Project Proposal

Josh Stabinsky

Problem Statement

Can we accurately predict box office success of films based on typical dimensions? (using a method like linear regression)

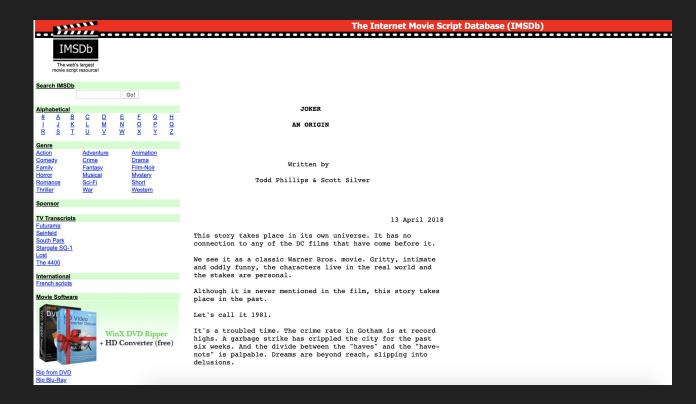
- Genre
- Actors
- Run time
- Budget
- ETC.

Can we accurately cluster / predict genre and/or box office success of films based on text from scripts? (using a natural language processing toolkit like nltk)

Data Sources

IMDB

IMSDB



Hypothesis / Success Metrics

Certain attributes do predict box office success:

- Dramas and Animated Films
- Longer run times
- Bigger budgets
- Certain actors

Certain keywords in scripts can predict movie genre at a "reasonable" rate:

If there are 16 genres, then > 6.25% of the time

Risks / Limitations

- There's a lot of nuance around what makes a film successful financially
- I'll have to be able to scrape the scripts off IMSDB in bulk
 - If I can't I'll have to figure out another method or data source

Feedback / Questions?