
# What to Watch Next?

A Team KIM Production

29 June 2022

# What to Watch Next?

Our goal was to build a fun movie recommendation system with the MovieLens 25M dataset

movielens

## Dataset Features

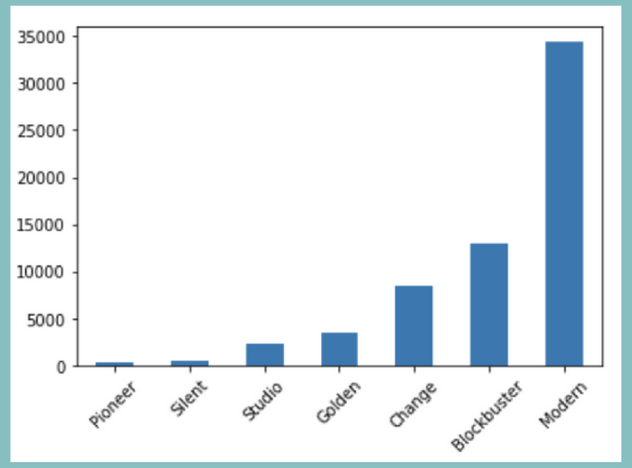
### Movies Data

- 62,423 Movie titles
- Released between 1874 and 2019
- 19 Genres

### Users Data

- Data from 162,541 users
- 25mn Ratings
- 1mn Movie tags
- Captured between 1995 and 2019

### Distribution of Movies by Era




### Dataset sample

| movieId | title | genres |
|---|---|---|
| 1 | Toy Story (1995) | Adventure\|Animation\|Children\|Comedy\|Fantasy |
| 2 | Jumanji (1995) | Adventure\|Children\|Fantasy |
| 3 | Grumpier Old Men (1995) | Comedy\|Romance |
| 4 | Waiting to Exhale (1995) | Comedy\|Drama\|Romance |
| 5 | Father of the Bride Part II (1995) | Comedy |
| ... | ... | ... |
| 209157 | We (2018) | Drama |
| 209159 | Window of the Soul (2001) | Documentary |
| 209163 | Bad Poems (2018) | Comedy\|Drama |
| 209169 | A Girl Thing (2001) | (no genres listed) |
| 209171 | Women of Devil's Island (1962) | Action\|Adventure\|Drama |

We had to address how to: 1) wrangle a huge dataset, 2) deliver prompt results to users through an intuitive user interface

# Feature Engineering

Combining contextual text into a supertext feature to be vectorized

## Genres were incorporated

| title | genres |
|---|---|
| Toy Story (1995) | Adventure\|Animation\|Children\|Comedy\|Fantasy |
| Jumanji (1995) | Adventure\|Children\|Fantasy |
| Grumpier Old Men (1995) | Comedy\|Romance |
| Waiting to Exhale (1995) | Comedy\|Drama\|Romance |
| Father of the Bride Part II (1995) | Comedy |
| ... | ... |
| We (2018) | Drama |
| Window of the Soul (2001) | Documentary |
| Bad Poems (2018) | Comedy\|Drama |
| A Girl Thing (2001) | (no genres listed) |
| Women of Devil's Island (1962) | Action\|Adventure\|Drama |

## User ratings were summarized and categorized

| userId | movieId | rating |
|---|---|---|
| 187 | 1 | 3.5 |
| 187 | 2 | 3.5 |
| 187 | 3 | 3.0 |
| 187 | 13 | 4.5 |
| 187 | 19 | 4.5 |
| ... | ... | ... |
| 162516 | 194947 | 3.5 |
| 162516 | 194951 | 4.0 |

| title | rating | ratingBins |
|---|---|---|
| Toy Story (1995) | 3.893708 | aboveAvg |
| Jumanji (1995) | 3.251527 | Avg |
| Grumpier Old Men (1995) | 3.142028 | Avg |
| Waiting to Exhale (1995) | 2.853547 | belowAvg |
| Father of the Bride Part II (1995) | 3.058434 | Avg |
| ... | ... | ... |
| We (2018) | 1.500000 | belowAvg |
| Window of the Soul (2001) | 3.000000 | Avg |
| Bad Poems (2018) | 4.500000 | aboveAvg |
| A Girl Thing (2001) | 3.000000 | Avg |
| Women of Devil's Island (1962) | 3.000000 | Avg |

## Movie release years extracted from title then categorized

| title | releaseYr | era |
|---|---|---|
| Toy Story (1995) | 1995 | Blockbuster |
| Jumanji (1995) | 1995 | Blockbuster |
| Grumpier Old Men (1995) | 1995 | Blockbuster |
| Waiting to Exhale (1995) | 1995 | Blockbuster |
| Father of the Bride Part II (1995) | 1995 | Blockbuster |
| ... | ... | ... |
| We (2018) | 2018 | Modern |
| Window of the Soul (2001) | 2001 | Modern |
| Bad Poems (2018) | 2018 | Modern |
| A Girl Thing (2001) | 2001 | Modern |
| Women of Devil's Island (1962) | 1962 | Change |

## User assigned tags were summarized and incorporated

| userId | movieId | tag |
|---|---|---|
| 3 | 260 | classic |
| 3 | 260 | sci-fi |
| 4 | 1732 | dark comedy |
| 4 | 1732 | great dialogue |
| 4 | 7569 | so bad it's good |
| 4 | 44665 | unreliable narrators |
| 4 | 115569 | tense |
| 4 | 115713 | artificial intelligence |
| 4 | 115713 | philosophical |
| 4 | 115713 | tense |
| 4 | 148426 | so bad it's good |
| 4 | 164909 | cliche |

| title | userTag |
|---|---|
| Toy Story (1995) | Owned imdb top 250 Pixar Pixar time travel chi... |
| Jumanji (1995) | Robin Williams time travel fantasy based on ch... |
| Grumpier Old Men (1995) | funny best friend duringcreditsstinger fishing... |
| Waiting to Exhale (1995) | based on novel or book chick flick divorce int... |
| Father of the Bride Part II (1995) | aging baby confidence daughter gynecologist mi... |
| Heat (1995) | imdb top 250 great acting realistic action sus... |
| Sabrina (1995) | remake chauffeur long island millionaire paris... |
| Tom and Huck (1995) | based on a book Mark Twain adapted from:book L... |
| Sudden Death (1995) | explosive hostage terrorist Jean-Claude Van Da... |
| GoldenEye (1995) | 007 Bond gadgets secret service sequel spies v... |

## Combined data ready for processing

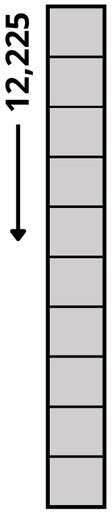
| movieName | text |
|---|---|
| Toy Story | Blockbuster aboveAvg Adventure Animation Child... |
| Jumanji | Blockbuster Avg Adventure Children Fantasy Rob... |
| Grumpier Old Men | Blockbuster Avg Comedy Romance funny best frie... |
| Waiting to Exhale | Blockbuster belowAvg Comedy Drama Romance base... |
| Father of the Bride Part II | Blockbuster Avg Comedy aging baby confidence d... |
| Heat | Blockbuster aboveAvg Action Crime Thriller imd... |
| Sabrina | Blockbuster Avg Comedy Romance remake chauffeu... |
| Tom and Huck | Blockbuster Avg Adventure Children based on a ... |
| Sudden Death | Blockbuster Avg Action explosive hostage terro... |
| GoldenEye | Blockbuster Avg Action Adventure Thriller 007 ... |
| American President, The | Blockbuster aboveAvg Comedy Drama Romance Roma... |
| Dracula: Dead and Loving It | Blockbuster belowAvg Comedy Horror dracula spo... |
| Balto | Blockbuster Avg Adventure Animation Children E... |
| Nixon | Blockbuster Avg Drama biography government his... |
| Cutthroat Island | Blockbuster belowAvg Action Adventure Romance ... |
| Casino | Blockbuster aboveAvg Crime Drama Mafia Mafia M... |
| Sense and Sensibility | Blockbuster aboveAvg Drama Romance chick flick... |
| Four Rooms | Blockbuster Avg Comedy anthology dark comedy f... |
| Ace Ventura: When Nature Calls | Blockbuster belowAvg Comedy detective childhoo... |
| Money Train | Blockbuster belowAvg Action Comedy Crime Drama... |
| Get Shorty | Blockbuster aboveAvg Comedy Crime Thriller Cri... |

# Architecture

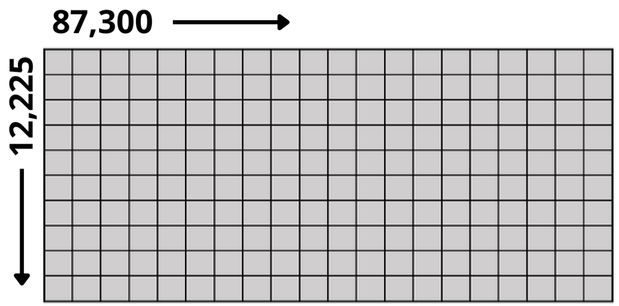
TF-IDF Vectorizer used to reshape inputs from which similarity scores are calculated

| text |
| --- |
| Modern Avg Johnny Depp pirates artifact carib... |
| Modern belowAvg Action Adventure Sci-Fi Batman... |
| Modern aboveAvg Action Adventure Sci-Fi afrofu... |
| Modern belowAvg Action Adventure Sci-Fi MCU Ma... |
| Modern Avg Action Adventure Sci-Fi brie larson... |
| Modern aboveAvg Action Adventure Sci-Fi busy p... |
| Modern aboveAvg Action Adventure Sci-Fi awesom... |
| Modern aboveAvg Action Adventure Sci-Fi Comedy... |
| Modern aboveAvg Action Adventure Fantasy comic... |
| Modern Avg dumbledore fantasy harry potter Jo... |
| Modern Avg Action Fantasy Sci-Fi action atlant... |
| Modern aboveAvg Action Adventure Fantasy DC Co... |
| Modern Avg acapella sequel singer acapella An... |
| Modern Avg Crime Drama Gangster 1920s 1930s ba... |
| Modern aboveAvg Drama Horror Thriller Claustro... |
| Modern aboveAvg Documentary Civil Rights human... |
| Modern aboveAvg Comedy Drama dog multiple stor... |
| Modern Avg Comedy middle ages nun servant midd... |
| Modern belowAvg Comedy high school high school... |
| Modern belowAvg Action Thriller gangster kidna... |

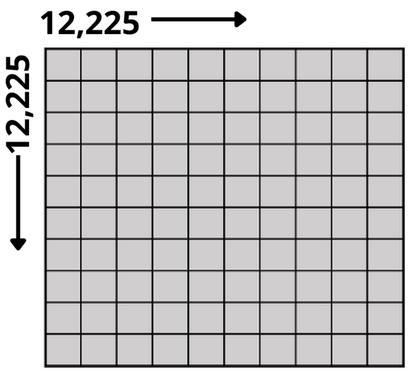
**1D Array**




**2D Array**




**2D Array**




```
--- Searched for: Ex Machina ---

Recommendations:
                          movieName  sim_scores
513                         Chappie    0.513092
4871                       Pass-Thru    0.471539
9100                              Ta    0.449297
2724                         Uncanny    0.424846
6728                      Amelia 2.0    0.423528
261              Autómata (Automata)    0.412459
7222                     Singularity    0.407655
5962                          Somnio    0.384374
7751                        AlphaGo    0.379739
9095             2036 Origin Unknown    0.375084
68                      Transcendence    0.372376
1304                            Debug    0.362378
721                              Vice    0.361531
10498                   A Crimson Man    0.342863
7779                 The God Question    0.307450
5677                          Teleios    0.290876
8492   Do You Trust this Computer?      0.290820
9026                          Upgrade    0.285853
34                        Interstellar    0.277124
3997                          Morgan     0.273019
```

**Preprocessing**

The text feature for all 12,225 movies to be included in our search tool is extracted into a 1D array for processing

**TD-IDF Vectorizer**

The text is processed with weights assigned to words based on significance, thus forming new feature vectors arranged in a 2D array

Each movie is represented by a vector of 87,300 weights

**Cosine Similarity**

The cosine similarity between vectors of two movies is calculated for all movies

This produces a "map" of shape 12,225 x 12,225 where the similarities of each movie to all other movies, including itself, can be referred from

# 3...2...1...Action!

Searching "Ex Machina" returns similarly themed AI/ Robotics  films

**Optimize Performance**
To minimize resource use and run times, we wrote the recommendation function as its own script, which loaded a pre-calculated cosine similarities map

**Enhance UX**
A user-facing app was designed in Streamlit and enhanced with the OMDB API that allowed movie metadata, such as posters, to be called in

TEAM KIM

Input your movie name here and press enter

Ex Machina


Genre:Drama, Sci-Fi, Thriller Year: 2014

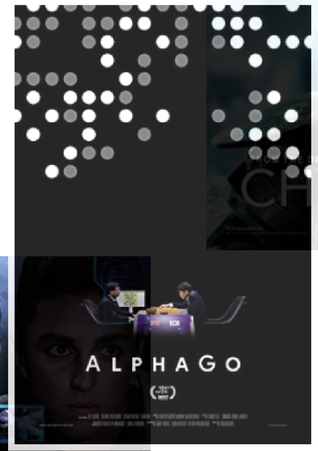
A young programmer is selected to participate in a ground-breaking experiment in synthetic intelligence by evaluating the human qualities of a highly advanced humanoid A.I.

Rating:7.7

**Recommended For You**

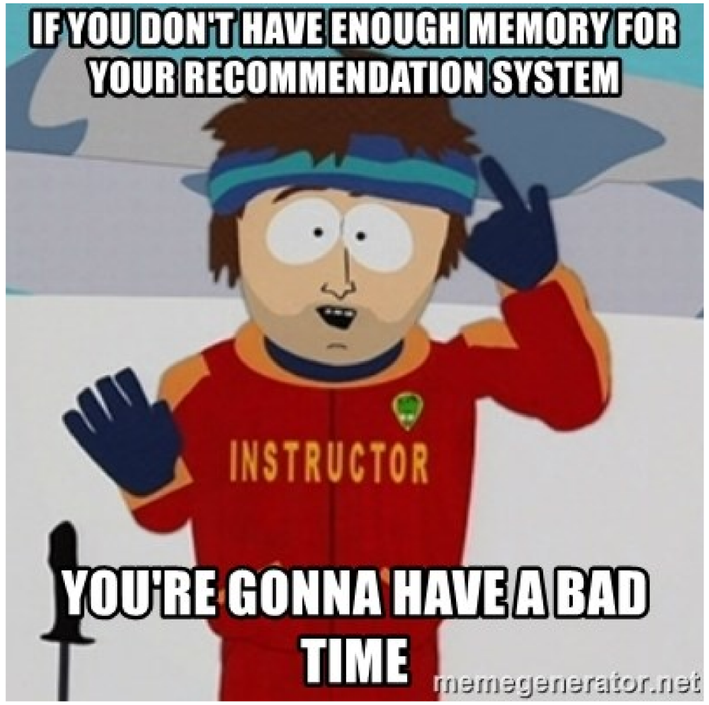
Chappie


AlphaGo




Singularity

# Challenges & Conclusions




## Large Datasets, Resource Limits

- We challenged ourselves to use a large dataset but found processing it was sometimes beyond our hardware capabilities (RAM!)
- This was particularly true running the computationally intensive linear kernel
- We also hit Streamlit's data capacity limits when implementing the user-facing app
- Ultimately, we downsized our dataset to five years for sake of running our search demo
- This followed an already aborted attempt to analyze an even larger Spotify dataset

## Recommendation Systems

- We did not find a satisfactory way of implementing collaborative-based filtering and keep to our goal of making the user experience as simple as possible
- An upgrade we discussed was to invite users to supply some movie ratings to boost recommended results
- Ultimately the demo is built on a content-based filtering model

## Upgrades with Neural Networks

- We experimented with but did not implement neural networks
- An idea was to use embeddings to vectorize our data but we deemed this too similar to the TF-IDF Vectorizer approach
- With more time we would explore ways to deploy NNs in place of linear kernel