Data mining on Friends quotes

Rajpreet Singh Samra, Yanzhong(Eric) Huang, Yongyi Tang

Executive Summary

This project explores the intricate relationships and emotional dynamics of the beloved characters from the iconic sitcom F.R.I.E.N.D.S. By leveraging advanced data mining techniques, we analyzed quotes from the show to uncover patterns of interaction, sentiment, and association among the six main characters. The analysis provides a unique perspective on the ensemble nature of the show and the emotional tapestry that binds its characters together. Key highlights include:

- **Missing Data Handling**: Addressed inconsistencies in sentiment dictionaries using cosine similarity to estimate and fill missing values, ensuring a robust foundation for analysis.
- **Sentiment Analysis and Clustering**: Quantified emotional tones of character quotes using sentiment lexicons, followed by PCA and k-means clustering to group characters based on emotional expressions, revealing distinct sentiment profiles.
- **Character Relationships via Association Rules**: Treated conversations as transactional data to uncover strong relationships between characters. Key dynamics include Chandler and Monica's romantic bond, Joey's role as a social connector, and the frequent ensemble appearances of all six main characters.

The findings highlight Joey Tribbiani as a central figure with strong associations across the group, reflecting his role as comic relief and a unifying presence. Romantic arcs like Ross and Rachel, and Monica and Chandler, are underscored by high mutual association. Phoebe Buffay's unique personality is evident in her relatively distinct interactions.

This project celebrates the enduring appeal of F.R.I.E.N.D.S while showcasing the power of data mining to uncover hidden patterns in cultural phenomena. The insights gained offer a deeper appreciation of the show's storytelling and character development, making it a valuable case study for fans and data enthusiasts alike.

Introduction

F.R.I.E.N.D.S is a popular sitcom that aired from 1994 to 2004. It follows the lives of six friends living in Manhattan, New York City. The show has become a cultural phenomenon and has a massive fan base around the world. In this report, we analyzed quotes from the show using data mining techniques to gain insights into the characters and their relationships.

The Data description seciton The dataset we used for this analysis contains quotes from the show, along with information about the characters who said them. The dataset was obtained from Kaggle. We also used sentiment analysis datasets from Kaggle to analyze the sentiment of the quotes.

Major achievements of this project include:

- **Missing data handling on sentiment analysis datasets**: the sentiment dictionaries are mismatched with each other. We used cosine similarity to fill the missing sentiment values.
- **Sentiment analysis based PCA and k-means clustering**: we used the sentiment dictionaries to analyze the sentiment of the quotes. Then using PCA and k-means clustering, we clustered the characters based on their sentiment.
- **Relationship analysis based on association rules**: we used association rules to analyze the relationships between the characters based on their quotes.

Data description

We are using five datasets in this project:

- `data/friends_quotes.csv`: sourced from Kaggle link
  - `author`: The character who said the quote
  - `episode_number`: episode number
  - `episode_title`: episode title
  - `quote`: the quote itself
  - `quote_order`: the order of the quote in the episode
  - `season`: the season number

Sentiment analysis datasets: Kaggle link

- `data/afinn.csv`: AFINN-111 sentiment analysis word list
  - `word`: the word
  - `value`: range from -5 (very negative) to +5 (very positive)
- `data/bing.csv`: Bing Liu's opinion lexicon
  - `word`: the word
  - `sentiment`: range from -1 (negative) to +1 (positive)
- `data/loughran.csv`: Loughran-McDonald sentiment word list
  - `word`: the word
  - `sentiment`: negative and positive

- `data/nrc.csv`: NRC Emotion Lexicon
  - `word`: the word
  - `sentiment`: the sentiment (e.g. anger, anticipation, disgust, fear, joy, sadness, surprise, trust)

Acknowledgements

The raw transcripts of every episode were originally scraped from here: Friends quotes. Additional work cleaning up the data and removing invalid rows was done by Jorge Nachtigall

Missing data handling on sentiment analysis datasets

Sentiment analysis

Relationship analysis based on association rules

If two characters have more converstations, they are more likely to have a closer relationship. If we could separate the quotes into "conversations", then we would be able to figure out who are in this conversation. Treat each conversation as an observation, and each character is a item. We can use association rules to find who have stronger association with other characters.

The steps are as follows:

1. Figure out the conversation boundaries.
2. Create a transaction table: each row is a conversation, and the value is a set of characters appeared in this conversation.
3. Use association rules to find the relationships between characters.

Figure out the conversation boundaries

We tested several approaches to figure out the conversation boundaries. First, we tried to use a fixed number of quotes as a conversation, typical length of a conversation for TV shows is around 15-30 quotes. However, this approach is not very accurate, as the number of quotes in a conversation can vary significantly. Second, we tried to use existing large language models such as ChatGPT to figure out the conversation boundaries, we still found that the results are way off. The model is not able to figure out the conversation boundaries accurately, as it is not trained on this specific task.

We finally decided to use "greetings" to determine the start of a conversation. The conversation is defined in:

0. Divide the quotes into episodes.
1. If a quote contains a greeting, then it is the start of a conversation.
2. For the following 5 quotes, if greetings appear again, then it is remaining in the same conversation.
3. Conversation continues until the next greeting appears.

Greetings include: hello, hi, howdy, yo, sup, morning, good morning. We also tried to use "hey", "greetings", "what's up", but they appear too often and not at the beginning of a conversation.

This approach is not perfect, since some conversations may not have greetings, it may result a low volume of conversations. It could further result more characters in a single conversation, raising the chance of false positives (Higher support and confidence level). However, our goal is to figure out who have the "strongest association" with other characters, it is a relative measure, absolute accuracy is not required. We believe this approach is good enough to figure out the relationships between characters, and indeed the results are quite interesting.

Here is a sample of the conversation, and the characters in this conversation:

Conversation	Characters
0	[chandler, joey, monica, phoebe, rachel, ross]
1	[chandler, joey, monica, phoebe, ross]
2	[chandler, joey, monica, paul, phoebe, rachel,...
3	[chandler, joey, monica, paul, phoebe, rachel,...
4	[chandler, joey, monica, on, paul, priest, rac...

The association rules

We used the apriori algorithm to find the association rules between characters. The minimum support are set to 0.05.

Most frequent characters:

	support	itemsets
0	0.785816	(chandler)
5	0.785816	(rachel)
6	0.785816	(ross)
1	0.764539	(joey)
2	0.761702	(monica)
4	0.727660	(phoebe)
8	0.680142	(chandler, monica)
7	0.669504	(chandler, joey)
22	0.663121	(rachel, ross)
11	0.661702	(chandler, ross)

The results show that Chandler, Rachel, and Ross are the most frequently appearing characters in conversations. Also the all six main characters (Monica, Joey, Chandler, Ross, Rachel, and Phoebe) are frequently appearing together in conversations. This obviously reflects the ensemble nature of the show.

Single antecedent

We filtered the results to show only the rules with a single antecedent. The results are as follows:

- **Joey Tribbiani**:

antecedents	consequents	support	confidence	lift
(joey)	(chandler)	0.669504	0.875696	1.114378
(joey)	(ross)	0.641844	0.839518	1.068339
(joey)	(rachel)	0.636170	0.832096	1.058895
(joey)	(monica)	0.631915	0.826531	1.085110
(joey)	(phoebe)	0.615603	0.805195	1.106554

- **Ross Geller**:

antecedents	consequents	support	confidence	lift
(ross)	(rachel)	0.663121	0.843863	1.073869
(ross)	(chandler)	0.661702	0.842058	1.071572
(ross)	(joey)	0.641844	0.816787	1.068339
(ross)	(monica)	0.634043	0.806859	1.059284
(ross)	(phoebe)	0.613475	0.780686	1.072872

- **Rachel Green**:

antecedents	consequents	support	confidence	lift
(rachel)	(ross)	0.663121	0.843863	1.073869
(rachel)	(monica)	0.643262	0.818592	1.074688
(rachel)	(chandler)	0.642553	0.817690	1.040562
(rachel)	(joey)	0.636170	0.809567	1.058895
(rachel)	(phoebe)	0.617021	0.785199	1.079074

- **Monica Geller**:

antecedents	consequents	support	confidence	lift
(monica)	(chandler)	0.680142	0.892924	1.136302
(monica)	(rachel)	0.643262	0.844507	1.074688
(monica)	(ross)	0.634043	0.832402	1.059284
(monica)	(joey)	0.631915	0.829609	1.085110
(monica)	(phoebe)	0.630496	0.827747	1.137547

- **Chandler Bing**:

antecedents	consequents	support	confidence	lift
(chandler)	(monica)	0.680142	0.865523	1.136302
(chandler)	(joey)	0.669504	0.851986	1.114378
(chandler)	(ross)	0.661702	0.842058	1.071572
(chandler)	(rachel)	0.642553	0.817690	1.040562
(chandler)	(phoebe)	0.627660	0.798736	1.097679

- **Phoebe Buffay**:

antecedents	consequents	support	confidence	lift
(phoebe)	(monica)	0.630496	0.866472	1.137547
(phoebe)	(chandler)	0.627660	0.862573	1.097679
(phoebe)	(rachel)	0.617021	0.847953	1.079074
(phoebe)	(joey)	0.615603	0.846004	1.106554
(phoebe)	(ross)	0.613475	0.843080	1.072872

The results are quite interesting. For example, we can see that:

- Joey has a strongest association with Chandler, it highlights their close friendship and frequent interactions.
- Both romantice relationships (Ross and Rachel, Monica and Chandler) have a strong association with each other. This reflects the romantic arcs in the show.
- Phoebe appears at the last place for all other characters, which is quite interesting. It may reflect her unique personality and the fact that she often stands out from the group.
- Joey has the highest average confidence level, which mean he probably has the most conversations with other characters. This is consistent with his character as the comic relief and the one who often interacts with others.

Other interesting rules

For Chandler and Monica couples, we could clearly see that they have a strong association with Joey, again Joey is the one who ofter appears in their conversations. This is consistent with the show, as Joey is often the one who brings them together and is involved in their relationship.

antecedents	consequents	support	confidence	lift
(chandler, monica)	(joey)	0.595035	0.874870	1.144310
(chandler, monica)	(ross)	0.587234	0.863399	1.098730
(chandler, monica)	(rachel)	0.581560	0.855057	1.088114
(chandler, monica)	(phoebe)	0.578723	0.850886	1.169347

Conclusion

Balabala