DSCI510 - Final Project - Big Bang Theory Rating

Qianmian Gai

Motivation

The Big Bang Theory is a well-known sitcom series with over 200 episodes. While the general rating for the series is good, whether each episode enjoys the same level of fame remains a question. On the other hand, what features actually contribute to the rating of each episode is also an interesting question to ask. In this project, I am trying to figure out the features contribute to the IMDb rating of each episode, and analyze the change of several features of the first 10 seasons of the series over time.

Data Source

In my project, I collect the data from the following three websites and got the features for each of the 279 episode:

1. Wikipedia (https://en.wikipedia.org/wiki/List_of_The_Big_Bang_Theory_episodes)

Wikipedia gives a brief summary of all episodes of the Big Bang Theory in regard of its production, episode length, writers, etc. I basically collected the following features by scraping the Wikipedia web page:

* overall number of the episode in the 279 episode
* Number of the episode in its season
* Director name
* Number of viewers (unit: millions) in the U.S.
* Number of its season
* story writer(s)
* teleplay writer(s)

2. Bigbangtrans (https://bigbangtrans.wordpress.com/)

Bigbangtrans is a fan-made website that contains the transcripts of every episode of the first 10 seasons of the Big Bang Theory. Because every line starts with the character who says it, it is easy to calculate how many lines one particular character says. I scraped the following features from this web page:

* The number of lines in the transcripts for each main characters, including Penny, Howard, Sheldon, Leonard, Raj, Bernadette, Amy (of the first 10 seasons).
* The average length of the lines in each episode (of the first 10 seasons).

NOTE: Because bifbangtrans only provides the transcripts of the first 10 seasons, the columns [Penny, Howard, Sheldon, Leonard, Raj, Bernadette, Amy, average length] in csv file were left empty for the 11 and 12 season. THIS IS NOT AN ERROR.

3. IMDb (API: SeasonEpisodes, instruction on this page: https://imdb-api.com/api/#UserRatings-header)

IMDb is the most commonly used movie/TV series rating website. I used its official api to collect the following data:

* How many people rated the episode on IMDb
** THE IMDB RATING OF THIS EPISODE (This is the dependent variable)
(16 features and 1 dependent variable in total.)

Since the object of each features is each episode separately, I used pandas data frame, and csv files to store the data.

Analysis

1. lineplot function

The lineplot() function provides the change of features ['No.overall', 'No. inseason', 'viewers(millions)', 'season', 'Penny', 'Howard', 'Sheldon', 'Leonard', 'Raj', 'Bernadette', 'Amy', 'Ave_length', 'imdb_rating_count', 'imdb_rating', 'story', 'tele'] over time. Among all features, 'No.overall', 'season' and 'No. inseason' do not need any analysis since those are the identity features for each episode; 'viewers(millions)' is the viewers in the US; 'Penny', 'Howard', 'Sheldon', 'Leonard', 'Raj', 'Bernadette', 'Amy', represents the number of lines each main characters in one episode; 'Ave_length' is the average length of all the lines from one episode; 'imdb_rating_count' is the number of people rated the eipsode on IMDb and 'imdb_rating' is the rating of the episode; finally, 'story', 'tele' represents how many writers contributed to the story or the teleplay seperately.

From the graphs of lineplot(), we can see that most characters' line has no pattern of increase or decrease when they are on the show (Amy, Bernadette joined the show later than the other five characters). The average length of all the lines from one episode basically remained the same. Also, there is no obvious pattern on the number of writers contributed to the story or the teleplay.

An interesting fact is that, while the general viewers increased in the first 7 season, and then gradually, slightly decreased for the later seasons, the number of people rated the series on IMDb keep decreasing since the beginning. This might reflect that less critics are paying attention to this series in the end. The rating itself was relatively high in the first several seasons, and decreased in the end (about the last three seasons). This is the symbol of a less interesting plot or production in the later several seasons and which lead to the cancellation of this series.

2. describe function

From the describe() function we can see the basic description of each feature (features explained in the first paragraph of the analysis.1), we can see Sheldon is the main character since he has the most lines overall (mean of 50.5). Penny and Leonard also have more lines among other characters. For the rating of the series, the lowest is 6.8 and the highest is 9.2, showing that this is generally a good show for those 10 seasons. In average, the story and teleplay are a result of 2 or 3 writers, which is the evidence of collaboration of play writers.

3. regression function

From the regression result (we do not consider the season number feature in this analysis), we can see the important features (p smaller than 0.01) is the following four features: number over all episode, number in seasons, the average length of lines, and how may people rated the episode. The coefficient of those four features are negative, positive, positive and positive. While we cannot assert any causality from the regression result, we can infer that the average length of lines and people rated the episodes has a positive relation with the rating. Over all, the show was the most highly evaluated at the beginning and in each season, the rating slightly increase from the first episode to the last one.

Conclusion

Since the analysis method is basic, we cannot arbitrarily draw any solid causal conclusion. However, some objective observation show that the rating of the series decreases over time, and the popularity of the show increased in the first 7 season and then decreased later on.