Digital Humanities Project: Sexism in Movies

Intermediary Report (4th February 2021)

Harshita Sharma
20171099

About the Project

The aim of this project is to understand and analyse how movies project sexism and how widespread it is over the course of time, in different genres etc. primarily by examining the dialogues from movie scripts using Machine Learning and NLP.

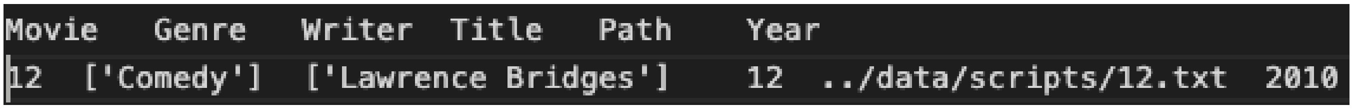
Project Milestones and Progress

	Tasks	Progress	Comments
1	Reading research papers related to the analysis of sexist texts, analysis of movie scripts etc.	Ongoing	
2	Data Collection	Complete	Web scarped 1137 movie scripts from IMSDb. Collected meta-information from IMDb and IMSDb.
3	Data Visualisation	Complete	Non-uniform distribution both by year and by genre - might affect the analysis.
4	Basic Pre-processing of scripts	Ongoing	Identifying different parts of scripts: Speaker, Narration, Dialogue
First (Intermediate) Meeting: 4th February			
5	Task 1: Number of dialogues by male/female		
6	Pre-processing for Task 2		
7	Task 2: Classifying dialogues as sexist or not		
8	...		

Data Collection

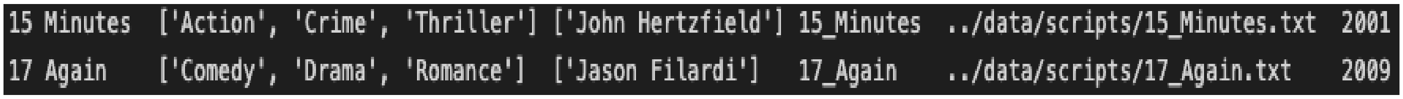
1. Collected 1210 scripts out of which ~1145 had a similar format and were available in HTML. On further examination of the scripts, ~8 were empty. Finally, 1137 scripts have been finalised(by the first intermediate meeting).
2. Meta-information regarding the scripts like Movie Title, Genre, Writers was collected from the same website.
3. The year of release was collected separately using IMDb library.
4. Missing or non-uniform information was checked and corrected manually.
5. The dataset ranges from the year 1915-2022. These years mark the year of release of the movie.
6. The dataset covers movies belonging to a total of 23 genres listed below:

   Comedy, Romance, Drama, Sci-Fi, Thriller, Adventure, Action, Crime, Horror, Mystery, Animation, Fantasy, Family, Musical, Western, War, Biography, Music, Film-Noir, History, Short, Sport, Action.Thriller, Horror.Mystery.
7. The meat-information of the movies in the dataset is organised is a CSV file like so:



```
Movie   Genre   Writer  Title   Path    Year
12  ['Comedy']  ['Lawrence Bridges']    12  ../data/scripts/12.txt  2010
```

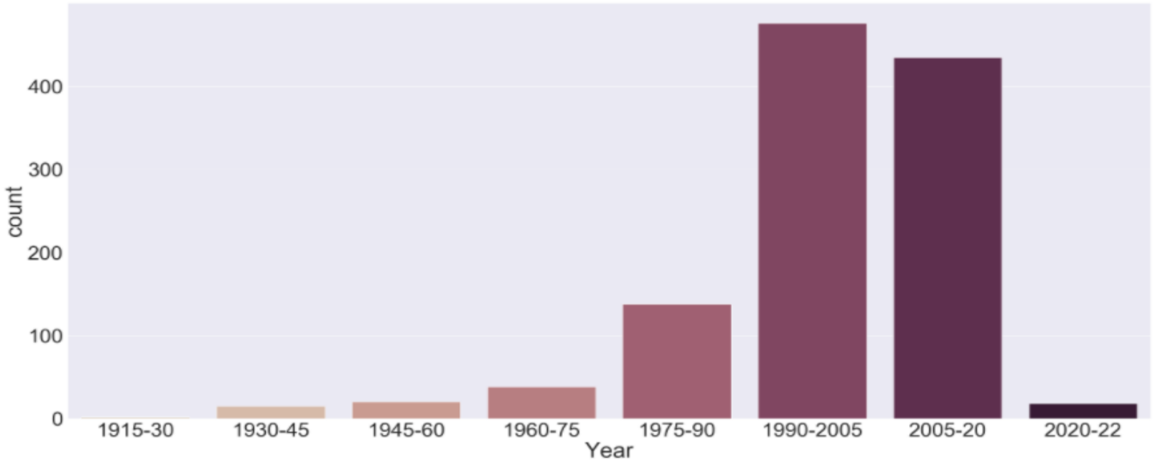
More examples show that a movie can be categorised in a number of genres:



```
15 Minutes  ['Action', 'Crime', 'Thriller'] ['John Hertzfield'] 15_Minutes  ../data/scripts/15_Minutes.txt  2001
17 Again    ['Comedy', 'Drama', 'Romance']  ['Jason Filardi']   17_Again    ../data/scripts/17_Again.txt    2009
```

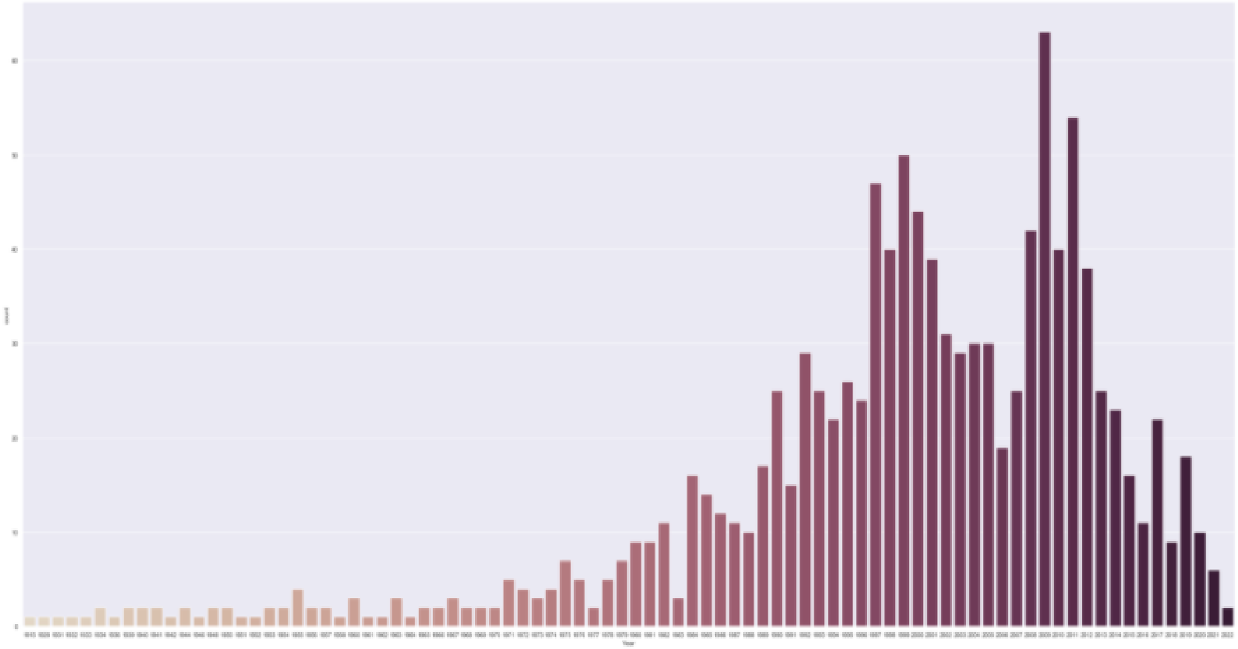
Data Visualisation

1. Visualising the collected data by year (groups of15 years): The dataset was divided into groups, grouped by year e.g. 1915-30, 1930-45…..2005-20, 2020-22.



Most of the movies in the dataset are from 1990-2020 counting up to 907 movies out of 1137 whereas 1915-30 has movies as little as 2. For more clarity on the distribution of the dataset, we look at the next visualization.

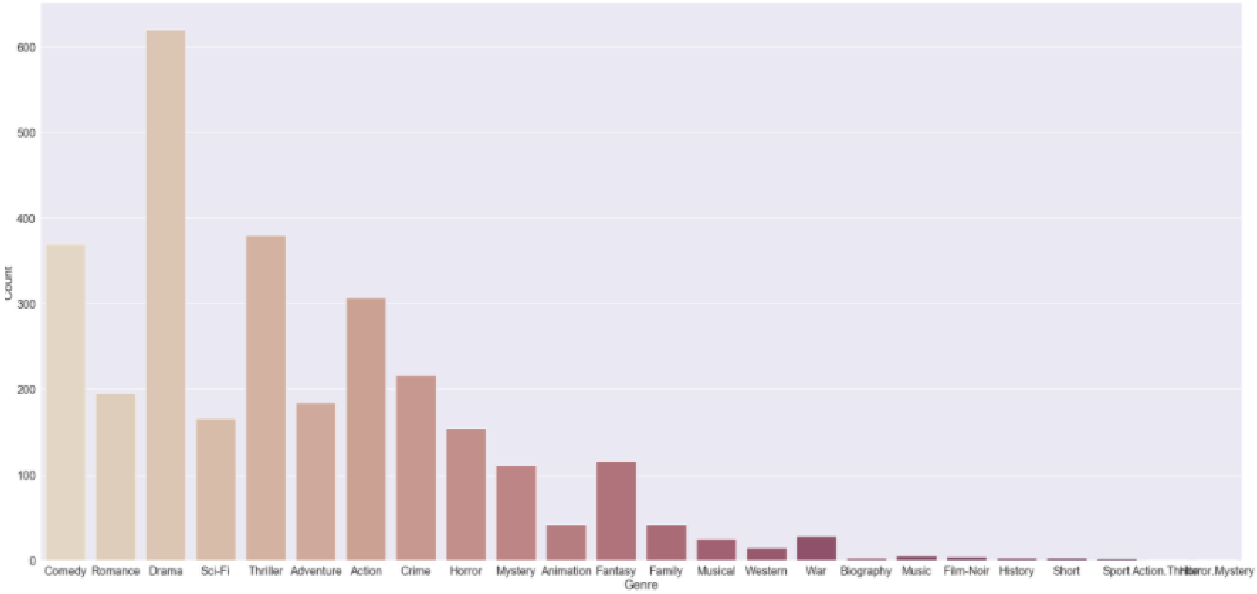
2. Visualising the collected data by exact year of release: The following figure shows the non-uniformity in the dataset if we look at the year of release of movies.



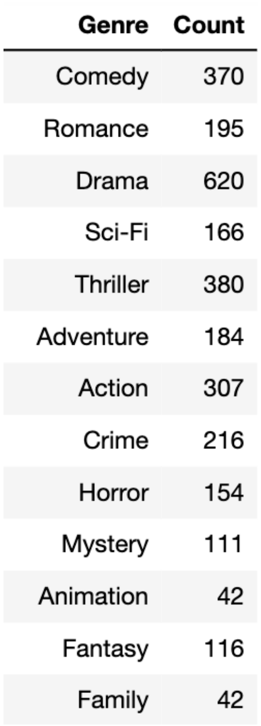
Some years in the dataset with the highest number of films:

2009	63
2011	54
1999	49
1997	47
2000	44
2008	42
2010	40
1998	40
2001	39
2012	37

3. Visualising the collected data by genre: Again, a non-uniformity can be seen in the distribution of movies over the different genres.



Genres and the number of movies in each genre:[1]



Genre	Count
Comedy	370
Romance	195
Drama	620
Sci-Fi	166
Thriller	380
Adventure	184
Action	307
Crime	216
Horror	154
Mystery	111
Animation	42
Fantasy	116
Family	42

[1] Each movie can have more than one genre so the sum of the counts will be greater than the number of movies