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

Sajujya Gangopadhyay | Grace Eunji Kim | Alex Yang

AGENDA

- VALUE
- DATASETS USED
- ANALYTICAL QUESTIONS
- FUTURE WORK

Netflix Management wants to increase…

Customer satisfaction

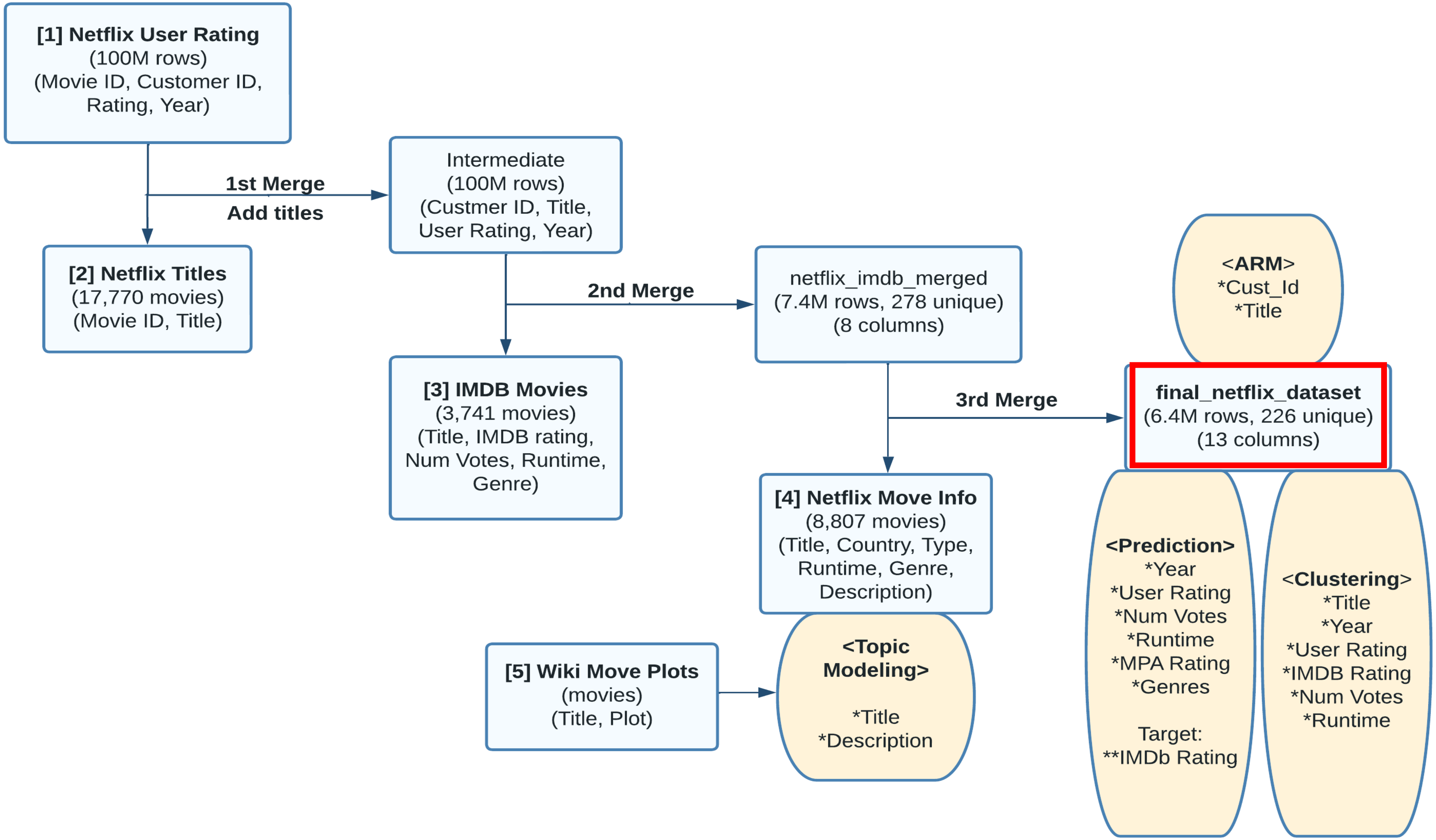
Quality

Customer retention

Netflix Management wants to know!

- Is the Netflix movie plot description sufficient for us to know the **topic** of the movie?
- What is the **IMDb rating** for a Netflix movie, given its information and rating from Netflix?
- What are the **features** of the movies which receive the highest and lowest ratings from our customers?
- What movies do users **frequently watch together**?

Netflix Data Merging Process & Tasks Performed



Q1. Is the Netflix movie description sufficient for us to know the topic of the movie?

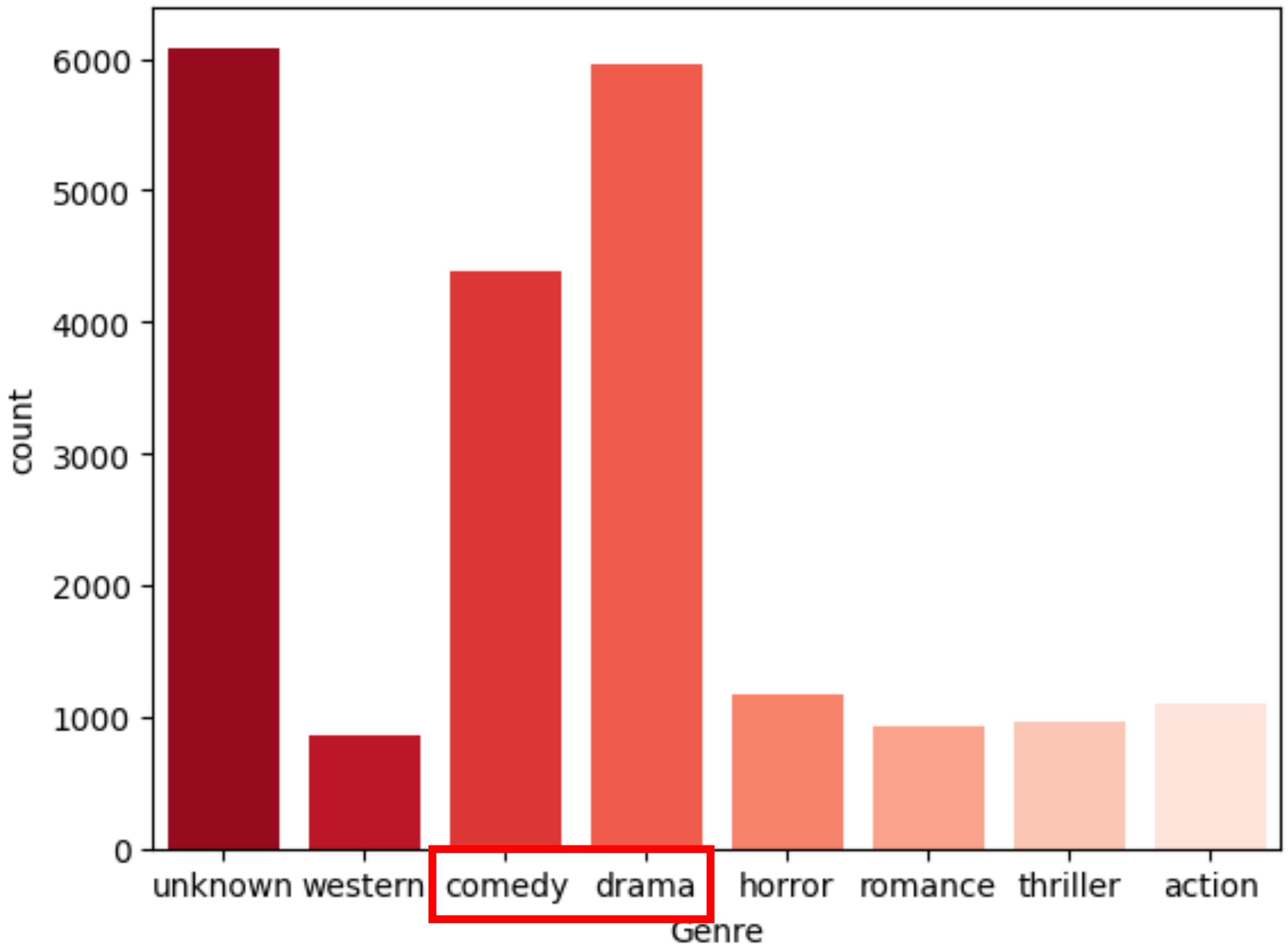
No, **Netflix description was too short (insufficient data)**

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?

Q1. 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

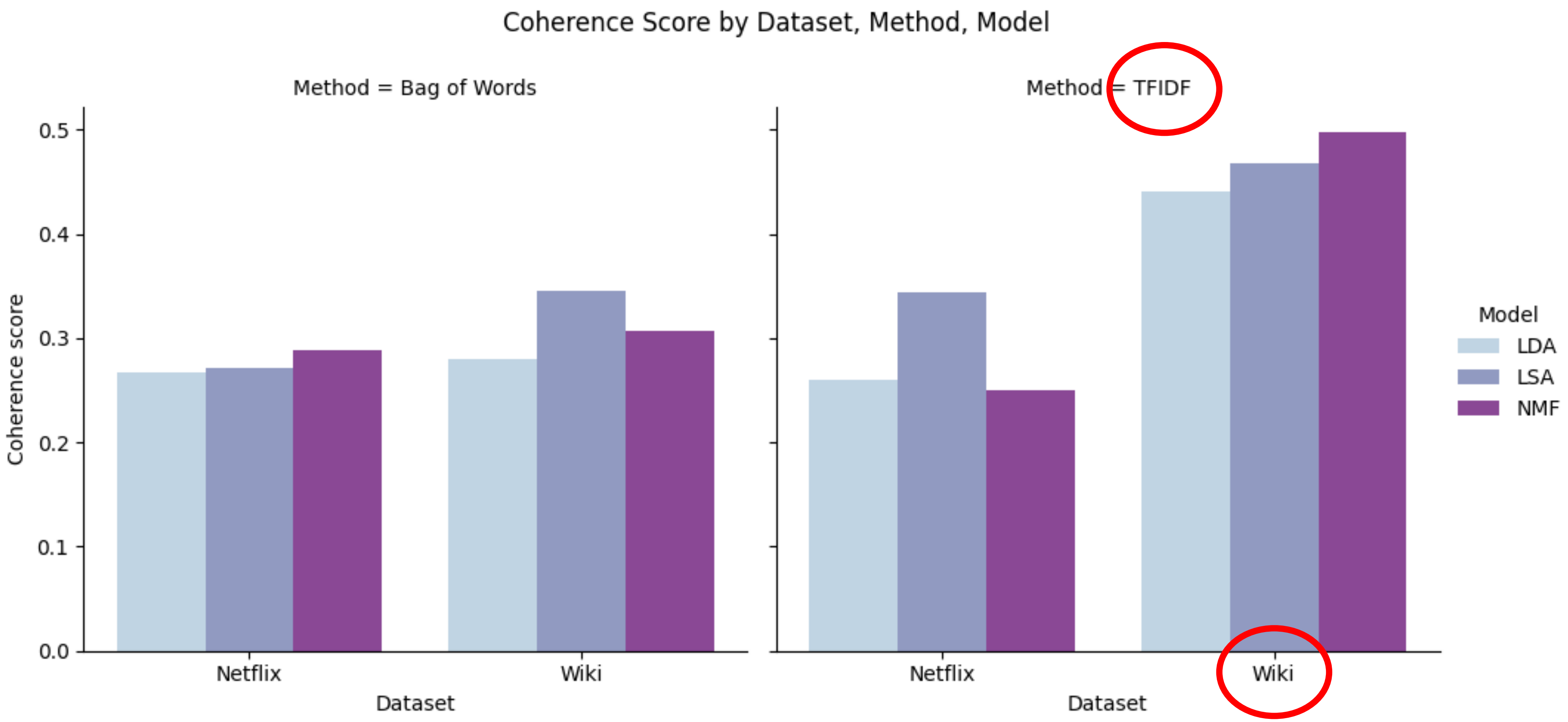
- # of topics

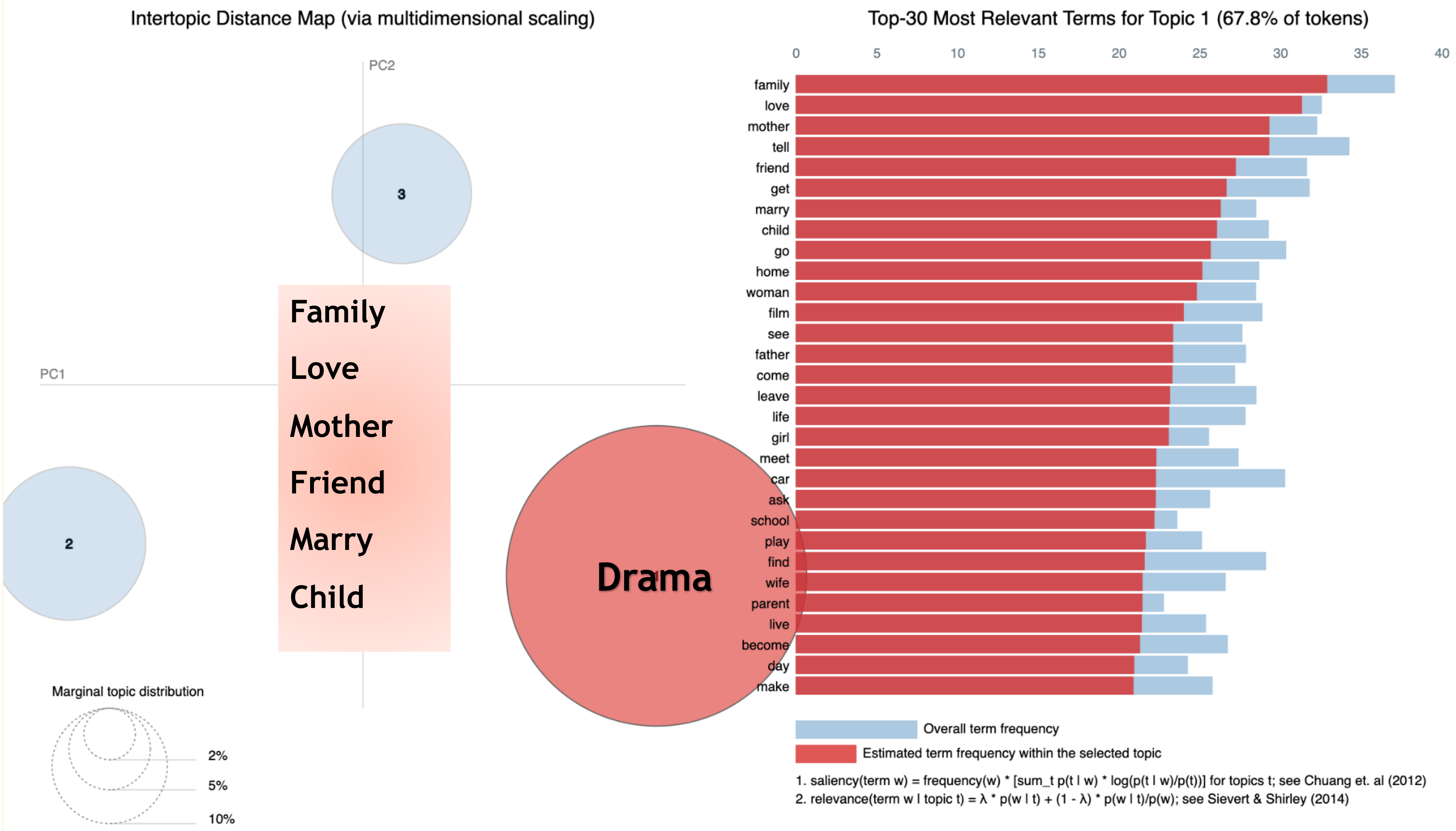
Evaluation

- Coherence score

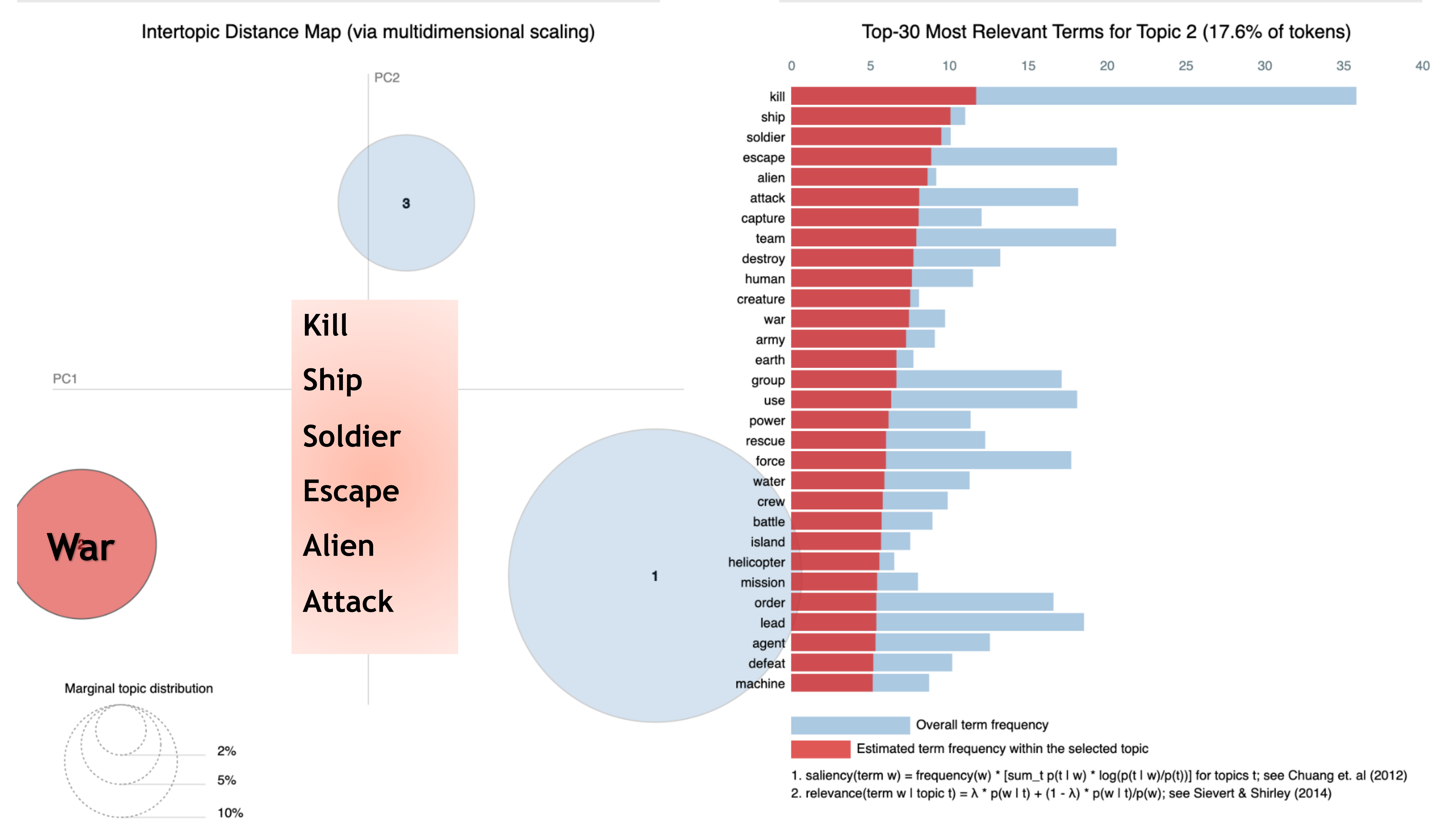
Model Evaluation

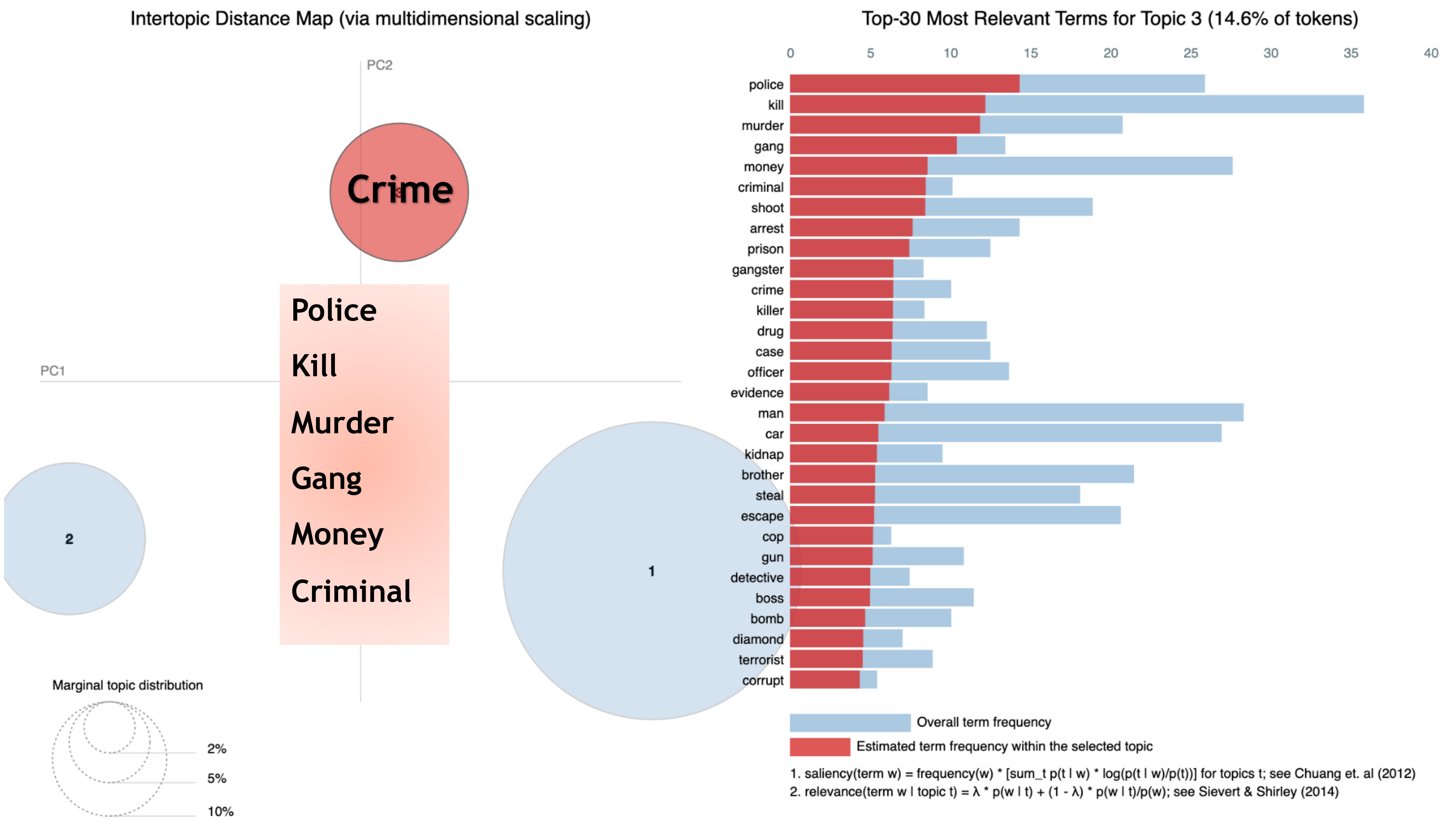
-Interpretable
-Coherent





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)

Q2. What is the **IMDb rating** for a Netflix movie, given its information and rating from Netflix?

Target Variable: **IMDb Rating**

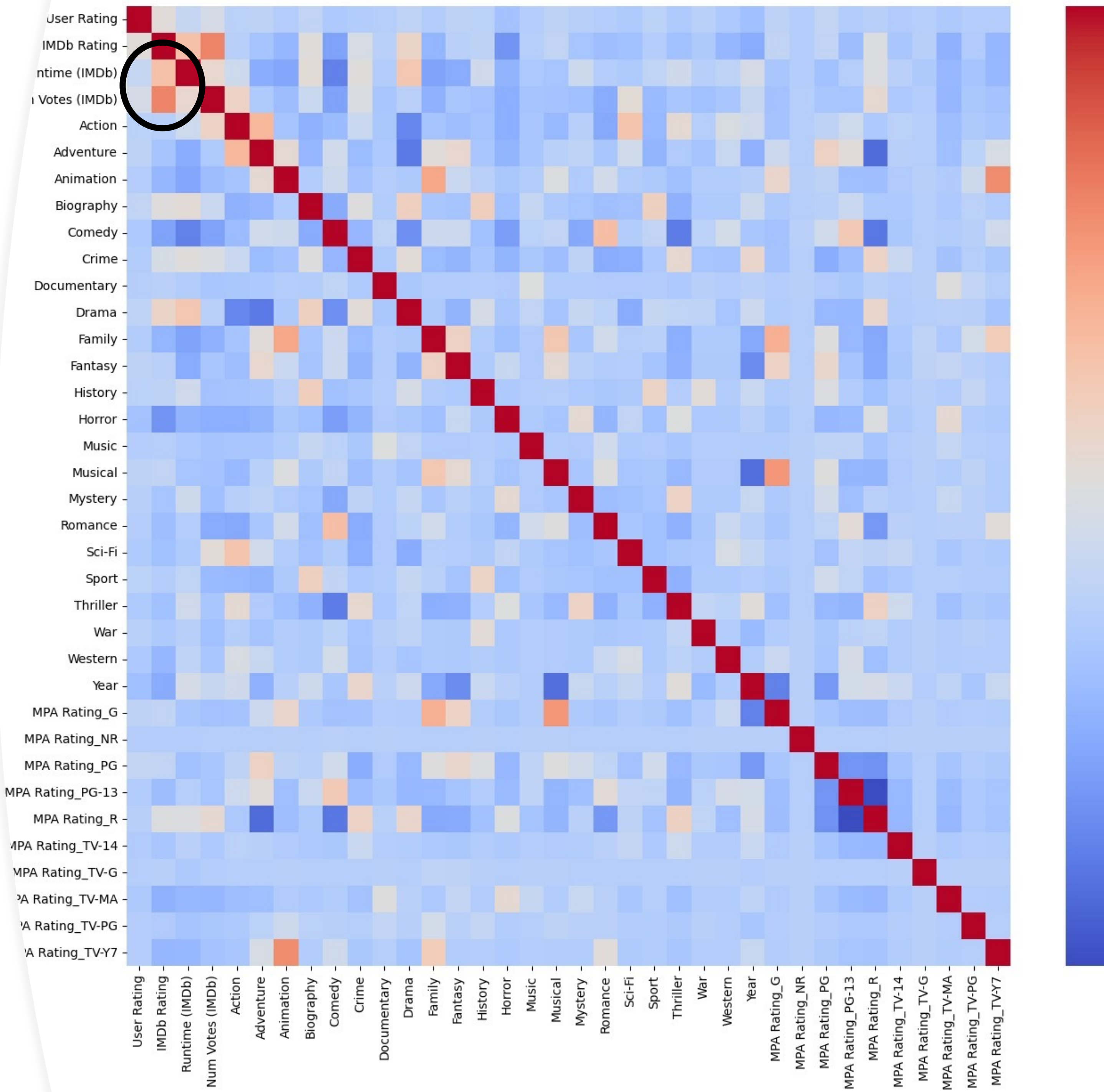
Guiding Algorithm

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

Feature Engineering

High correlations
-Number of Votes
-Runtime

No collinearity otherwise



Results from Predictive Modeling

- Baseline – mean of training data (RMSE 0.923)

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

GridSearchCV (Parameters Used)

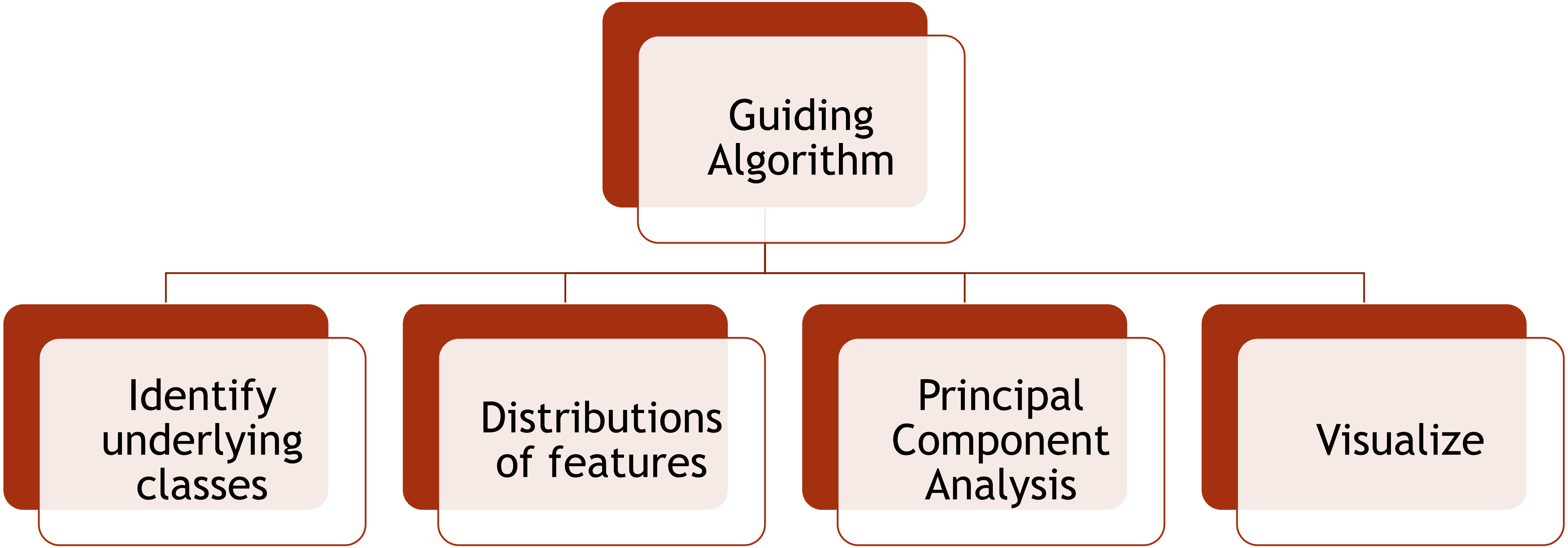
- Number of Estimators
- Minimum Samples Leaf
- Maximum Leaf Nodes

[OOB Score = True] for Validation

Best Model:
Random Forest Regressor

- Hyperparameters
  - max_leaf_nodes = 100
  - min_samples_leaf = 2
  - n_estimators = 300
- RMSE = 0.069
- Test R^2 = 0.994

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



Models Explored

KMeans Clustering

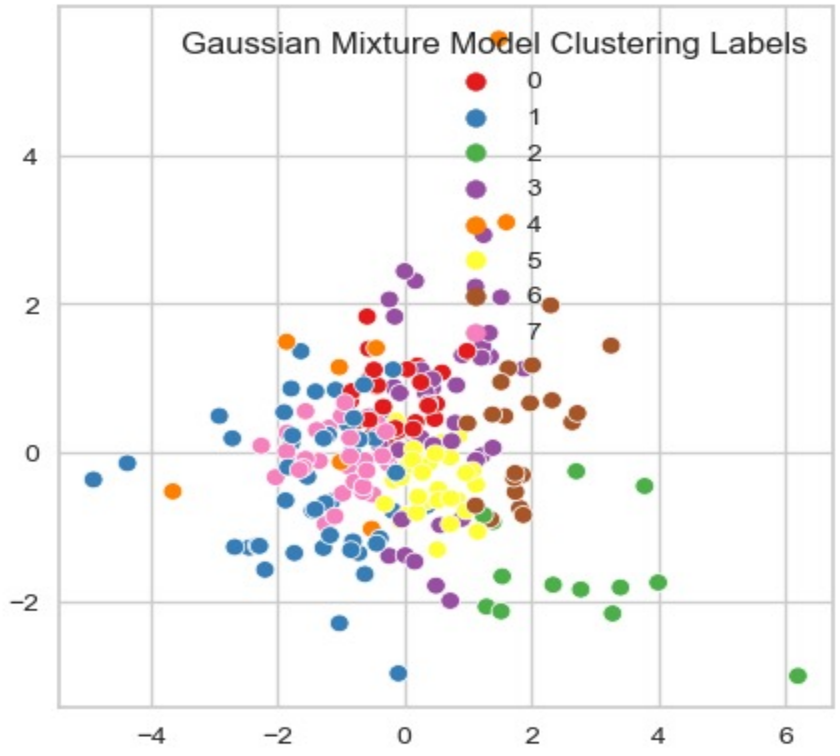
Agglomerative Clustering

Gaussian Mixture Methods

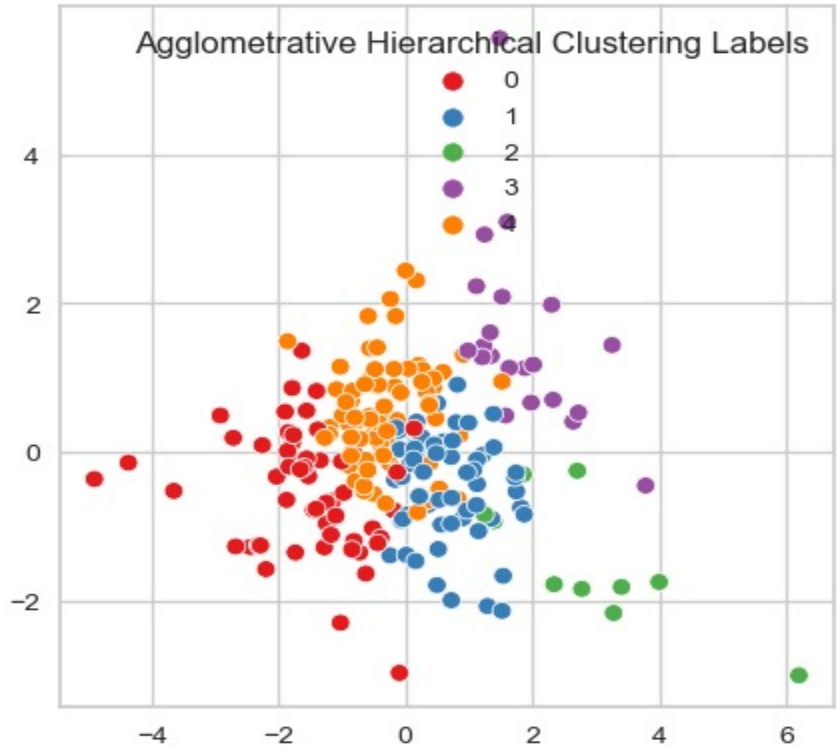
Spectral Clustering

Results – The *Bad* Models

Gaussian Mixture Model

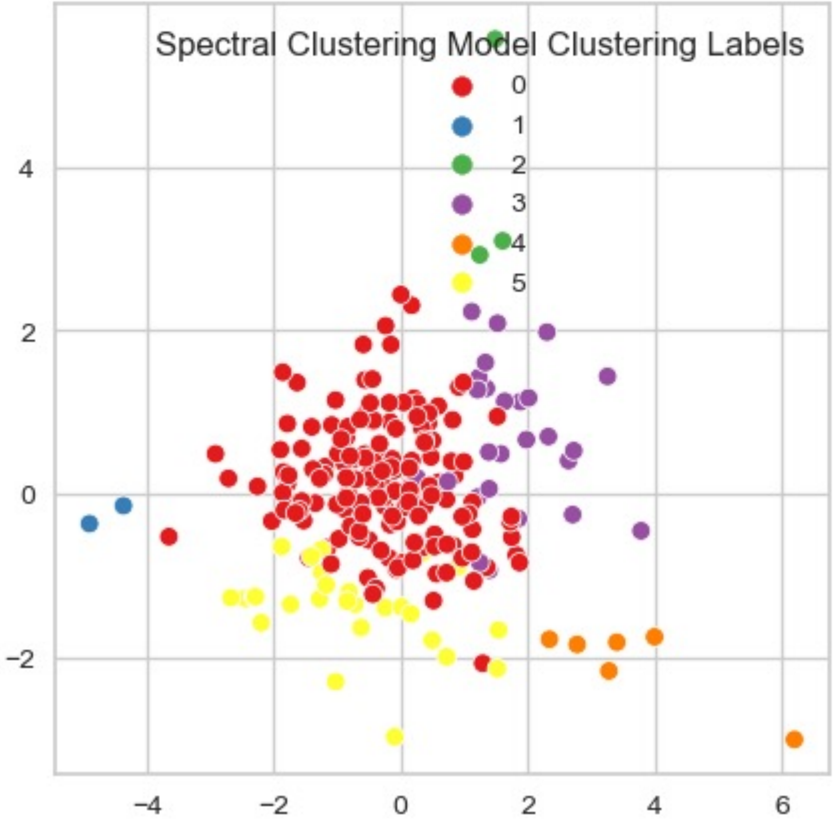


Agglomerative Clustering

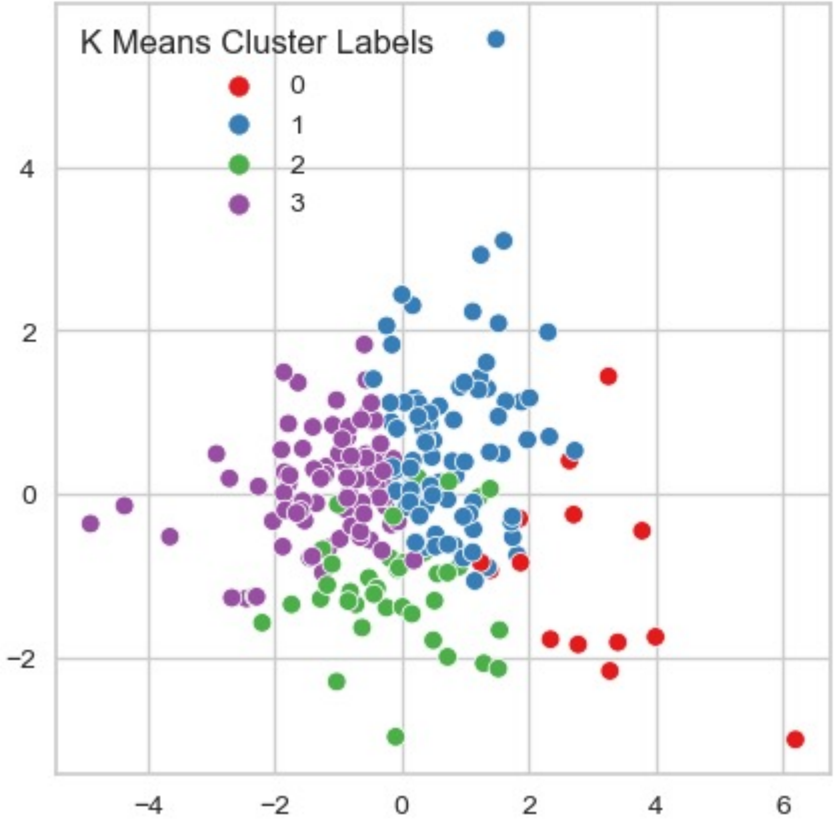


Results – The *Good* Models

Spectral Clustering Model



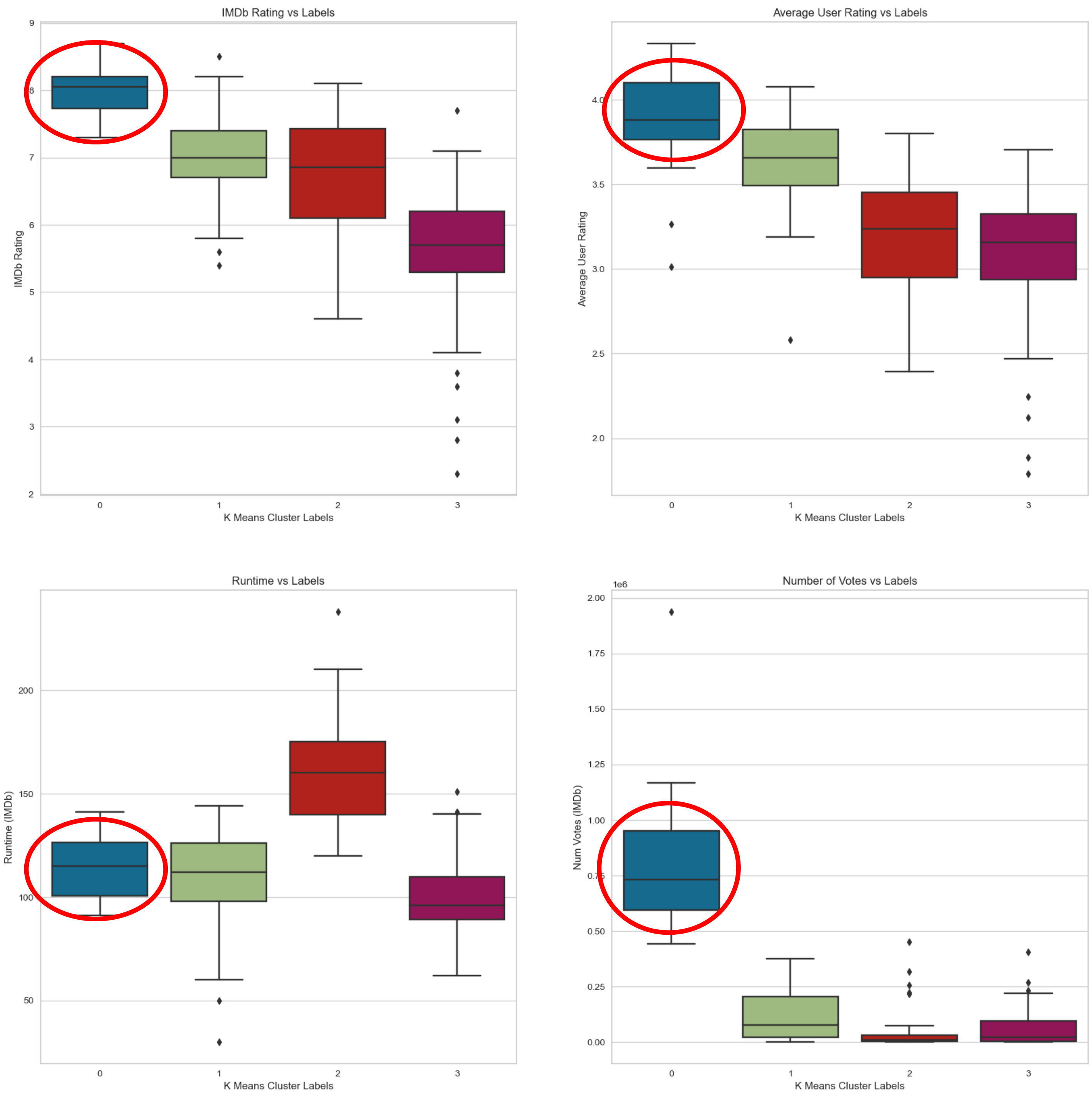
KMeans Clustering Model



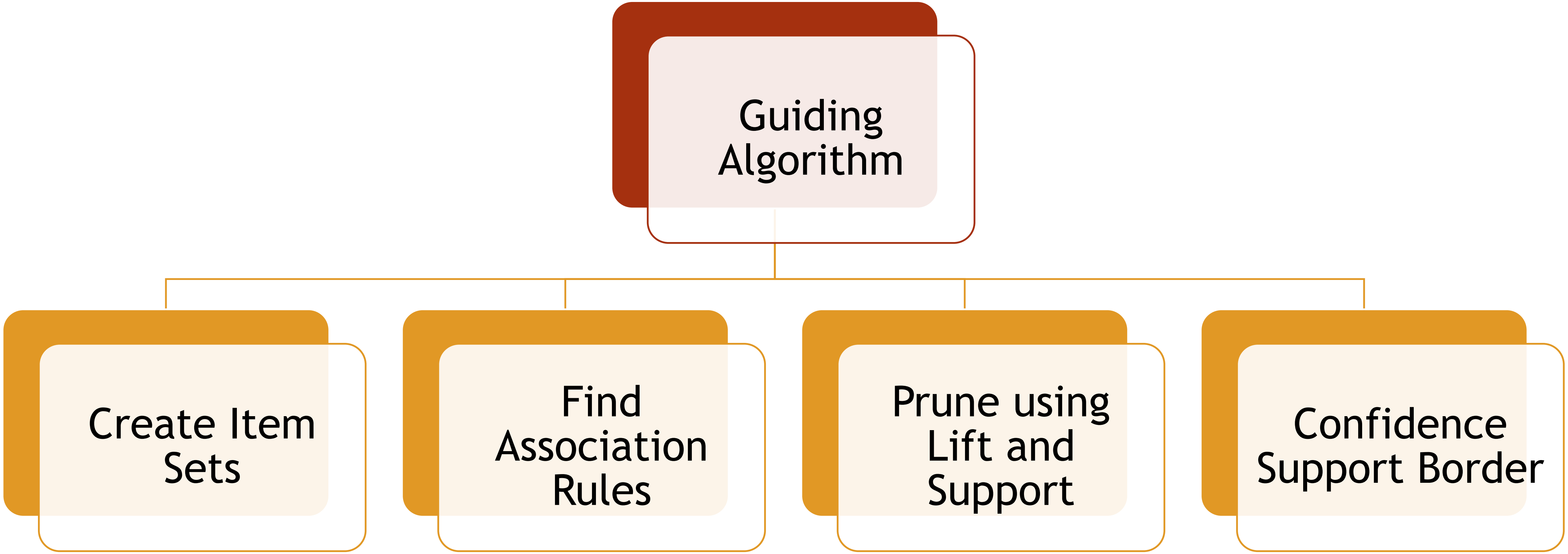
Note: PC1 and PC2 just explains 60%

Best Model:

KMeans Clustering



Q4. What movies do users frequently watch together?



Models Explored

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

Results – Top 10 Rules

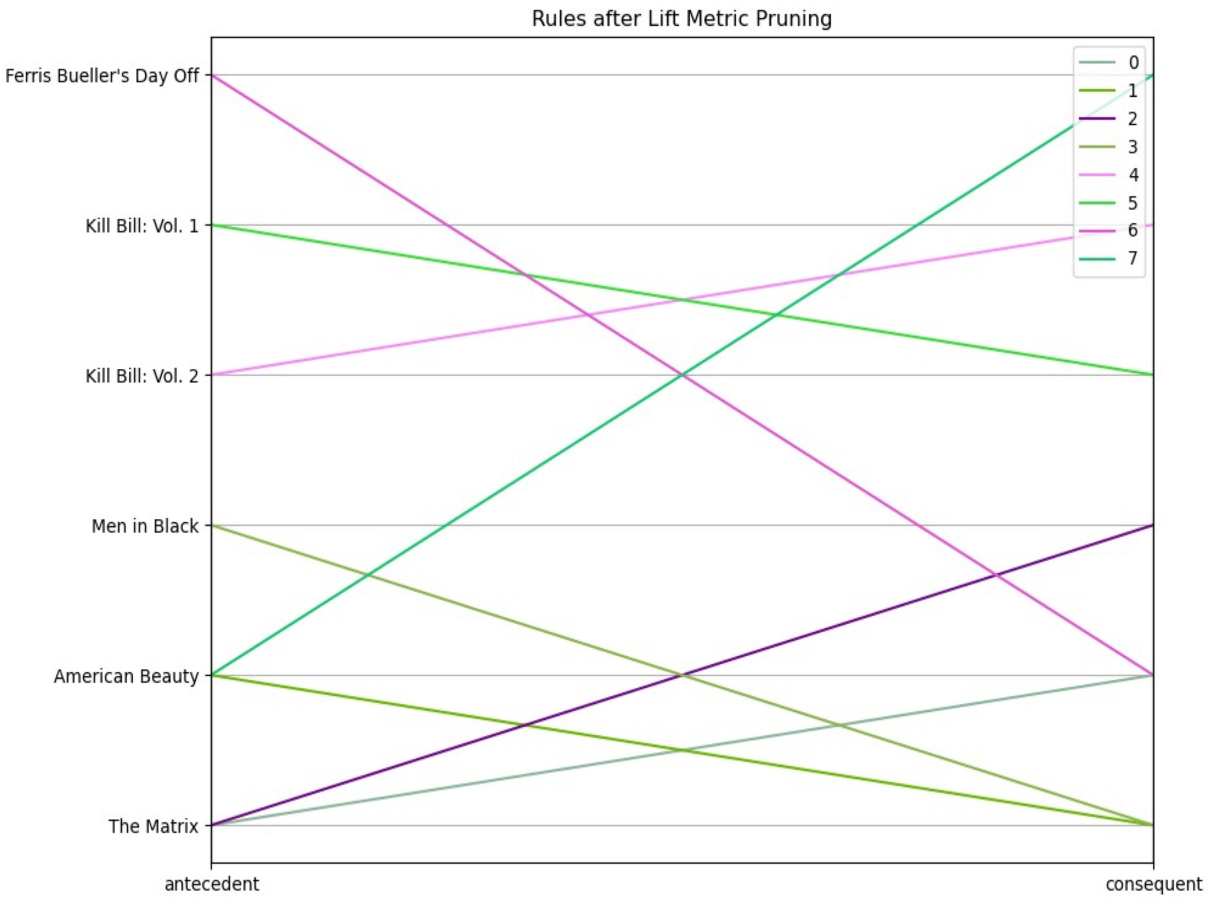
Support Metric Pruning

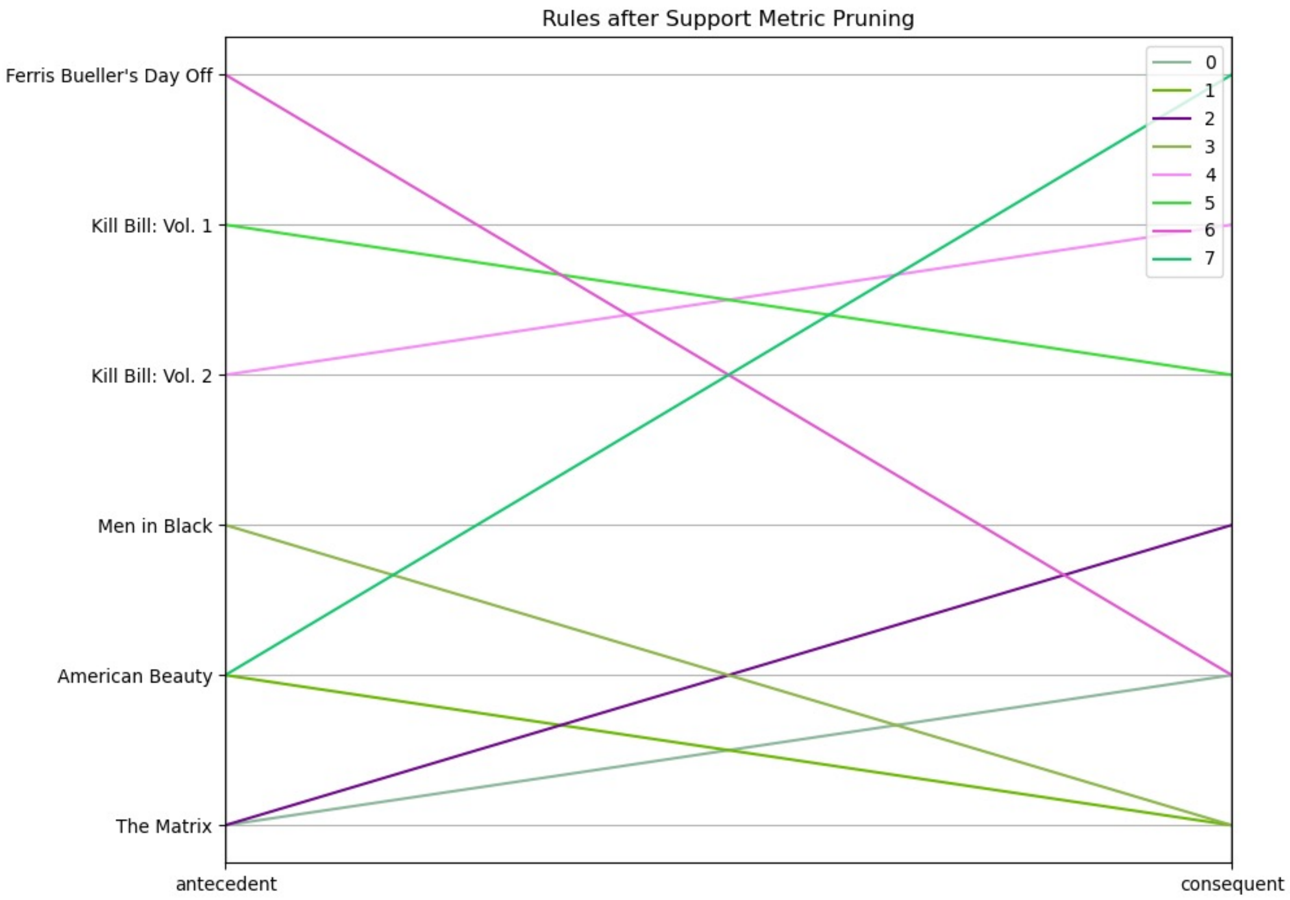
	antecedents	consequents
	(Kill Bill: Vol. 1)	(Kill Bill: Vol. 2)
	(Kill Bill: Vol. 2)	(Kill Bill: Vol. 1)
	(The Matrix)	(American Beauty)
	(American Beauty)	(The Matrix)
	(Men in Black)	(The Matrix)
	(The Matrix)	(Men in Black)
	(Ferris Bueller's Day Off)	(American Beauty)
	(American Beauty)	(Ferris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

Lift Metric Pruning

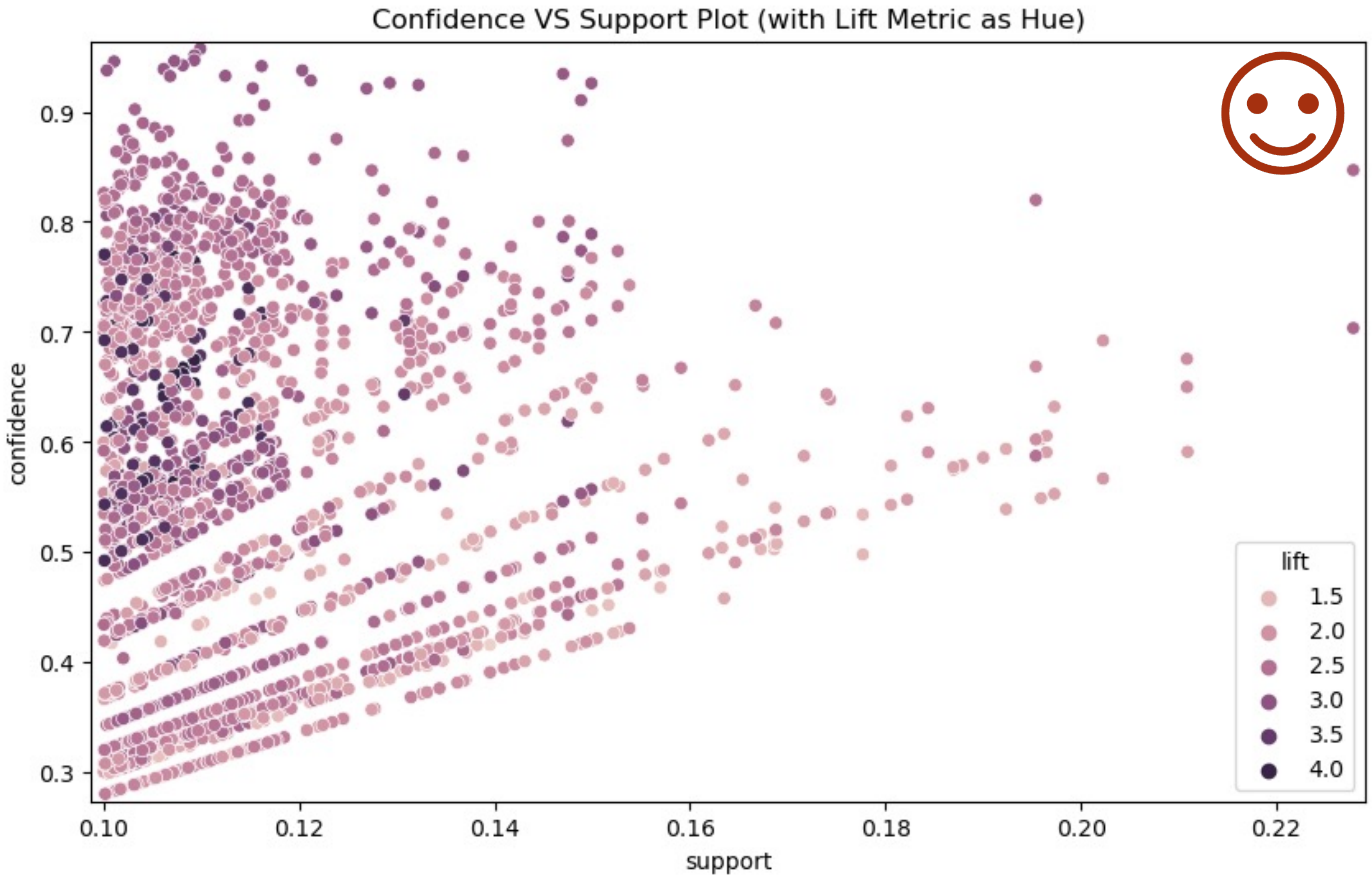
	antecedents	consequents
	(Kill Bill: Vol. 2, Men in Black)	(Kill Bill: Vol. 1, The Matrix)
	(Kill Bill: Vol. 1, The Matrix)	(Kill Bill: Vol. 2, Men in Black)
	(Kill Bill: Vol. 2, The Matrix)	(Kill Bill: Vol. 1, Men in Black)
	(Kill Bill: Vol. 1, Men in Black)	(Kill Bill: Vol. 2, The Matrix)
	(Kill Bill: 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...

Results – Parallel Plots





Confidence vs Support Border of Association Rules



Future Work

Develop a **recommendation system** based on our work on association rule mining and LDA

Incorporate **customer demographic data** to better our predictive model

Any Questions?