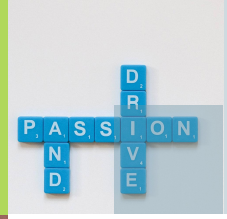
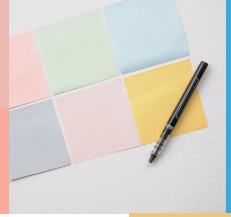





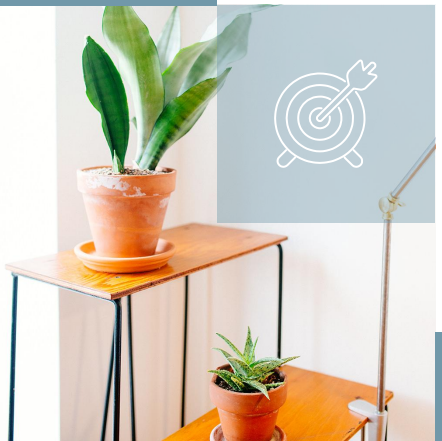
Sphinxi

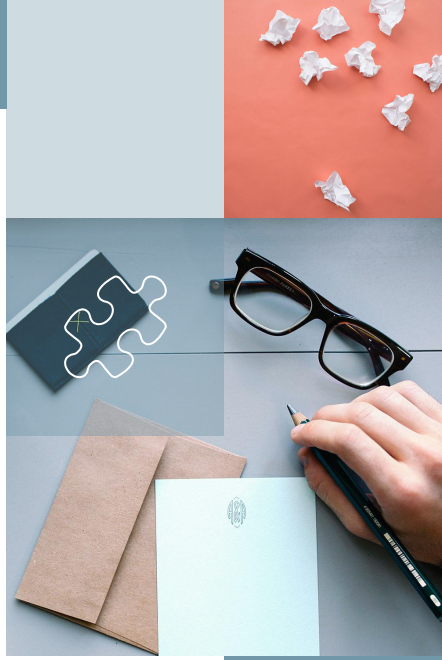
Word Similarities Game

By Teng Mao

Data Science Fellow from General Assembly







Agenda

- Problem Statement
- Methodology
- Corpora
- Exploratory Data Analysis
- Model
- Bonus: Challenges
- Conclusion & Next Steps

The Riddle of the Sphinx

"What is the creature that walks on four legs in the morning, two legs at noon and three in the evening?"

Man

Problem Statement:

How do you build a game using natural language processing with a neural net?

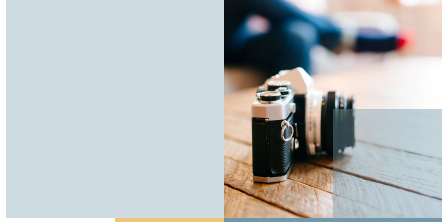
Methodology

Find the Data, Clean and Explore

Train the Model

Generate Prompts

Launch the Model



Corpora

Google News Vectors

300-dimensional vectors for 3 million words and phrases trained on Google News dataset of about 100 billion words.

Wikipedia, UMBC, and statmt.org

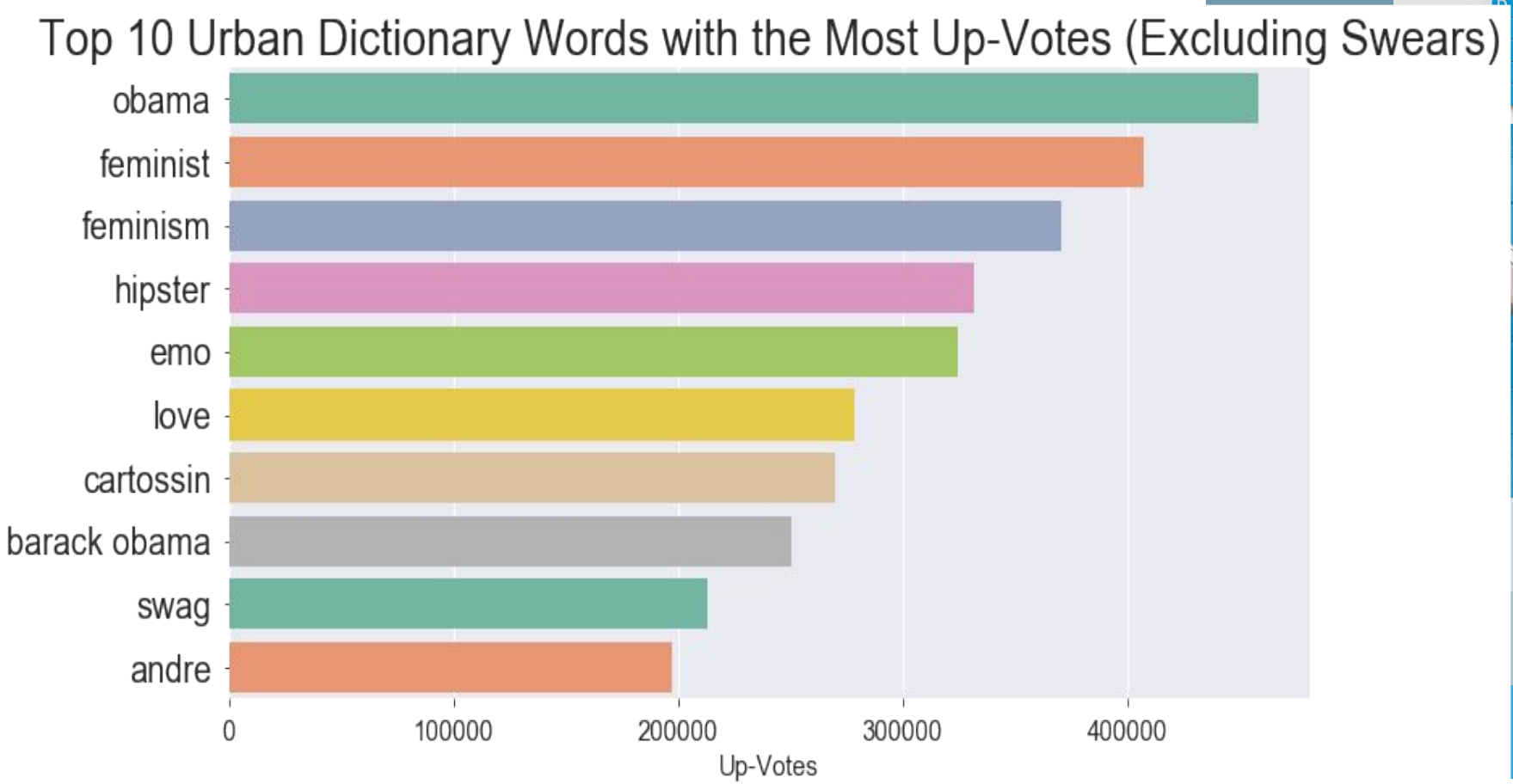
300-dimensional vectors for 1 million words trained on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset (16 billion words).

Urban Dictionary

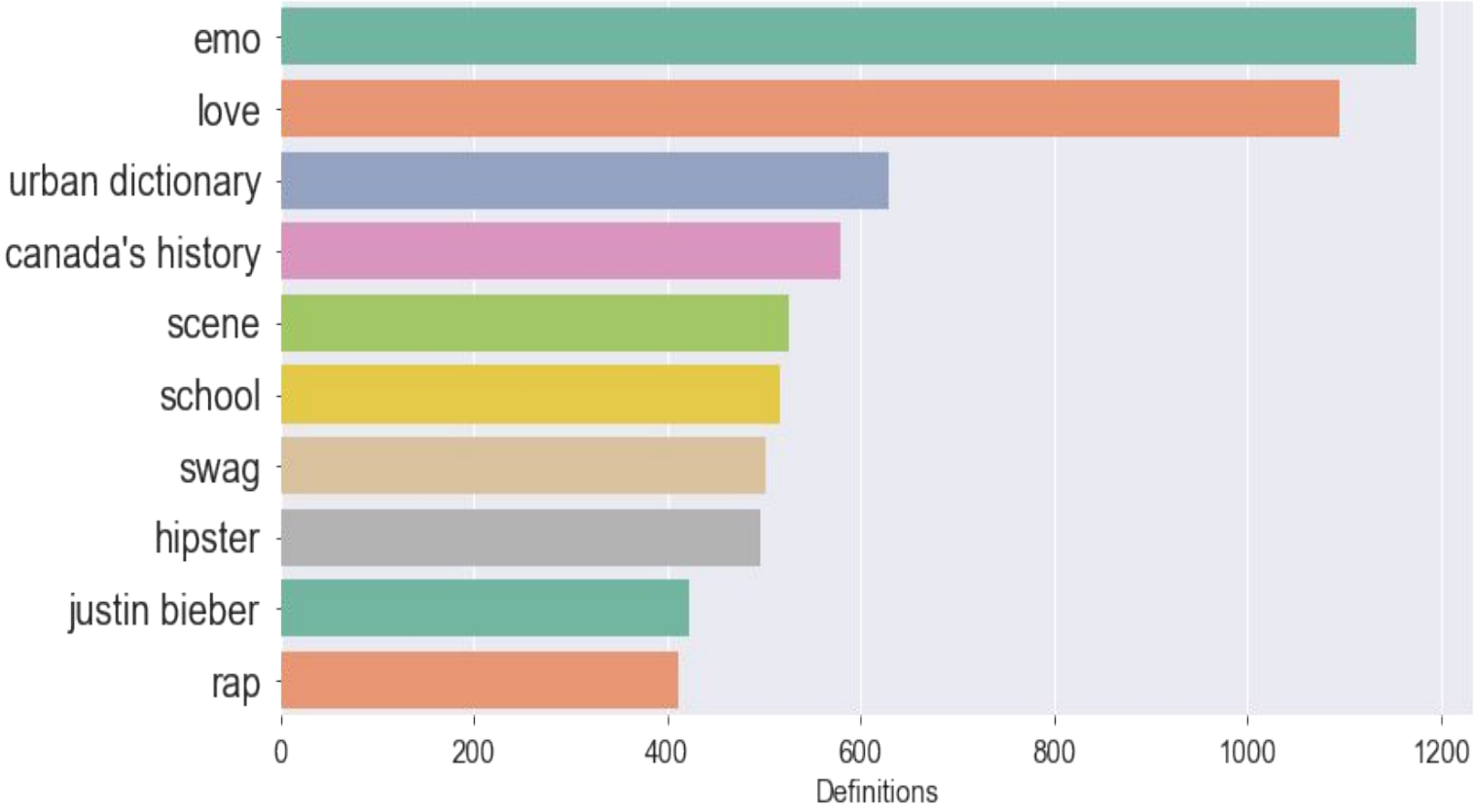
2.6 million words (1.6 million unique), authors, up-votes and down-votes, as well as definitions were scraped and put on kaggle for a competition.

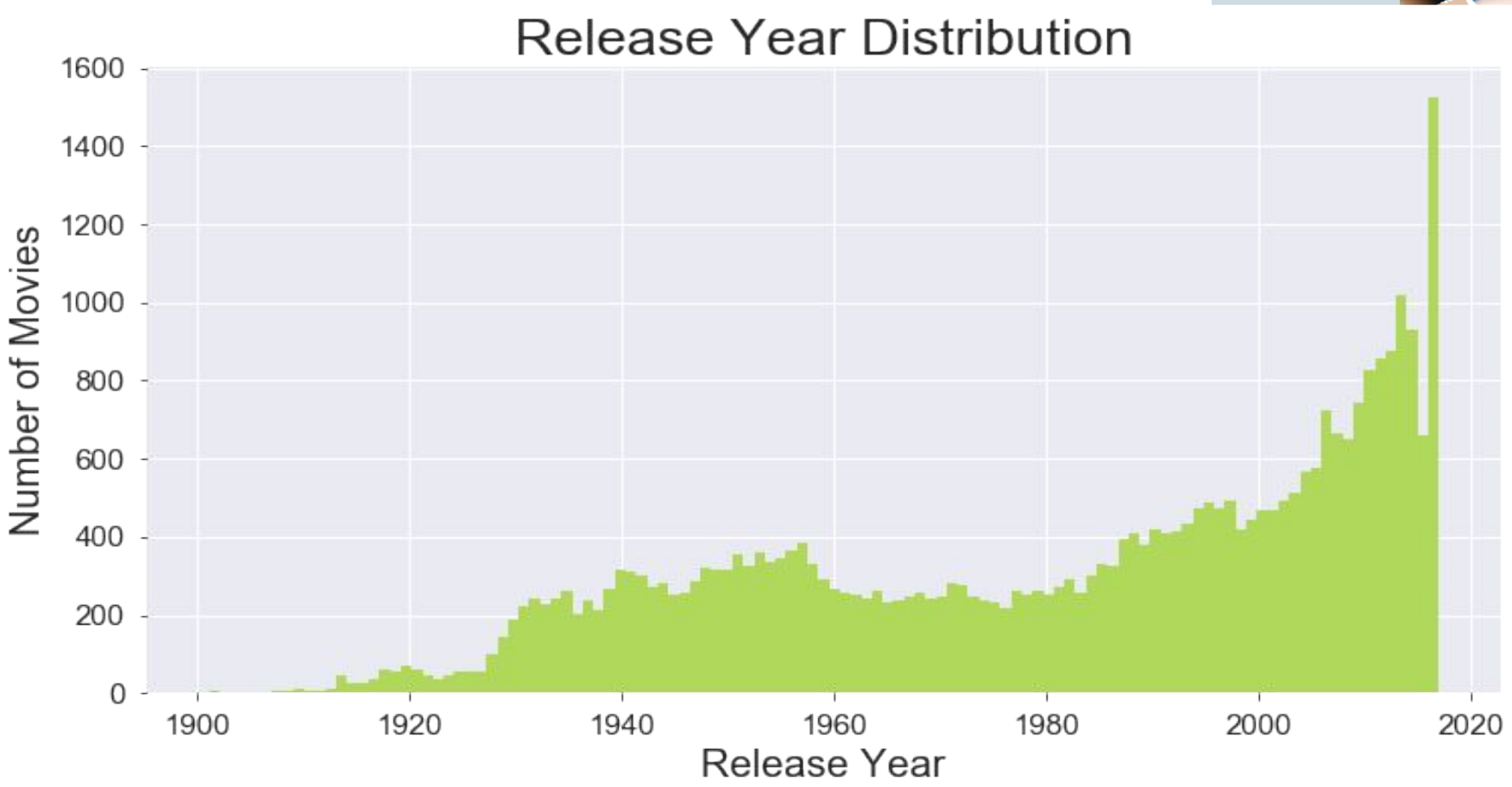
Wikipedia Movie Plots

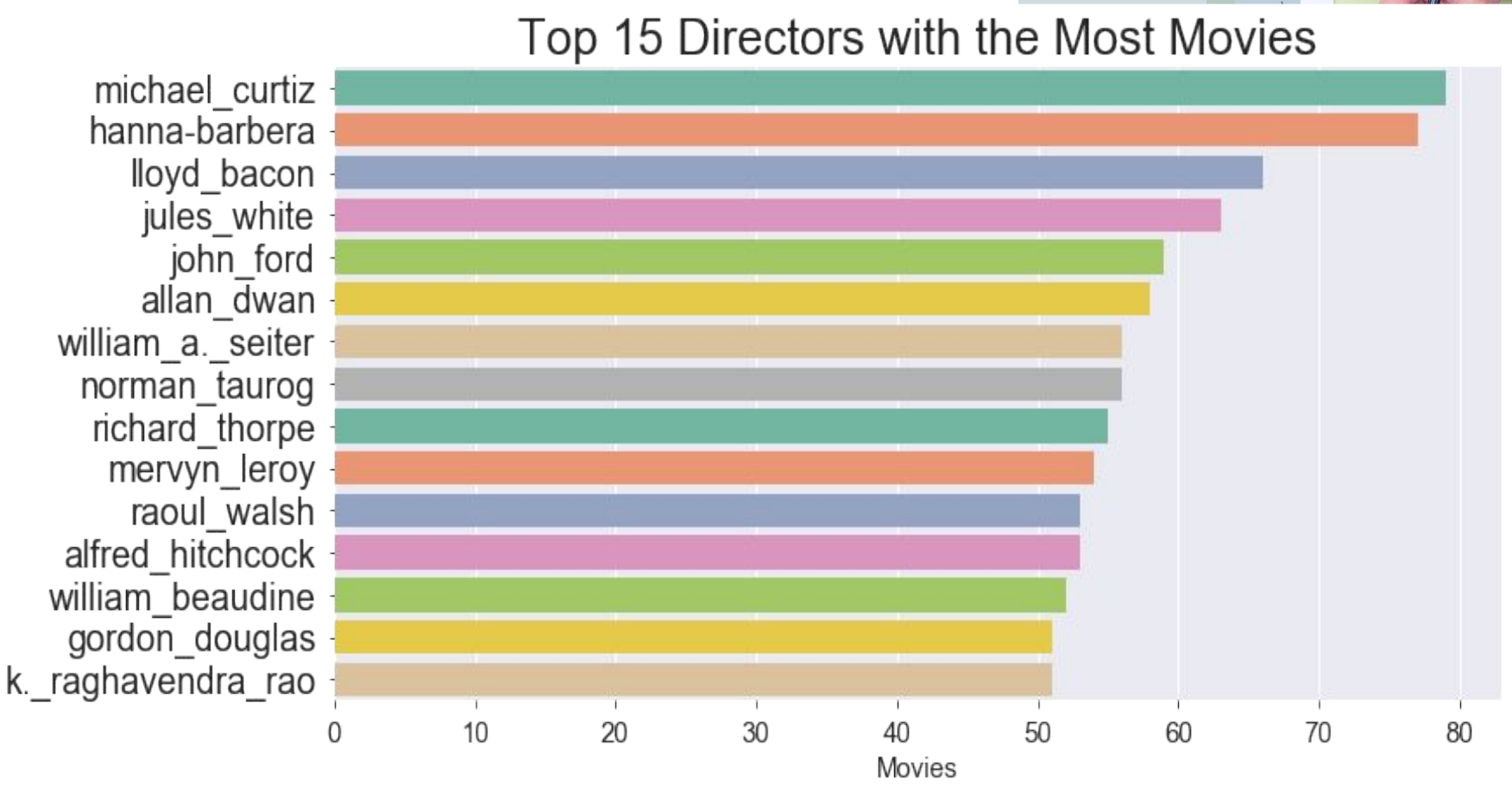
34,886 movies were scraped from Wikipedia and put on kaggle for a competition. Data includes release year, title, origin/ethnicity, director, cast, genre, and plot.

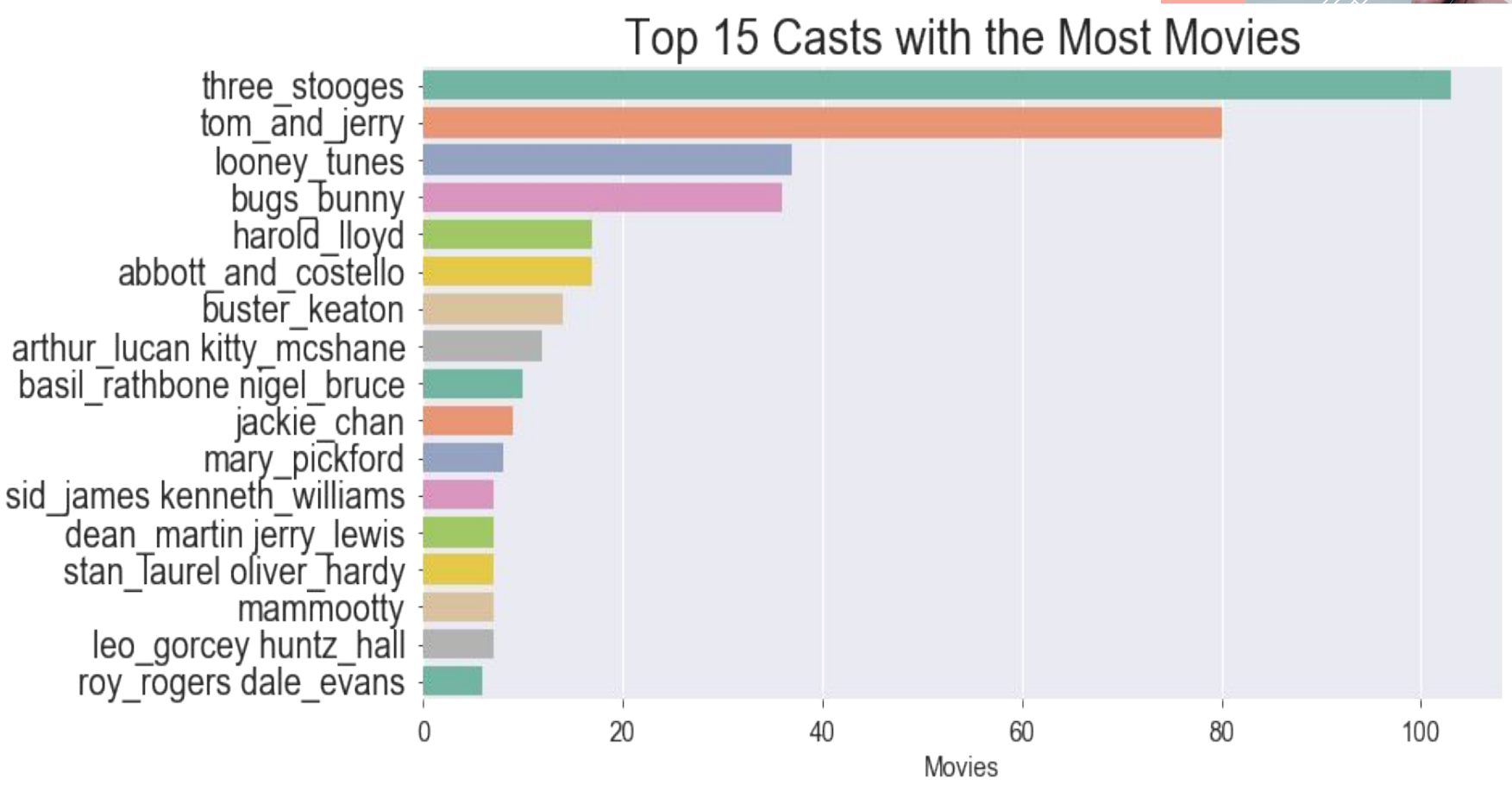

Top 10 Urban Dictionary Words with the Most Up-Votes (Excluding Swears)
obama
feminist
feminism
hipster
emo
love
cartossin
barack obama
swag
andre
0
100000
200000
300000
400000
Up-Votes

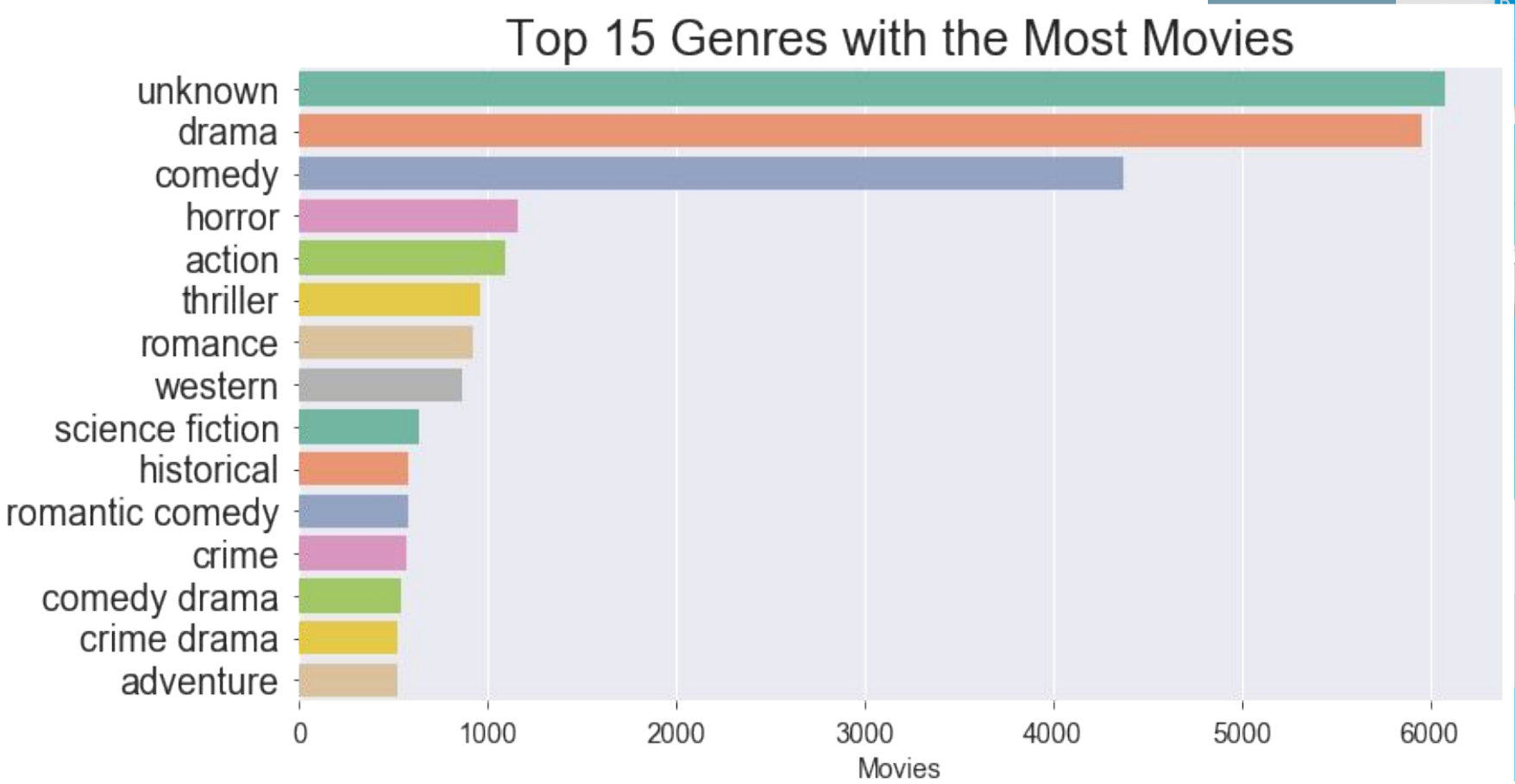
Top 10 Urban Dictionary Words with the Most Definitions (Excluding Swears)

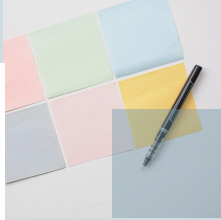



Release Year Distribution
Number of Movies
1600
1400
1200
1000
800
600
400
200
0
1900
1920
1940
1960
1980
2000
2020
Release Year


Top 15 Directors with the Most Movies
michael_curtiz
hanna-barbera
lloyd_bacon
jules_white
john_ford
allan_dwan
william_a._seiter
norman_taurog
richard_thorpe
mervyn_leroy
raoul_walsh
alfred_hitchcock
william_beaudine
gordon_douglas
k._raghavendra_rao
0
10
20
30
40
50
60
70
80
Movies


Top 15 Casts with the Most Movies
three_stooges
tom_and_jerry
looney_tunes
bugs_bunny
harold_lloyd
abbott_and_costello
buster_keaton
arthur_lucan kitty_mcshane
basil_rathbone nigel_bruce
jackie_chan
mary_pickford
sid_james kenneth_williams
dean_martin jerry_lewis
stan_laurel oliver_hardy
mammootty
leo_gorcey huntz_hall
roy_rogers dale_evans
0
20
40
60
80
100
Movies

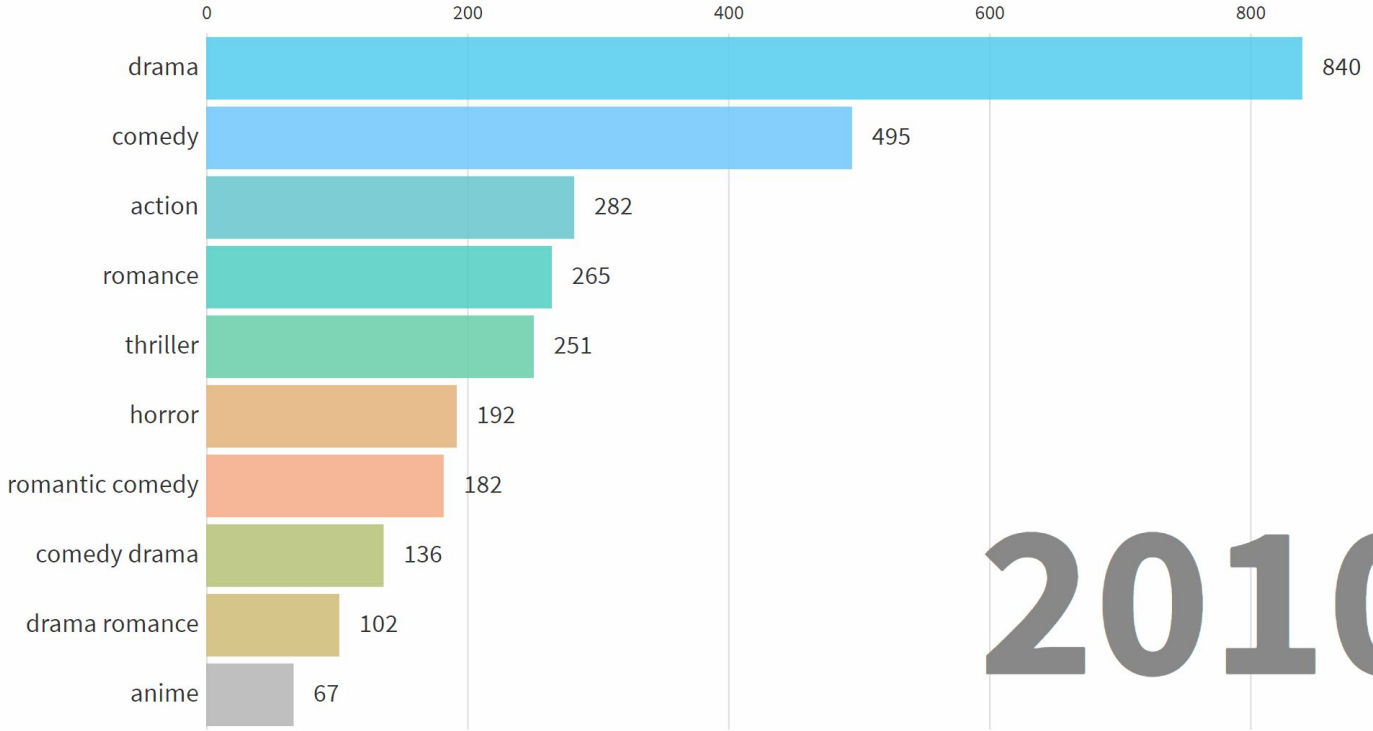

Top 15 Genres with the Most Movies
unknown
drama
comedy
horror
action
thriller
romance
western
science fiction
historical
romantic comedy
crime
comedy drama
crime drama
adventure
0
1000
2000
3000
4000
5000
6000
Movies



Top 10 Movie Genres by Decade (1900-2017)

from Wikipedia Movies Plots

Replay



Word2Vec Model

horror

'horrors','vacui', 'Horror', 'terror', 'horror-', 'dread', 'shock', 'disgust', 'revulsion', 'carnage'

Patriots

'Pats', 'England_Patriots', 'Chargers', 'Colts', 'Ravens', 'Jaguars', 'Raiders', 'Seahawks', 'Bengals', 'Belichick'

turnt

'flipped_up', 'f*cked', 'krunked', 'geeked', 'fubared', 't_u', 'ttgt', 'troud', 'schywasted', 'crunked'

horror

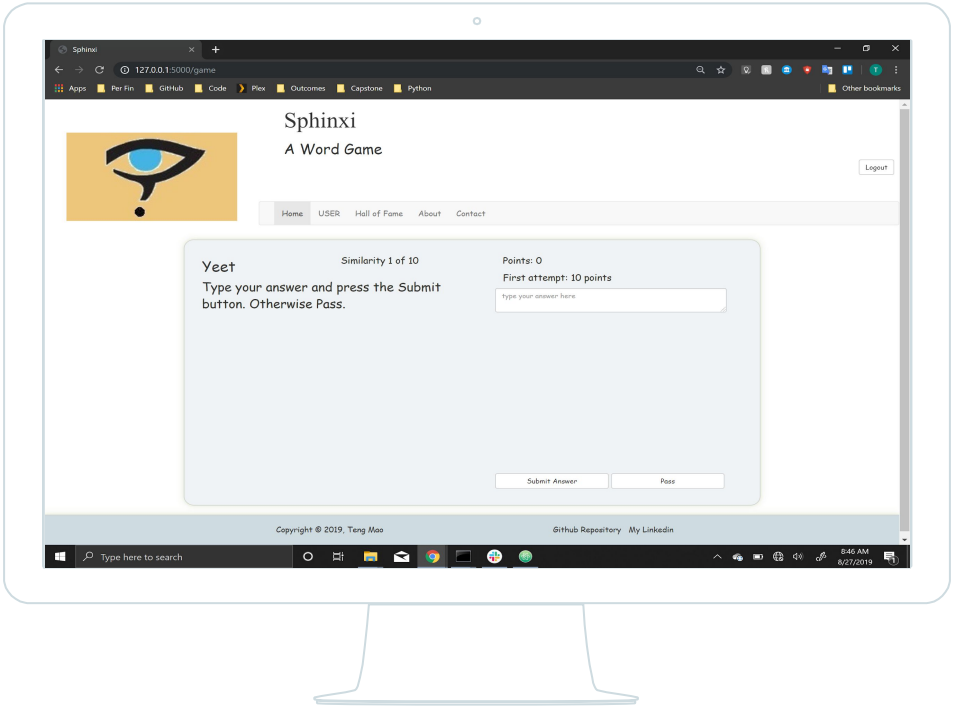
'thriller', 'fantasy', 'zombi_kampung_pisang', 'jangan_pandang_belakang', ‘the_creeping_flesh', 'Suspense', ‘diamond_is_unbreakable_chapter_i', 'susan_denberg', 'the_skull', 'freddie_francis'

Netflix and chill

'netflix_and_cuddle', 'netflix_an_chill', 'hulu_and_hangout', 'imax_and_climax', 'youtube_and_blow', 'hulu_and_chill', 'netflix_and_chill_alone', 'hulu_and_hang', 'p**nhub_and_chill', 'netflix_and_jill'

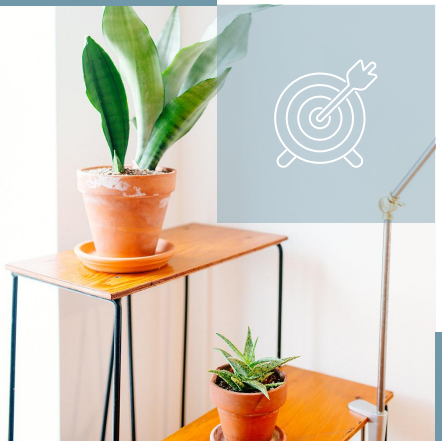
Sphinxi on the Desktop

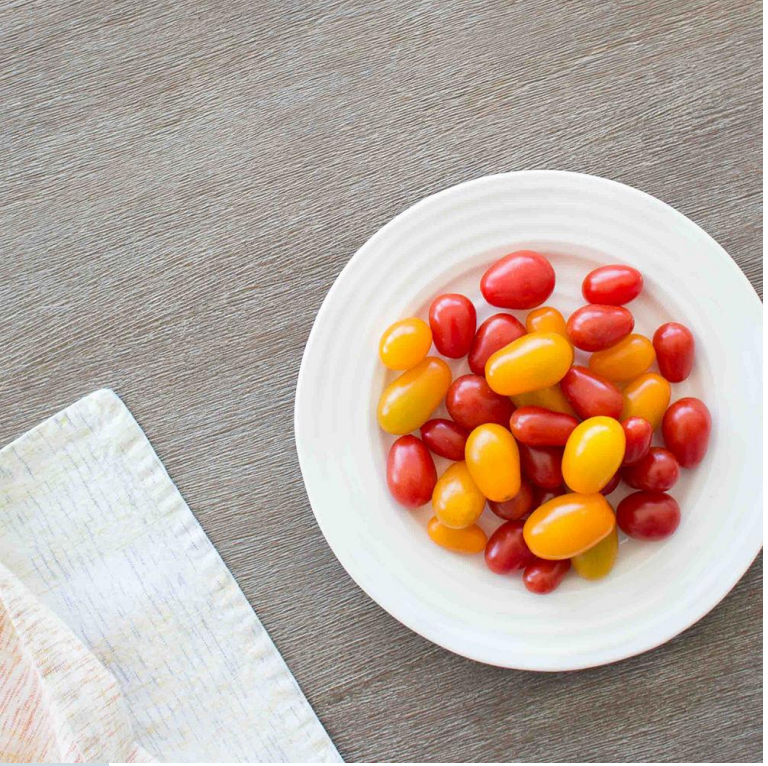
Using Flask, Sphinxi is able to be launched in a browser.



Challenges

- Deciding on a capstone
- Size of the data
- Offensive content
- Scope of the project
- Deciphering someone else's code
- Momentum





Conclusions

- Natural Language Processing - powerful, but flawed

Next Steps

- Deploy model to AWS
- Finish the game in Flask
- Generate more prompts
- Create analogies option
- Gather more data

Thanks!

Any questions?

You can find me at @ github.com/TengCXXI/ and linkedin.com/in/tengmao