Movies Dataset

Jonathan Thomas, Conner Hsieh, Grady Ta, Athish Kumar

Objective

- Look for growth rate across multiple areas of film, such as actors, directors, amount of movies made, number of cast members and the years tied to that data.

Trends that we looked for

- If any growth plateaued
- If how movies grew across the years

Meta data - Movie Dataset

Abstract: This data set contains a list of over 10000 films including many older, odd, and cult films. There is information on actors, casts, directors, producers, studios, etc.

The central file (MAIN) is a list of movies, each with a unique identifier. These identifiers may change in successive versions. The actors (CAST) for those movies are listed with their roles in a distinct file. More information about individual actors (ACTORS) is in a third file. All directors in MAIN are listed in a fourth file (PEOPLE), with a number of important producers, writers, and cinematographers. A fifth file (REMAKES) links movies that were copied to a substantial extent from each other. The sixth file (STUDIOS) provides some information about studios show in MAIN. This documentation file provides supplementary information, and is an essential part of the database. Some images are also available, but not on-line now. There are many cross-linking names throughout the files, so many in fact that web browsers have choked when trying to make them live. Currently we simply try to be careful in naming films and people consistently. More detail follows below. Counts have been made using emacs `count-matches'.

MAIN: 11400 entries

PEOPLE: 3290 entries

CASTS: 46000 entries

ACTORS: 6800 entries

REMAKES: 1278 entries

STUDIOS: 200 entries

Query for top 10 directors who directed the most movies

grep '<td> D:' main.html | awk '{print $2}' | sort | uniq -c | sort | tail

Using grep got rid of all lines not involving directors, and then awk prints the column I want.

queriedData > directors.txt

```
1    34 D:M.LeRoy
2    36 D:DeMille
3    36 D:Griffith
4    36 D:Hiller
5    36 D:J.Ford
6    40 D:Lubitsch
7    41 D:Curtiz
8    53 D:Cukor
9    53 D:R.Stevens
10   80 D:Hitchcock
11
```

Query for amount of movies remade each year

```
awk -F"<td>" '{
    for(i=1; i<=NF; i++) {
        if($i ~
/^[[:space:]]*[0-9]+[[:space:]]*$/) {
            printf "%s\n", $i
        }
    }
}' remakes.html | sort -n | uniq -c >
movieRemakeYears.txt
```

This command splits up the file with the field separator and then prints the column containing the year because it's the only column with numbers.

```
36  1930
36  1956
36  1994
37  1937
38  1951
38  1959
38  1976
38  1990
39  1931
39  1957
40  1992
41  1941
41  1944
41  1955
42  1953
44  1948
44  1993
46  1936
47  1933
47  1949
49  1974
50  1946
52  1950
54  1954
59  1934
59  1938
62  1935
62  1940
73  1932
77  1939
```

Query for actors with the most movies

Command used:

sort casts.html | grep ‘<tr><td>’ | awk -F’<’ ‘{print $5}’ | awk -F’>’ ‘{print $2}’ | sed -e ‘s/^ //’ -e ‘s/ a//’ | sort | uniq -c | sort -n >> actors_with_most_movies.txt

Top 5:

56 Victor Prince

56 Humphrey Bogart

52 James Stewart

50 Henry Fonda

50 Gary Cooper

```
 46 John Carradine
 47 Burt Lancaster
 48 Cary Grant
 50 Gary Cooper
 50 Henry Fonda
 52 James Stewart
 56 Humphrey Bogart
 56 Vincent Price
```

Sorts then greps all lines with actor data, then awk file to get actors, sed to remove extra spaces and unwanted strings. Sort file then get count of each actor occurrence

Query for movies with the largest cast

```
awk '{if(substr($1,0) ==
"<tr><td>") print $0}'
casts.html | awk -F'>'
'{print $4}' | awk -F':' '{print
$2}' | awk -F'<' '{print $1}' |
sed -e 's/^ //' | uniq -c |
sort -r
```

Used awk to find the right string header then used different delimiters to cut down to the movie names. Finally used sort and uniq -c to get the count

```
46 War and Remembrance
46 Around the World in 80 Days
34 The Longest Day
32 The Cannonball Run
31 Variety Girl
31 The Player
31 On Her Majesty's Secret Service
28 Coming to America
26 The Man in The Gray Flannel Suit
25 The Right Stuff
25 Star Trek
24 The Roaring Twenties
24 Nashville
24 Foreign Correspondent
23 The Naked Gun 2 1/2
23 Maid of Salem
23 Jamaica Inn
23 How the West Was Won
23 From the Earth to the Moon
23 Deep Impact
23 Blood on the Sun
23 Black Legion
22 Voyage of the Damned
22 Henry V
22 David Copperfield
21 Wrong is Right
21 Those Magnificent Men in Their Flying Machines
21 The Big Broadcast of 1936
21 Tales of Manhattan
21 Hollywood Canteen
21 Deep in My Heart
21 Casablanca
20 Tora! Tora! Tora!
```

Query for top years with the most movies

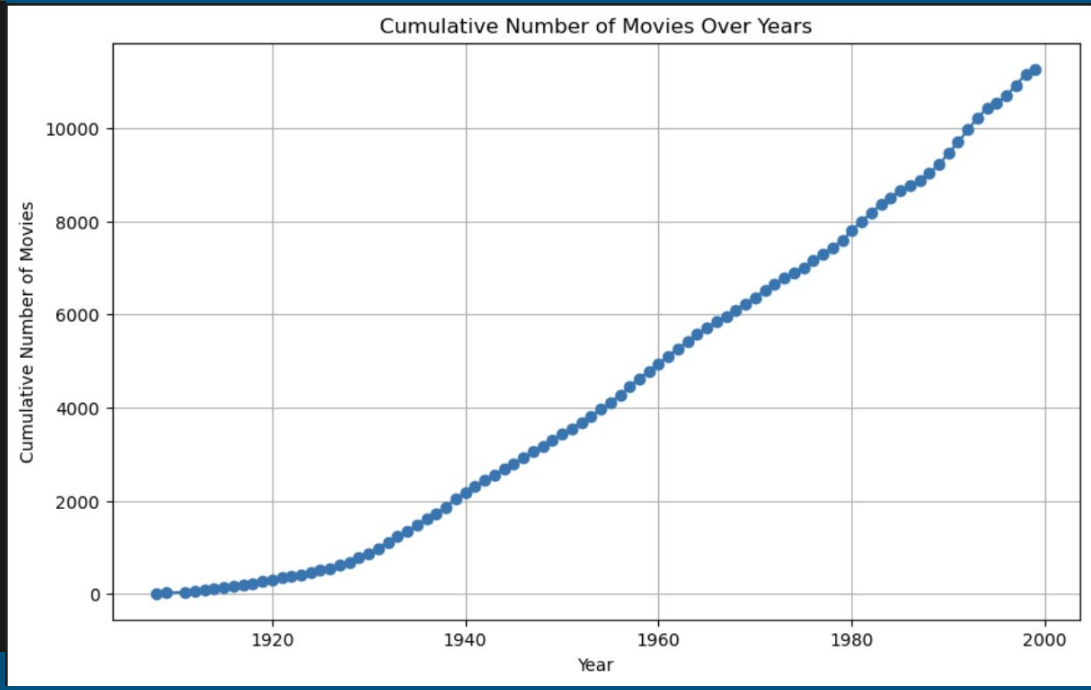
```
awk '{if(substr($1,0) == "<tr><td>") print $0}' main.html | awk -F':' '{print $2}' | sed -e 's/<td>//g' -e 's/.$//' -e 's/^ //' | awk '{print $NF}' | sed 's/[a-z]//g' | sort | uniq -c | awk '{if(substr($2,0,1) == 1) print $0}' | sort -r
```

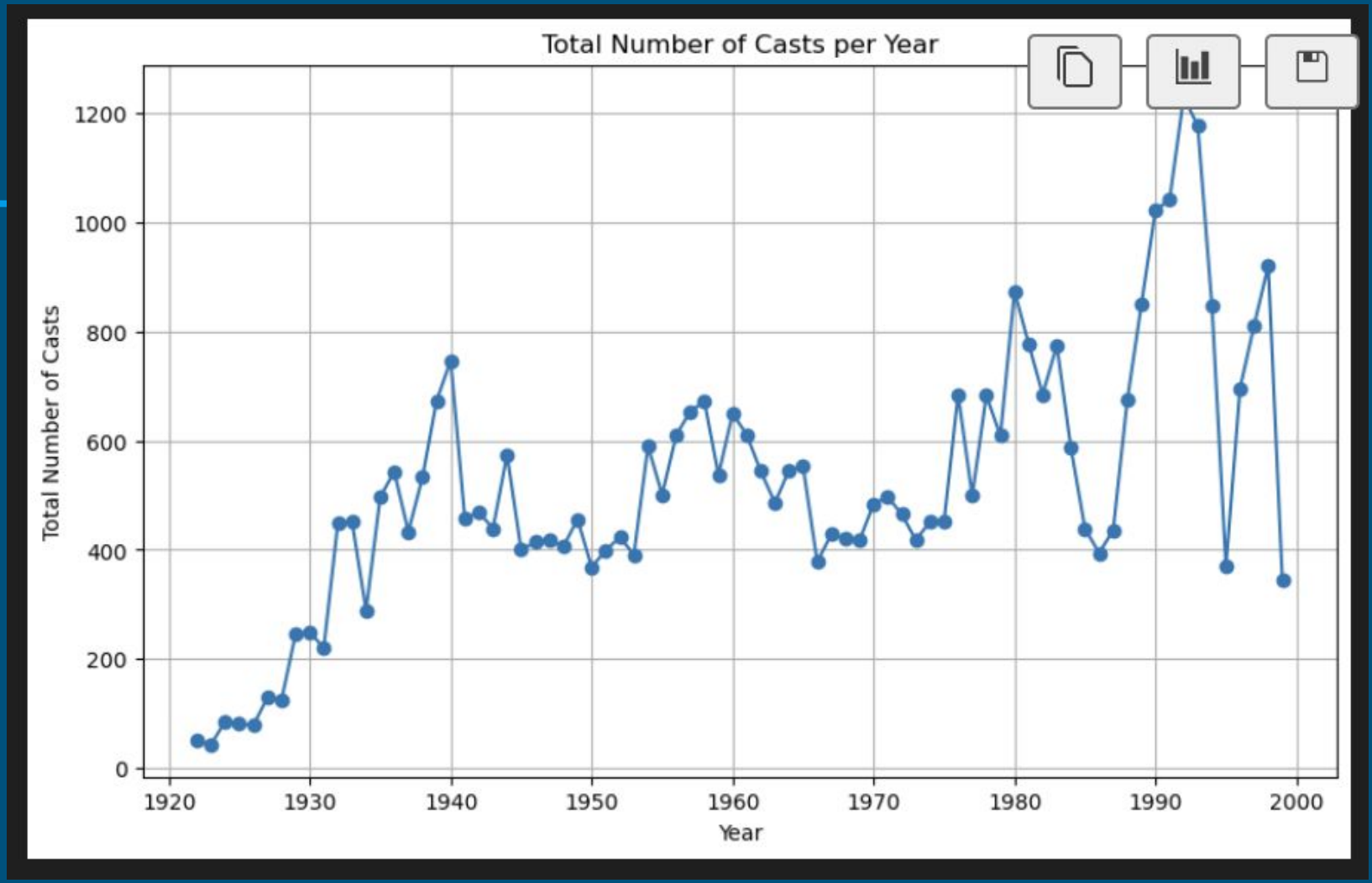
Used awk to find the heading to get the lines with the correct info. Then used awk and sed to cut down the information into just the years that movies came out. Final sorted and used uniq -c to get the count

```
1     264 1992
2     260 1993
3     249 1998
4     245 1990
5     239 1991
6     224 1980
7     207 1997
8     206 1994
9     187 1981
10    186 1989
11    184 1983
12    184 1982
13    184 1957
14    181 1996
15    179 1958
16    167 1961
17    166 1960
18    166 1939
19    164 1962
20    158 1964
21    158 1940
22    156 1976
23    156 1956
24    153 1978
25    152 1988
26    149 1965
27    148 1971
28    148 1955
29    147 1985
30    147 1963
```

Visualizations

```python
import matplotlib.pyplot as plt

# Read data from file
data = []
with open('numMoviesPerYear.txt', 'r') as f:
    for line in f:
        if line.strip():  # Ignore empty lines
            num_movies, year = map(int, line.split())
            data.append((year, num_movies))

# Sort data by year
data.sort(key=lambda x: x[0])

# Calculate cumulative sum of number of movies
cumulative_movies = [0]
for year, num_movies in data:
    cumulative_movies.append(cumulative_movies[-1] + num_movies)

# Plotting
plt.figure(figsize=(10, 6))
plt.plot([x[0] for x in data], cumulative_movies[1:], marker='o', linestyle='-')
plt.title('Cumulative Number of Movies Over Years')
plt.xlabel('Year')
plt.ylabel('Cumulative Number of Movies')
plt.grid(True)
plt.show()
```




Total Number of Casts per Year
Total Number of Casts
Year
0
200
400
600
800
1000
1200
1920
1930
1940
1950
1960
1970
1980
1990
2000

Conclusion

We found that the growth trends did not plateau and instead continued increasing over time. This could be due to rapid development of technology making it easier to produce movies.