

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



The web's largest
movie script resource!


Search IMSDb

Alphabetical
A B C D E F G H
I J K L M N O P Q
R S T U V W X Y Z

Genre
[Action](#) [Adventure](#) [Animation](#)
[Comedy](#) [Crime](#) [Drama](#)
[Family](#) [Fantasy](#) [Film-Noir](#)
[Horror](#) [Musical](#) [Mystery](#)
[Romance](#) [Sci-Fi](#) [Short](#)
[Thriller](#) [War](#) [Western](#)

Sponsor
TV Transcripts
[Futurama](#)
[Seinfeld](#)
[South Park](#)
[Stargate SG-1](#)
[Lost](#)
[The 4400](#)

International
[French scripts](#)

Movie Software

**WinX DVD Ripper
+ HD Converter (free)**
[Rip from DVD](#)
[Rip Blu-Ray](#)

The Internet Movie Script Database (IMSDb)

JOKER

AN ORIGIN

Written by
Todd Phillips & Scott Silver

13 April 2018

This story takes place in its own universe. It has no connection to any of the DC films that have come before it.

We see it as a classic Warner Bros. movie. Gritty, intimate and oddly funny, the characters live in the real world and the stakes are personal.

Although it is never mentioned in the film, this story takes place in the past.

Let's call it 1981.

It's a troubled time. The crime rate in Gotham is at record highs. A garbage strike has crippled the city for the past six weeks. And the divide between the "haves" and the "have-nots" is palpable. Dreams are beyond reach, slipping into delusions.

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?