TV Shows Muster and Analysis

Aditya Bhagwat
Binghamton University
abhagwa1@binghamton.edu

Abhimanshu Mishra
Binghamton University
amishr11@binghamton.edu

Sharvari Joshi
Binghamton University
sjoshi14@binghamton.edu

Vinit Bhosale
Binghamton University
vbhosal1@binghamton.edu

ABSTRACT

Online entertainment has become a crucial part of life in the modern age. In recent times, the amount of content and the number of content producers has increased exponentially. With such an increased consumption of content, everyone has opinions about different TV shows and water cooler discussions have become commonplace. People take to different social media forums to express what they think and connect with the world at large. This provides data scientists and researchers a unique opportunity to study the pulse of the public about a specific show easily. Analysis of TV shows based on various factors can provide insights about the influence of shows on age demographic, whether their reviews have a positive or negative impact on the viewers, their popularity based on trends and in a particular location.

1 INTRODUCTION

Our project aims to collect data about TV Shows from different social media platforms like Twitter and IMDb like tweets and ratings on a large scale and draw patterns about the popularity of a show, it's influence on users from different age groups, the amount of engagement it receives based on retweets, likes and replies as well as how the show trends in different parts of the world. The data sources we have selected are IMDb and Twitter. IMDb is a movie/tv show rating and review website which is very popular around the world and is trusted in most circles for recommendations of content. Twitter is one of the largest social media forums in existence and has become the platform of choice for people to express their feelings and broadcast it on a universal scale. Based on the data collected from these two platforms we would perform appropriate sanity checks and analyze the data to reach our goal. Furthermore, we propose an exploration of sentiment analysis techniques using this collected data.

2 DATA SOURCES

2.1 Twitter

Twitter is a micro-blogging and social networking service where users post and interact with messages, "tweets," restricted to 280 characters. Information can be shared using photos, videos and links as and when they're happening, enabling insightful and real-time search results. Tweets will be fetched using the Official Twitter API which allows up to 50 requests per hour. Tweets collected are provided in JSON format which needs to be stored on our remote storage. The tweets collected often contain some fields that are irrelevant to the analysis to be performed , so the process of data cleaning and removal of these irrelevant (to the analysis) fields needs to be done before it can be sent for storage. We do so by storing the tweets locally in a python list where they're stored in a dictionary format thus creating a list of dictionaries where each entry in the list (a dictionary) represents an individual tweet received. The JSON response (tweet) consists of various fields like text field containing text of status update, user field containing information of user who posted the information, coordinates field containing Geographic information of the user, retweet count field containing number of times tweet has been retweeted, favorite count field containing number of times tweet has been liked by any twitter user. Based on the extracted field values, analysis can be made to decide whether a TV show has a good review or not based on the keywords in the tweet, the number of times it has been retweeted, the likes and replies it received and also the number of times it trended at a particular location in the world. This information also helps in determining how a negative or positive tweet can easily influence other social media users in reviewing a TV show.

2.2 IMDb

IMDb (Internet Movie Database) offers an extensive database of information related to films, television program, video games, and streaming content online – including crew details, plot summaries, trivia, fan and critical reviews.IMDb currently contains approximately 6.5 million titles (which includes movie and TV show titles), 10.4 million personalities as well as 83 million registered users. We will be finding information of the TV show by using OMDB API which allows maximum 1000 requests per day. Information collected from OMDB API contains data about TV shows like their ratings, genre it belongs to, release dates and other such information. For the user reviews we will be parsing the IMDb website using XPath queries. The extracted information will help us to analyze whether a show is the most liked or least liked among TV fanatics as well as other useful knowledge about the ratings.

3 DATABASE

3.1 MongoDB

To store the data collected using the Twitter and OMDB API we will be using MongoDB, an open source document-oriented database system. It is part of the NoSQL family of database systems. Classical relational database stores data in tables but MongoDB stores structured data as JSON-like documents with dynamic schemas known as BSON format. As our data will be in JSON format it will be easier to manage and integrate it using Mongo DB. Information collected using OMDb API, Twitter API and IMDB website will be stored in a single MongoDb database where the collection name will match the name of TV show.

4 DATA COLLECTION

We start by running a Daemon script on the virtual machine Written using cronjob which runs the python script every 15 minutes of an hour. This python script uses Twitter API to fetch the tweets of various TV shows in JSON format. The hashtags provided by us to the API results in the tweets to be fetched and will be stored in a list which creates list of dictionaries. Certain fields are extracted from this twitter data and those fields are stored into MongoDB into a single database with multiple collections. Each collection represents a different TV show in the database. OMDB API[1] will be used to fetch user ratings and IMDB [2] will be used to parse user reviews as specified above. Some upcoming and recently launched shows were selected and corresponding data is collected using Twitter [3]and OMDB API and stored in MongoDB. Based on the launch of TV shows on worldwide streaming platforms, the popularity of these shows may vary with time which is one major [4] factor of consideration for the task at hand. In the process of data collection, if we notice a decrease in a show's ratings, we might replace it with another show with a better popularity score for that week.

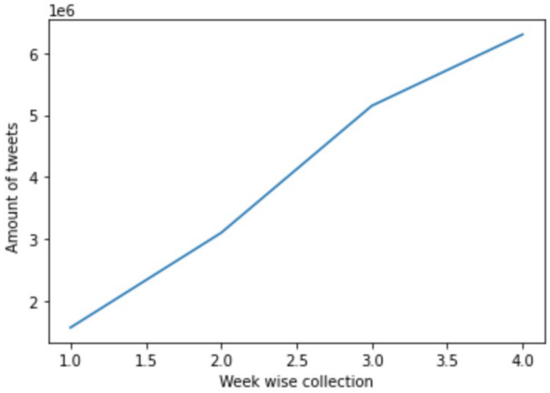

Figure 1: A graph showing the collection of data weekly using the Twitter API.

The above graph plots the number of tweets we collected using the Twitter API against a particular week. We expect to collect approximately 1 Million tweets from Twitter from which we estimate around 150 lakh tweets to be meaningful. We would be able to get a fairly justifiable sample at hand to develop insights using this information.

The above graph plots the number of reviews we collected from IMDb against a particular week. We expect to collect approximately 10,000 reviews from the IMDb website from which we estimate around 5k-6k to be meaningful. We would be able to obtain a fair graph from sample based on these estimations

5 CONCLUSION

Based on the collected data using the Twitter API and the OMDb API and IMDB website, we will be able to study and develop insights about the influence of shows on age demographic, positive or negative impact of reviews on the users, trending shows based on the popularity, region wise popularity of the shows and also the deep insight of a review considering retweets and likes as the parameters.

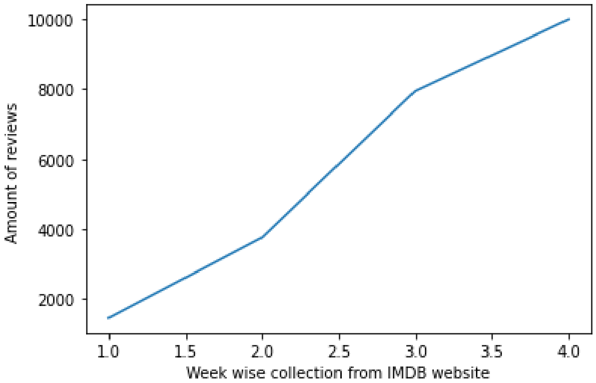

Figure 2: A graph showing the collection of data weekly from the IMDB website.

REFERENCES

[1] [n.d.]. http://www.omdbapi.com/
[2] [n.d.]. https://www.imdb.com/
[3] [n.d.]. https://developer.twitter.com/en/docs/twitter-api
[4] 3Nikhil Kulkarni Pranil Nalawade Prof.Aniket M. Junghare Veeresh Belgur, Aniket Karande. 2017. Statistical Analysis on Movie Reviews and Ratings. In *International Journal of Science, Engineering and Technology Research (IJSETR)*. IJSETR, MH, 2278 –7798.