Youssof Mohammad Attallah

Big Data Project Report

CU1900127

Table of Contents

[Introduction: 2](#_Toc103445751)

[Task 1: 2](#_Toc103445752)

[Task 2: 5](#_Toc103445753)

[Task 4: 8](#_Toc103445754)

[Task 6: 13](#_Toc103445755)

[Conclusion : 14](#_Toc103445756)

# Introduction:

The field of data science has been growing bigger and bigger each year, this is due to the fact that the amount of data that we have has been growing at an exponential rate, and the more data you have, the more information you gain, and the more you gain, the more you are able to benefit from these insights which fuels more investment into tools to harvest more data and so on and so forth. In this project, we will start by creating our dataset through scrapping, then we will use tools to store these massive amounts of data and help us work on them efficiently, finally we will begin to examine the data up close and infer some new information or extrapolate some insights. We will be doing these operations on a dataset regarding anime, which is animated movies or tv series that is mainly from Asian countries like Japan, to understand more about it.

# Phase 1 (Data Collection):

The first, and most crucial, step of any big data project is the data collection. It is important because all the analysis that will come after will dependent upon the data’s attributes, format, and size. So we start by finding the data source which will be the website “9anime.to”. This website was chosen because it had the greatest number of animes on it compared to other websites, all the anime names were written in their English translated versions not In Japanese, which will come in handy in the sentiment analysis phase and having plethora of information regarding each anime such as views, rating, country, studio, and many more.

After the data source was chosen, a scrapping algorithm was written to automatically fetch all the data from the website. The layout of the website had a page were all the animes were listed in alphabetical order, each page contained a number of anime and had displayed the name and number of episodes in each anime. After clicking on the anime, it will direct it to its personal page where the user can watch it. This page will contain all the information we will want for our data analysis. An algorithm was written using the library of beautifulsoup to scrap the data, by first opening the alphabetical list page, taking the name and number of episodes and saving them, after that it will take the link to each anime and pass it to another function that will scrap the individual data, returning the remaining scrapped fields. It repeats this process until all animes in the page have been scrapped, then it will change the URL to go to the second page and repeat the process until all the animes have been collected. After that, it will use the pandas library to store all of the incoming data in a data frame before extracting it to a csv file to be operated on.

After all animes and their individual data has been collected, I did a quick sort based in number of views and got the top 30 animes, the list had duplicate animes, referencing the subtitled and dubbed version for the animes, or having the anime be represented multiple times for every season. So to avoid having these duplicates a manual selection for the top 10 viewed animes where listed. These then were used to scrap tweets that mention them. Using a library called Twint, each of these animes had 10,000 tweets extracted so we can review them later on in the sentiment analysis phase. The only exception was “Bleach”, as some tweets resulted were talking about the detergent thus invalidating a lot of the tweets.

Now, we have all our numerical data, and textual data, which allows us to perform exploratory data analysis, and sentiment analysis.

# Phase 2 (Hadoop):

In this phase, we are going to upload the data we have to Hadoop. Hadoop is a collection of tools that helps us process large amounts of data in a quick and efficient way. To start witht the setup, we created a Docker Image that hosts Hadoop, creating the docker-compose.yml file which will have the location of ther data files we have collected. Then, HQL files were created for each csv file, the HQL files define the tables, attributes, column names, and column data types for the files. After they are all set up and added to Hadoop. The files can be accessed and interacted with, allowing the creation of querues and such.

# Phase 3 (Sentiment Analysis):

In this phase, we will now begin working with the data we have collected and start processing it to obtain information. This information will come from the tweets that we collected regarding the top 10 animes. We will import the csv file that contains the tweets and begin to clean these tweets. The tweets need to be edited in this way to minimize un-useful data. We will be using the NLTK library to help us clean and work on these tweets. To clean the tweet, we will run functions that will remove unnecessary data such as symbols and trailing white spaces, after that we will remove all other symbols keeping only values in the alphabet. And finally, we will de-contract the mostly used contractions to have all the words be clear. Now that the tweet is purely a set of words, we will remove all the stop words from them. Stop words are a collection of words provided by the NLTK library that are not beneficial for the sentiment analysis, words such as : ( in, out , to , at, or, by, are) , so they are removed from the tweets.

After the tweets are now cleaned and ready to be explored, we start by creating a word cloud of the most frequent words that pop up in the tweets. We exclude the names of all the animes themselves and this is the result:

Text

Description automatically generated

If we look at the most common words, we find that anime is of course is the top word, think is also a top word implying that a lot of these tweets were about fans giving their opinions on each show. Some character names like Sasuke or Luffy make theb list, and othe anime abbreviants like AoT(standing for Attack on Titan) and jjk (stands for jujutsu kaisen) are present on her as well. Some anime specific terminologios are represented like manga, which is the material source animators use to animate any series, and amp (short fro aplificaation) and used to indicate that a character has gotten stronger.

Onto the next step which is analysaing the sentimanilty of a tweet, if that tweet is positive, negative or neutral. This is done by giving each word a score, and adding the total score for each tweet. If its positive than the tweet is considered to be a good tweet and vise versa. We start calculating the number of tweets in each class and ehre is the output:

Graphical user interface, text

Description automatically generated

And here represented as a pie chart:

Chart, pie chart

Description automatically generated

we infer from this ghraph, that the tweets overall are positve, but since the negative tweets are worth 27% of the total, this means that 1 out of every 4 tweets will be a negative one.

we will again use the wordcloud to display the most common words, but in the tweets we deemed postive alone, and for the tweets we deemed negative alone.

For the positive side:

Text

Description automatically generated

Here, the highlight words are positive words, such as good, great, and better. Which is what was expected.

Ontot he negative tweets:  
Text

Description automatically generated

Common words that are neutral, like anime or think, are still the most common, but in here a lot of profanities are now included which is expected since these words have a negative value.

Finally, we have a list of the most common:

Most frequent words: [('attack titan', 4317), ('https', 4188), ('naruto', 2972), ('tokyo revenger', 2934), ('one piece', 2436), ('demon slayer', 2145), ('jujutsu kaisen', 2074), ('boruto', 1888), ('black clover', 1695), ('bleach', 1681), ('one', 1382), ('shit', 1253), ('anime', 1169), ('think', 849), ('even', 848), ('hunter hunter', 838), ('know', 807), ('got', 799), ('people', 759), ('fight', 668), ('amp', 667), ('still', 663), ('bad', 640), ('time', 611), ('really', 609), ('hate', 605), ('character', 580), ('see', 576), ('say', 568), ('fuck', 560), ('make', 552), ('way', 543), ('death note', 538), ('manga', 524), ('literally', 498), ('thing', 493), ('ass', 486), ('good', 477), ('want', 474), ('show', 470), ('said', 455), ('need', 454), ('much', 454), ('blackclover twt', 448), ('never', 447), ('look', 425), ('serie', 419), ('going', 410), ('bro', 401), ('naruto shippuden', 384)]

As we can see, most of the common words are the anime names themselves, and what surprising is that a lot of profanities are present in that list, even more than the good positive words.

# Phase 4 (EDA):

The final phase of the project is the EDA (exploratory data analysis) in here we will take aspects from the data we have collected (mainly the anime’s individual data) and explore them to find some interesting insights.

As stated before, anime is mainy produced by Asian countries, but which countries contribute more. I thought that Japan will lead the way, then Korea behind them, then China. After counting the values and plotting the graphs the output is:



Chart, pie chart

Description automatically generated

It was expected that Japan would be leading the way, but I did not expect that huge difference. In addition, the absence of Korean based animes was a surprise as Korean mangas (original source material for anime) are quite popular.

Next we have types of content, animes can be classified into normal TV Series and Movies as normal, but other classifications exist, those are ONA, OVA, and Special. These new types are special extra episodes or a promotional episode for an event. I believed that TV series types will dominate the graph percentage, while all the special types will be very little. Here is the output for the graph:

Text

Description automatically generated with low confidence

Chart, pie chart

Description automatically generated

As expected, Series were the most popular but not by much as I thought it would be, and on the other hand, the special types did turn out to have a much more relevance than I thought.

Another interesting aspect to look at is the release season of the anime, I believed it it will be an equal divide but that wasn’t the case:

A picture containing chart

Description automatically generated

Chart, pie chart

Description automatically generated

I believed if any season had an advantage, it would be summer since there is no school thus the kids are free, so releasing an anime at that time is would be optimal from a business perspective. However, upon research about the Japanese school system, it unfolds that they have 3 semesters, and that it starts in April and end in March, basically having an all year around school negating the extra free time in the summer.

Now, we discussed the but season, but lets discuss the debut year:

Chart, histogram

Description automatically generated

As this graph indicates, there are animes that have debut back in 1911, but what’s more interesting is the exponential increase in anime productions as seen in the constant growth across each 10 year period.

And to deeper analyze the most heavy on production years, we will see all the years that had over 200 animes debut in its year:

Chart, bar chart

Description automatically generated

As shown, most of these years lie in the 2000-2020 era, with 2016 taking the crown as the year with the most animes to debut.

Now onto something more on the speculation side, is there a correlation between views and scores? Or number of reviewers and scores?

For the views and score:



For the number of reviewers and score:



The very low number for views and score indicate that there is probably no correlation between them. However, for the number of reviewers, it isn’t a necessarily high value to conclusively say anything, but there might be something worth researching more.

Finally, we are to give a count on the number of dubbed animes. A dubbed anime is one that has enbglish dubbing instead of the traditional Japanese voices with subtitles:



That means that out of the 13470 on the sight, 3644 are dubbed.

# Conclusion :

To conclude, getting to work on this project was a really rewarding experience, as it allowed me to learn new skills, languages, and concepts, while also learning about a field that I am really interested in and thinking about pursuing as a career. Not only that, but also it allowed me to dive into a subject that I have been interested in and learn more about it through experimentation of graphs and numbers and uncover more about that topic.