Upcoming TV Series Episode Rating Prediction

Course BIA-660-C

Instructor: Rong Liu

Submitted by: Group 4

Manas Kulkarni

Allan Shivji

Diksha Pancholi

Shubham Heda

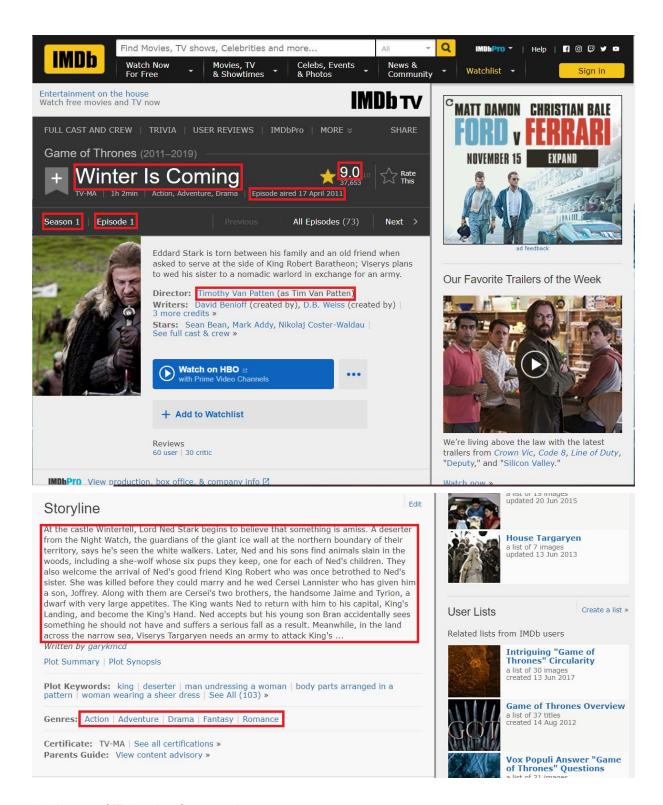
Objective

- The prediction of upcoming episode of a TV series can be predicted using the ratings storyline, cast and the title of the previous episode. In this Web Mining project, we propose a method to predict the ratings of future episode based on the storyline, title and the star cast that is present on IMDB's website.
- In our project, we will take data from IMDB's website such as TV series' storyline, title, rating, air date and the star cast which we will scrape using Web Scraping methods. Then we will predict the rating of upcoming episode purely based on previous episodes rating using Logistic Regression.
- After that we will predict the rating of upcoming episode by adding other features like title, storyline, cast, airdate etc. using Sentimental Analysis, Natural Language Processing and Text Mining, then compare the accuracy of the model for both the predictions.
- In this project, we are set to create an analysis which helps the workforce
 keep up with the trends around them and develop the skills which cannot be easily
 replaced by automation.

We will consider any one TV Series and then we will first train a model which will predict the rating of the upcoming episode purely based on previous episodes' ratings. Then we will try to find out if features such as title, storyline, cast, air date etc. gives us any insights to predict the rating of the next episode. For example, if for a particular TV series, the ratings for the previous episodes were 8.5, 8.0, 7.6

IMDB Website

- IMDB is a website where you can find all the information related to films, TV series, viewers rating, plot summaries etc.It uses data science to evolve the human capital management space, with options that assist employers find, employ and manage fantastic people.
- The TV Series dataset will include 'rating' which is a combined rating of all the viewers, 'title'which is the name of the episode, 'storyline' which is a short summary of the episode, 'air-date' is the date on which the episode is aired, 'cast' are the actors/director features in the episode.



Attributes of TV series for scraping:

- 1. Title, Director
- 2. Rating, Airdate
- 3. Season, Episode number
- 4. Storyline, Genres
- 5. Cast, Character

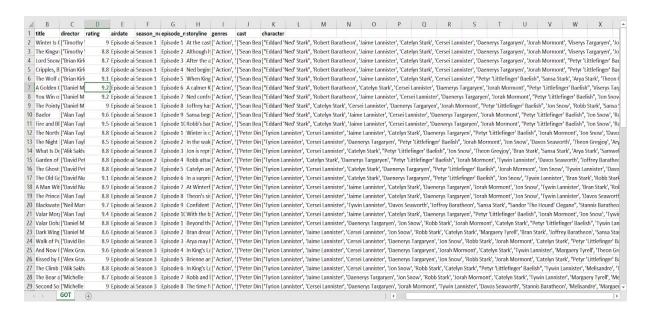
Web Scraping

- We scraped two TV series named Game of Thrones and Breaking Bad
- Game of Thrones consists of eight seasons & total of seventy-three episodes and Breaking Bad consists of five season & total of sixty-two episodes.
- We used BeautifulSoup library and scrapped the attributes as mentioned in above section

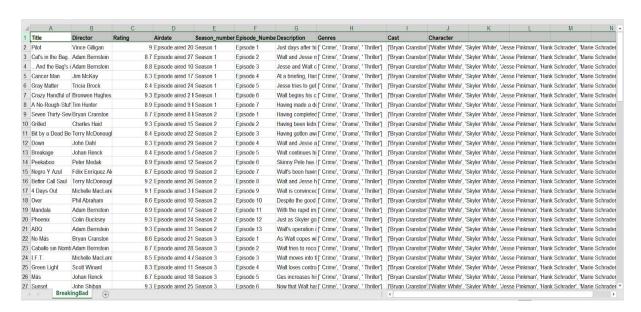
Code snippet for scrapping:

```
In [3]: 1 data=[]
2 for i in range(len(name)):
                    #print(z)
sublink="https://www.imdb.com"+link[i]
                    sublink="https://www.imdb.com"+link[i]
page = requests.get(sublink)
#page = requests.get("https://www.imdb.com/title/tt1480055/?ref_=ttep_ep1")
if page.status_code==200:
    soup = BeautifulSoup(page.content, 'html.parser')
                          soup.prettify()
                          title=soup.find('div', attrs={'class':["titleBar"]}).find("h1").get text().strip()
                          \textit{\#director=soup.find('div', attrs=\{'class':["credit\_summary\_item"]\}).find('a').get\_text()}
                          directorArrays=[]
                          director=soup.find('div', attrs={'class':["credit_summary_item"]}).findAll('a')
for x in director:
    directorArrays.append(x.get_text())
                          rating=soup.find('div', attrs={'class':["imdbRating"]}).find("strong").get_text()
           22
23
24
25
26
                          airdate=soup.find('div', attrs={'class':["subtext"]}).find('a', title="See more release dates").get text().strip()
                          season_number=soup.find('div', attrs={'class':["bp_heading"]}).get_text().split('|')[0].strip()
           30
31
32
                          episode\_number=soup.find('div', \ attrs=\{'class': ["bp\_heading"]\}).get\_text().split('|')[1].strip()
                          #print(episode_number)
                          storyline = soup.find('div', \ attrs=\{'class': ["inline \ canwrap"]\}).find('span').get\_text().strip()
                          #print(storyline)
                          p=soup.find('div', attrs={'class':["see-more inline canwrap"]}).find('h4').get text()
                          if p=='Plot Keywords:':
    genresAnchors=soup.findAll('div', attrs={'class':["see-more inline canwrap"]})[1].findAll('a')
           40
                               genres=[]
                                for a in genresAnchors:
```

Scrapped data from IMDB for Game of Thrones:



Scrapped data from IMDB for Breaking Bad:

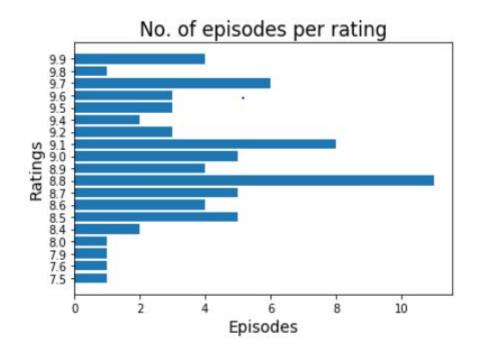


Exploratory data analysis

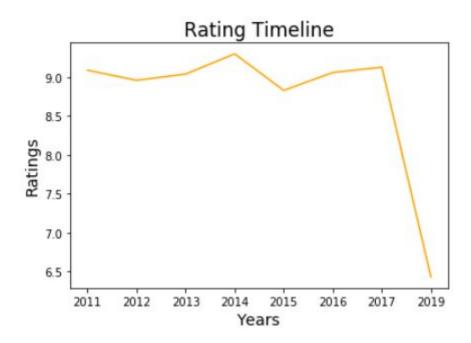
- Since we had a lot of data that was scraped from IMDB website as seen in the Excel files. So we did perform some Exploratory data analysis to understand the relations between the ratings and the other columns.
- As seen in the below graphs first we performed total number of episodes per their rating so we understood that there were some episodes which had less rating.
- Next we showed the rating timeline graph per year and we could see that there were some ups and downs in rating per year and it was never a constant value.
- Later we showed a bar graph of Directors per rating and we could observe that the rating of the episodes depends on who directed the TV series of that particular season.

Plots:

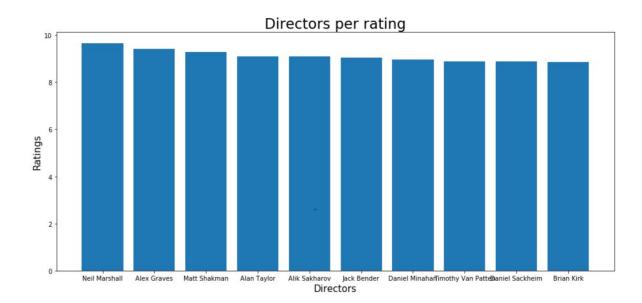
1. Number of episodes per rating



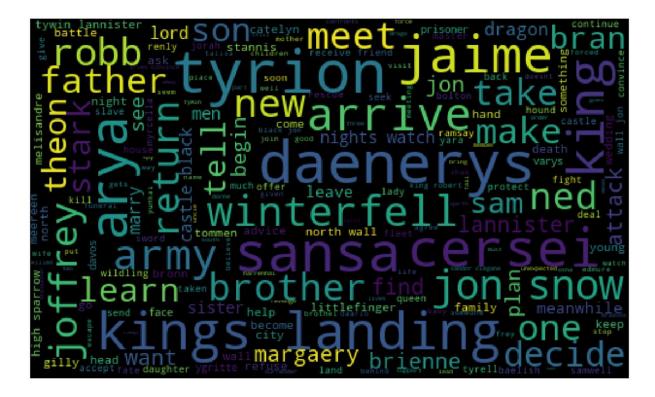
2. Trend of rating over the years



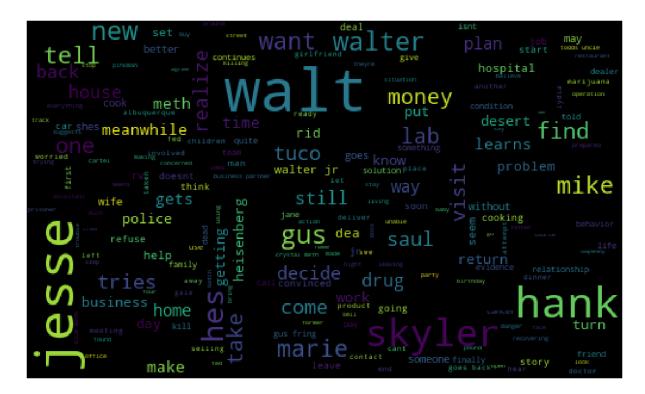
3. Directors per rating



Word Cloud based on storyline tokens of Game of Thrones:



Word Cloud based on storyline tokens of Breaking Bad:



In this we have tokenized the plot and removed the stop words from the plot and then created a WordCloud which shows most appeared word in the entire plot of a particular series.

Steps completed till now in the project along with the team member is given below:

Task completed	Member Allocated	
Scraping of TV series from IMDB	Shubham and Diksha	
Data Preprocessing	Manas	
Exploratory Data Analysis	Allan	
Documentation	Manas, Allan, Shubham	

Action Plan

- We will predict the rating of upcoming episode by adding features like title, storyline, cast, airdate etc. using Sentimental Analysis, Natural Language Processing and Text Mining, then compare the accuracy of the model for both the predictions.
- As per the results obtained we will find features that will define the ratings of upcoming episodes.

References:

Research Paper 1: In You-Jin Kim, Jung-Hoo Lee and Yun-Gyung Cheong (2019). Prediction of a Movie's Success From Plot Summaries Using Deep Learning Models. Proceedings of the Second Storytelling Workshop, Association for Computational Linguistics

Research paper 2:

http://usir.salford.ac.uk/id/eprint/18838/1/Wessex_movie.pdf:public A data mining approach to analysis and prediction of movie ratings. Authors Saraee, MH, White, S and Eccleston, J. Type Conference or Workshop Item. Published Date 2004