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 [1] 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 when the data fetched is from a percent of sampled stream of the total tweets collected in real-time. We use the bearer token for authentication, provided by twitter API when requested using the developer account. We filter and scale the search request according the fields necessary to perform analysis and the user related data. Tweets collected are provided in JSON format which is collected and stored on our remote storage. Tweets collected may contain some noise or unnecessary data which can be easily removed before 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. Since the python script to perform this data collection is to be performed everyday at regular intervals of time. We use cron-job to achieve this task. This cron-job runs every hour with an interval of 10 minutes everyday. As the cron-job schedules the script to run after every 10 minutes, the stream is again refreshed so no duplicates arrive while fetching. At an approximate , it can be seen that around 600-700 meaningful tweets arrive after every call. So on an hourly basis we achieve upto 3000-3500 tweets per hour, which approximates to be about 80,000+ tweet per day. We consider this to be meaningful data since the data and noise cleansing has already been done. Even if the total meaningless data collected per day approximates to be around 10-15 per cent of the total collected per day, we are on track to achieve our estimated goal of achieving around 1 million+ meaningful tweets.

2.2 IMDb

IMDb (Internet Movie Database) [2] offers an extensive database of information related to films, television programs, 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. Based on the launch of TV shows on worldwide streaming platforms, the popularity of these shows may vary with time which is one major factor of consideration for the task at hand [4]. The

collection of TV shows data is done by using OMDb API and scrapping IMDB website. Each call of OMDb API depends on either TV show title or the imdbID, we call the api based on Tv show title and in response we get all the attributes "Title", "Year", "Rated", "Released", "Genre", "Writer", "Plot", "Language", "Ratings", "imdbId" and others. While collecting data of TV show we saw OMDb API failed to provide user reviews for TV shows, so to over come this problem we decided to scrape the imdb website. Scrapping the whole website was a challenging task, as it contains a lot of HTML tags, so we used xpath queries to solve this problem. Xpath query are used to query data from XML documents, we first convert the HTML code in a tree representation with tags as node by using html.fromstring() and then use xpath to selects node to extract data. By inspecting the webpage of imdb we searched those tags body which contains the user reviews of the TV show and with the help of xpath we then extract the text from the HTML code of that website. Once we extract the reviews we append those with the OMDb api [3] result and create a JSON collection object of each TV show which then get inserted into IMDb collection of MongoDB database on remote server and every TV shows associated data can be retrieved on demand for analysis. While performing the scraping logic we encounter the "Load more" button in the website which loads other user reviews for the TV shows, that's the only part of the scraping logic remaining so that we will be able to collect more user reviews for the TV shows.

3 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.

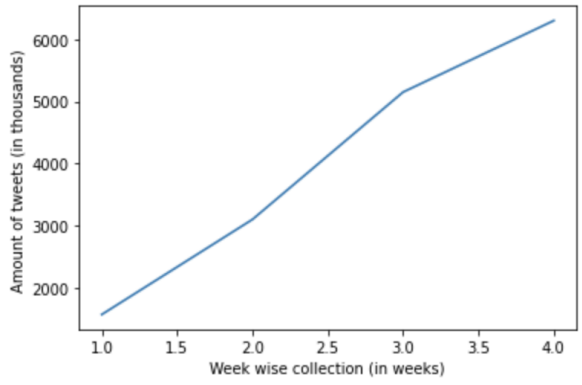

Figure 1: A graph showing the collection of data weekly using the Twitter API.

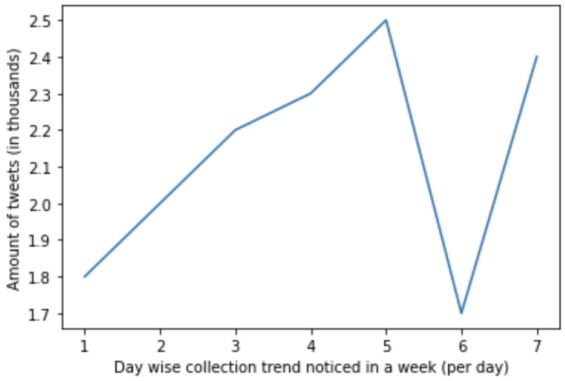

Figure 2: A graph showing the collection of data day-wise using the Twitter API.

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.

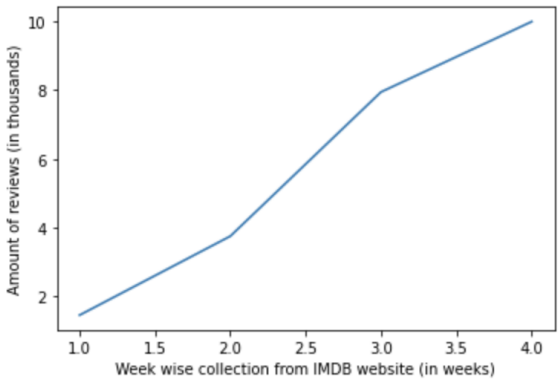

Figure 3: A graph showing the collection of data weekly from the IMDB website.

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

4 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.

REFERENCES

[1] [n.d.]. https://developer.twitter.com/en/docs/twitter-api
[2] [n.d.]. https://www.imdb.com/
[3] [n.d.]. http://www.omdbapi.com/
[4] Adam Nyberg. 2018. Classifying movie genres by analyzing text reviews. *CoRR* abs/1802.05322 (2018). arXiv:1802.05322 http://arxiv.org/abs/1802.05322