Finding Patterns in the Stream: A Machine Learning Analysis of Netflix Movie Data

Team 8

AGENDA

- Context
- Datasets Used
- IMDb Rating Prediction
- Topic Modeling of Netflix Movie Plots
- Clustering for Movie Classes of Interest
- Association Rule Mining for Commonly Watched Movies
- Future Work

Netflix Management wants to know!

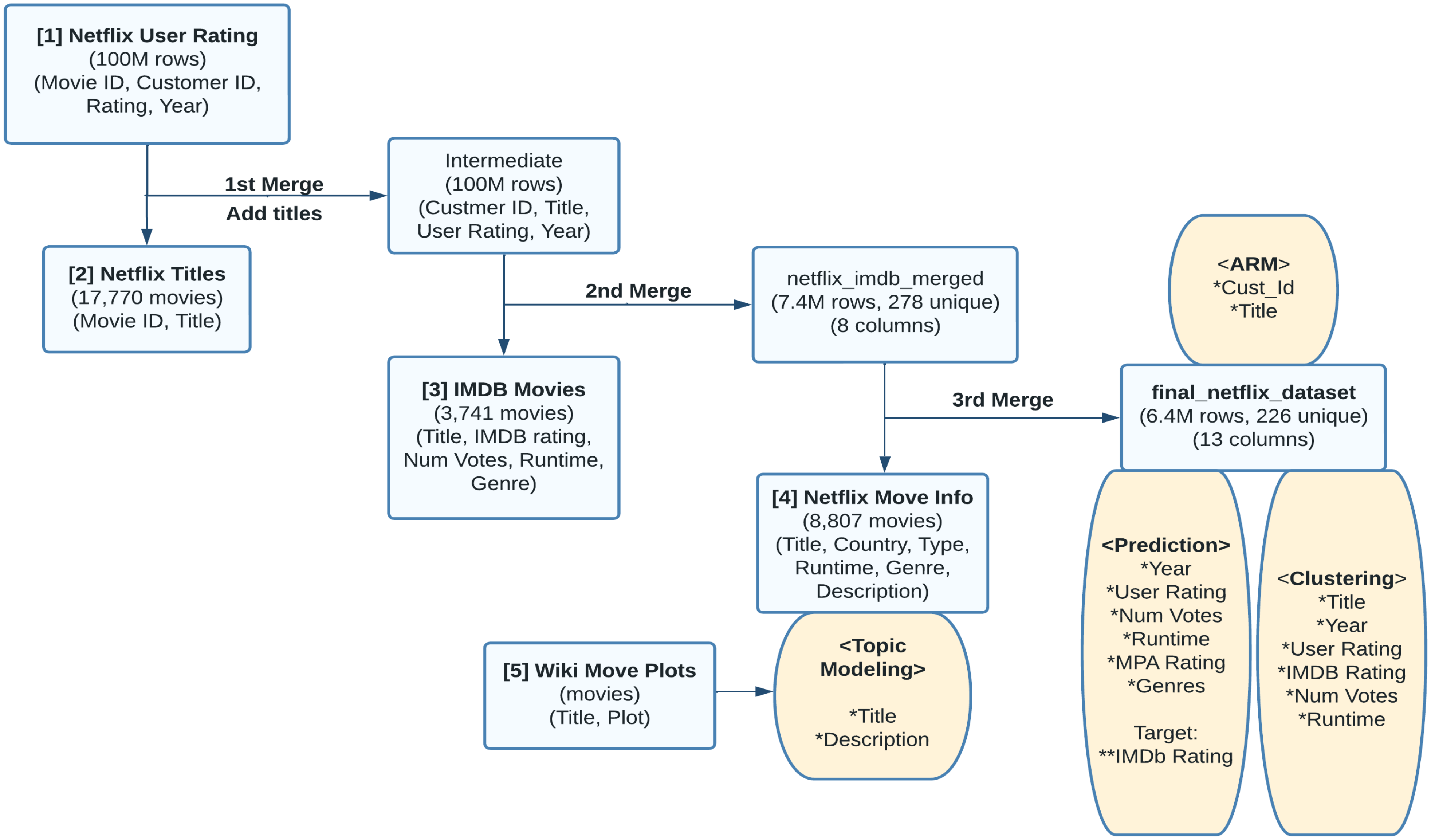
1. What is the IMDb rating for a Netflix movie, given its information and rating from Netflix?

2. Is the Netflix movie description sufficient for us to know the topic of the movie?

3. What are the features of the movies which receive the highest and lowest ratings from our customers?

4. What movies do users frequently watch together?

Netflix Data Merging Process
& Tasks Performed



Q1. What is the IMDb rating for a Netflix movie, given its information and rating from Netflix?

Target Variable: **IMDb Rating**

Guiding Algorithm

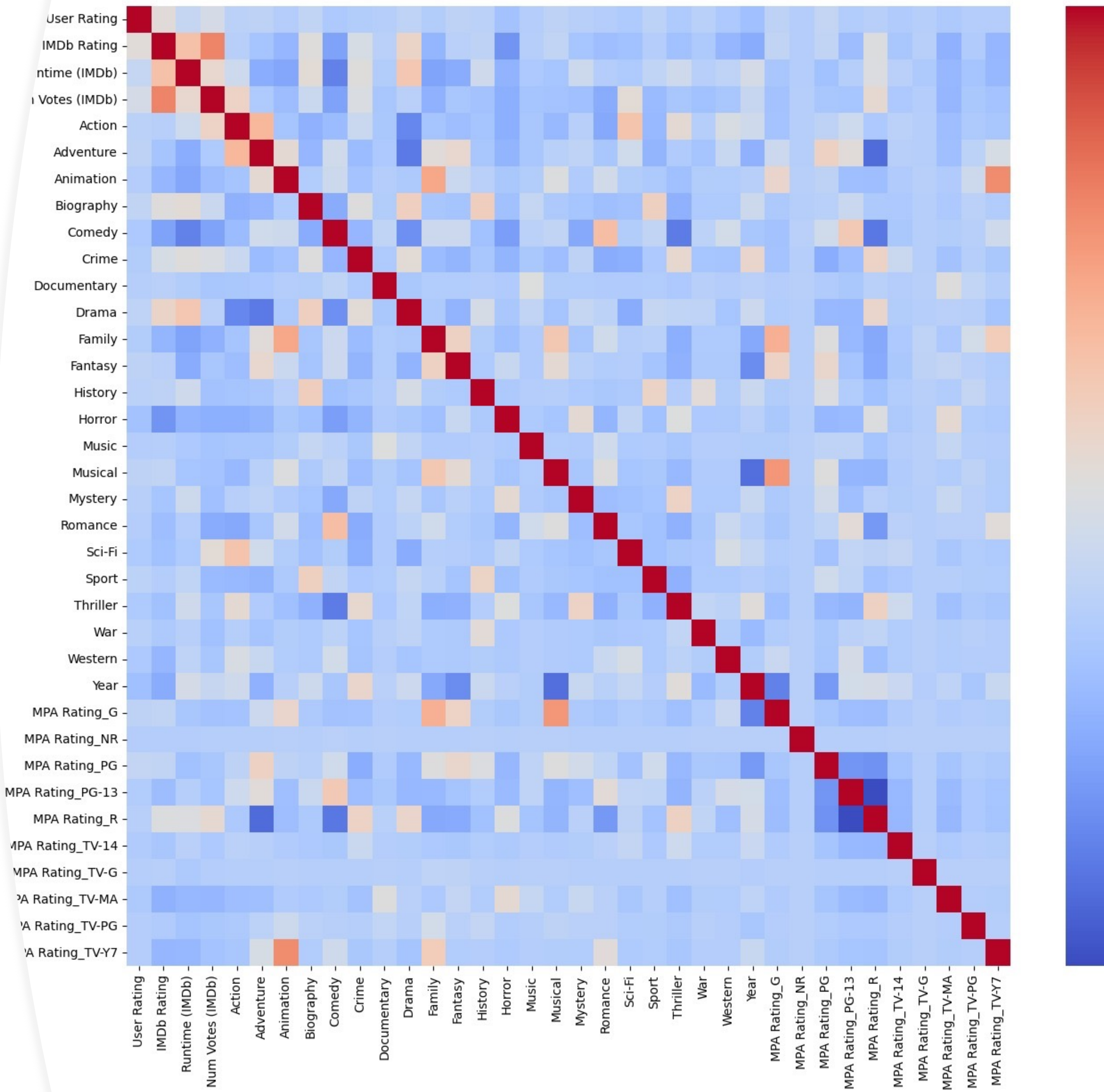
- Exploratory Analysis
- Feature Engineering
- One-Hot & Ordinal Encoding
- Scaling
- Splitting
- Modeling
- Model Evaluation

Feature Engineering

Number of Votes and Runtime have high correlations with the Target Variable

Might play a big role in this Prediction Model

No other sources of collinearity, so we don't remove any other features



Baseline Value – Mean

- Baseline Value of Target Variable:
  - Mean of the Training Data
- Root Mean Squared Error in Baseline
  - 0.923

Results from Predictive Modeling

Model	RMSE
Decision Tree Regressor	$1.004 * e^{-11}$
Random Forest Regressor	$3.77 * e^{-12}$
Gradient Boosting Regressor	0.183
XGBoost Regressor	0.001

Hyperparameter Tuning

Using GridSearchCV

Parameters Used

- Number of Estimators
- Minimum Samples Leaf
- Maximum Leaf Nodes

OOB Score = True for Validation

Best Model: Random Forest Regressor

- RMSE = 0.069
- Test $R^2 = \mathbf{0.9943}$
- Hyperparameters
    - max_leaf_nodes = 100
    - min_samples_leaf = 2
    - n_estimators = 300

Q2. Is the Netflix movie description sufficient for us to know the topic of the movie?

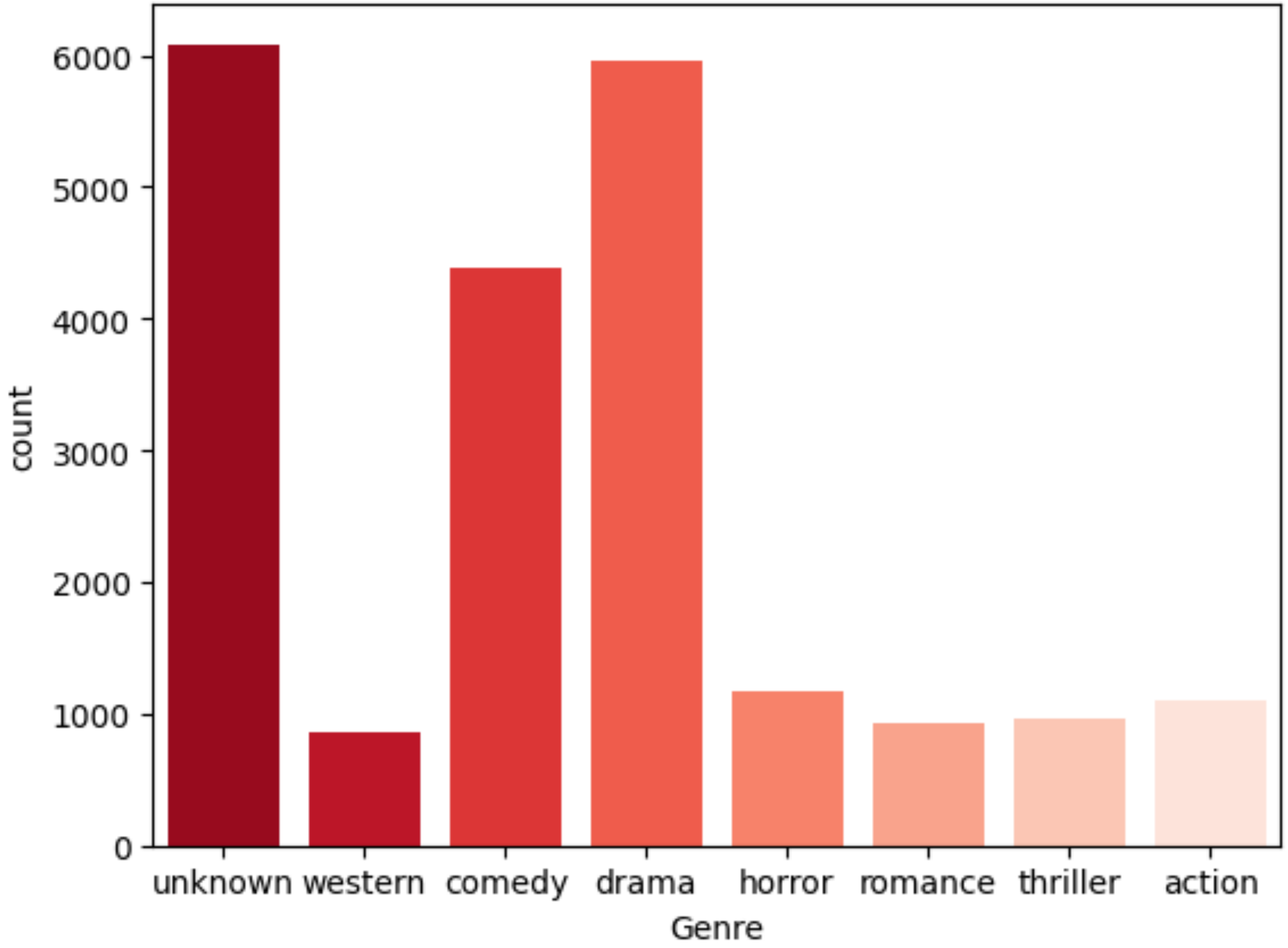
No, Netflix description was too short (insufficient data)

As her father nears the end of his life, filmmaker Kirsten Johnson stages his death in inventive and comical ways to help them both face the inevitable.

Maybe… Wikipedia plot description?

Q2. Is the Netflix movie description sufficient for us to know the topic of the movie?

Countplot of genres



Methods

Text Pre-Processing

- Stop Words
- Bigrams
- Lemmatization

Modeling

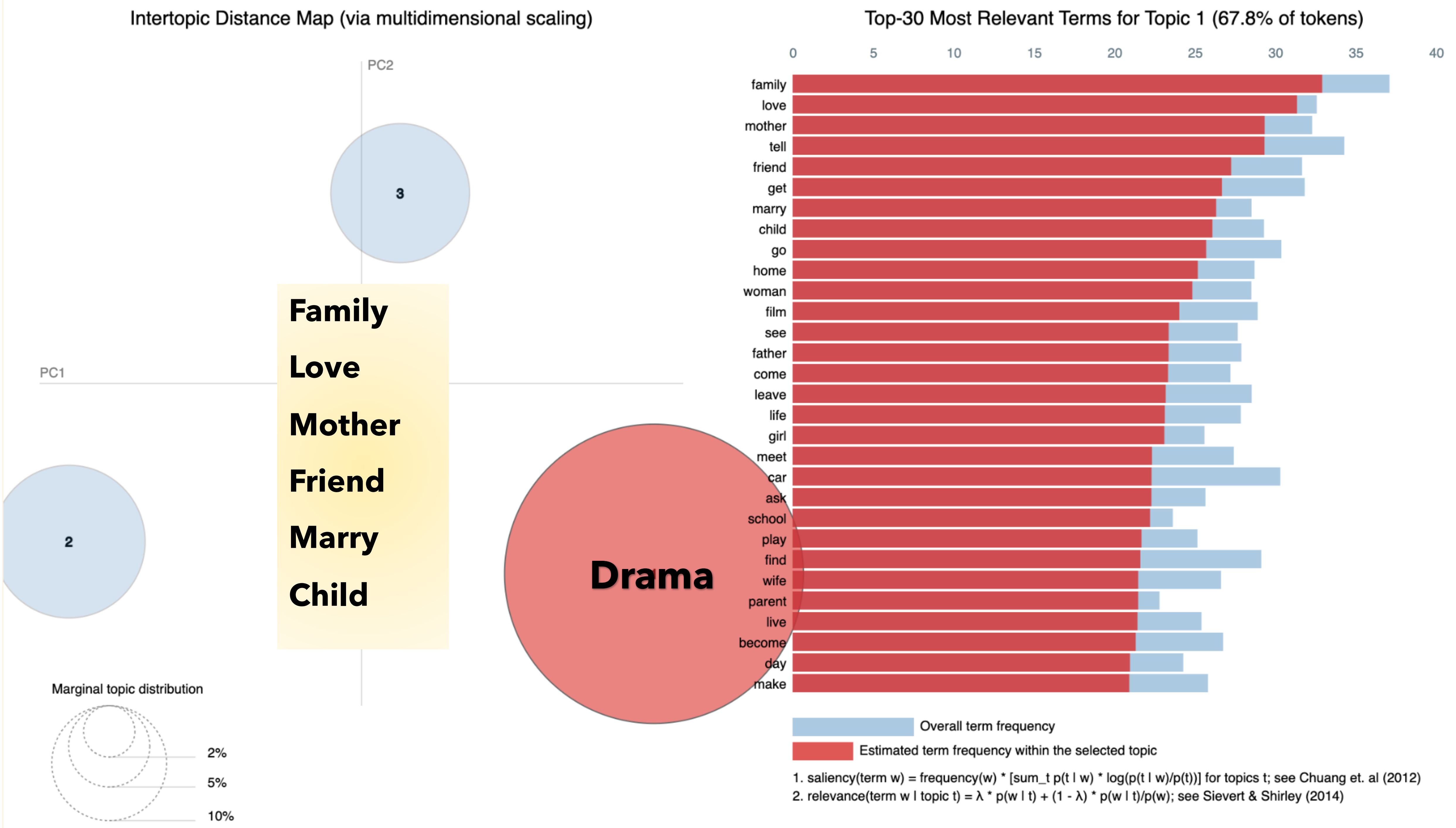
- LDA
- LSA
- NMF

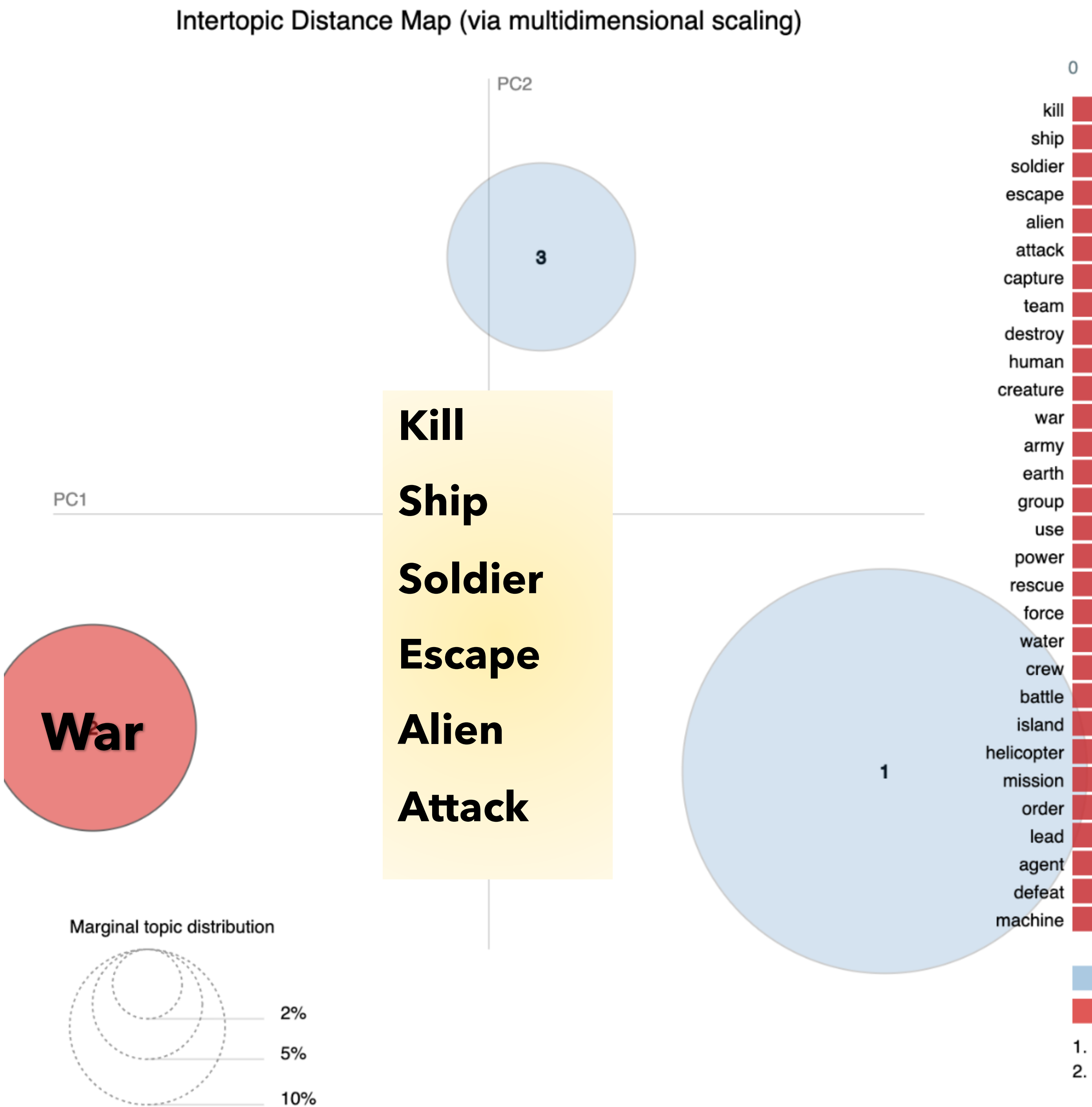
Tuning

- # of topics

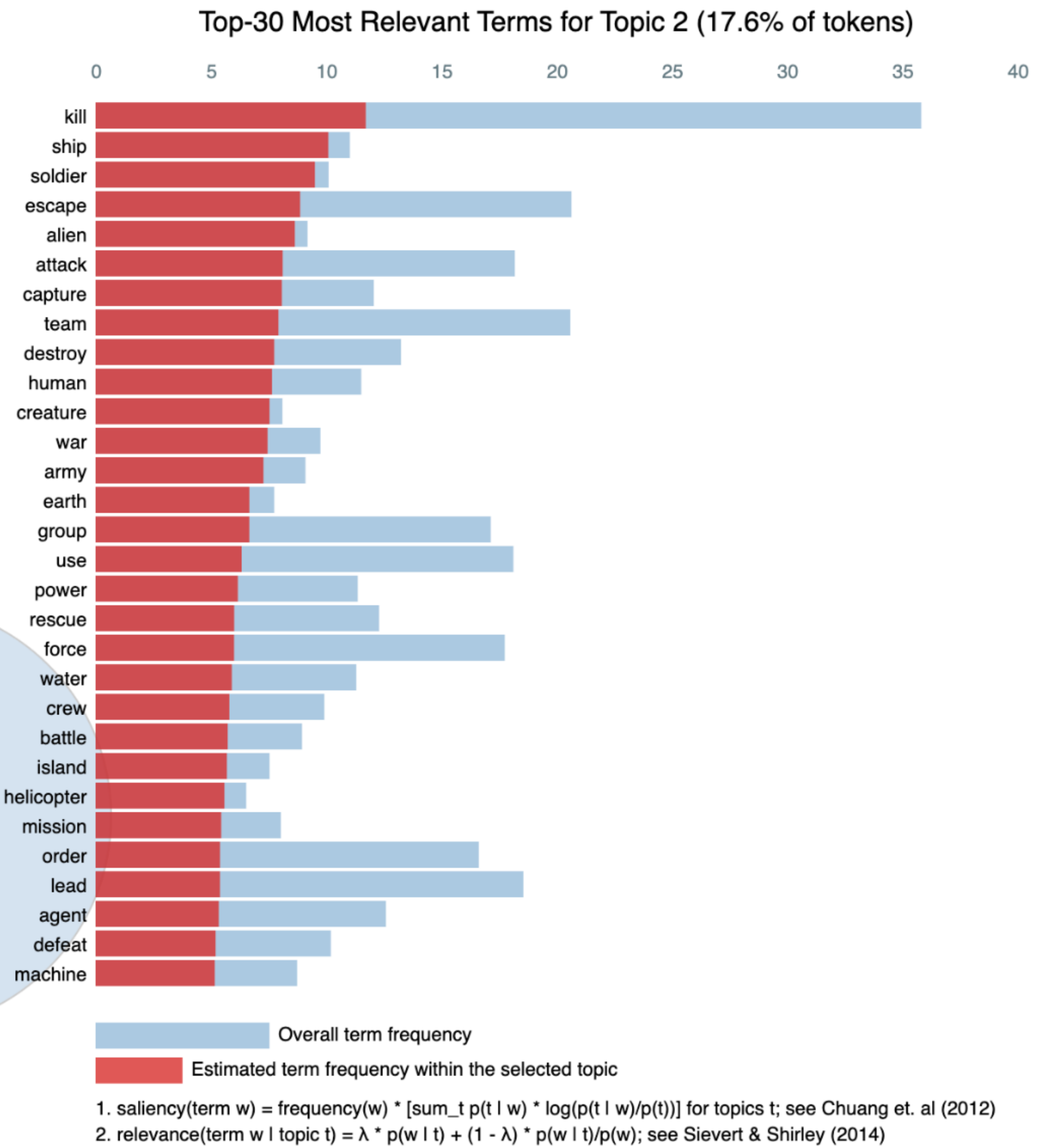
Evaluation

- Coherence score

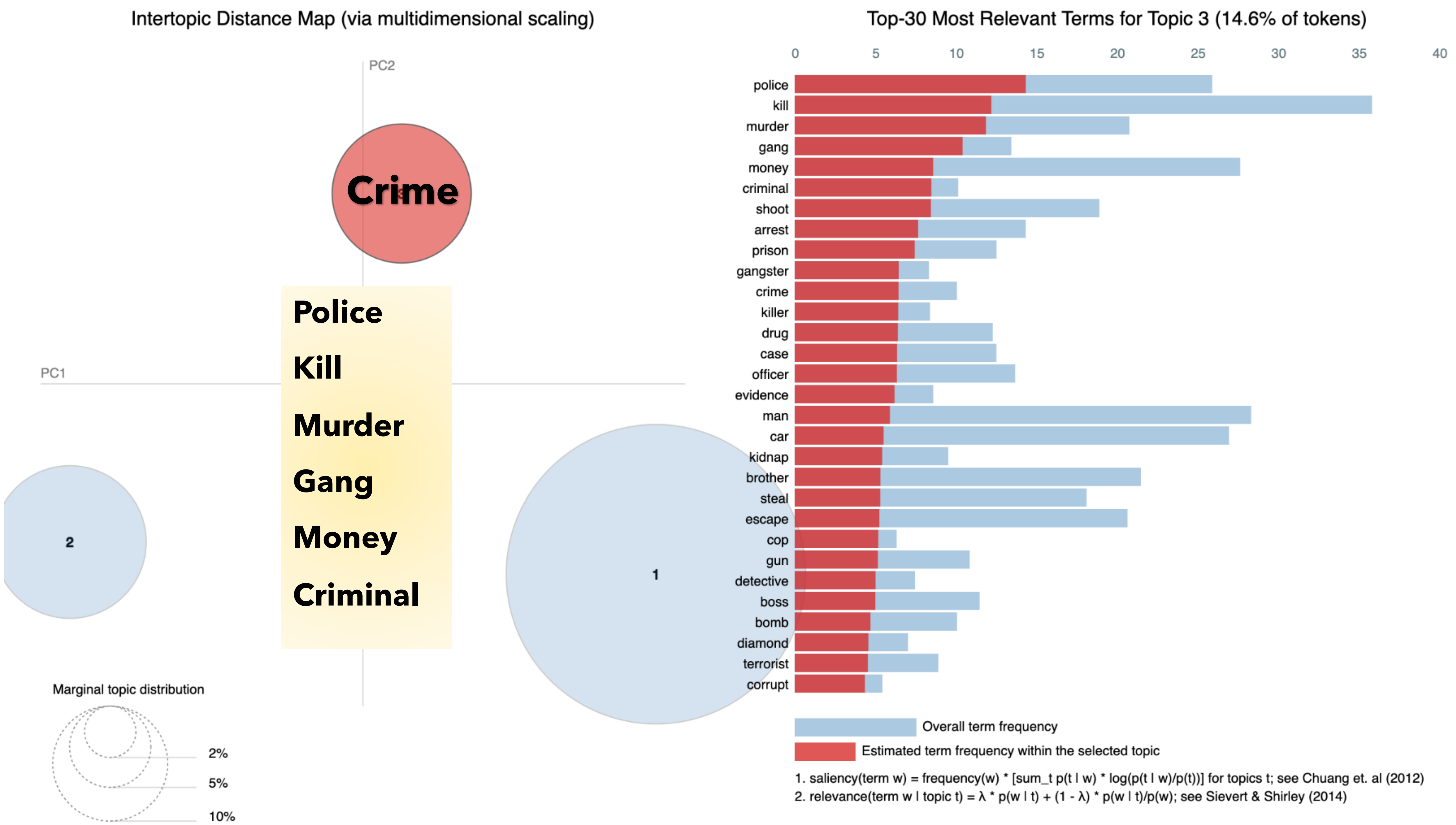

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
3
2
Family
Love
Mother
Friend
Marry
Child
Drama
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (67.8% of tokens)
0
5
10
15
20
25
30
35
40
family
love
mother
tell
friend
get
marry
child
go
home
woman
film
see
father
come
leave
life
girl
meet
car
ask
school
play
find
wife
parent
live
become
day
make
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)



Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
3
1
Kill
Ship
Soldier
Escape
Alien
Attack
War
Marginal topic distribution
2%
5%
10%



Top-30 Most Relevant Terms for Topic 2 (17.6% of tokens)
0
5
10
15
20
25
30
35
40
kill
ship
soldier
escape
alien
attack
capture
team
destroy
human
creature
war
army
earth
group
use
power
rescue
force
water
crew
battle
island
helicopter
mission
order
lead
agent
defeat
machine
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Crime
Police
Kill
Murder
Gang
Money
Criminal
2
1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 3 (14.6% of tokens)
0
5
10
15
20
25
30
35
40
police
kill
murder
gang
money
criminal
shoot
arrest
prison
gangster
crime
killer
drug
case
officer
evidence
man
car
kidnap
brother
steal
escape
cop
gun
detective
boss
bomb
diamond
terrorist
corrupt
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Q3. What are the features of the movies which receive the highest and lowest ratings from our customers?

Unsupervised ML Model

Guiding Algorithm

- Mine for underlying classes in the data
- Look at distributions of features across the labelled classes for sanity check
- Conduct Principal Component Analysis
- Look at Clusters across PC1 and PC2

Models Explored

KMeans Clustering

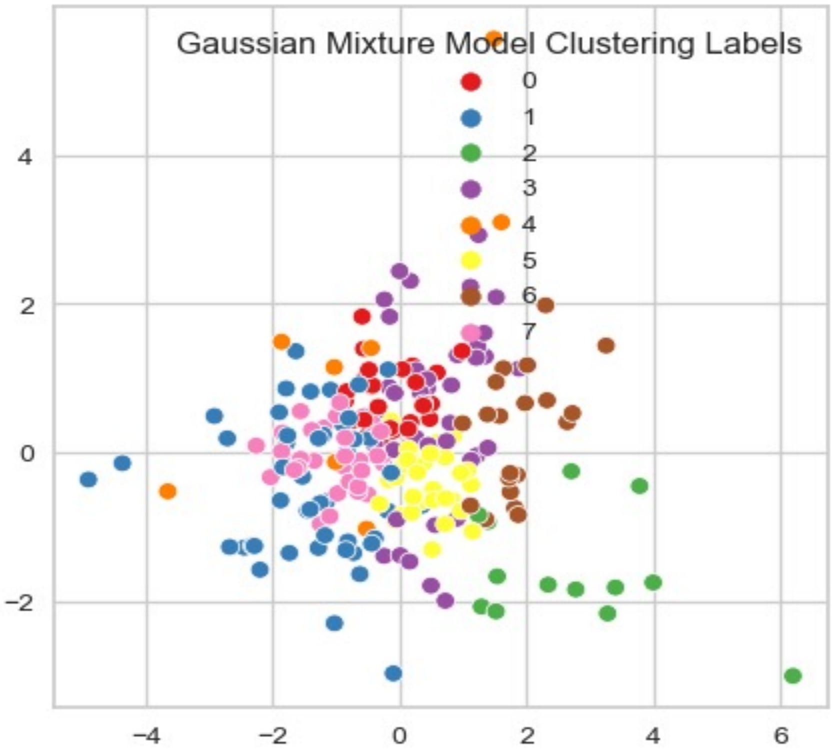
Agglomerative Clustering

Gaussian Mixture Methods

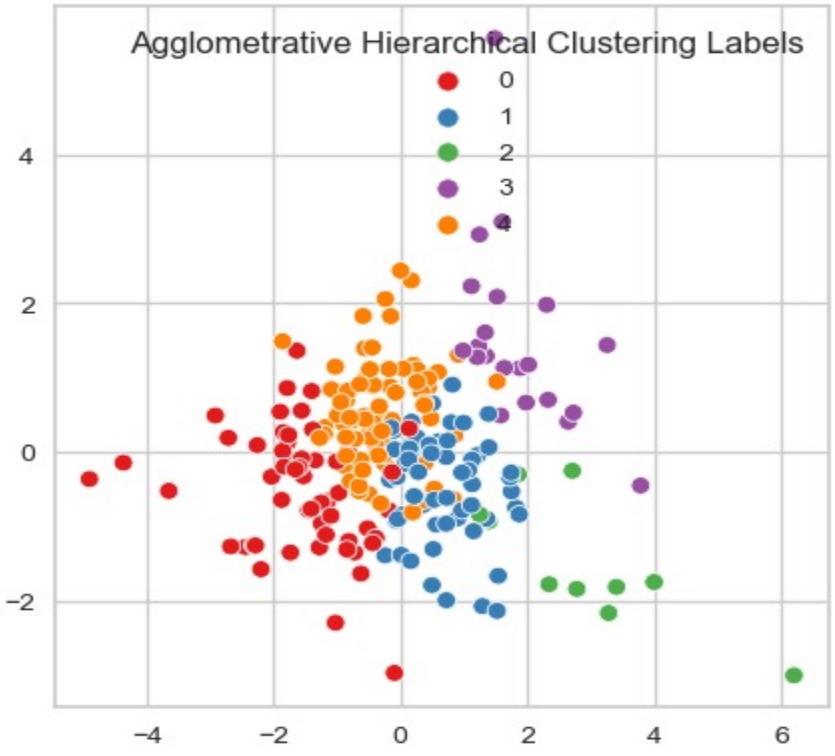
Spectral Clustering

Results – The *Bad* Models

Gaussian Mixture Model

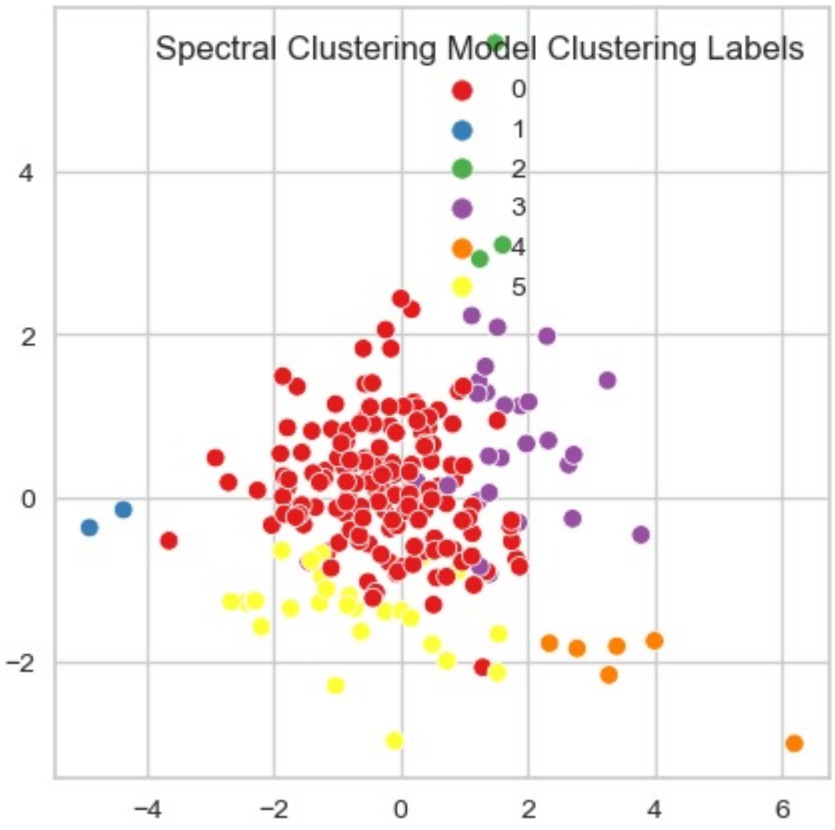


Agglomerative Clustering

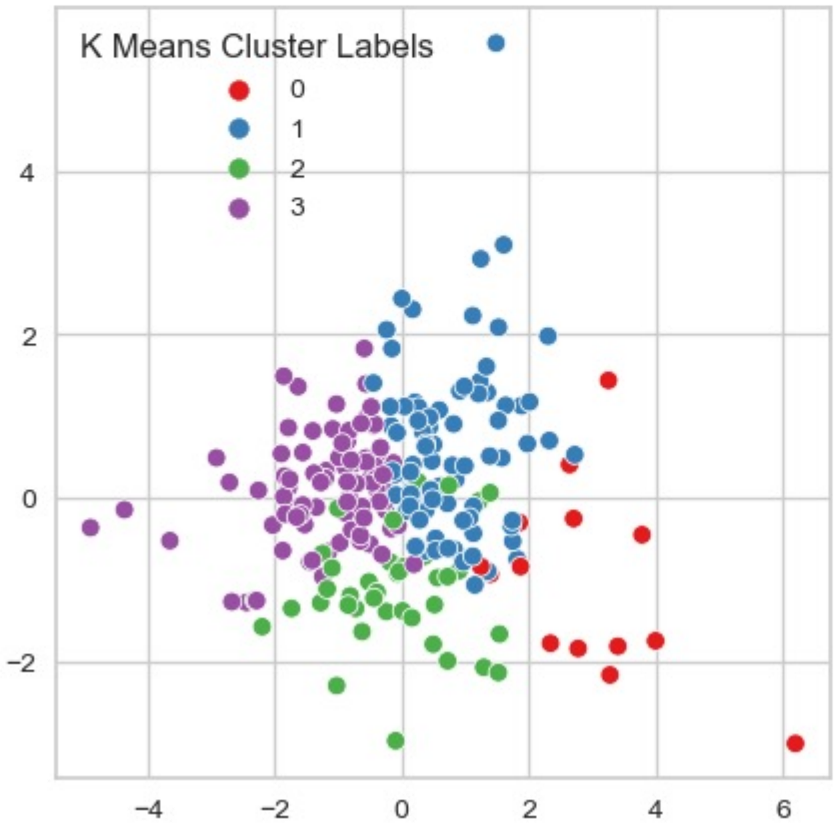


Results – The *Good* Models

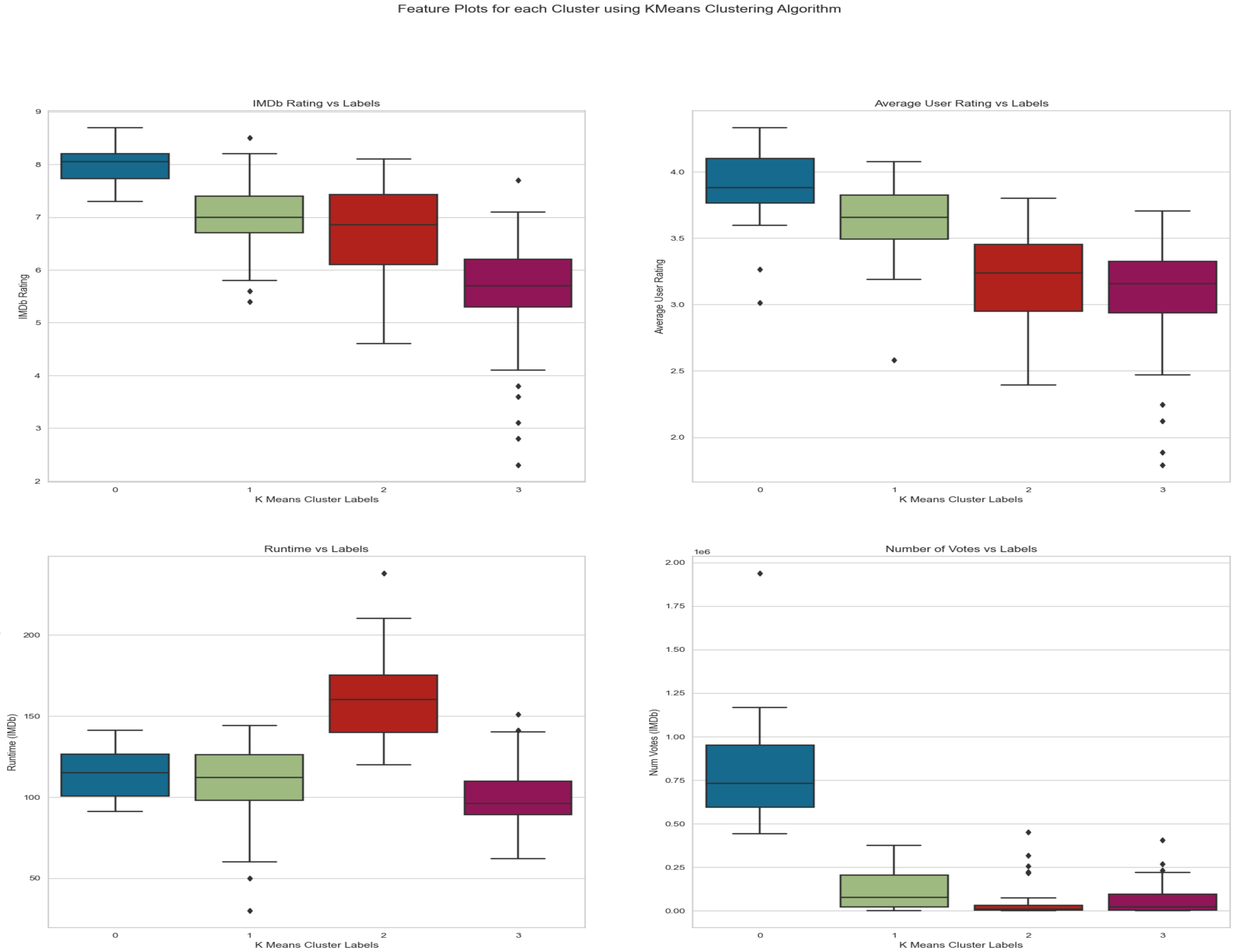
Spectral Clustering Model



KMeans Clustering Model



Best Model



Feature Plots for each Cluster using KMeans Clustering Algorithm
IMDb Rating vs Labels
IMDb Rating
K Means Cluster Labels
Average User Rating vs Labels
Average User Rating
K Means Cluster Labels
Runtime vs Labels
Runtime (IMDb)
K Means Cluster Labels
Number of Votes vs Labels
1e6
Num Votes (IMDb)
K Means Cluster Labels

Q4. What movies do users frequently watch together?

Unsupervised Machine Learning Model

Guiding Algorithm

- Create Item Sets, i.e., Most Frequent Antecedents
- Find Association Rules
- Prune Using Lift and Support Metrics
- Look at Confidence Support Frontier

Models Explored

- Apriori Algorithm
- FPGrowth Algorithm
  - Faster processing Time
  - Less Computationally Intensive

Results – Top 10 Rules

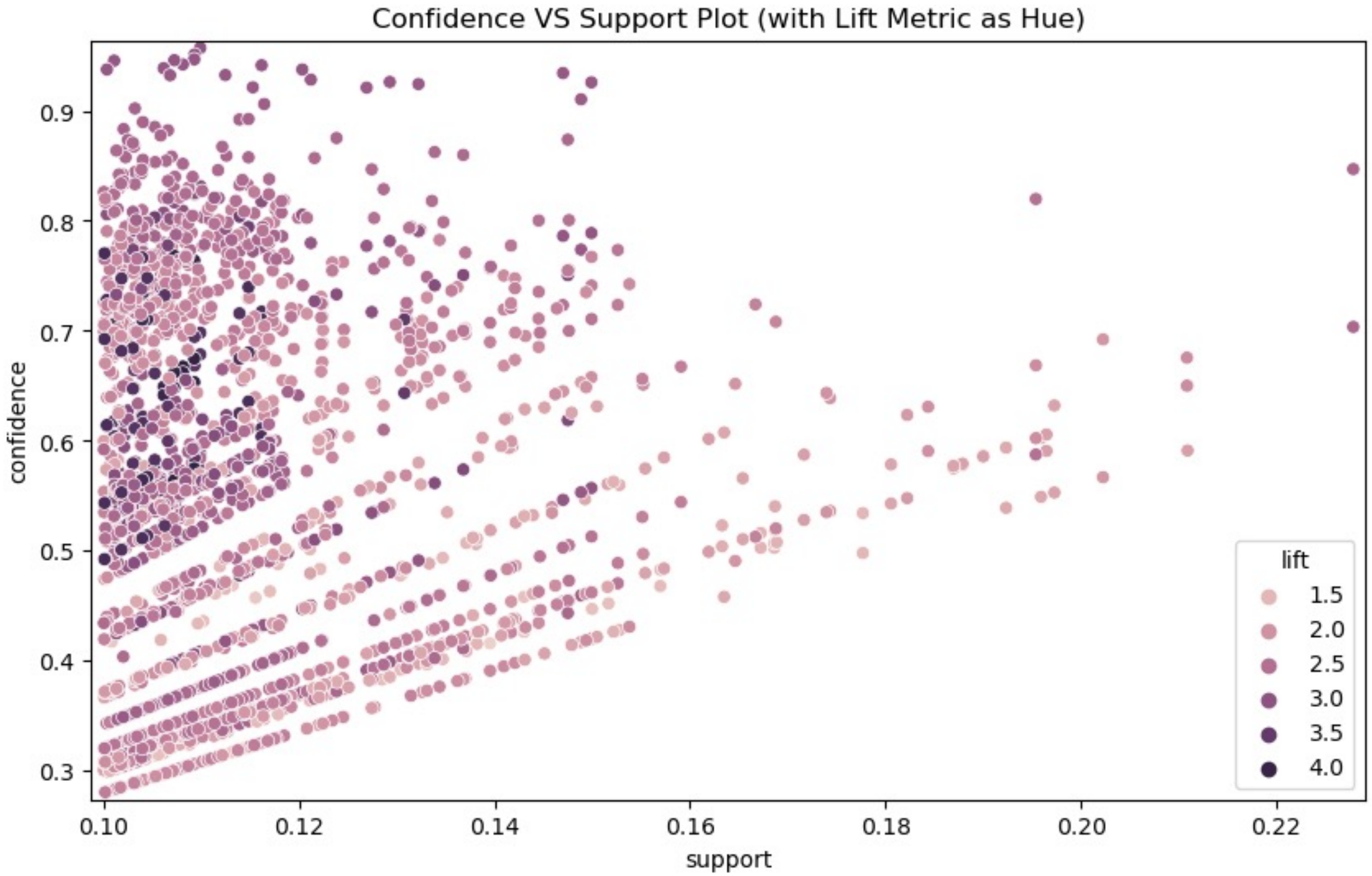
Support Metric Pruning

antecedents	consequents
(Kill Bill: Vol. 1)	(Kill Bill: Vol. 2)
(Kill Bill: Vol. 2)	(Kill Bill: Vol. 1)
(The Matrix)	(American Beauty)
(American Beauty)	(The Matrix)
(Men in Black)	(The Matrix)
(The Matrix)	(Men in Black)
(Ferris Bueller's Day Off)	(American Beauty)
(American Beauty)	(Ferris Bueller's Day Off)
(Men in Black)	(American Beauty)
(American Beauty)	(Men in Black)

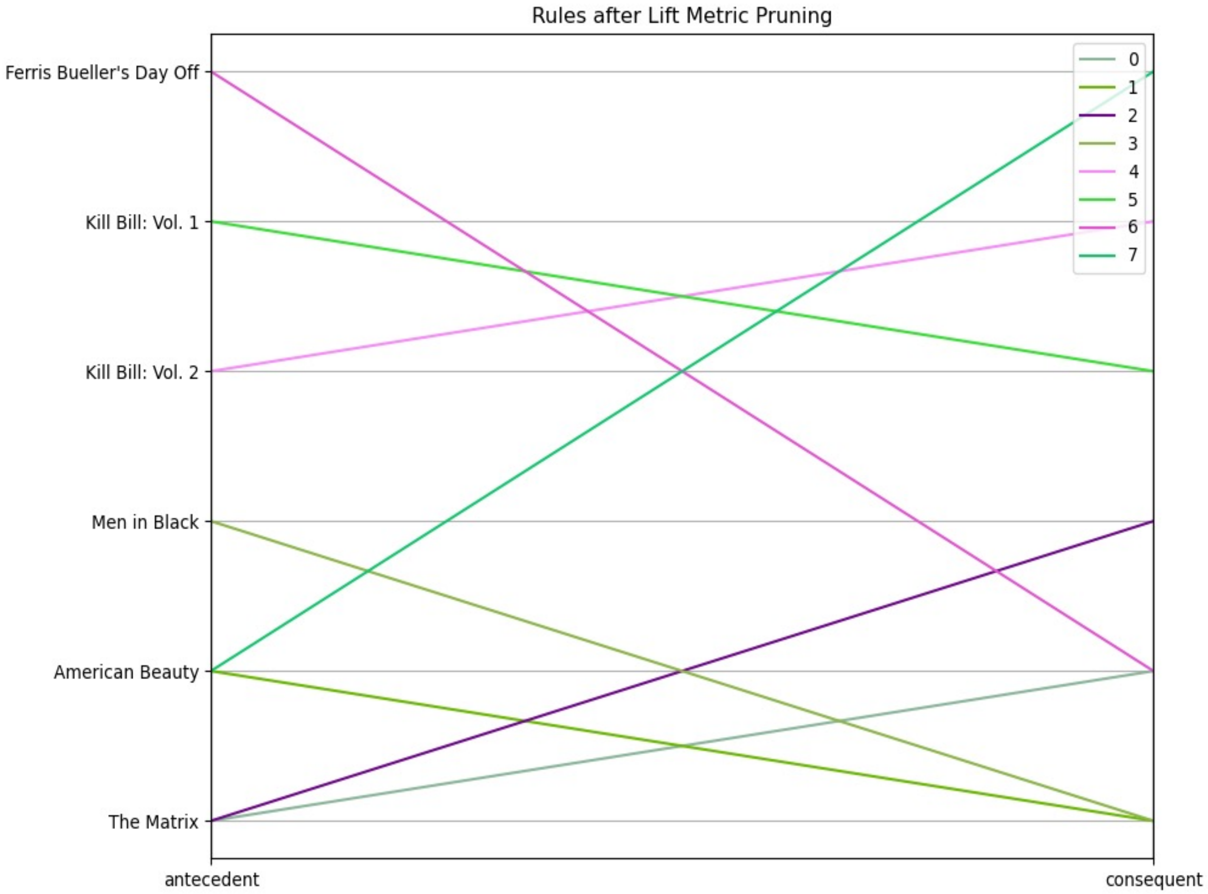
Lift Metric Pruning

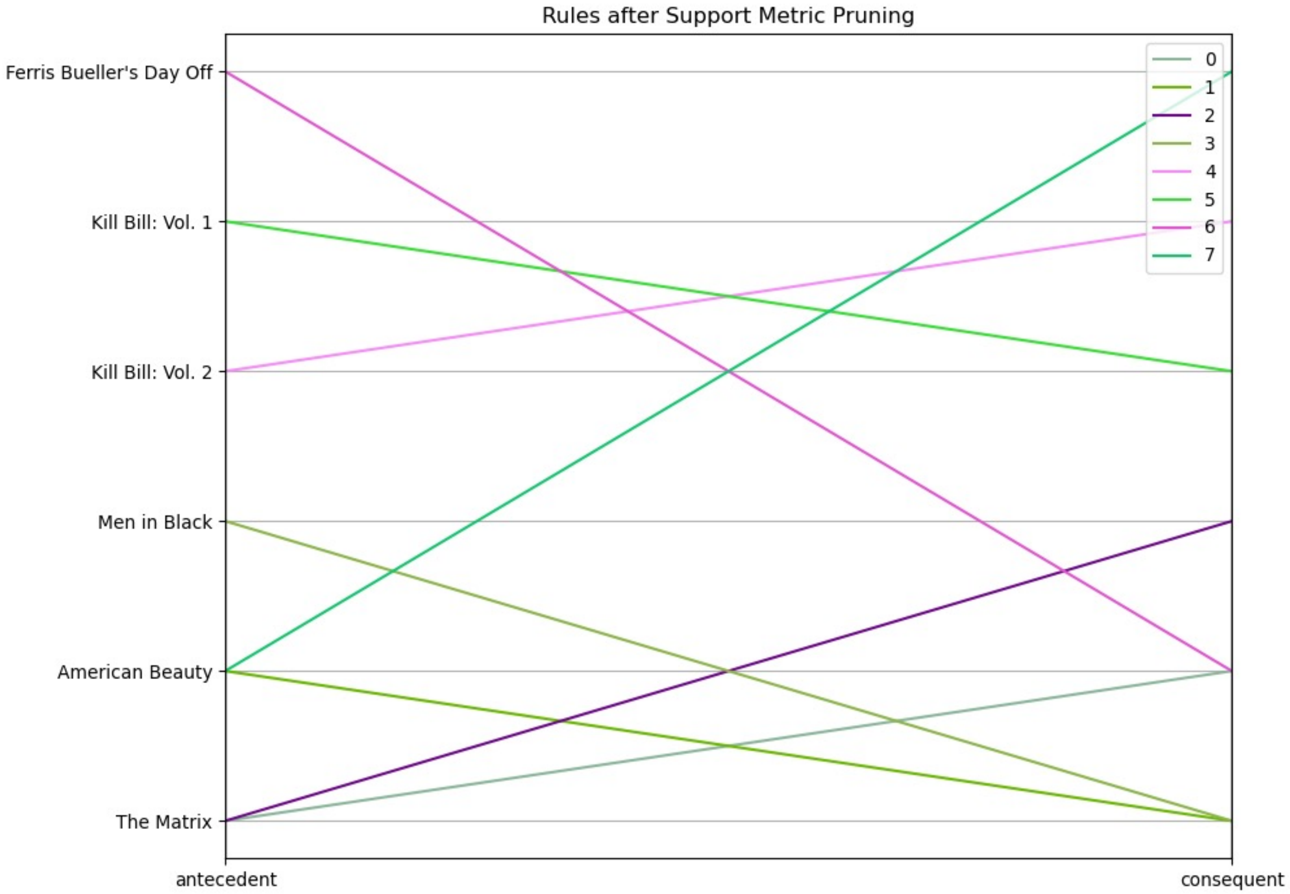
antecedents	consequents
(Kill Bill: Vol. 2, Men in Black)	(Kill Bill: Vol. 1, The Matrix)
(Kill Bill: Vol. 1, The Matrix)	(Kill Bill: Vol. 2, Men in Black)
(Kill Bill: Vol. 2, The Matrix)	(Kill Bill: Vol. 1, Men in Black)
(Kill Bill: Vol. 1, Men in Black)	(Kill Bill: Vol. 2, The Matrix)
(Kill Bill: Vol. 1, The Matrix)	(Kill Bill: Vol. 2, Ferris Bueller's Day Off)
(Kill Bill: Vol. 2, Ferris Bueller's Day Off)	(Kill Bill: Vol. 1, The Matrix)
(Kill Bill: Vol. 2, The Matrix)	(Kill Bill: Vol. 1, Indiana Jones and the Last...
(Kill Bill: Vol. 1, Indiana Jones and the Last...	(Kill Bill: Vol. 2, The Matrix)
(Kill Bill: Vol. 2, Indiana Jones and the Last...	(Kill Bill: Vol. 1, The Matrix)
(Kill Bill: Vol. 1, The Matrix)	(Kill Bill: Vol. 2, Indiana Jones and the Last...

Confidence – Support Border of Association Rules



Results – Parallel Plots





Future Work

1. Develop a recommendation system based on our work on association rule mining and LDA
2. Incorporate customer demographic data to better our predictive model

Any Questions?