

MOVIES & NLP

Prathap Rajaraman



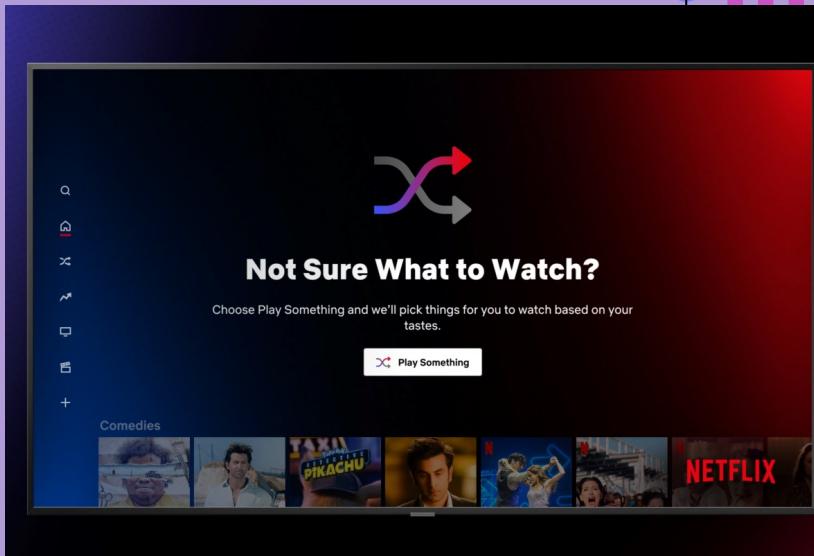
BACKGROUND

- **Motivation**

- Movie streaming is on the rise
- With a robust recommendation system, streaming services keep you engaged with their apps and you are further incentivized to continue your subscription

- **Objective**

- Using NLP, build recommendation system to predict what movies a user likes based on their previously liked films



+++

DATA & METHODOLOGY

DATA

- **Source**
 - TMDB API
 - ~3,000 movies with summaries, cast/crew, budget, release date, rating, number of votes
- **Feature Engineering: Key words get appended to summary as a tag if certain condition is met***
 - Leading Actors/Director as single string
 - Inflation-adjusted budget over \$220M
 - Over 11,000 votes
 - Rating over 8.0
 - Released prior to 1950



*values based off 95th percentile of data point, with exception of release date, which was 1st percentile

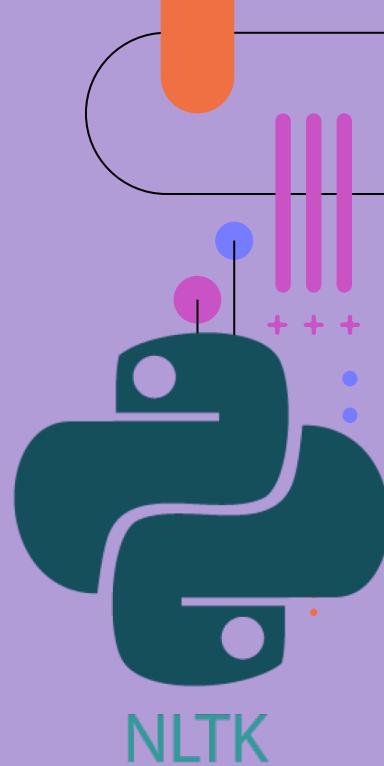
METHODOLOGY

- **Preprocessing**

- Stop words removed, text was cleaned via NLTK
- TFIDF vectorizer for Document-Term Matrix

- **Dimensionality Reduction**

- NMF, PCA, and CorEx tested
- NMF won out due to intuitiveness and generalized topics
- 18 topics chosen based on coherence score optimization



METHODOLOGY

- **Recommendation System**

- Content based recommendation system
- Calculate **Cosine Similarity** between input and all other movies, return N smallest movies
- Single and multiple input recommendation system
 - Single: Locate smallest cosine distance
 - Multiple: Locate smallest average ranked cosine distance. Example Below:

Movie	Ranked Distance to Input 1	Ranked Distance to Input 2	Total Score
A	1	2	3
B	2	3	5
C	3	1	4

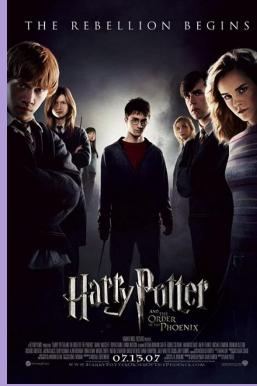
RESULTS

TOPICS

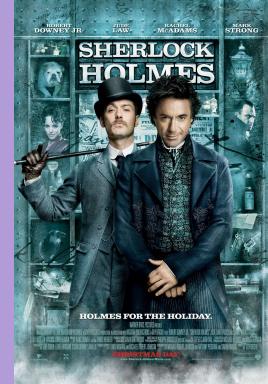
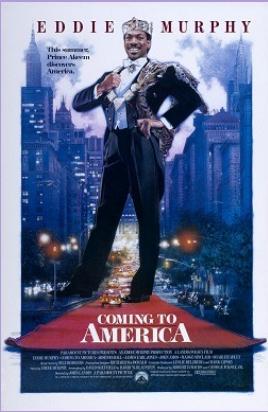
- Corpus broken down by 18 topics via NMF
- Examples:

Topic	Top Terms	High Scoring Movies
Award Winning War Drama	STORY, TRUE, BASED, HIGHLYRATED, WAR	Schindler's List, Defiance
Big Budget Franchises	POPULARMOVIE, BIGBUDGET, EVIL, BATTLE, HARRY	Star Wars: The Last Jedi, Avengers: Endgame
Comedy	FRIEND, BEST, MAKE, NIGHT, ADVENTURE	The Hangover, No Strings Attached

SINGLE RECOMMENDATION SYSTEMS



MULTIPLE RECOMMENDATION SYSTEMS

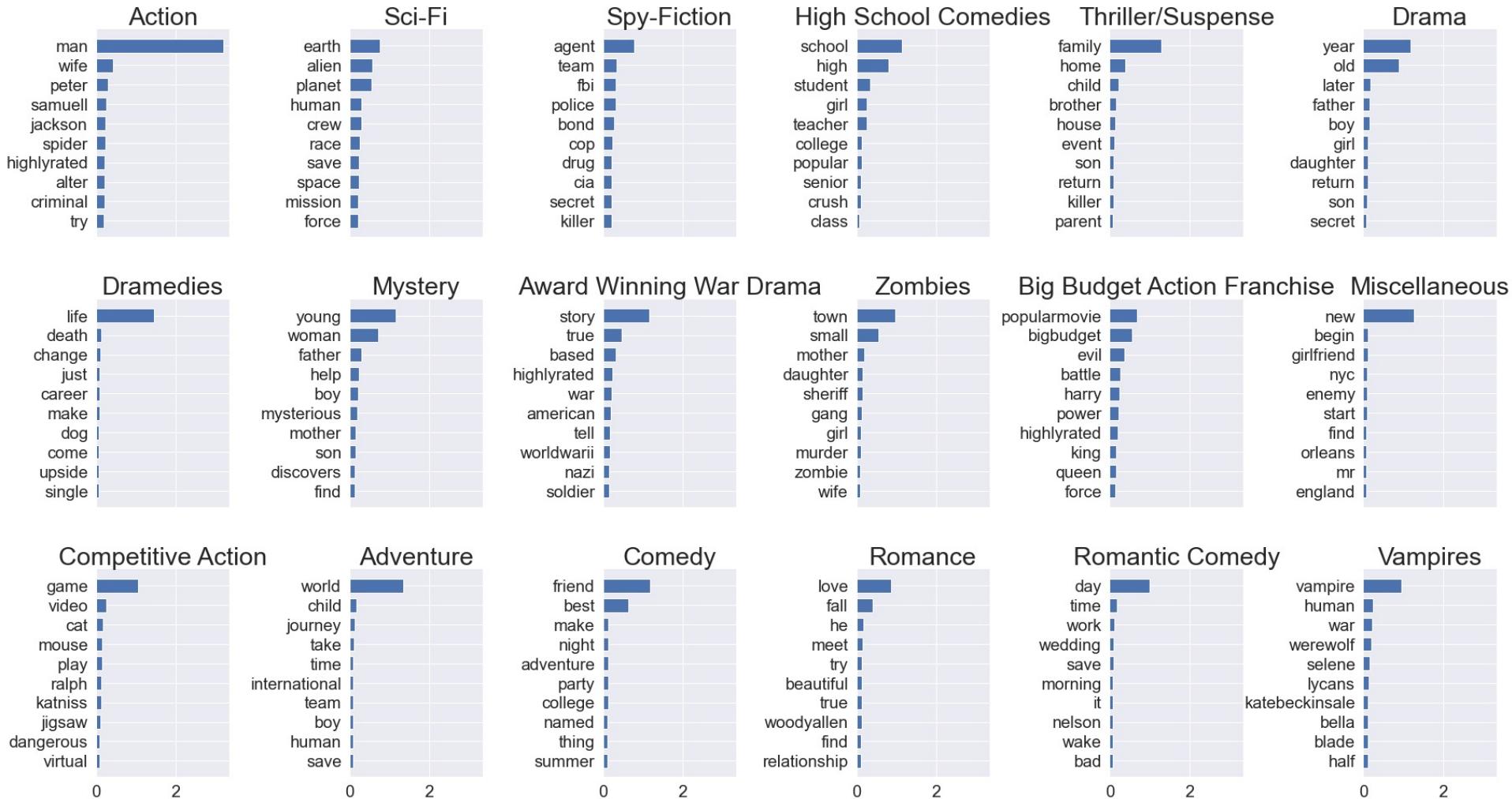


CONCLUSION/FUTURE WORK

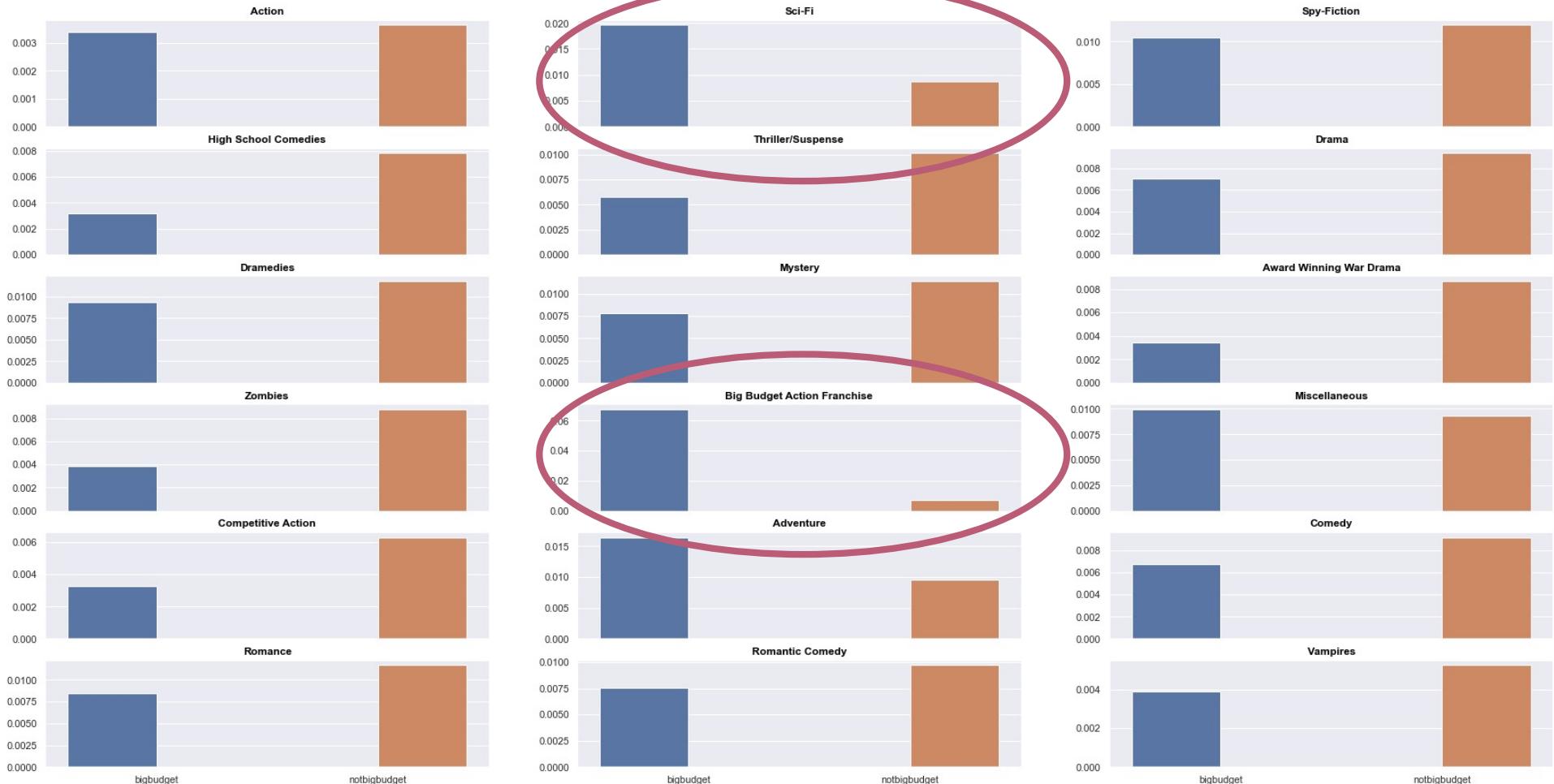
- **Insights**
 - NMF breaks movies down into intuitive, easy to understand topics
 - Recommender finds similar movies based on users given tastes
 - Topics are better defined for action/drama than comedy
- **Future Work**
 - Layer in more data, such as franchise tags or sequel tags
 - Add collaborative filtering measures to boost model for comedies
 - Streamlit App

APPENDIX

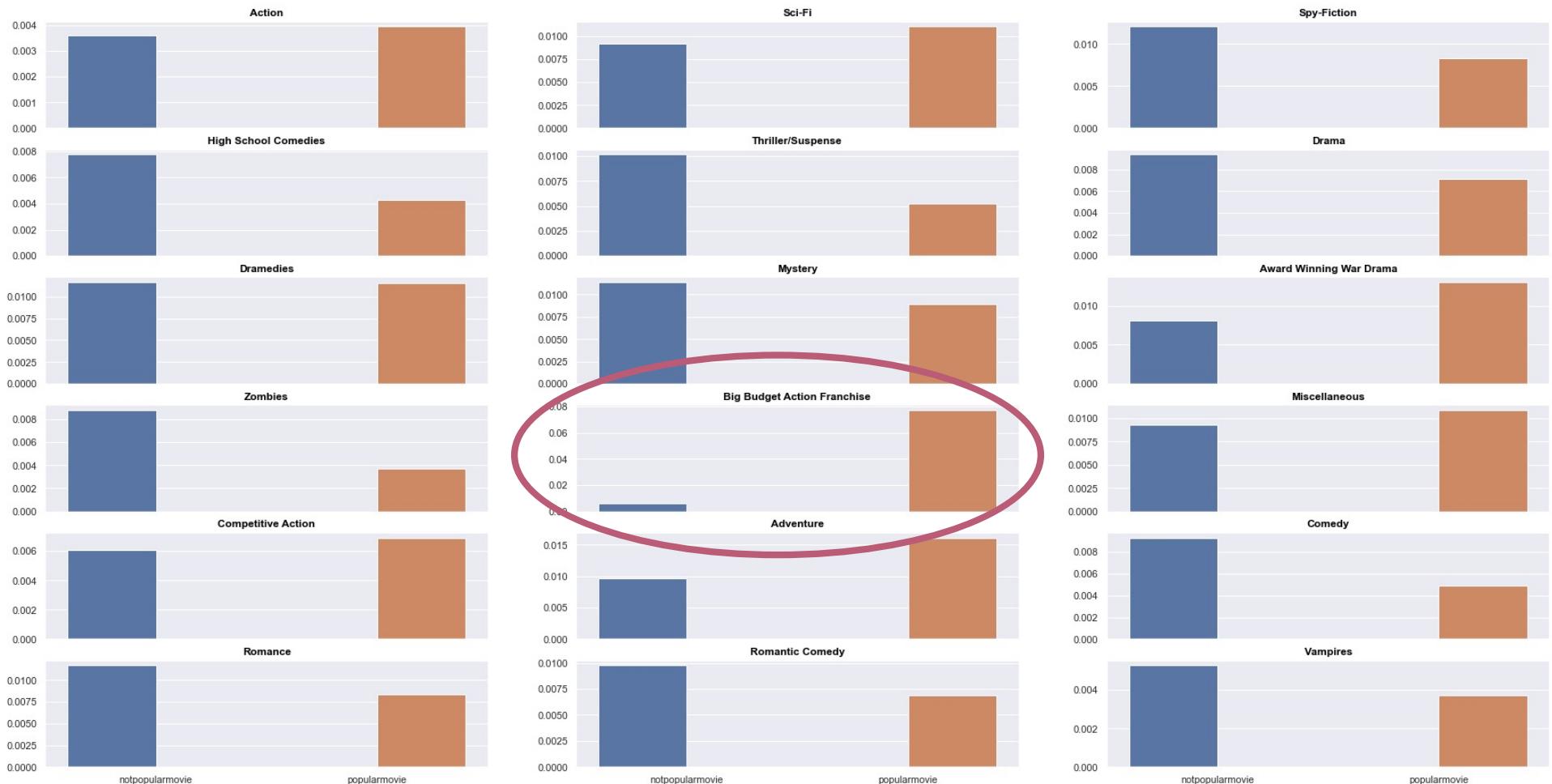
Topics in NMF Model



Topic Strength by Bigbudget Indicator



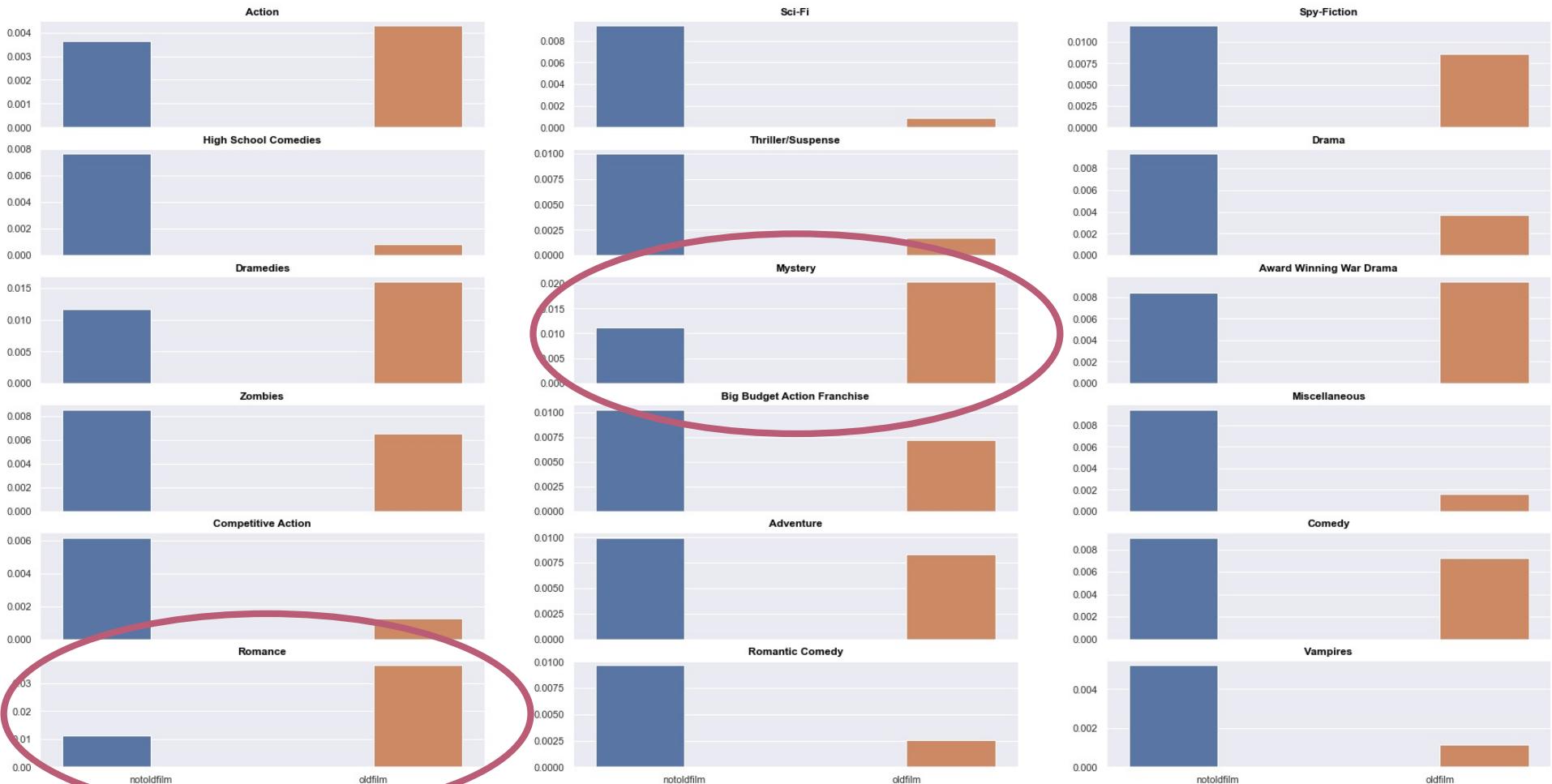
Topic Strength by Popularmovie Indicator



Topic Strength by HighlyRated Indicator



Topic Strength by Oldfilm Indicator



TSNE Topic Clustering Map

