IMDb.com:
BI in Movie Review Platforms

Text Mining and Sentiment Analysis on Movie Reviews

CONTENTS

"Cinema is the ultimate pervert art. It doesn't give you what you desire - it tells you how to desire."

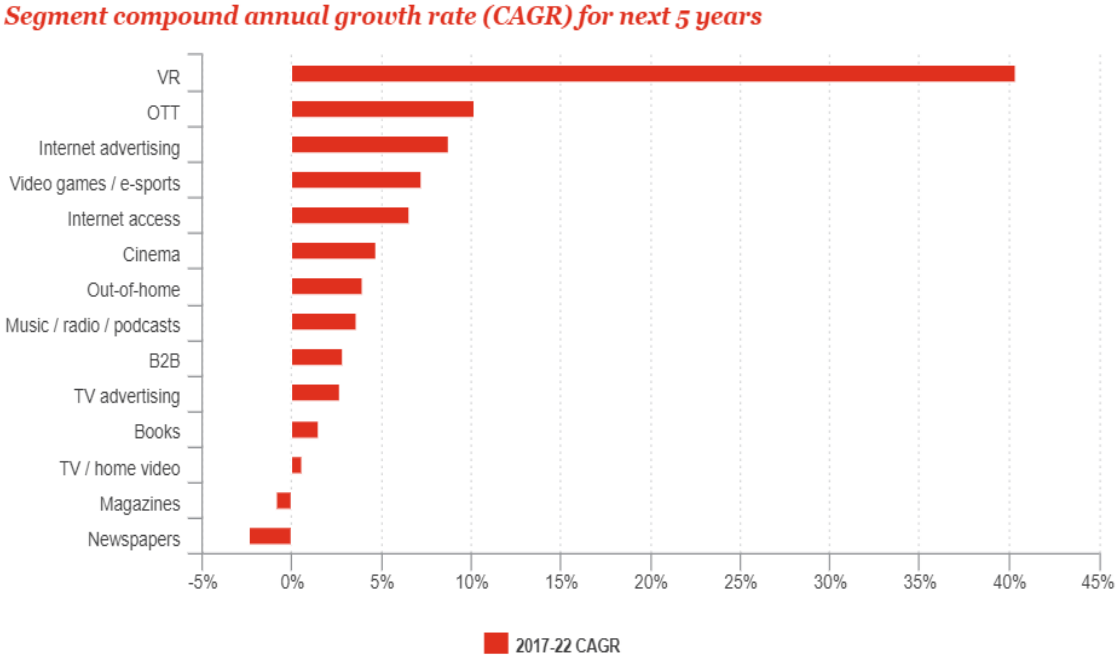
— Slavoj Žižek

INTRODUCTION

Project Background

Business Intelligence Questions

Project Background



Source: PwC Global Entertainment & Media Outlook: 2018–2022. PwC. Ovum

- ★ Movie industry is profitable, especially in US movie market
- ★ Imponderable future of Movie Industry
- ★ Visual Reality technique is about to offer an innovation
- ★ Digital/online movie consumption is taking over traditional consumption mode in Movie Industry

Internet Movie Database:

- ★ #1 movie and visual media database in the world
- ★ Fan-operated website from 1990, and become a subsidiary of Amazon.com since 1998
- ★ 5.3 million movie and episode titles, 9.3 million of personalities, and about 83 million registered users (as of October 2018)
- ★ Users' contribution on final score of a movie
  - Each of the 'Regular Voter' is given a weight (not disclosed)
  - Use Bayesian posterior mean to calculate weighted rating

$$W = \frac{R \cdot v + C \cdot m}{v + m}$$

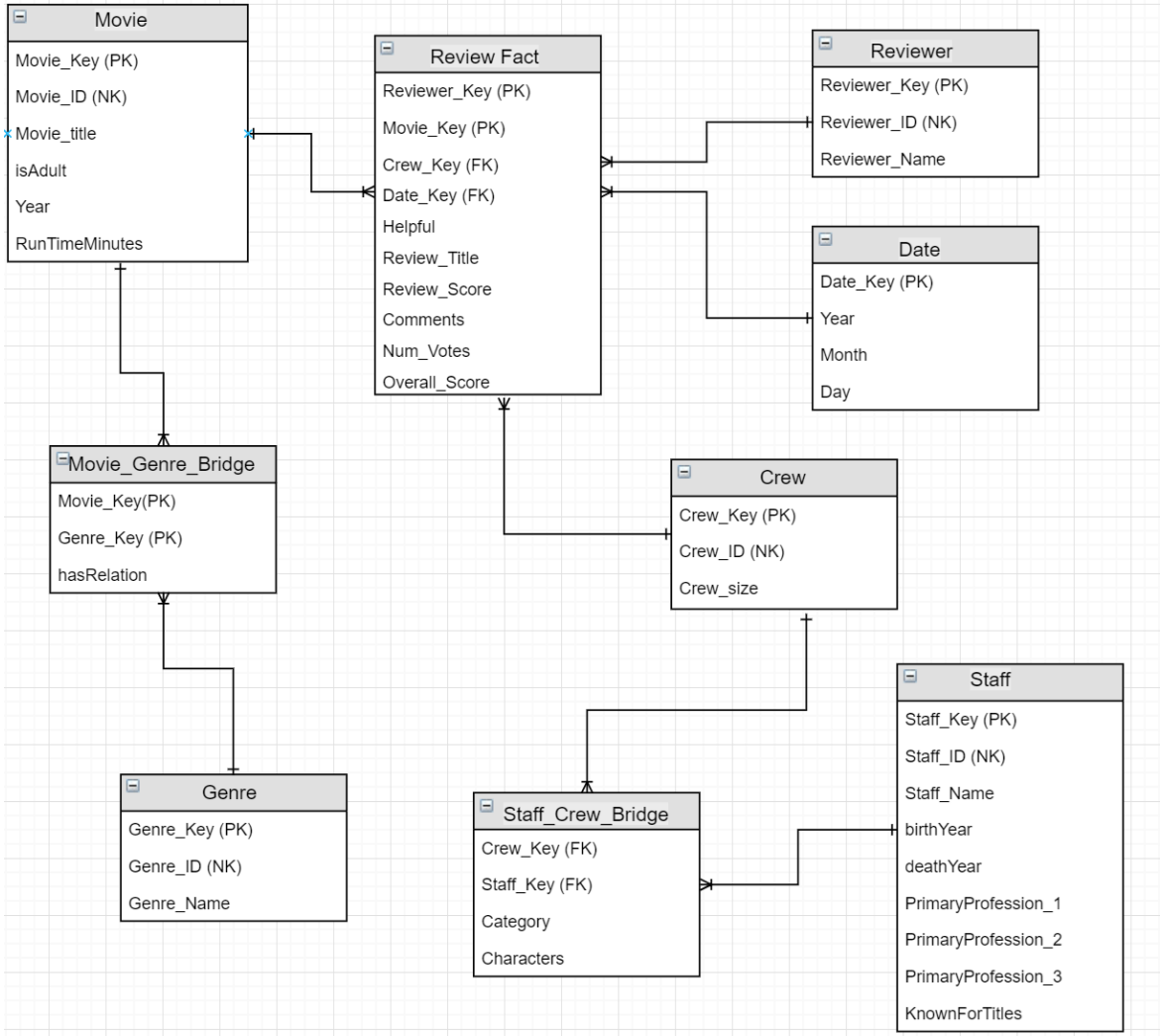
Business Intelligence Questions

- What are the most popular movies in recent three years?
- For each type of genre, what are the top ranking movies?
- How many movies were produced each year in different genres?
- What is the average weighted rating for each genre in the recent three years?
- Which movie genre is the most popular recently?
- What aspect of a movie do users talk about most and pay most attention to?
- What is the average score a user will give based on a positive/neutral/negative review?
- Is there any associations among different genres?

DATA WAREHOUSE DESIGN

ETL and data quality control

Star Schema

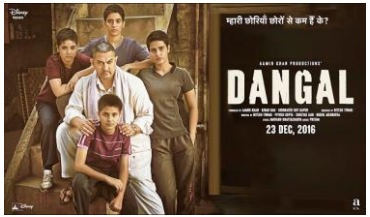

Movie
Movie_Key (PK)
Movie_ID (NK)
Movie_title
isAdult
Year
RunTimeMinutes
Review Fact
Reviewer_Key (PK)
Movie_Key (PK)
Crew_Key (FK)
Date_Key (FK)
Helpful
Review_Title
Review_Score
Comments
Num_Votes
Overall_Score
Reviewer
Reviewer_Key (PK)
Reviewer_ID (NK)
Reviewer_Name
Date
Date_Key (PK)
Year
Month
Day
Movie_Genre_Bridge
Movie_Key(PK)
Genre_Key (PK)
hasRelation
Crew
Crew_Key (PK)
Crew_ID (NK)
Crew_size
Staff
Staff_Key (PK)
Staff_ID (NK)
Staff_Name
birthYear
deathYear
PrimaryProfession_1
PrimaryProfession_2
PrimaryProfession_3
KnownForTitles
Genre
Genre_Key (PK)
Genre_ID (NK)
Genre_Name
Staff_Crew_Bridge
Crew_Key (FK)
Staff_Key (FK)
Category
Characters

ETL Process

- **Extraction**
  - IMDb Public Dataset
  - Web Scraping - Movie reviews
- **Transformation**
  - Splitting up fields (genres, professions)
  - Format changes (Date)
  - Deduplication
- **Loading**
  - Dimensions - Create surrogate keys
  - Fact table - No null for foreign keys
- **Data Quality Control**
  - Referential Integrity Constraints
  - Missing values
  - Anomaly Detection

ONLINE ANALYTICAL PROCESSING

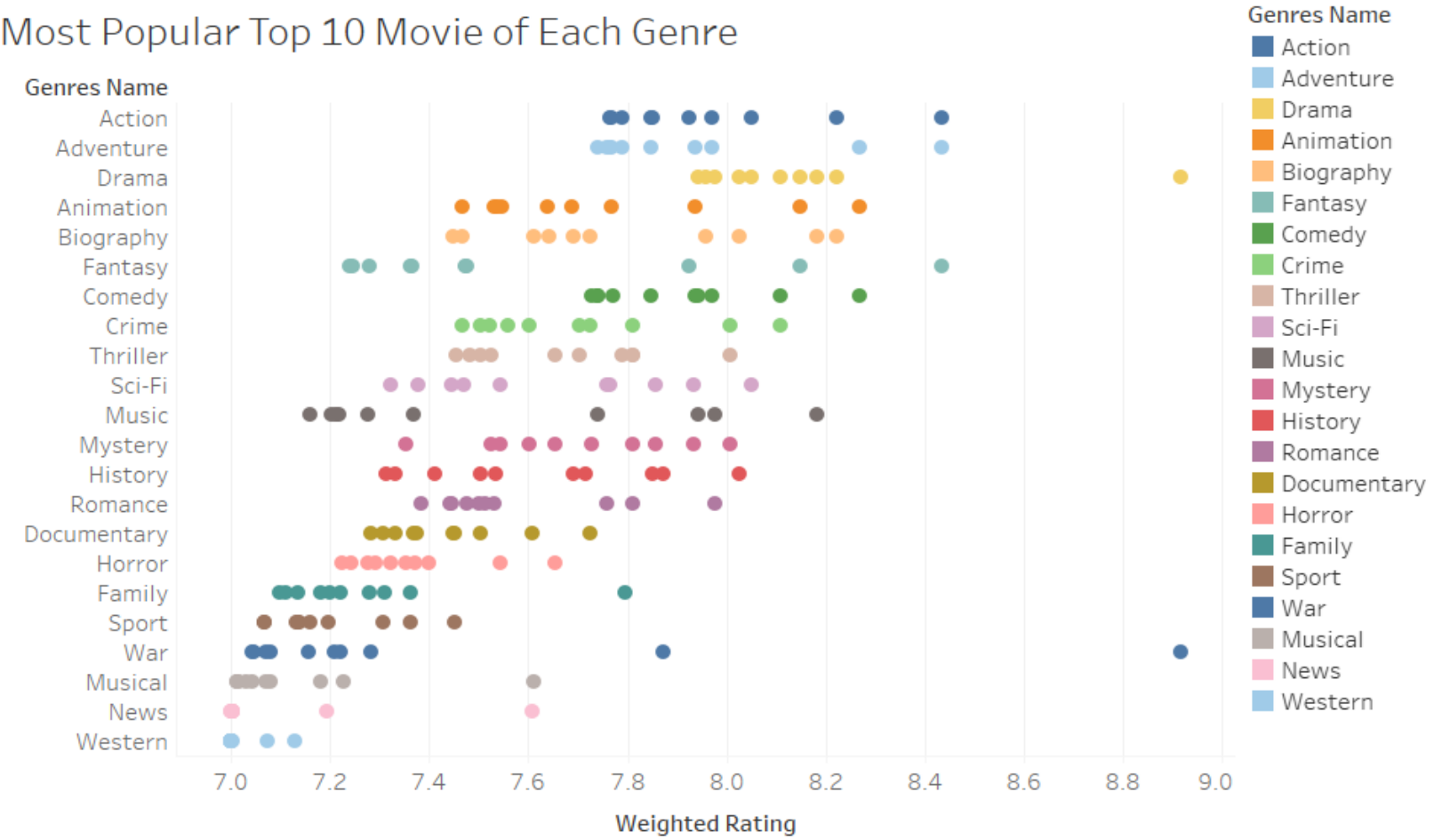
Data visualizations on OLAP
Exploratory Data Analysis



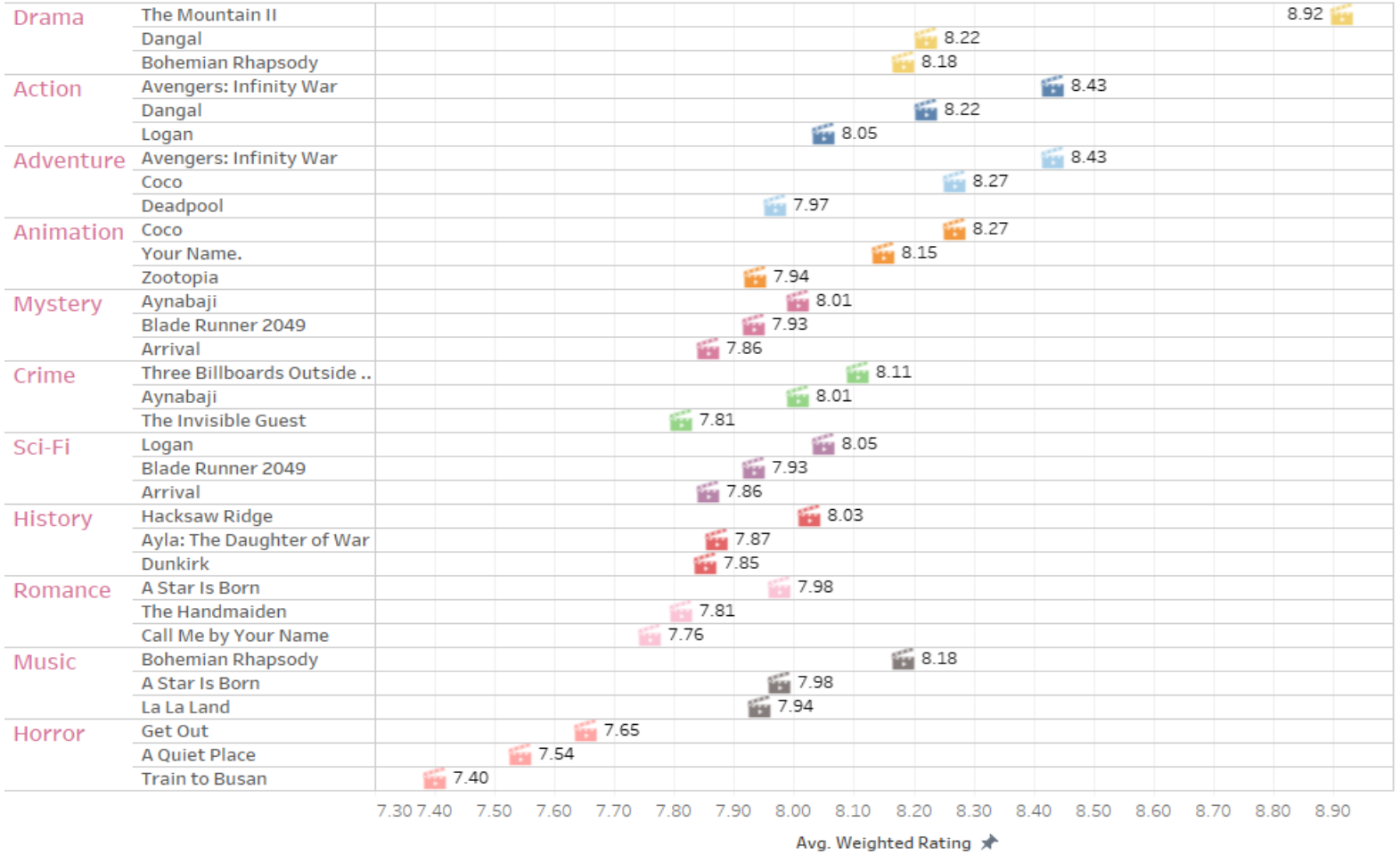
Top 10 Movies
Ranked by Weighted Rate

1	The Mountain II	8.92
2	Avengers: Infinity War	8.43
3	Coco	8.27
4	Dangal	8.22
5	Bohemian Rhapsody	8.18
6	Your Name.	8.14
7	Three Billboards Outside Ebbing, Missouri	8.11
8	Logan	8.05
9	Hacksaw Ridge	8.03
10	Aynabaji	8.01

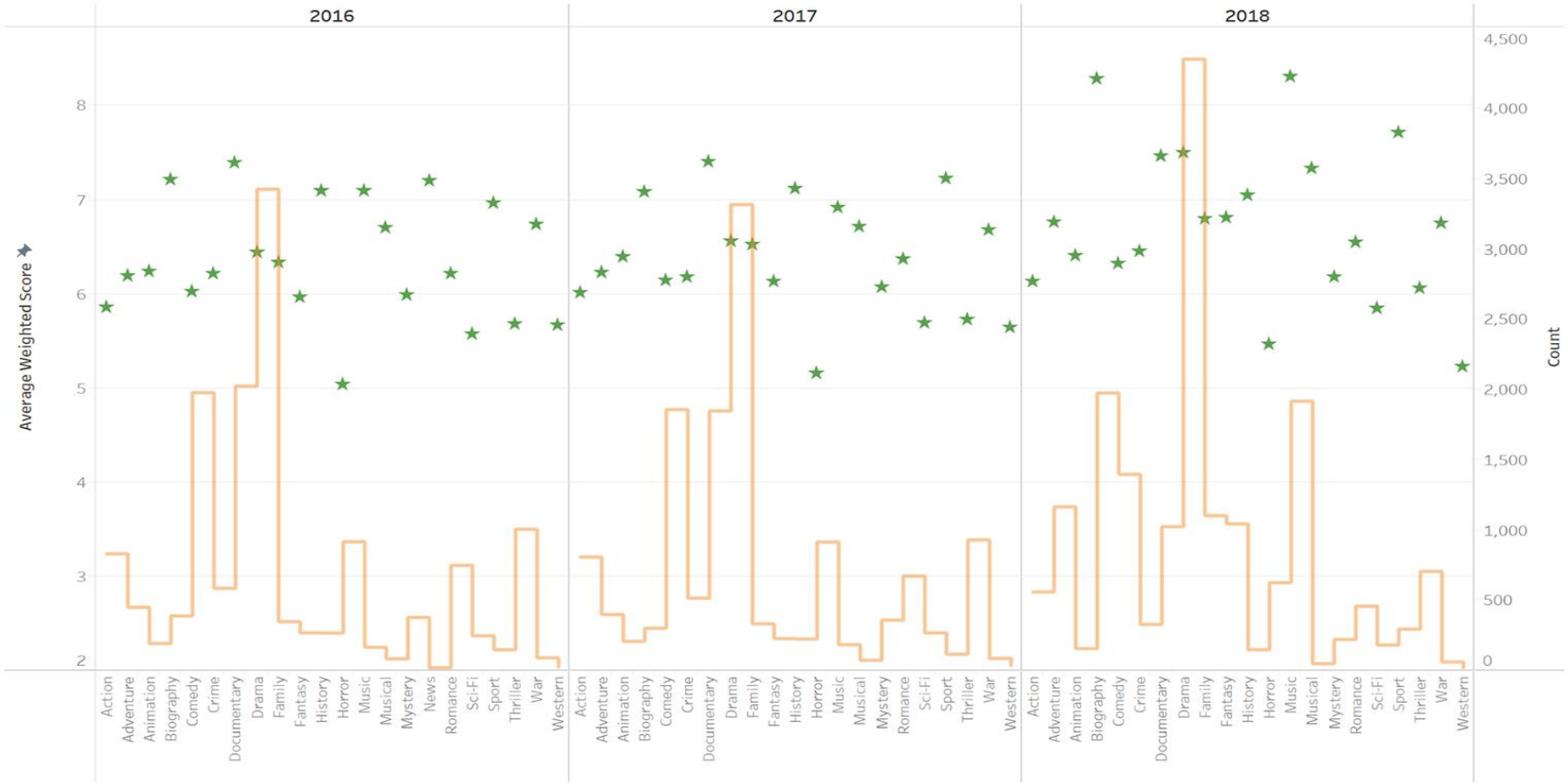
Most Popular Top 10 Movie of Each Genre



Top 3 Movies By Genres



Movie Performance In Recent Three Years



DATA/TEXT MINING

Data mining

Text mining and Sentiment Analysis

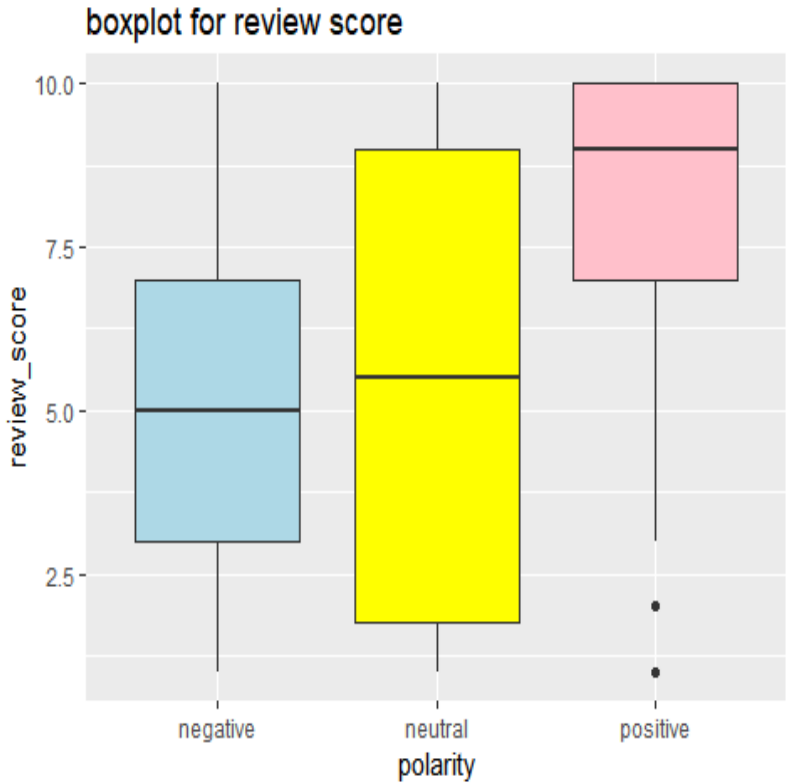
Association Rule Mining

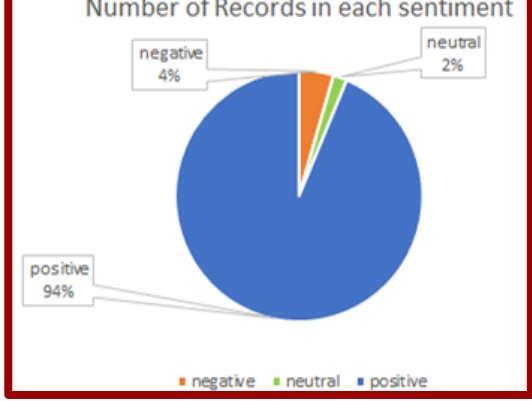
min support = 0.05
min confidence = 0.2

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift ↓	Convicti...
7	Adventure	Action	0.011	0.289	0.973	-0.067	0.009	4.100	1.307
10	Romance	Comedy	0.021	0.349	0.964	-0.098	0.011	2.045	1.274
17	Biography	Documentary	0.027	0.648	0.986	-0.057	0.013	1.892	1.868
16	History	Documentary	0.018	0.605	0.989	-0.041	0.008	1.765	1.663
15	Music	Documentary	0.012	0.538	0.990	-0.033	0.004	1.571	1.423
14	Romance	Drama	0.031	0.518	0.973	-0.088	0.011	1.525	1.370
13	Crime	Drama	0.023	0.491	0.977	-0.071	0.007	1.447	1.298
12	Mystery	Drama	0.014	0.427	0.982	-0.052	0.003	1.258	1.153
11	Family	Drama	0.014	0.390	0.979	-0.057	0.002	1.147	1.082
9	Thriller	Drama	0.027	0.298	0.942	-0.152	-0.004	0.877	0.941
8	Comedy	Drama	0.049	0.289	0.896	-0.292	-0.009	0.850	0.928
6	Action	Drama	0.019	0.263	0.952	-0.122	-0.005	0.776	0.897
5	Biography	Drama	0.011	0.260	0.970	-0.074	-0.003	0.765	0.892

Text Mining

Sentiment Analysis for all reviews



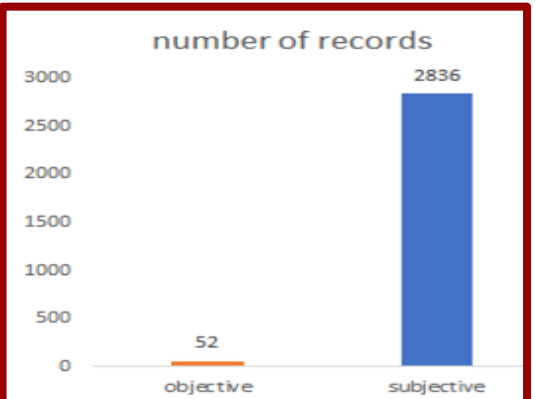


Most positive

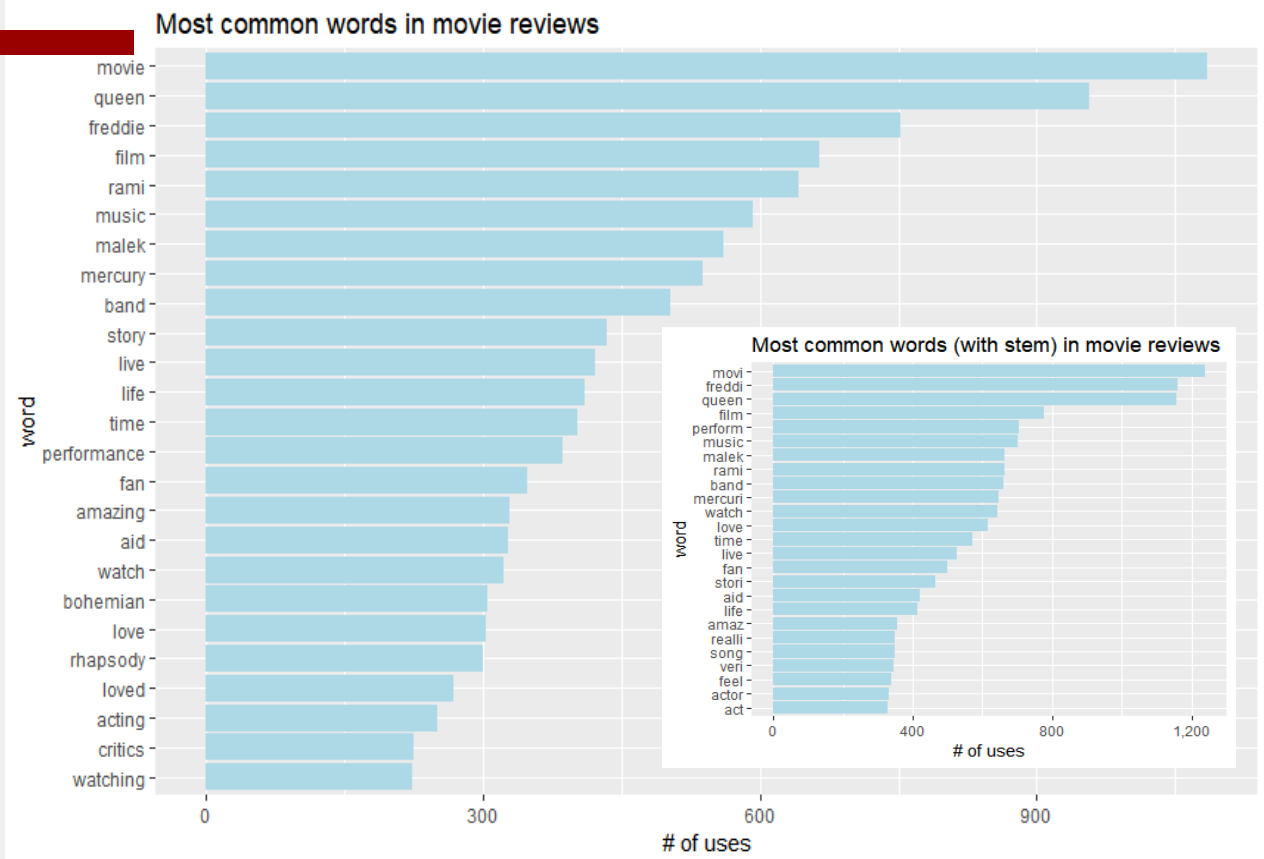
Most subjective

Positive: >= 9

Negative: <= 5



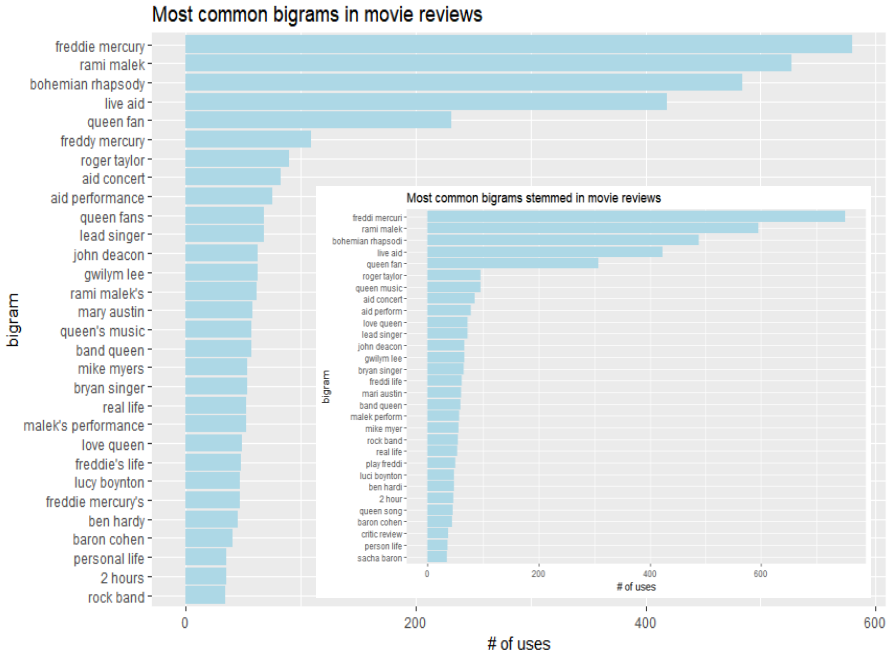
Keywords and Topic Extraction: movie 'Bohemian Rhapsody'

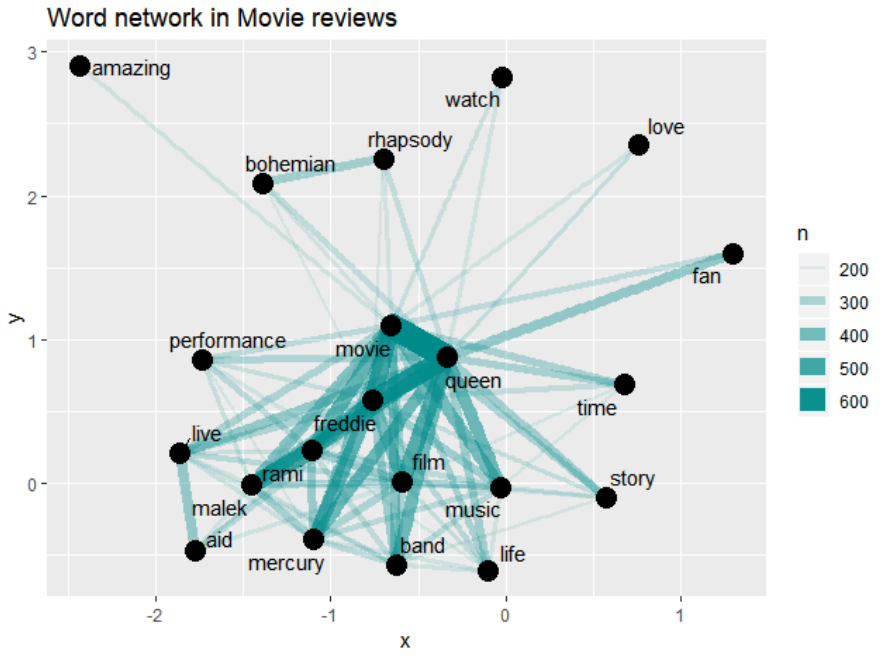


Common words:

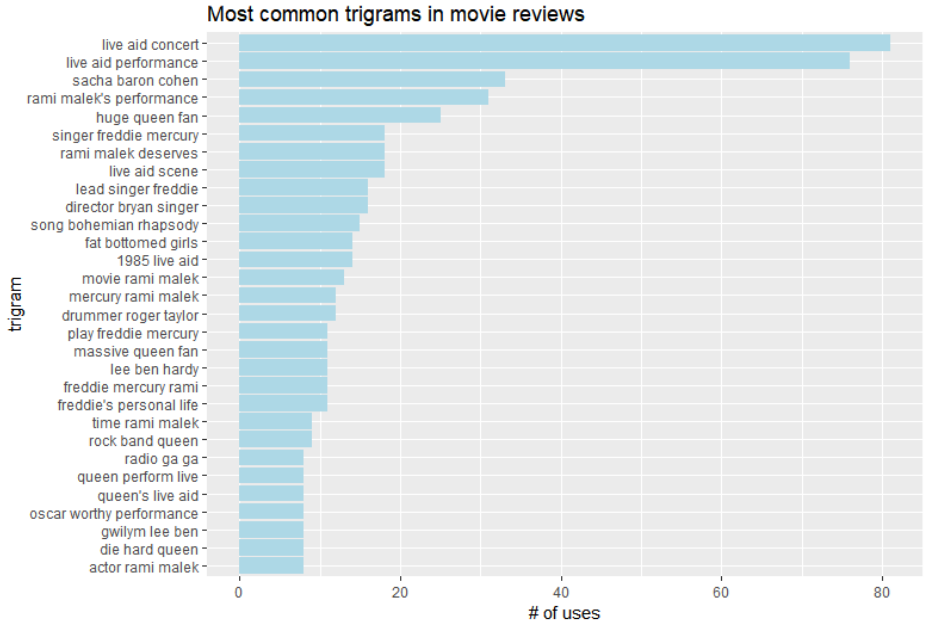
(1) movie, film - background
(2) music, band - topic
(3) queen, freddie mercury, malek, life, story - content

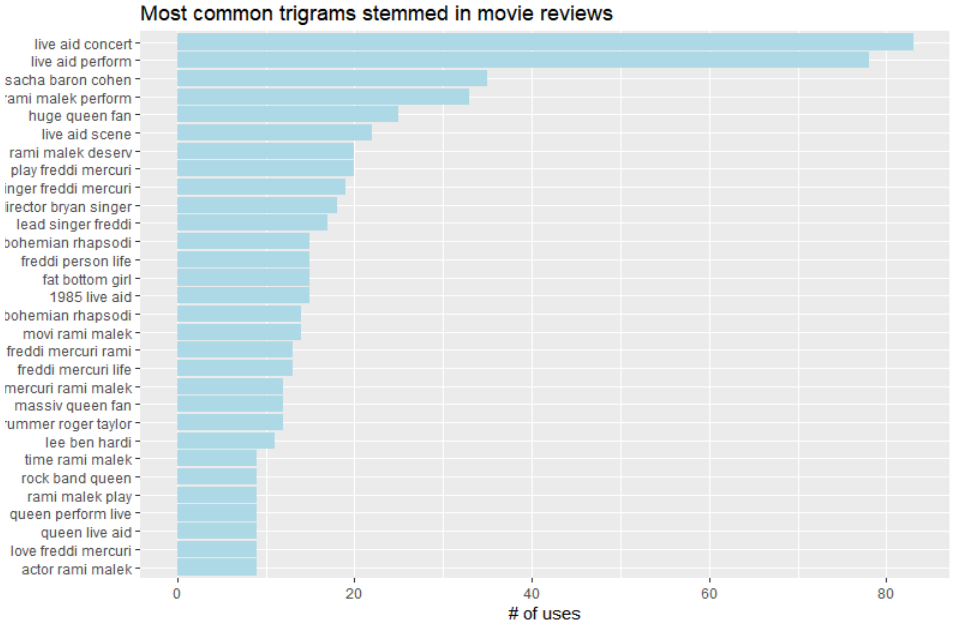
N-grams Analysis





Character name, Actor name, Movie name/song name, Concert name, Target: fan of the band, Music, biography, Band type: rock





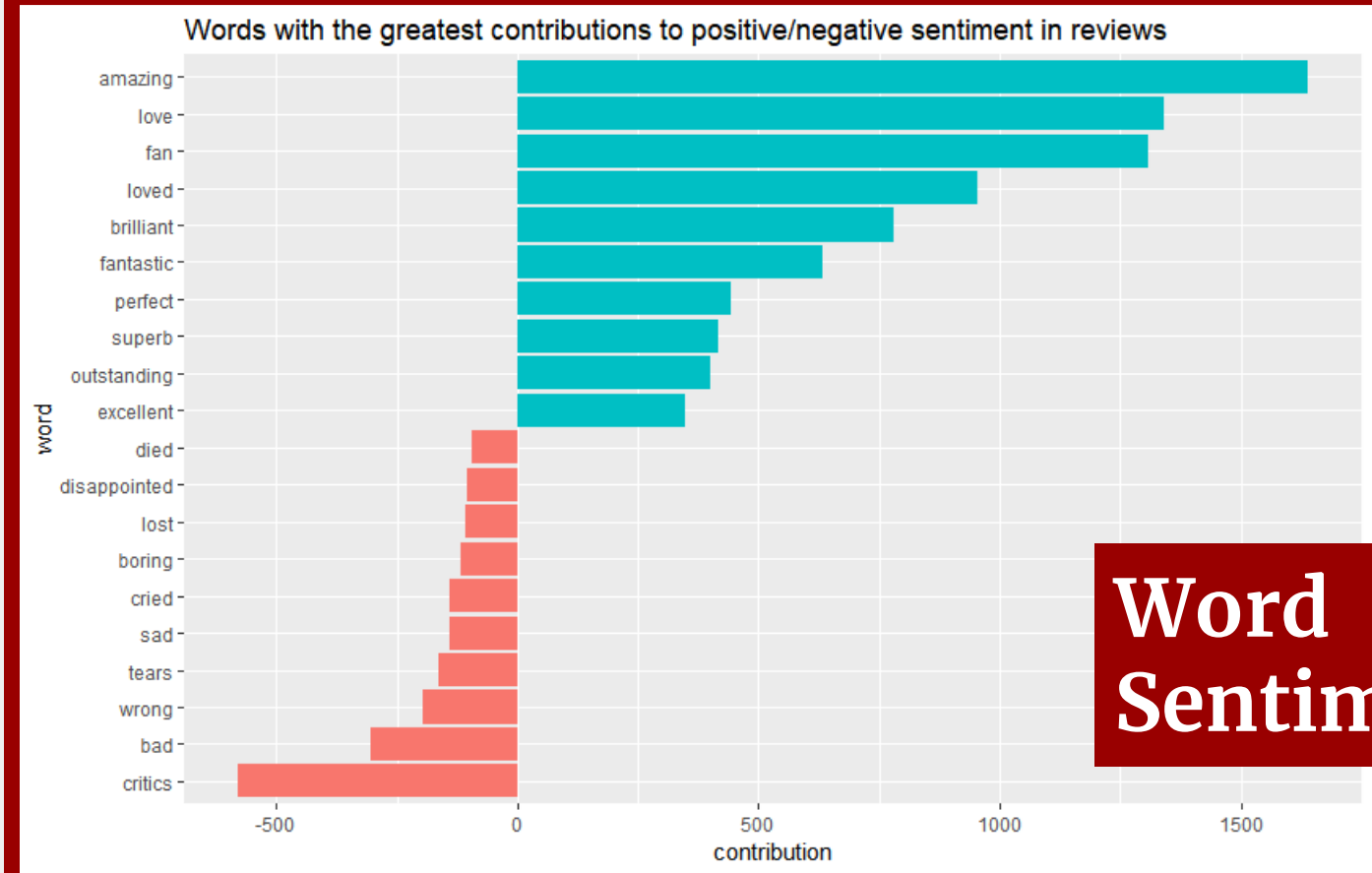
Role of each players in the band: eg. drummer

Concert: 1985 live aid

Movie: Oscar

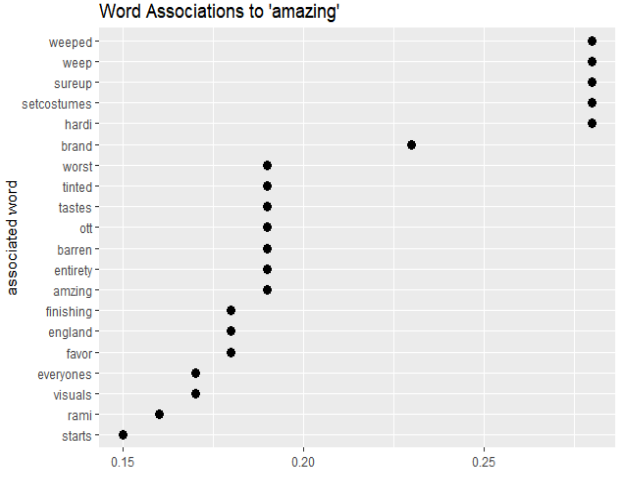
This movie reviews are mainly focused on the content. This is in general a good movie, no apparent negative words.

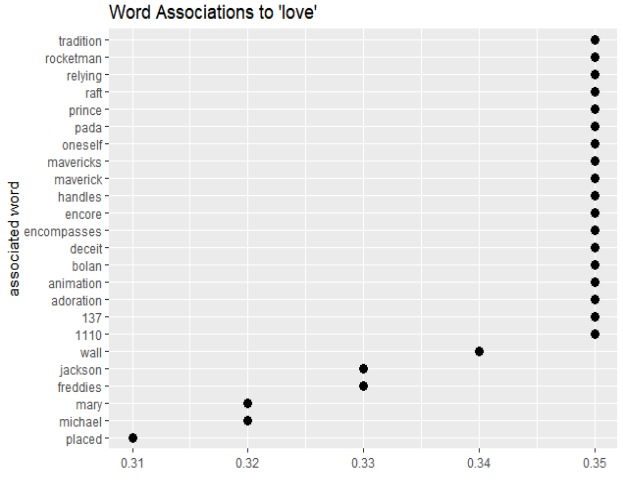
N-grams Analysis

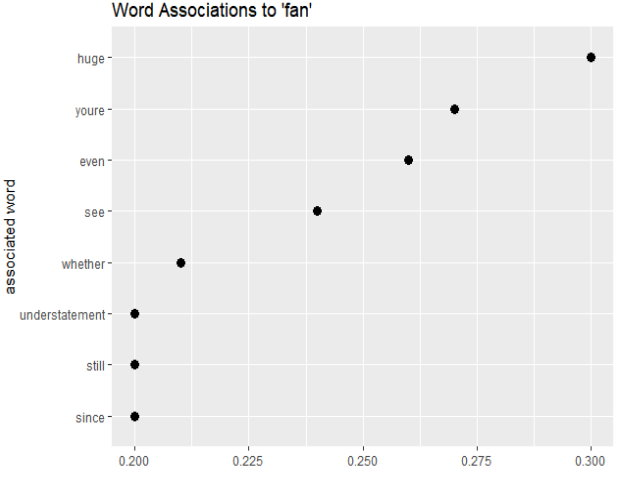


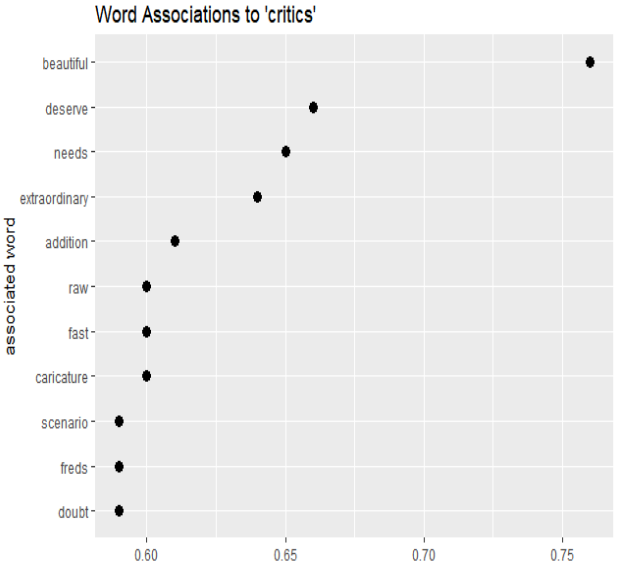
Word Sentiment

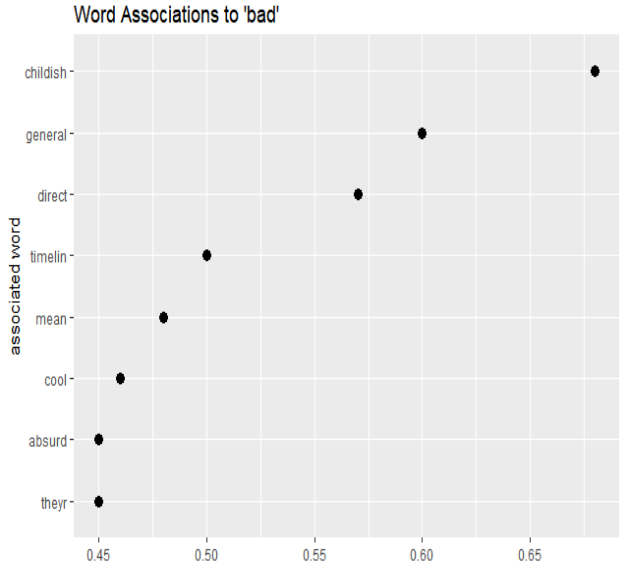
Word associations: for good movie (>=9)

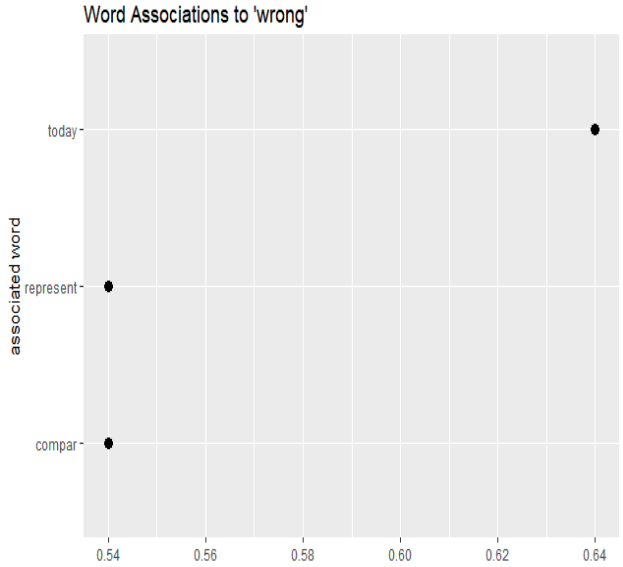








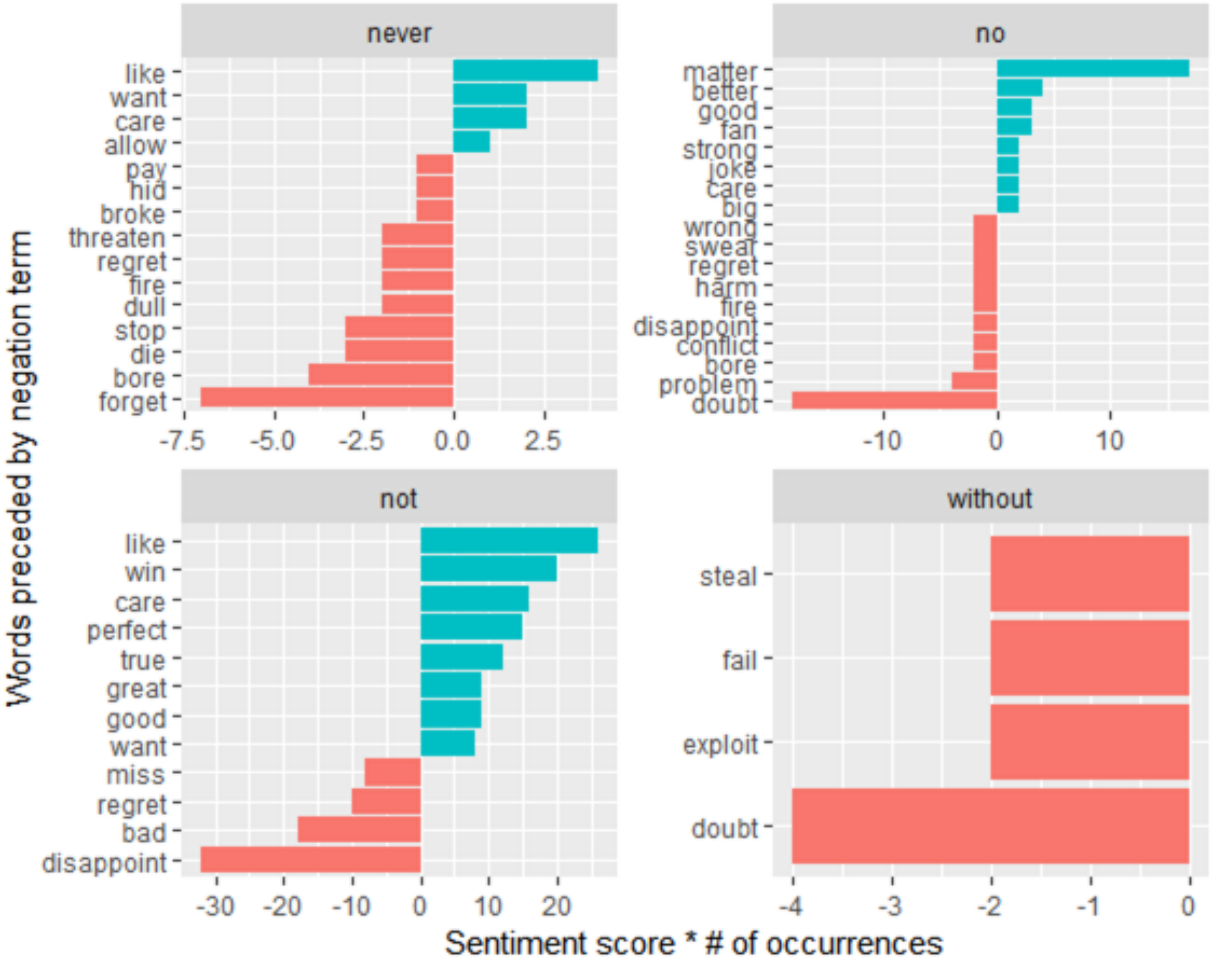




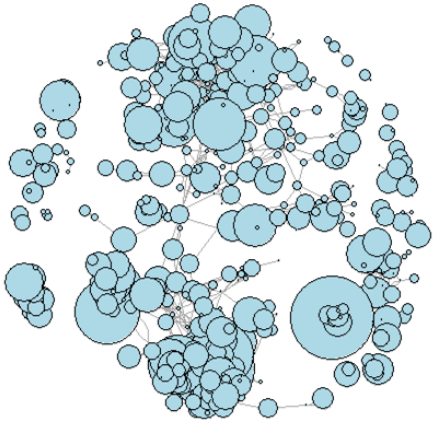
Word associations: for bad movie (<=5)

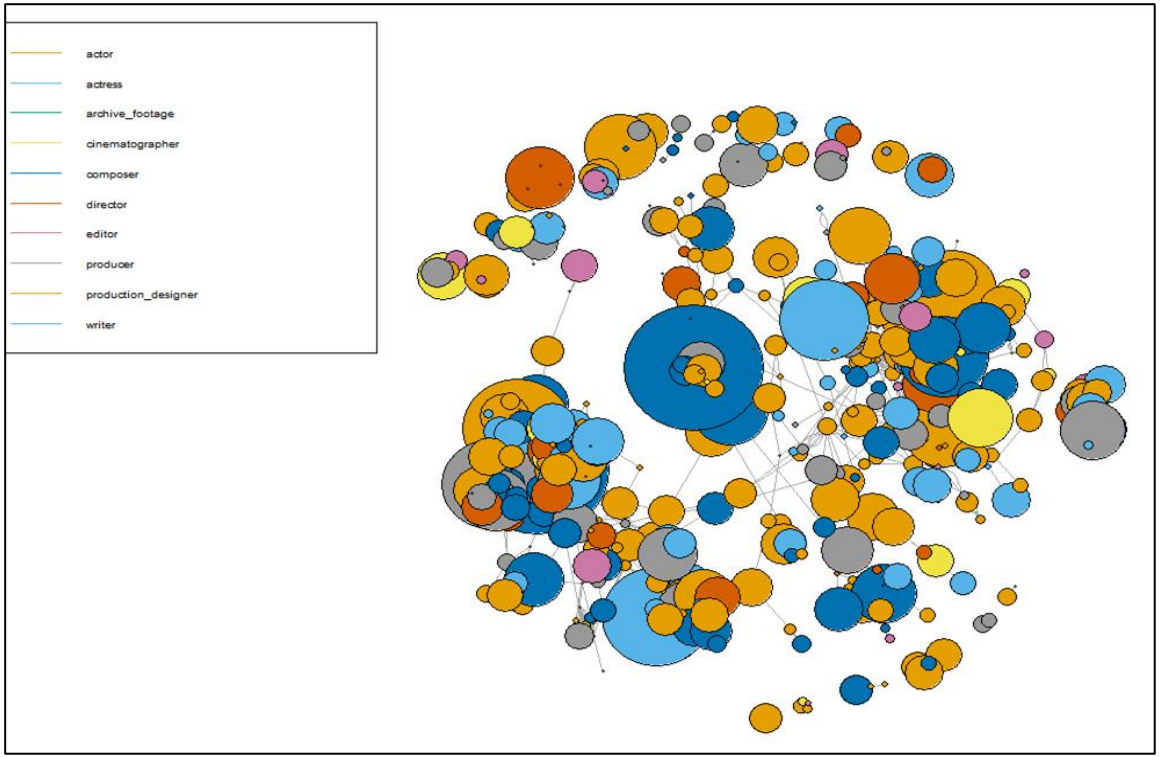
How negative words influence sentiment

The most common positive or negative words to follow negations, such as "no", "not", "never", and "without"



Social Network Analysis





CONCLUSION

Results and Conclusion
Future Development

Conclusion

- Drama, Action, Adventure - Most Popular Genres
- Increase in Biography, Music and Musical Genres with High Rating Scores
- Adventure always comes with Action; Drama can apply to various genres.
- Negative: score is high

Recommendation for IMDB:

- Display keywords and topics
- Reviews on several aspects

Recommendation for movie maker:

- Future movie type
- Find good story
- Find good partner

Future Development

- ★ **Data Diversity**
  - Extract more movie related data from different movie review platforms
  - Extract more relative information: box-office, budget, etc.
- ★ **Unsolved BI Questions:**
  - Popular directors/actors
  - May require more information on personalities, but we focused on movie
- ★ **Text mining:**
  - Too many keywords that are related to scenarios
  - Think about getting rid of spoiler reviews or episode highly-related reviews

Applications:

- ★ movie review platforms like IMDb.com
- ★ TV series platforms, social media platforms, and also the most popular short-form mobile video apps, such as Tiktok and Douyin