

Figure 1: Structure of Augmented Amazon X-Ray Dataset.

are difficult to access, and publicly available scripts often represent early drafts, necessitating extensive preprocessing and cleaning efforts. Moreover, existing screenplays and subtitles data often lack detailed or accurate scene-level character data, particularly regarding inclusivity for non-speaking roles, highlighting a significant gap between scripts and final on-screen content.

To overcome these challenges, we harness advancements in movie metadata documentation through the extraction of extensive longitudinal data from Amazon X-ray. X-Ray is a unique Prime Video feature that provides granular movie data unlike any other streaming platform[12]. It is also recognized as a gold standard for truthful and up-to-date movie data in academia[13, 14]. This feature provides viewers with a dropdown menu displaying a list of cast members currently on-screen, information about the song playing in the background, and interesting trivia about the filming process. Integrated with Internet Movie Database (IMDb) at high accuracy through a partnership, Amazon X-ray facilitates the mapping of actors with their IMDb IDs, resolving challenges faced by previous datasets.

Our dataset comprises this parsed scene-level breakdown extracted from Amazon X-ray data, complemented by subtitles and screenplays sourced from seven script website providers. This combination offers unprecedented depth and breadth for movie analysis, providing a robust tool for cultural scholars and researchers.

2 METHODS

Get Movie Listing from Amazon US

Defining retrieval scope. To start scraping X-Ray data from the Amazon Prime platform, we first established a scope and got a complete list of available videos. Due to intellectual property laws, Amazon Prime Video makes different

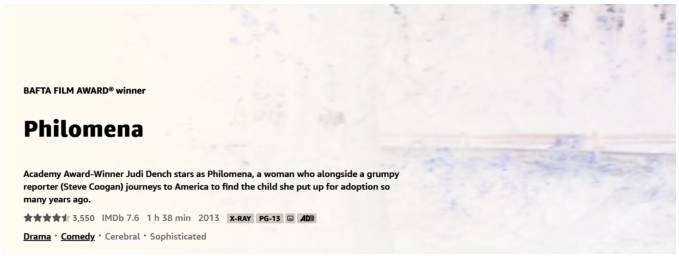

Figure 2: Example of Amazon Prime Video page for a movie <Philomena>.

selections of movie and television titles available in different regions. We provide this data resource for the US market, collecting data from the Amazon US website during the month of August, 2023. At the time of collection, the US Amazon website presented a catalog of movies and TV series under the Prime Video category. We chose to collect only movies that were bundled with Prime, and therefore would not incur an additional cost to users above their Prime subscription. This corpus of movies was then available to the widest audience.

Data retrieval. We used the selenium-wire[15] browser automation library for data collection, an extension of the selenium library[16] with added features to inspect requests and subsequent responses from the browser.

The Amazon US website limited result pagination, allowing browser navigation up to page 400. To circumvent this limitation, we used a filtering approach to access all movies in successive cohorts. Firstly, we collected the titles (with Content-Type "Movies" and "Included with Prime" filters) without applying any additional filtering to our search. We then augmented this initial collection in batches based on their year of release using a decade-based filtering to keep the pages under the 400 limit: movies before 2010, between 2010-2020, and after 2020. During data collection, we found that Amazon's filtering isn't 100% accurate. Therefore, the first scrape increased our data recall.

Processing entries & Duplicates Removal. After processing the entries, we removed the duplicates from the batches we collected. Through manual inspection of the data, we noticed that filtering out duplicates based on title was not adequate. There is a possibility of multiple movies sharing the same title and one movie having multiple titles. Hence, to remove duplicates, we utilized each entity's title and a portion of the unique URL attached to Amazon Prime Video — as visualized in Figure 4. The entries were identified as equivalent if they share the same title portion of the URL and title and distinct if not. Utilizing this heuristic allowed us to remove almost all of the duplicates in our dataset. There were some cases where the same movie would have slight title variations which

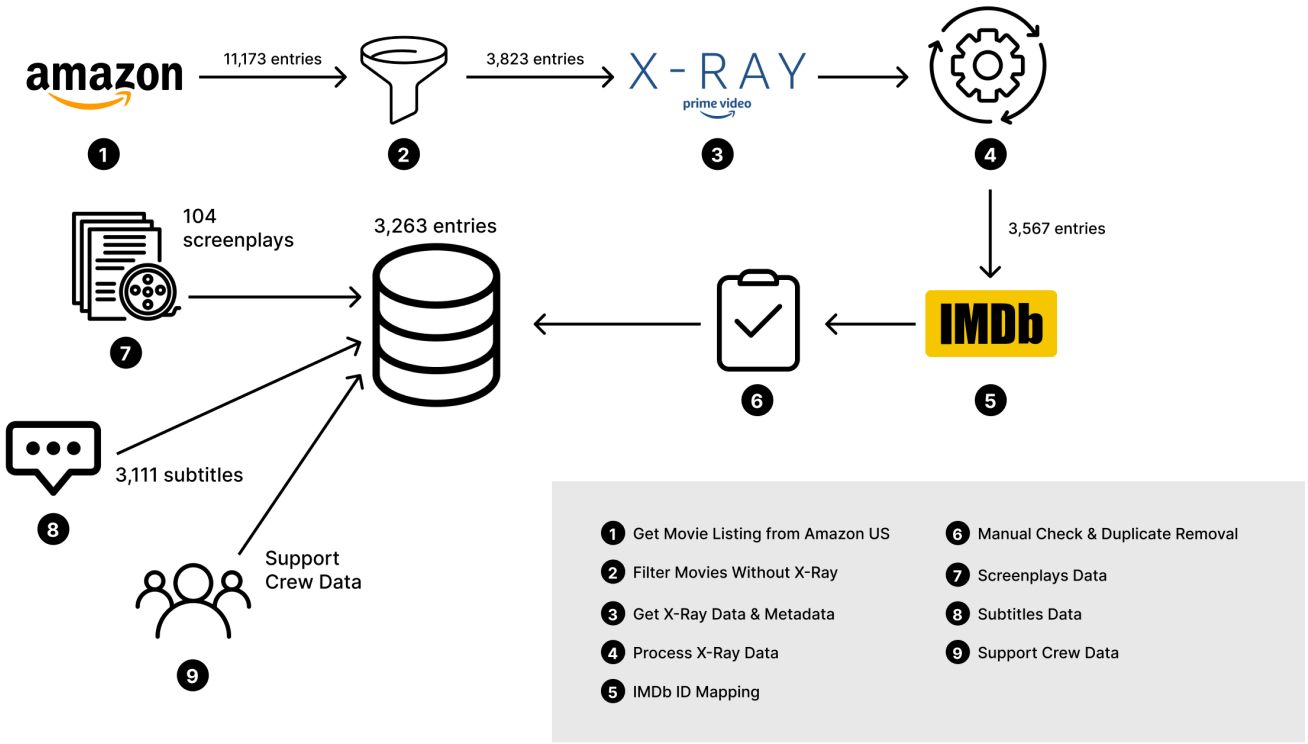

Figure 3: Data Collection Pipeline for Augment Amazon X-Ray Dataset.

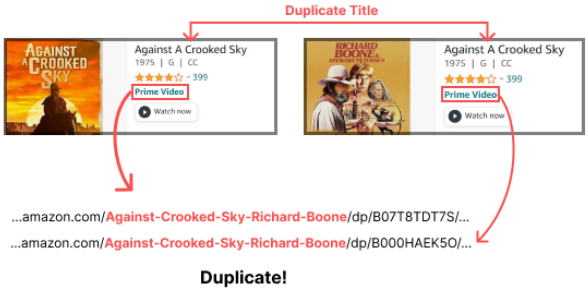

Figure 4: Duplicate Removal Based on Portion of URL.

weren't removed. However, we remove potential duplicates after we finish collecting the unique IMDb id for each movie, as discussed in a later section. In total, we were left with 11,173 entries with links to their respective prime video pages. For each of the entries, we create an identifier using the title of the movie collected at this step. This identifier is used as a directory name to uniquely specify data for each movie; thus, we clean the title thoroughly. We preprocess the title to remove any non-alphanumeric characters with the unidecode library[17] to map any non-ASCII characters to ASCII format and replacing spaces with underscores. We also prepend each movie with the index of the movie in each batch. Therefore, a movie with the title "12 Days with God" maps to "1265_12_Days_with_God".

Filter Movies Without X-Ray

Not all the collected listings on Amazon Prime Video have X-ray data. After extracting the entries, we visited the prime video page of each movie with the browser automation tool to collect additional metadata about X-Ray availability. As shown in Figure 2, the prime video page includes details of each movie, including the title, description, and tags. The "X-RAY" tag denotes whether the movie has X-ray data. After processing such pages with BeautifulSoup[18], we dropped the movies without the "X-Ray" tag and were left with 3,823 entries.

Get X-Ray data and metadata

We collect the details of each entry in our final list by intercepting two network requests. These requests are triggered by clicking on the play button for each movie. Firstly, we intercept the request for 'PlaybackResource'. The JSON response contains metadata for each entry that includes title, entity type (movies), runtime in seconds, synopsis, rating

count on Amazon, rating, subtitle types (subtitle, narrative, or SDH), description, image links, and links to subtitles in possibly multiple languages. The JSON response also has additional information like the rating of the movie, supported audio tracks, and additional metadata unique to the prime video platform. Second, we intercept the request for "X-Ray", the JSON response which contains timestamp information for characters in different scenes. This approach is adapted from a previous work[19]. We save the responses into two JSON files: PlaybackResources.json and Xray.json.

Process X-Ray data

We extract metadata of all movies from their respective PlaybackResource files and put it in a single file. We also parse the X-Ray files into three separate files for each movie. More details about each of the files are available in the Data Records section.

(1) "people.csv" includes a list of actors, corresponding characters, and IMDb tags for actors
(2) "scene.csv" includes a list of scenes and the start and end timestamps of each scene
(3) "people_in_scenes.csv" includes a list of scenes with IMDb ID of people who appear in that scene along with the scene start and end timestamps

IMDB ID Mapping

Linking each retrieved X-Ray movie to its associated (IMDb) ID augments our data with background cast, alternative titles, user ratings, crew and cast information, awards, nominations, quotes, and more. Unfortunately, PlaybackResource information doesn't include IMDb IDs. To match each movie with its corresponding IMDb entry, we devised an algorithm, a simpler adaptation of Ramakrishna et. al. (2017)[20]. Since we have accurate data on the actors playing in a movie along with their IMDb profiles from X-ray data, we can match movies based on this information. We used cinemagoer, a Python package, to retrieve data from IMDb[21]. We first make a search using the title of the movie available in our existing metadata. This results in several matches. We get the top 5 movie matches and starting from the top, we look at the top 5 cast from the cast listing of each movie match. The cast list is ordered either by appearance or alphabetically based on the movie[22]. Previous work has used IMDb cast order as a measure for a cast's relative importance[23, 24]. To add, the cast list from X-Ray data is sourced from IMDb in itself[25], so the top 5 cast provides good matching information. If there's at least one matched actor profile on this top-5 cast list with the list of actors we got from X-ray, we store the movie ID as a match and move on to the next movie title needing an IMDb ID. We also store the matching proportion out of 5 for later validation checks. Among the 3567 movies, we collected IMDB IDs for all movies with varying matching errors except for 441 movies. For such entries, we perform manual matching. We perform additional sanity checks and filtration to mitigate possible errors leaving us with the final dataset of 3,263 movies. We test the IMDb matching accuracy of our dataset by randomly sampling 120 movies and validating their IMDb IDs. We find great results with only 2 errors out of 120 random samples. Details of the validation process is available in the Technical Validation section.

3 DATA RECORDS

Metadata File

- **title**: Title of the movie
- **movie_id**: IMDb ID of movie
- **file**: Unique indicator of a movie in the dataset
- **dir**: Name of batch that data was collected in (*com*, *before2010*, *in2010s*, *after2020*). Data was collected in batches using decade-based filtering. The *com* dataset is the initial collection of data without using any filters. Data such as the subtitles, xrays, and associated metadata for any movie are stored in their respective batch-named directories.
- **synopsis**: Brief synopsis of movie collected from Prime Video
- **year**: Movie's year of release
- **link**: Link to Prime Video page of movie
- **screenplay**: Boolean value indicating whether movie has an associated screenplay
- **screenplay_file**: Unique file name of screenplay if exists else null
- **subtitle**: Indicator of subtitle availability.
  - **SDH**: Movie has English SDH data
  - **SDH_EN**: Movie has non-English SDH data and English non-SDH subtitle
  - **EN**: Movie only has English non-SDH subtitle
  - **Null**: Subtitle data not available
- **origin**: Language of origin of movie, sourced by TMDb, null if data unavailable
- **is_English**: Indicator whether movie is English origin
- **country_origin**: Country of origin of movie, null if data unavailable
- **is_USA**: True if country_origin is United States, False if else or country_origin unavailable
- **tmdb_id**: TMDb ID of movie

Support Crew Data

- **movie_id**: IMDb ID of movie
- **file**: Unique indicator of a movie in the dataset
- **role**: Role of individual in movie, like "cast", "director", etc.
- **person_id**: IMDb ID of person

- **name**: Name of person as listed on IMDb
- **long_canonical_name**: Name in canonical format
- **headshot**: Headshot of person if available

Parsed X-Rays data

Each valid movie has the following three data files:

All People Data.

- **id**: ID of person in X-Ray data in the format (/name/nm<person_imdb_id>/<character_name>)
- **person**: Name of person
- **character**: Name of character in movie

All Scenes Data.

- **scene**: Unique scene identifier in format (/xray/scene/<scene_number>)
- **start**: Scene start timestamp in milliseconds
- **end**: Scene end timestamp in milliseconds

People in Scenes Data.

- **Scene**: Unique scene identifier in format (/xray/scene/<scene_number>)
- **start**: Scene start timestamp in milliseconds
- **end**: Scene end timestamp in milliseconds
- **person_id**: ID of person in X-Ray data in the format (/name/nm<person_imdb_id>/<character_name>)
- **timestamp**: Timestamp of character's first appearance in scene in milliseconds

Screenplays Data

We gather screenplays using the Movie Script Database repository[26]. More information about the tool and generated files are available in the repository.

Parsed Screenplays. There are 3 main files generated about screenplays for each movie. The data is structured as follows:

(1) Character-utterances data:
<Character name>: <no. of utterances>
(2) Dialogue data:
<Character name>=><Dialogue>
(3) Tagged screenplay data:
Screenplay with each line tagged with following markings: S = Scene, N = Scene description, C = Character, D = Dialogue, E = Dialogue metadata, T = Transition, M = Metadata
<Tag>: <Screenplay line>

Example:
C: CHASTITY
D: No.
C: BIANCA
D: You might wanna think about it

Screenplay Metadata & IMDb Matching Stats. After collecting IMDb IDs manually for each movie, we also try matching each character in the screenplay with the character entry on the respective IMDb listing using fuzzy string matching[27]. Along with metadata about the screenplay, we report various matching scores we used to filter out movies with low matching. The final set of movies have a "high_matches" score > 0.6 and "coverage" > 0.7. We pick these numbers through a manual inspection to balance between quantity and quality.

Metadata for movies with high coverage/match score:

- **Imdb**: IMDb id of movie
- **High_matches**: Portion of cast with high matching score (>80). Value ranges from 0 to 1. Calculated as: high_matches = (number of cast with match score > 80) / total_cast
- **Avg_score**: Average match score of all cast in a movie. Value ranges from 0 to 100.
- **Coverage**: Portion of total dialogues left in movie after filtration of erroneous, low-utterance characters. Value ranges from 0 to 1.
- **Title**: Title of movie
- **Script_url**: Link to the screenplay of movie
- **Filename**: Unique filename of the movie
- **Char_fname**: Filename for Character-utterances data for the movie.

Metadata for each character:

- **Imdb**: IMDb id of movie
- **personID**: Matched IMDb ID of character in the movie
- **Name**: Real name of actor/actress
- **Imdb_character**: Name of character on IMDb
- **Script_character**: Name of character in the screenplay
- **Utterances**: Utterances by the character in the screenplay
- **Total_utt**: Total number of utterances in the whole movie
- **Score**: Fuzzy string matching score of the character in the screenplay with the most likely IMDb character (this is the character with the highest matching score)
- **High_matches**: Portion of cast with high matching score (>80) for the movie
- **Avg_score**: Average matching score of all cast in a movie
- **Coverage**: Portion of total dialogues left in movie after filtration of erroneous, low-utterance characters

4 TECHNICAL VALIDATION

4.1 Manual Check and Duplicate Removal

We investigate the unmatched 441 movies to ascertain the primary causes. We noticed that some PlaybackResources

and X-Ray files didn't match the actual movie. Potentially, because of some caching issue, multiple movies shared the same X-Ray and/or PlaybackResources file. We checked if X-Ray files were duplicated based on the cast list extracted from each file. Since the IMDb matching procedure relies on the cast list from X-Rays, a movie won't be matched if the associated X-Ray file is wrong. We removed 94 instances where the IMDb matching was wrong due to this erroneous X-Ray data duplication. We also found 28 duplicated PlaybackResource files by comparing the unique movie ID created at the start of the data pipeline to the movie title collected from the PlaybackResources file. We process the title from the PlaybackResources file similarly to how we processed the unique movie IDs and compare them. In cases where PlaybackResources file is a duplicate, we wouldn't be able to match the movie since the title we get from it is wrong. We manually checked each entry with this problem to ensure that the X-Ray files and subtitles were correct. We fixed the metadata for these movies manually. After removing movies with erroneous X-Ray data and fixing metadata because of faulty PlaybackResources file, we matched the remaining unmatched movies manually. We make use of metadata like the title, description, year of release, and cast to get the IMDb IDs. In this manual process, we also removed entries that weren't necessarily traditional movies, like stand-up comedy specials and anthologies. Such entries are not movies and have a different narrative structure than movies which are our primary focus. After the IMDb ID collection and manual revision, we removed duplicates based on the IMDb IDs and were finally left with a dataset of 3,263 movies.

4.2 Coverage Assessment

The robustness and applicability of a dataset significantly hinge on its coverage and representativeness. To evaluate this, we extracted the top 100 movies by popularity from IMDb for each decade, serving as a benchmark for assessing the inclusiveness of our dataset. Initially, coverage within our dataset was sparse, particularly in earlier decades with only select years like 1931, 1932, 1936, and 1939 represented in the 1930s. However, from the 1950s onward, our dataset includes movies from every year within each decade. The coverage of top movies varies, ranging from 1% to 11%, and generally shows a progressive improvement over the decades.

This level of coverage, while moderate, confirms that our dataset aligns reasonably well with audience preferences, paving the way for potentially insightful research in both academia and industry using the augmented Amazon X-Ray dataset.

To further enhance the dataset's coverage, especially for more recent decades, we plan to implement periodic updates. These updates will involve scraping additional data as it becomes available and refining our collection methods to focus more on recent productions. Additionally, exploring partnerships with movie databases and production companies could provide better access to recent, high-quality metadata and screenplays. This proactive approach will ensure our dataset remains a dynamic and valuable resource for cultural analysis and movie studies.

Table 1: Top 100 movies by decade coverage in Augmented Amazon X-Ray Dataset

Decade	Years covered in dataset	Total movies in dataset	Dataset movies covered in top 100 list
1930s	1931, 1932, 1936, 1939	6	1
1940s	1940, 1941, 1944, 1945, 1947, 1948, 1949	10	2
1950s	All years	24	7
1960s	All years	33	5
1970s	All years	54	11
1980s	All years	91	3
1990s	All years	145	4
2000s	All years	327	10
2010s	All years	1400	9
2020s	All years	1171	7

Comparing coverage rate against screenplays in Table 2, we find that screenplays cover much more of the top-100 movies of each decade. This is a reasonable conclusion since popular movies gain more attraction from movie fanatics leading to screenplays being released or transcribed by third parties. Similar to X-Ray dataset, we notice that fewer screenplays are available for older movies while data from recent movies are available in abundance. Still, X-Ray dataset covers a larger ground and because of the automated nature of X-Ray data generation, X-Ray data for new movies is readily available once it releases on Prime Video. Prime Video, as a platform, strategically chooses content to be kept or removed from the platform based on audience preferences. Thus, data from such a source acts as a good representation of viewer proclivity.

5 CONCLUSION

The field of cultural analysis has long been crucial in uncovering societal norms and nuances. Previous work has mainly relied on data sources like music and books to perform a range of possible analyses like stylistic, historical context, narrative structure, and impact analysis. Movies have also been a popular choice of answering cultural questions, but

Table 2: Top 100 movies by decade coverage in Screenplays Dataset

Decade	Years covered in dataset	Total movies in dataset	Dataset movies covered in top 100 list
1930s	1930, 1931, 1932, 1933, 1934, 1938, 1939	11	7
1940s	1940, 1941, 1943, 1944, 1945, 1946, 1948, 1949	22	9
1950s	All years	24	18
1960s	1960, 1961, 1962, 1963, 1964, 1965, 1967, 1968, 1969	23	11
1970s	All years	59	25
1980s	All years	124	25
1990s	All years	333	49
2000s	All years	390	50
2010s	All years	362	46
2020s	2020, 2021, 2023	19	4

availability of rich data is scarce. The current methodologies rely on screenplays, subtitles, or tedious manual work. Especially screenplays and subtitles are inconsistent, unavailable, and inaccurate. To fill in the current gap in academia of rich movie data to answer more nuanced questions, we propose the Amazon X-Ray Dataset. Our dataset includes not only Amazon X-ray data (timestamped character/actors) but also mapped screenplays with IMDb ID. This dataset hopes to inspire more nuanced questions not only about movies but cultural dynamics as a whole by contributing a major resource.

REFERENCES

[1] Griswold, Wendy. "American character and the American novel: An expansion of reflection theory in the sociology of literature." American Journal of Sociology 86.4 (1981): 740-765.

[2] Rideout, Walter B. The Radical Novel in the United States 1900–1954: Some Interrelations of Literature and Society. Columbia University Press, 1992.

[3] Martin, Peter J. Sounds and society: Themes in the sociology of music. Manchester University Press, 1995.

[4] Hesmondhalgh, David. Why music matters. John Wiley & Sons, 2013.

[5] Longhurst, Brian. Popular music and society. Polity, 2007.

[6] Okamura, Hitomi. "Movies and women: an overview of women's position in the changing society in the 1920s." Doshisha literature 32 (1986): 173-195.

[7] Al-Zoubi, Suhail Mahmoud, and Samer Mahmoud Al-Zoubi. "The portrayal of persons with disabilities in Arabic drama: A literature review." Research in developmental disabilities 125 (2022): 104221.

[8] Elkins, Katherine, and Jon Chun. "Can Sentiment Analysis Reveal Structure in a Plotless Novel?." arXiv preprint arXiv:1910.01441 (2019).

[9] Kiritchenko, Svetlana, and Saif M. Mohammad. "Examining gender and race bias in two hundred sentiment analysis systems." arXiv preprint arXiv:1805.04508 (2018).

[10] Gao, Xiaotong. "Anti-war Movies and Its Negative Effects on Society." 2022 3rd International Conference on Mental Health, Education and Human Development (MHEHD 2022). Atlantis Press, 2022.

[11] Charlesworth, Annemarie, and Stanton A. Glantz. "Smoking in the movies increases adolescent smoking: a review." Pediatrics 116.6 (2005): 1516-1528.

[12] Amazon.com Press Release. For the First Time Ever, X-Ray for Movies and TV Shows Now Available Directly on Your TV — Answer the Classic MovieWatching Question "Who's That Guy?" with Your Amazon Fire TV.

[13] Kagan, Dima, Thomas Chesney, and Michael Fire. "Using data science to understand the film industry's gender gap." Palgrave Communications 6, no. 1 (2020): 1-16.

[14] Bober, Miroslaw, et al. "BRIDGET: An approach at sustainable and efficient production of second screen media applications." (2015): 9-9.

[15] Selenium Wire. Retrieved August, 2023. https://pypi.org/project/selenium-wire/

[16] Selenium. Retrieved August, 2023. https://www.selenium.dev/

[17] Unidecode. Retrieved August, 2023. https://pypi.org/project/Unidecode/

[18] "Beautiful Soup" n.d. https://beautiful-soup-4.readthedocs.io/en/latest/

[19] Poggel, Lisa, and Frank Fischer. 2022. "Automatic Extraction of Network Data from Amazon Prime Videos (Using '1917' as an Example)." Weltliteratur.net. February 16, 2022. https://weltliteratur.net/extracting-network-data-from-amazon-prime-videos/

[20] Ramakrishna, Anil, Victor R. Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. "Linguistic analysis of differences in portrayal of movie characters." In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1669-1678. 2017.

[21] Cinemagoer. https://cinemagoer.github.io/

[22] IMDb (2019a) How are cast credits ordered? why don't the main stars appear at the top of the cast? Accessed 25th April, 2024. https://help.imdb.com/article/contribution/filmography-credits/how-are-cast-credits-ordered-why-don-t-the-main-stars-appear-at-the-top-of-the-cast/G39K5N4YYV2QJ4GR?ref_=helpsect_pro_3_4#

[23] Kagan, Dima, Thomas Chesney, and Michael Fire. "Using data science to understand the film industry's gender gap." Palgrave Communications 6, no. 1 (2020): 1-16.

[24] Tran, Quang Dieu, and Jai E. Jung. "CoCharNet: Extracting Social Networks using Character Co-occurrence in Movies." J. Univers. Comput. Sci. 21, no. 6 (2015): 796-815.

[25] "Introducing 'X-Ray for Movies,' Powered by IMDb and Available Exclusively on the All-New Kindle Fire Family." 2012. September 6, 2012. https://press.aboutamazon.com/2012/9/introducing-x-ray-for-movies-powered-by-imdb-and-available-exclusively-on-the-all-new-kindle-fire-family

[26] Saha, Aveek. "Movie Script Database." Last modified July 2021. GitHub. https://github.com/Aveek-Saha/Movie-Script-Database

[27] TheFuzz. Accessed April 25, 2024. GitHub. https://github.com/seatgeek/thefuzz