Data Mining Project

Analysing Popularity of TV Shows

Group ID **: G11**

Team Members **:**

1) Samarth Mittal
2) Siddhant Jain
3) Rattanjot Singh
4) Dhruv Jogi

OBJECTIVE :

- Data collection for various TV Series using Tumblr API.
- Learn about Tumblr API.
- Analysing popularity of TV Shows episode wise as well as season wise.
- Finding the most enthusiastic bloggers of various seasons in different TV series.
- How does the behaviour of these enthusiastic bloggers vary throughout the season.

Learning Outcomes :

- Learned how to collect data from tumblr
- Learned how to collect more than 20 posts from tumblr
- Learned how to use sql connector
- Learned how to make a database with relationships
- Learned how to store data collected from an api in sql
- Found relationships between various attributes
- Learned how to transform raw data into an understandable format that is consistent with certain behaviour or trends.
- Learned how to approach to analyse datasets to summarise their main characteristics with visual methods.
- Performed EDA to see what the data couldn't tell us through formal modelling.
- Gained insights into the data via Exploratory Data Analysis.

Data Overview :

Various types of data available on Tumblr include Text, Photo, Chat, Quote, Link, Audio, Video, Answer. However, we collected the data which belonged to Text, Photo, Chat, Quote or Answer type. We discarded Audio, Video as well as Link types of posts since these types of data were not relevant for the type of EDA which we performed. The attributes that we required to perform the EDA include post_id, blog_name, post_url, timestamp, reblog_key and note_count. These attributes remain common throughout all the types of posts. The image type posts include the important caption attribute while there is title and body in the case of text type posts.

The TV series for which data was collected include Game of thrones, Homeland, Sherlock and Big Bang Theory. Game of Thrones consisted of around 7000 text data and 17000 image data. Sherlock consisted of around 5000 text and image data. Homeland consisted of around 1500 Text data and 4000 image data. While Big Bang Theory consisted of around 3300 text data and 11000 image data. The data corresponding to other categories(Answer, Quote, Chat etc..) was comparatively very less.

Inferences :

While performing EDA we performed total 9 queries on our dataset and plotted relevant graphs using line charts and bar charts. From these graphs we were able to collect and consolidate information, and then manipulating and analysing it to uncover pattern, trends and relationships.

1) We judged the episode-wise popularity of TV Shows by plotting the count of posts posted in particular intervals of time. In Game of Thrones, the line charts of season 1 and season 7(1s1got.png and 1s7got.png) show that the popularity increases as the season progresses. In case of Sherlock, during season 3, episode 1 has the highest popularity and during season 4, episode 2 has the highest popularity. However, no regular trend can be estimated. In case of BBT, the popularity doesn't vary much as can be seen in the graphs(1s6_BBT,1s10_BBT). At the time of trailer in season 1 many posts were there but they fell drastically down as it actually started airing though it gained momentum as the show progressed so we can say that it didn't start so well but ended on a popular note . In season 6 as the TV series progressed the popularity kept on increasing though there was a huge dip during episode 3 implying that episode not being upto the mark but it ended very good .(1s1home.png and 1s6home.png).
2) The popularity of series episode-wise was estimated using the number of positive and negative posts(The positivity and negativity of posts was decided using TextBlob) . As can be inferred from the graphs(4s1got.png and 4s7got.png), in case of Game of Thrones, the popularity in general increases as the season progresses. In case of Sherlock, again the highest popularity occurs at episode 1 in season 3 and episode 2 in season 4. In case of BBT, the popularity of season 6 episode 17 dips drastically as can be seen form chart(4s6_BBT.png). While in season 10, episodes 18 and 21 saw quite low reviews which indicate dip in popularity. Started with lot of positive reveiws but they reduced a lot as it progressed negative review increased during episode 9 so episode 9 was not liked much . In season 6 episode 6 got more negative reveiws than positive reviews showing it being not liked much . Season finalle gathered most comments both positive as well as negative .
3) Note_counts indicate the popularity of posts and in turn the popularity of corresponding TV Show. We have used this fact to analyse the the popularity of TV Shows. In case of Game of Thrones, from the line charts(7s1got.png and 7s7got.png), it is quite clear that the popularity of the show was at its peak when season episode 5 was aired and season 1 episode 2 was released. In Sherlock, there were highly negative reviews just after the trailer during season 3 and season 4 episode 2 has the highest popularity. In BBT, season 6 episode 15 had the highest popularity and again the popularity starts dipping. Season 10 episode 21 saw drastically reduced popularity with least total reviews after which the show again gets going. Season 1 largerly had positive reveiws gathering lots of likes and reblogs while negative comments were largerly unsupported with the exception of episode 9 that had more negative reviews than positive reviews .In season 6 Episode 2 had a lot more positive reveiws than negative ones and episode 8 had slightly more negative reviews than positive ones .

Episode 11 had the most number of likes and reblogs both positive as well as negative meaning lot of people saw it but had mixed reveiws .

4) Taking the median and maximum of note_counts for various types of posts season-wise, we have tried to judge the popularity of a type of post for different TV Series. In Game of Thrones, from the graphs of median and max(9s1median_got.png, 9s7median_got.png, 9s7max_got.png and 9s1max_got.png), in all the cases Photo type post always has a higher attraction since the number of note_counts is much larger than other types of posts. The same has been observed in case of Sherlock, BBT and Homeland. From the above inferences, we can say that people like to post and share posts that have images which tells us about the high perceptual and vision intelligence of people in the social media.
5) We tried to find die hard fans of Game of Thrones that actively blogs throughout the series but we can see (q2resp_got.txt) the most 10 active blogger changes till the series end. In case of Sherlock From this we can infer that a03feed-johnlock is a die hard fan of Sherlock apart from this many new bloggers joined the Sherlock fandom while no previous blogger apart from a03feed-johnlock made to the top bloggers of season 4 . This shows growing popularity of the fandom (q2resp_sher.txt). There were no bloggers from season6 who made it to the top bloggers of season 10 . Though many new bloggers joined the big bang theory fandom .(q2resp_BBT)There were no bloggers from season 1 who made it to the top bloggers of season 6 . Though many new bloggers joined the homeland fandom and even the number of posts by them increased a lot .
6) Jonerys, jon X dany tag got high negative reviews.We can say that the couple Jon Snow and Daneyrs wasn't liked by people. A Dance with dragons got high positive reviews we can say people liked fictional characters like dragons in the show. (q5resp_got.txt)Mycroft Holmes got high negative reviews which can be related to the fact that he used to trouble the main character sherlock holmes which wasn't liked by people. Moriarty got high positive reviews so despite him playing anegative character he still managed to be a fan favourite which speaks volume about the acting ability of his actor . The sign of three got more positive reviews then the other 2 episodes the empty hearse and the last vow .(q5resp_sher.txt)In both the seasons shamy remained the most positive sentiment collecting tag which is the fan name of couple scheldon and Amy which shows the popularity and liking of this couple . Even Scheldon Cooper and Amy Farah Fowler have largerly positive sentiment showing that people like them not only as a couple but individually as well .(q5resp_BBT) Claire danes got lot of positive reveiw in season 1 but became less popular in season 6 . In season 6 Peter Quinn gathered lots of positive reveiws . Carrie Mathison who had largerly positive reveiws in S1 gathered mixed responses in Season 6 .
7) We have also tried to analyse popularity using sentiment of tags and text posts. In case of Homeland, in season 1 a lot of tags and body had different sentiment they largerly disagreed . In episode 7 mostly the tags agreed meaning they give the smae sentiment mostly . In episode 12 the sentiment by various tags was the most . In season 6 episode 1,7 and 12 had largely sentiment of tags and body agreeing with each other meaning they implied the same sentiment . While episode 4,5,9 and 10 had tags and body largely disagreeing .

Results and Conclusions :

- The data mainly consisted of Photos and Texts while the number of remaining types of posts were very limited. Hence, most of the conveying about TV Series happens through images and text posts as compared to any other type of post in Tumblr.
- The Popularity of TV Shows in general tend to increase with the progress of the season as we saw in the case of Game of Thrones, Sherlock as well as Homeland.
- We have observed that the top bloggers for a TV Series keep on changing from season to season.
- We have found that people relate to photo post far better than text post and in general a photo post gains much more popularity as compared to a text post which is clearly evident from the graphs(max and median graphs for every season of every episode)