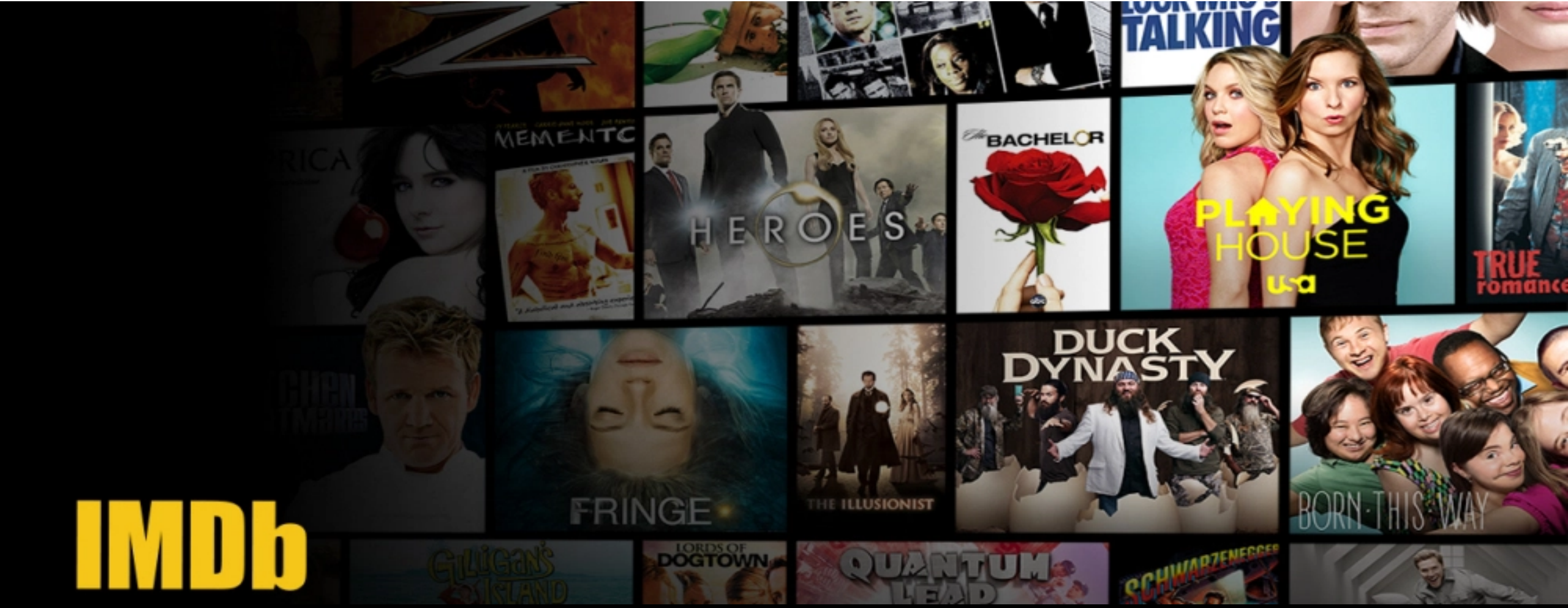


Movie Reviews

Nianzhe Li | Hongfa Huang| Yadi Gong
Yiping Liu | Shuang Liang

Agenda

- Business Problem
- Data Description
- Logistic Regression
- K-means Clustering
- Hierarchical Clustering + LDA Clustering
- Business Insights

Business Problem

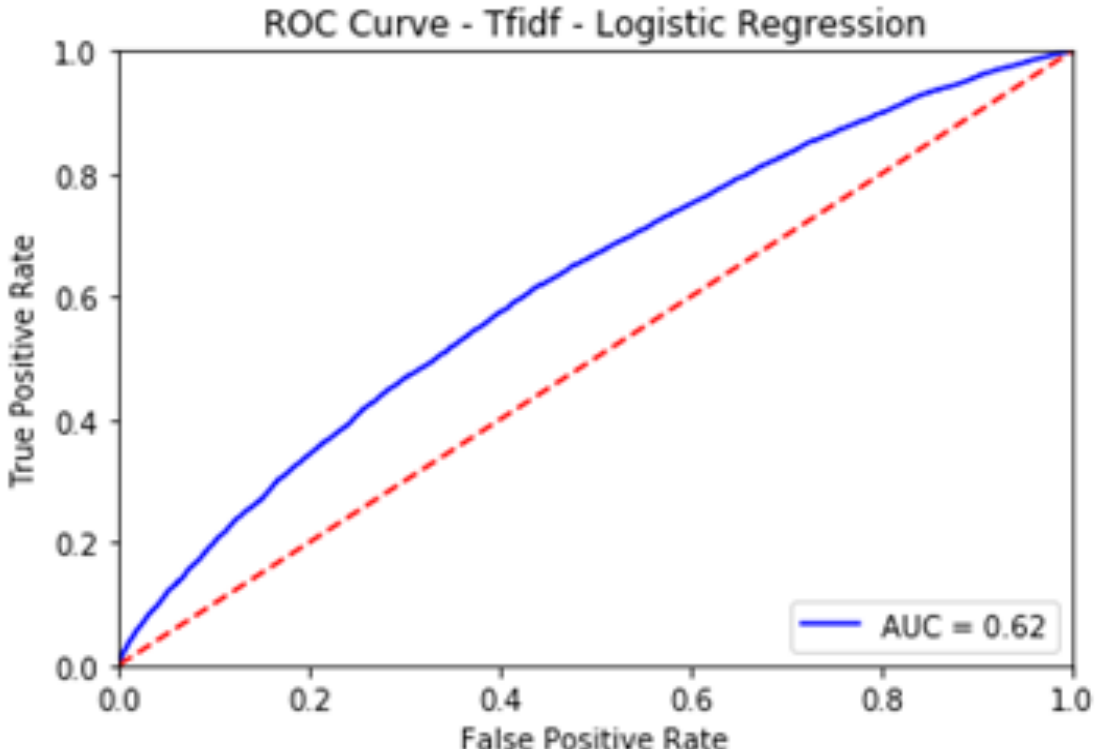
Classification Model - Identify positive and negative movie reviews

User Profile - Catch key words in different review clusters

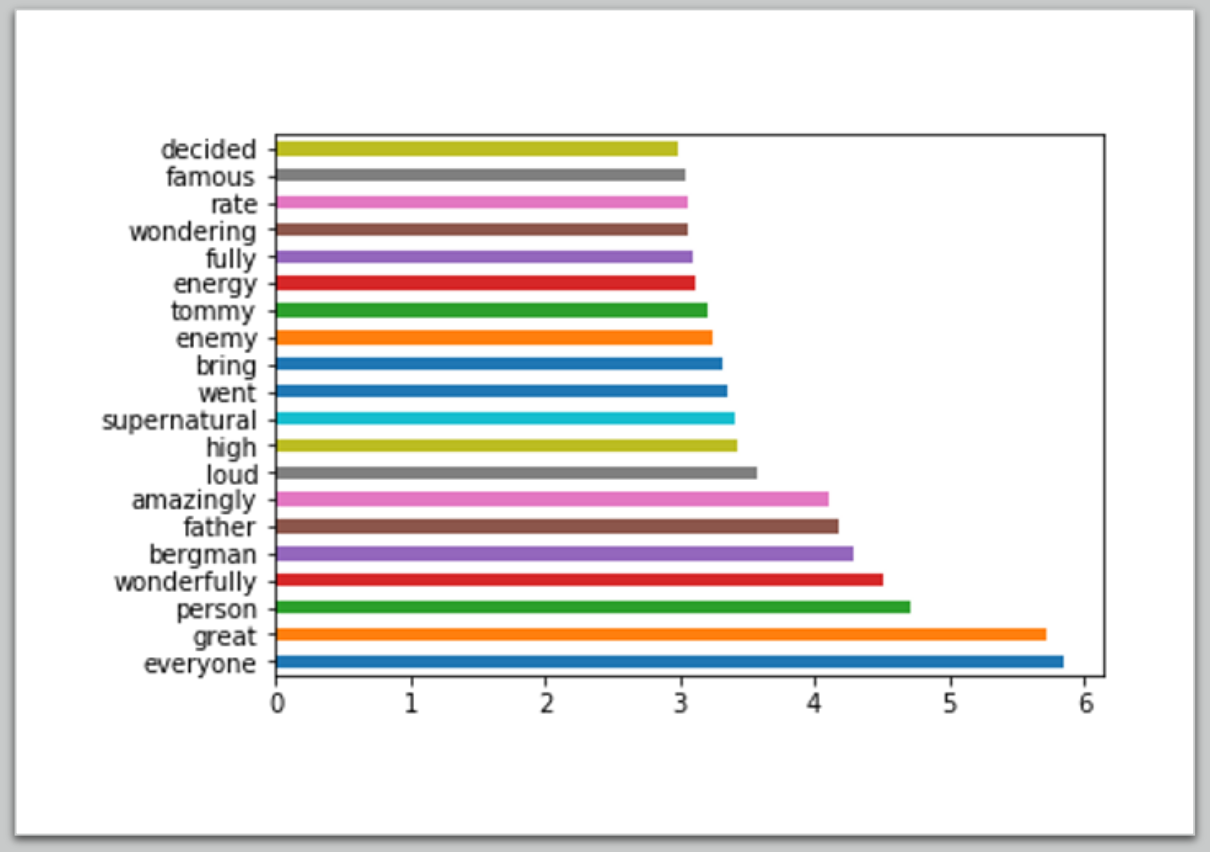
Data Description

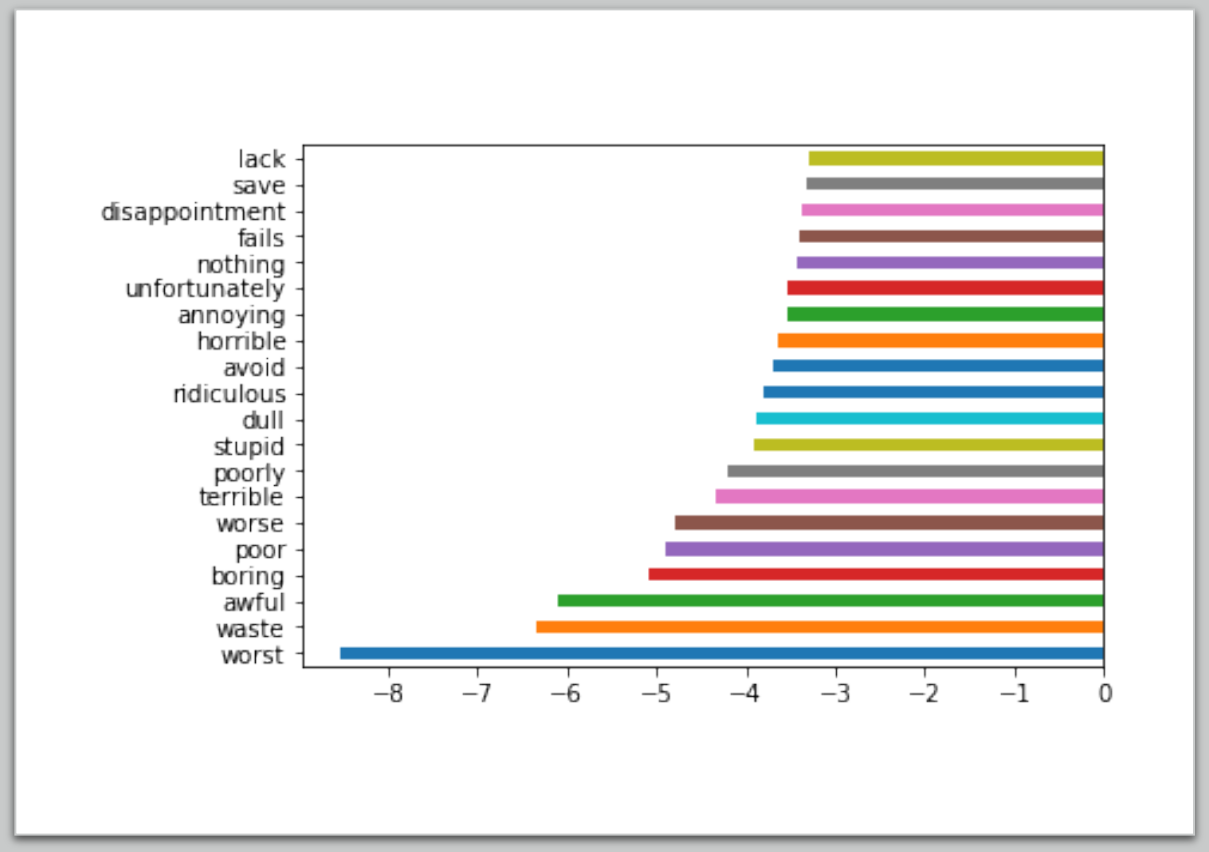
- IMDb Large Movie Review Dataset
- 50k reviews separated into balanced 25k training and balanced 25k testing
- Train and test set are reviews from disjoint set of movies
- Positive reviews: $\text{score} \geq 7$
- Negative reviews: $\text{score} \leq 4$
- Data preprocessing:
  - Remove irrelevant words/numbers/symbols
  - TFIDF Vectorizer

Logistics Regression



- Use Training and Test dataset
  - Half training and Half test set
  - Training and test set from the disjoint movies
- Fit model with Logistics Regression
  - AUC= 0.62





Top 20 Important Features

Clustering

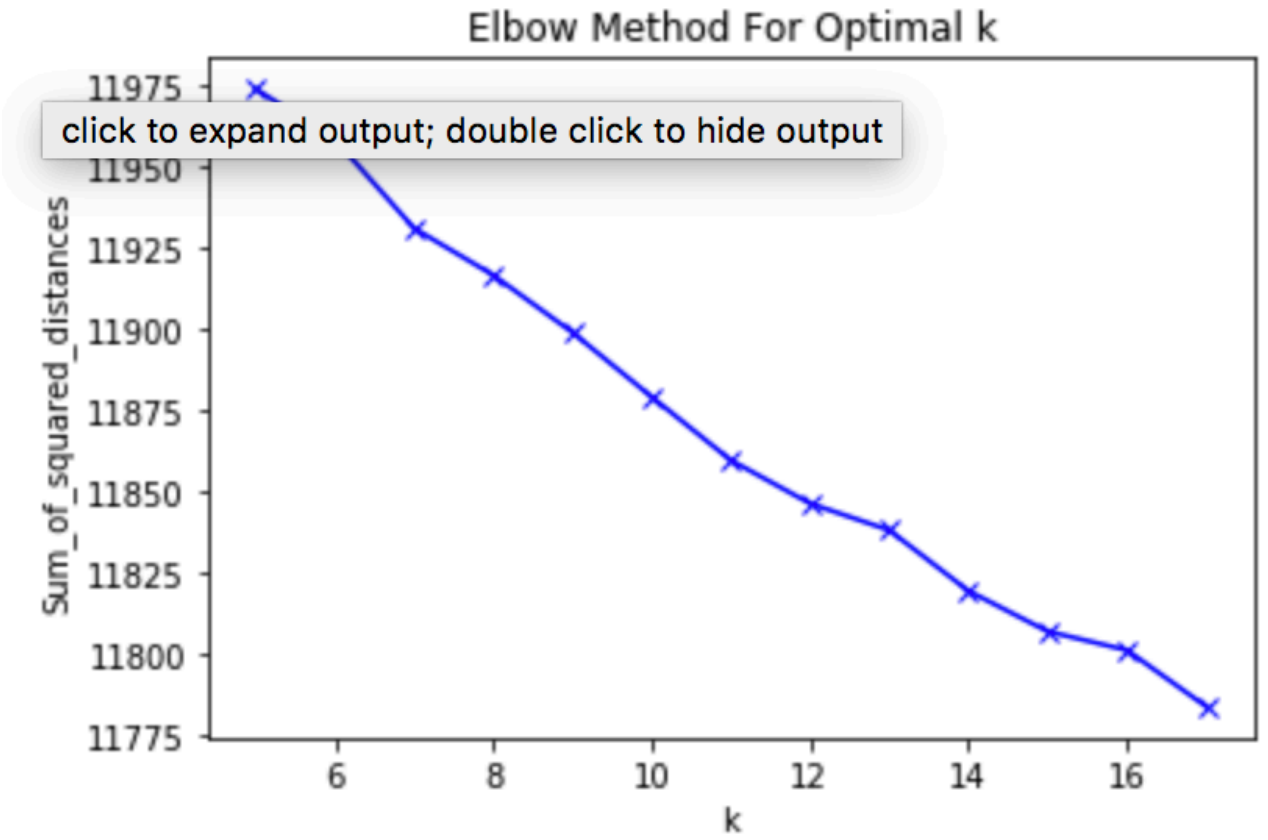
- K-Means
- Hierarchical + LDA

K-means clustering

- **Select Optimal K:**

  High Silhouette Coefficient

  Elbow method
  (lower point in the curve)



❖ Positive clusters

Animation	Family Story	Comedy	Music
animation	family	comedy	performance
disney	father	funny	music
animated	mother	laugh	musical
cinderella	child	hilarious	actor
story	brother	watch	character
character	young	romantic	scene
bakshi	story	character	dance
voice	character	funniest	story
child	parent	really	oscar
little	daughter	scene	wonderful

❖ Positive clusters

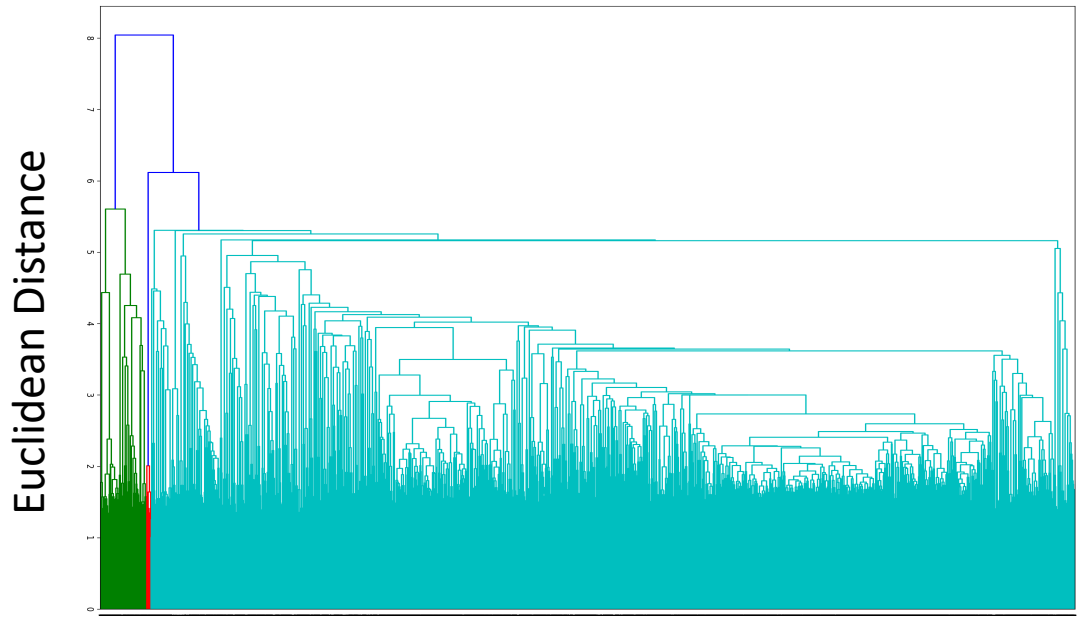
Animation	Family Story	Comedy	Music
animation	family	comedy	performance
disney	father	funny	music
animated	mother	laugh	musical
cinderella	child	hilarious	actor
story	brother	watch	character
character	young	romantic	scene
bakshi	story	character	dance
voice	character	funniest	story
child	parent	really	oscar
little	daughter	scene	wonderful

Hierarchical Clustering: Predicting Subgroups

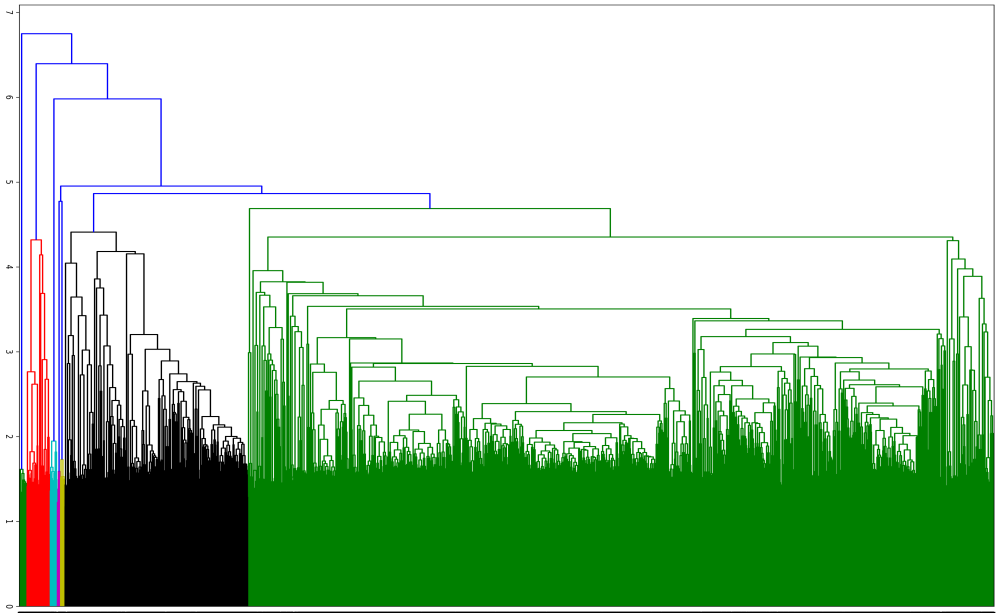
Groups = 3

Groups = 7

Subgroups of Positive Reviews



Subgroups of Negative Reviews



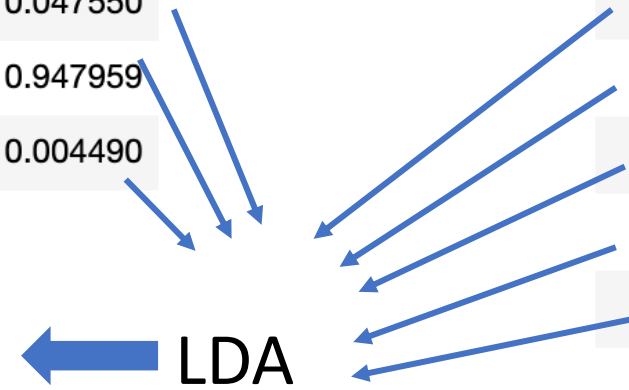
Hierarchical Clustering

Positive Clusters

cluster	review	sentiment	percentage
0	593	593	0.047550
1	11822	11822	0.947959
2	56	56	0.004490

Negative Clusters

cluster	review	sentiment	percentage
0	9522	9522	0.765988
1	91	91	0.007320
2	2327	2327	0.187193
3	97	97	0.007803
4	55	55	0.004424
5	298	298	0.023972
6	41	41	0.003298



Topic Words for each group

Latent Dirichlet Allocation

- Topic Modeling: a repeating pattern of co-occurring terms in a corpus

“health”, “doctor”, “patient”, “hospital” for a topic – Healthcare

“farm”, “crops”, “wheat” for a topic – “Farming”

- Bag of words: sequence of words in sentence is not important
- Unsupervised
- Data Processing: remove stopwords, lemmatize, tokenize, remove punctuations, len(token)>4

LDA – Output(Positive)

Mario-spiderman-superhero-game

['game', 'first', 'spiderman', 'graphic', 'comic', 'show', 'played', 'level', 'time', 'payne', 'contestant', 'superhero', 'spider-man', 'horror', 'though', 'something', 'person', 'hour', 'better', 'expect']

['level', 'game', 'graphic', 'james', 'mario', 'great', 'world', 'think', 'weapon', 'first', 'enemy', 'voice', 'bowser', 'gameplay', 'character', 'every', 'quite', 'secret', 'around', 'system']

['game', 'mission', 'great', 'story', 'character', 'still', 'played', 'scene', 'final', 'chess', 'every', 'time', 'control', 'super', 'first', 'mario', 'graphic', 'series', 'playing', 'around']

Positive:

Star Trek

['spock', 'planet', 'kolchak', 'enterprise', 'earth', 'episode', 'family', 'mccoy', 'captain', 'character', 'vulcan', 'series', 'year', 'time', 'three', 'abigail', 'first', 'edmund', 'james', 'death']

Positive

Musical-Woman-Taylor-Comedy

['story', 'character', 'people', 'world', 'young', 'woman', 'scene', 'first', 'performance', 'family', 'film', 'father', 'great', 'director', 'year', 'never', 'human', 'still', 'taylor', 'think']

['great', 'character', 'story', 'scene', 'think', 'watch', 'first', 'people', 'little', 'film', 'still', 'never', 'actor', 'acting', 'funny', 'better', 'watching', 'comedy', 'performance', 'though']

['great', 'film', 'scene', 'musical', 'story', 'performance', 'director', 'play', 'character', 'stewart', 'comedy', 'number', 'little', 'first', 'murder', 'actor', 'woman', 'young', 'played', 'dance']

Negative

World War European-Japenese

['american', 'world', 'people', 'political', 'documentary', 'propaganda', 'dream', 'stone', 'german', 'soldier', 'london', 'salman', 'audience', 'british', 'black', 'family', 'seberg', 'japanese', 'history', 'niven']

Musical-feminism-taylor

['character', 'story', 'woman', 'performance', 'novel', 'actor', 'version', 'seems', 'scene', 'played', 'young', 'director', 'play', 'great', 'rather', 'year', 'taylor', 'musical', 'richard', 'look']

Batman-tv series

Vampire-magic-horror

Slasher-horror-scarecrow-bloody

Business Insights – User Profile

- **The Reviewer Clusters Giving Positive Reviews:**
  - Animation, comedy, family story tends to receive more positive reviews.
  - Startrack Fans
  - Superhero/ Spiderman/Game/Mario

    …….
- **The Reviewer Clusters Giving Negative Reviews:**
  - Batman-tv series
  - Vampire-magic-horror
  - Slasher-horror-scarecrow

Business Insights – Classification Model

- **Sentiment Analysis:**
  - Our classification model is built on reviews with clear rating (IMDB)
  - Applied in classifying comments without score (Social Media)
  - Explore public reaction to a certain movie or event.

Return on Investment

- **Costs:** 2 Data Scientists
  (Long term project - maintain database and update model)
  - $ **202K per year**

- **Benefits:** More profitable movie project investment
  - **30%** improvement of **profit**

- **Risks:** Sudden change of the market preference that we might not capture by simply using our model

- **Implementation Roadmap**:
  Social media movie comments data
  -> Classification model
  -> Clustering
  -> Improve future production and project evaluation