Isaac Lake
3-7-2025
CSC 369

Letterbox Analysis

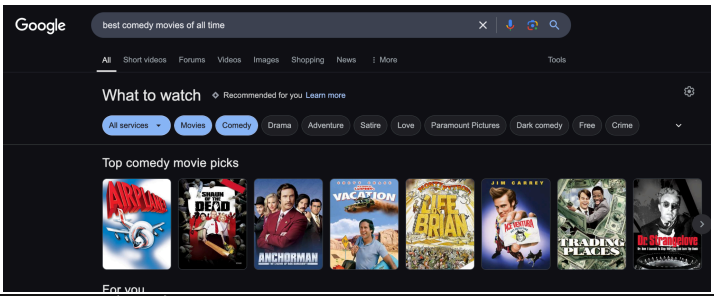
Introduction:

I initially took upon the task of exploring the letterbox dataset with the intention of unraveling the biases present in my previous movie search engine of choice, Google. In the process I determined that my initial theory of Google being biased towards certain studios was untrue but realized a couple of other biases, both in Google's results and the letterbox dataset itself.

Data Acquisition:

My initial plans involved scraping thousands of results from google and comparing them to results of my sql queries to find any discrepancy. Unfortunately after a couple of hours trying this I realized scraping google was more trouble than it was worth, and a bit outside of my wheelhouse. In the end I had to manually copy results from my browser and was thus limited to 50 films total, the top 10 from what I felt were the most relevant genres (Action, Drama, Horror, Comedy, Romance).

The Google side involved a single query: best *genre* movies of all time. And then recording the top 10 results. Afterwards I took the movie titles and ran them through a quick sql query to populate the results with more data like year, studio, and rating.




```
1. The Thing (1982) - Rating: 4.35 - Studio: Universal Pictures
2. The Silence of the Lambs (1991) - Rating: 4.33 - Studio: Orion Pictures
3. Psycho (1960) - Rating: 4.32 - Studio: Shamley Productions
4. Alien (1979) - Rating: 4.27 - Studio: 20th Century Fox
5. The Cremator (1969) - Rating: 4.25 - Studio: Filmové studio Barrandov
6. Cure (1997) - Rating: 4.24 - Studio: Daiei Film
7. The Shining (1980) - Rating: 4.24 - Studio: Warner Bros. Pictures
8. Kwaidan (1964) - Rating: 4.21 - Studio: Ninjin Club
9. Rosemary's Baby (1968) - Rating: 4.21 - Studio: William Castle Productions
10. Twin Peaks: Fire Walk with Me (1992) - Rating: 4.19 - Studio: New Line Cinema
Average Rating of Top 10 Movies: 4.261
```

The letterbox section was a fair bit more involved. First I had to filter out what a movie is by my own biases. As the dataset includes everything from stand up to tv shows to short films I had to include a ‘definition’ of what qualifies as a movie. For my case a movie is any piece of media with a full theatrical release that has a runtime between 1 and 4 hours long. Then I printed the top 10 results for each genre of choice sorted by rating.

Results:

While I was wrong about Google having a studio bias I did find some interesting biases in both its algorithm and the letterbox rating system. On the letterbox side there is an interesting bias that I admittedly was not expecting, and that is that there is a type of survivorship bias to movie ratings. People tend to only rate movies that they watch which means certain movies that are only going to be watched by fans, like movies for popular anime, are going to have more inflated ratings than my biases tell me they should have. To Illustrate my point I have included the results from the action section below.

***Action*:**

#	**Google Top 10 (Avg Rating 3.76):**	**Letterbox top 10 (Avg Rating 4.5):**
1	The Fugitive (1993) - Rating: 3.9	Harakiri (1962) - Rating: 4.69
2	Speed (1994) - Rating: 3.67	Seven Samurai (1954) - Rating: 4.6
3	Léon: The Professional (1994)	Ran (1985) - Rating: 4.5
4	The Rock (1996) - Rating: 3.55	The Lord of the Rings: The Return of the King (2003) - Rating: 4.5
5	Oldboy (2003) - Rating: 4.39	Attack on Titan: Chronicle (2020) - Rating: 4.48
6	Point Break (1991) - Rating: 3.73	The Dark Knight (2008) - Rating: 4.47
7	Die Hard: With a Vengeance (1995) - Rating: 3.69	Neon Genesis Evangelion: The End of Evangelion (1997) - Rating: 4.47
8	The Bourne Identity (2002) - Rating: 3.72	Spider-Man: Across the Spider-Verse (2023) - Rating: 4.45
9	Blade (1998) - Rating: 3.42	Spider-Man: Into the Spider-Verse (2018) - Rating: 4.42
10	GoldenEye (1995) - Rating: 3.54	The Empire Strikes Back (1980) - Rating: 4.41

While Attack on Titan and Evangelion are good shows with decent movies, they definitely don't belong this high in the rankings as some of the greatest movies of all time.

Interestingly Google's algorithm heavily favors movies from the 1990s and early 2000s. This is likely due to nostalgia optimization, engineering biases, or even potentially personalization based on the data google has collected from me.

One of Google's biggest flaws is its lack of foreign films—out of 50 movies I checked, most were American, a few British, one French (Léon), and two German (Nosferatu and Downfall)—likely reflecting American viewing habits.

Letterboxd data also failed me; I started this Google vendeta after seeing Parasite as a top comedy on Google, and my analysis confirmed it. I have a couple of thoughts on why this is the case. Comedy films tend to rate lower than dramas, artists commit genre fraud for award show eligibility, and my queries assume the provided data is accurate—so while I don't consider Parasite a comedy, I guess it's the funniest movie ever.

Conclusion:

I found some very interesting biases in both the letterbox and Google's top 10 movies from but ultimately my analysis is limited by my understanding of both internet scraping and machine learning. Also interestingly, despite specifically asking for the greatest movies from each genre, the average ratings tended on the lower than expected side. While it is not fair to compare it to movies directly sorted by ratings I definitely expected at least some of the top rated movies to show up in Google's results. I think I am going to keep this repo on my computer and add to it over time to try to refine myself a proper query engine once I learn some machine learning. I feel like the themes table could be invaluable in filtering our genre fraud.

Repo: https://github.com/IsaacLake03/LetterboxAnalysis