Streaming Media Analysis

Assignment 2

Rahul Kumar Gupta (s3635232)
Monika Vurigity (s3675394)
Megha Mohan (s3598762)

Table of Contents

Introduction 3
Methods 3
Data Collection 3
Data Pre-processing 3
Analysis 5
Sentiment Analysis 5
Word Distribution 6
Top Tweets by Retweet Count 8
Topic Modelling 10
Network Model 12
Limitation 13
Conclusion 14
References 14

Introduction

In the 21st century live streaming media has become one of the major sources of entertainment. We can observe the trend which is changing from the cable televisions to online streaming. This includes media, video and music streaming. This report analyses on which channel grabs most of the user's attention in the recent times. For this we used different analysis techniques as sentiment analysis, topic modelling and network analysis. The data is collected from twitter.com from May 12th to May 20th in the year 2018 by using some keywords. The data is then pre-processed in order to do the different analysis.

Edited by Monika

Methods

Data Collection

Data is collected from Twitter using Python API, Tweepy. Total 5000 tweets are collected in sequence manner. Collecting historical tweets are currently restricted from Twitter APIs. Maximum of last 7 days data can be collected.

Based on our study we have queries data for multiple keywords. Below are the generic keywords related to streaming services

- streaming *service*
- streaming *movies*
- streaming *series*
- streaming *media*

Tweets are present from 12-May-2018 to 20th May 2018. Out of 5000 tweets, only 1999 tweets are unique, most of the analysis are performed on unique sentiments.

Code are implemented in Jupyter Notebook, Python 3. Pandas, Matplotlib and seaborn are used for visualization purposes. Most of the code is using python in-built data structures such as lists and dictionaries.

Data Pre-processing

NLTK is a NLP library for python. It provides corpora and lexical resources with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

NLTK is used to pre-process data and create sentiments. It has well defined functions for tokenize, removing stop words, and regular expression cleaning. Sentiments are made using Opinion lexicon corpus, which contains a list of positive and negative words created over the time.

Topic modelling feature extraction is done using Count Vectorizer function [1]. Latent Dirichlet allocation (LDA) helped to create generative statistical model too explains why some parts of the data are similar.

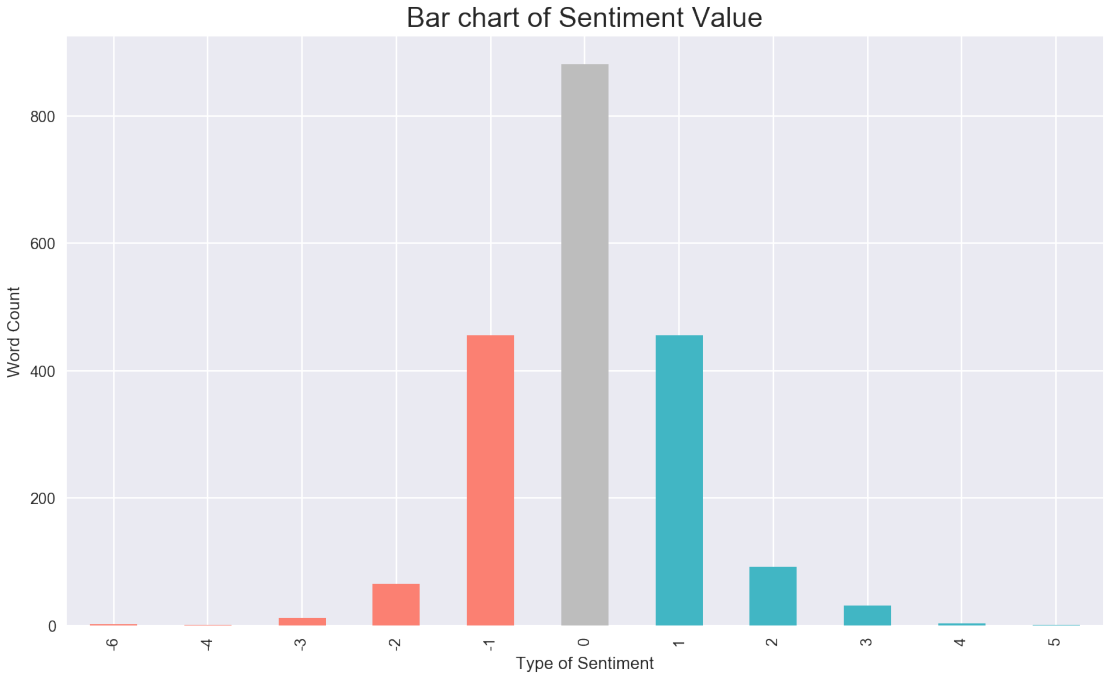
Analysis

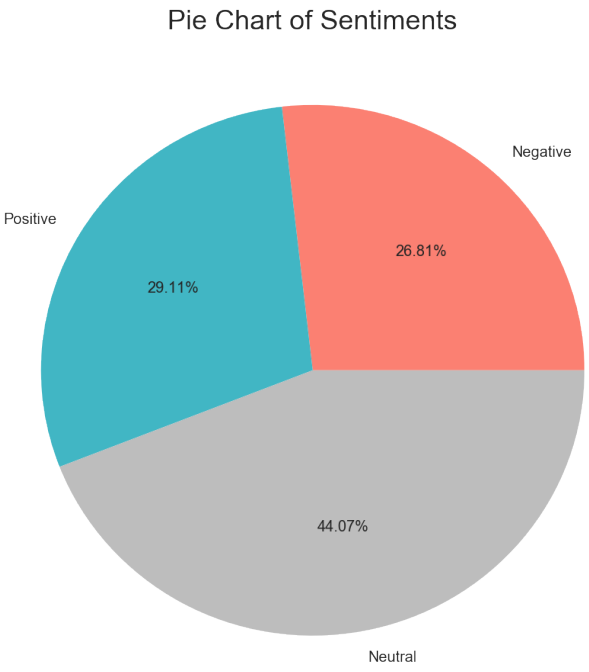
After gathering and pre-processing the tweets, sentiment analysis is performed to get general overview.

Sentiment Analysis

Below is the visualization of direct word match sentiments. Negative value in red colour indicated negative tweets count, 0 is neutral and green is positive sentiment count.

Distribution looks symmetry in our case.

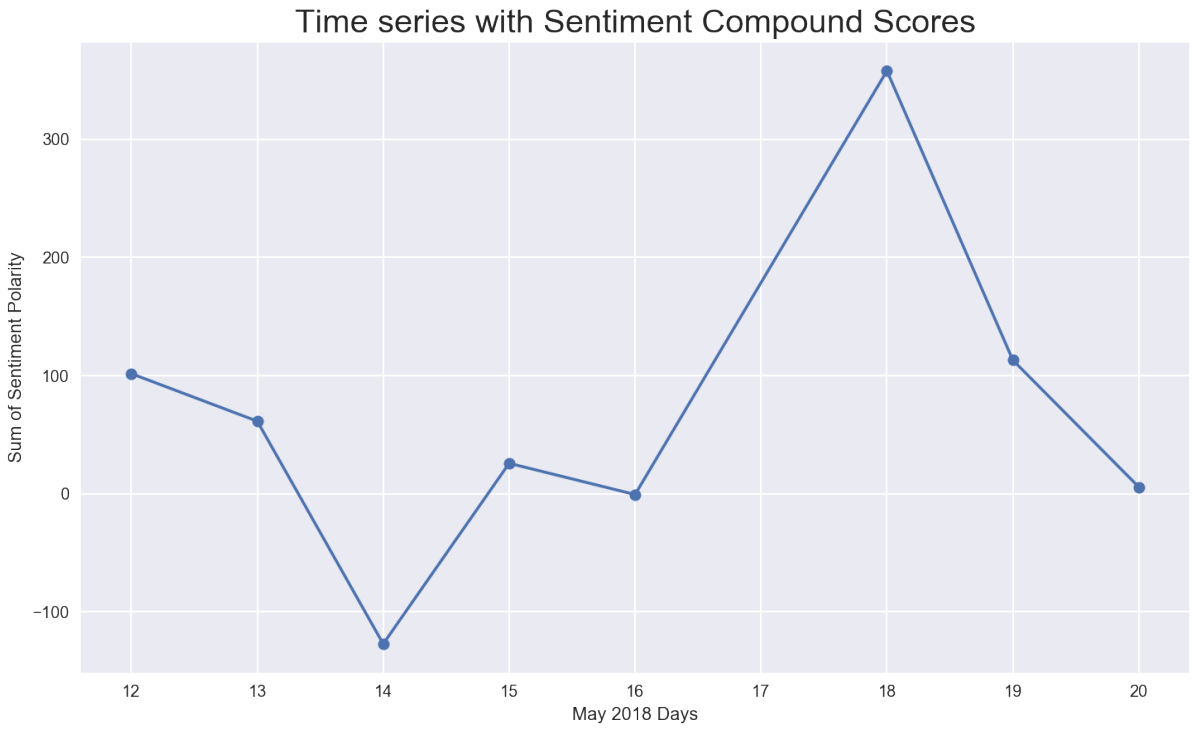




Pie chart shows that negative and positive tweets are almost equal. Almost 45% of the tweets are neutral.

We investigated sentiment polarity using NLTK sentiment analyser module to get the compound scores. Dates are group by and score are summed to create a time series plot to get trend.

The next visualization represents shows time series plot of compound polarity scores per day. We can notice that the variation in series is too high. On 14th May, most of the tweets were highly negative, on the other hand 18th May has high positive values. Also, the data is missing in 17th may, this is due to tweepy restrictions of 7 days and maximum of 3200 tweets in a single run.



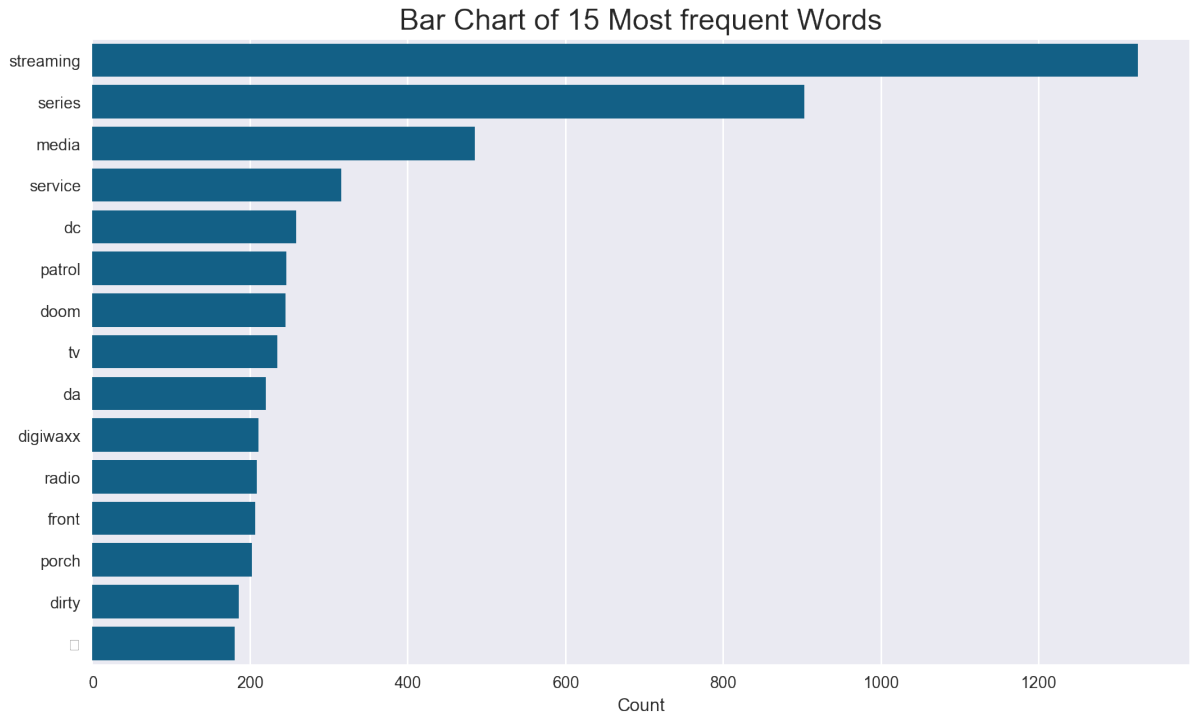
Next, we'll analyse the frequent words used in streaming media tweets. This will give us little picture of what kind of topics and keywords are used in the conversions.

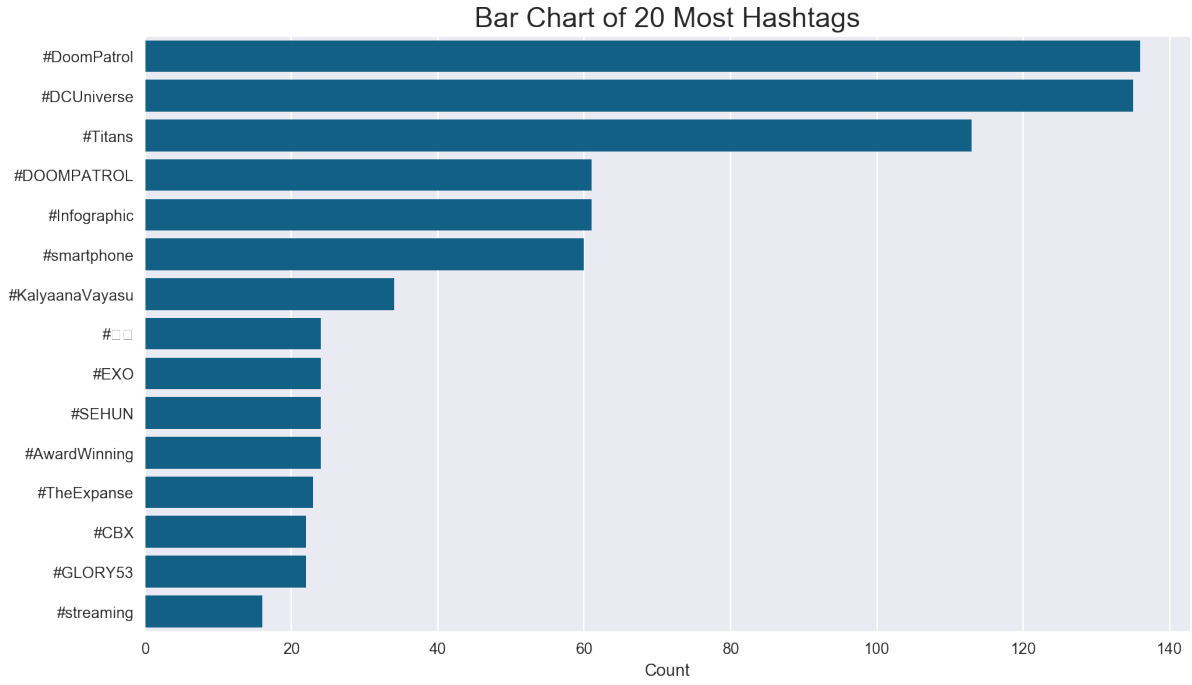
Word Distribution

Frequency distribution of word has streaming series as most popular term. Also, we noticed that there is huge marketing about new DC comics series streaming services. Below are the key findings from these count charts:

- Marketing by DC comics are quite high in dataset
- Hashtag DoomPatrol is quite high in counts. The Doom Patrol is a superhero from DC Comics
- kalyaana vayasu word has high values. On further analysis it was found that it's a local Indian song.
- Few mentions of Spotify and Youtube songs like #cbx

Over these keywords are not providing enough data and insights about streaming media. This is highly affected by viral marketing and daily influences.





Below are few highly positive tweets from our datasets

```
What are you doing now to ensure that your 2018 grantmaking will be eve
n more effective, interesting and rewarding… https://t.co/KVP4yQxmSJ

RT @JDBVoteUpdateTR: Billboard throwback 2011. He won six awards on Bil
lboard 2011.

-Top New Artist
-Top Streaming Artist
```

-Top Digital…

Huge congratulations to @VanessaKirby on her Best Supporting Actress wi
n at #BAFTAs2018 for @TheCrownNetflix 🏆🍾 Ver… https://t.co/eUNS0NJvay

RT @ManaByte: Disney is going to want a Marvel series on the streaming
service that will make Marvel fans want to pay for the service, like…
How are you a work in progress?

'Work In Progress,' our new original comedy series, now streaming for F
REE only on… https://t.co/qx827dt0JE

Here we observe that sentiment values are highly affected by individuals and marketing strategies. To get more robust values, we need to filter data using retweet counts and no of followers.

Top Tweets by Retweet Count

text	retweet_count
RT @netflix: .@Logic301 @Rapsody @2chainz @tip @artisthbtl @daveeast @justblaze @G_Eazy @nas and @Netflix Their Words. Their Way. Rapture,Ä¶	14155
RT @cwtvd: See Elena,Äôs return on the series finale of #TVD, streaming now on The CW App: https://t.co/hXUCSvJKWV https://t.co/6fs1SGxYKS	3499
RT @choi_bts2: K media reported the streaming site Spotify which has the most users advertises the Come back by @BTS_twt on their billboard,Ä¶	3292
RT @cwtvd: See one last battle with #TVD,Äôs greatest villain on the series finale, now streaming on The CW App: https://t.co/hXUCSvJKWV http,Ä¶	2700
RT @exo_schedules: DAILY SCHEDULE □üói 180510 | 10th May | #Ïóë ÏÜå #EXO #CBX Schedule □üíª Netflix series [#SEHUN] □üé§ Concert + Meet & Greet [#CBX_Ma,Ä¶	1701
RT @netflix: .@Logic301 @Rapsody @2chainz @tip @artisthbtl @daveeast @justblaze @G_Eazy & @nas Their Words. Their Way. @RaptureNetflix, an,Ä¶	1611

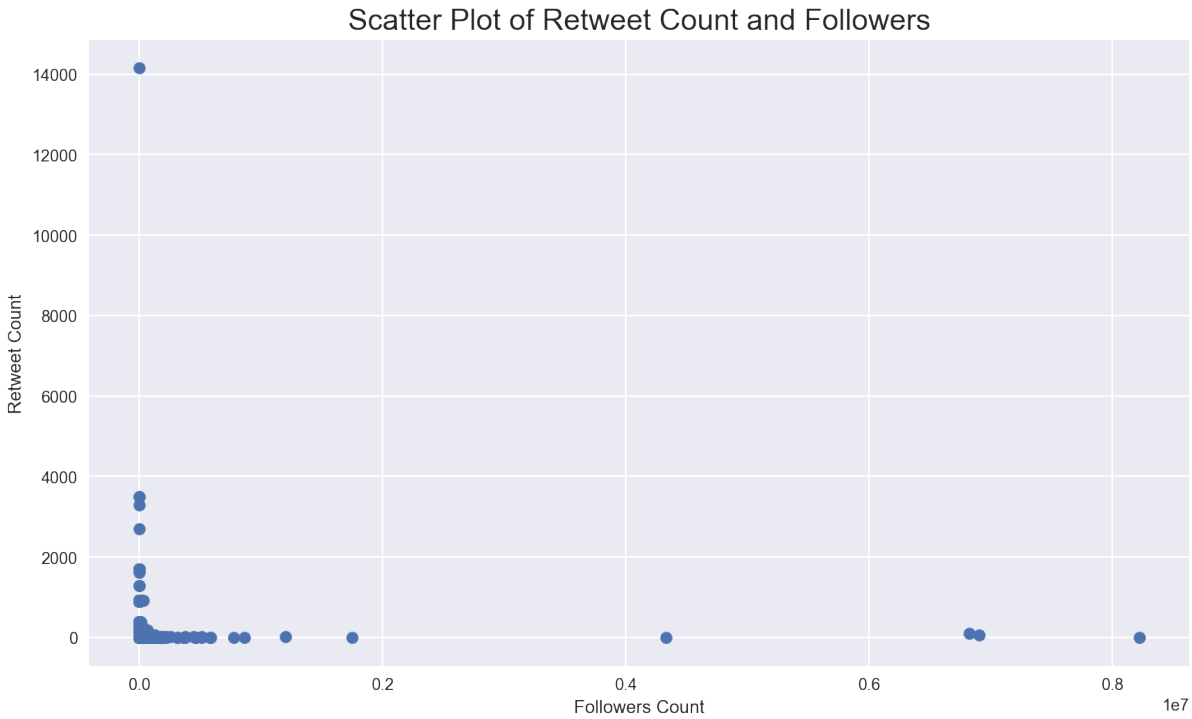
RT @exo_schedules: DAILY SCHEDULE □üói 180504 | 4th May | #Ïóë ÏÜå #EXO Schedule □üéâ Youth Day festival [#LAY] □üíª Netflix series [#SEHUN] □üì∫ TV Appear,Ä¶	1290
RT @shadow_twts: Armys while you are streaming Fake Love in Spotify, click the share button and share Fake Love's Spotify link on social me,Ä¶	931
Armys while you are streaming Fake Love in Spotify, click the share button and share Fake Love's Spotify link on so,Ä¶ https://t.co/mv5rsOvwNg	931
RT @shadow_twts: Friendly Reminders: - When the MV drops, stream only on ibighit. Turn on CC - When streaming on Spotify, share it on so,Ä¶	906

Results sorted by retweet count gives the following information

- Netflix is the biggest source of online streaming channel
- Spotify for music is quite popular than YouTube red, which is quite new in markets
- Some other sources of entertainment channel are also popular. Tweet by @exo_schedules Youth day festival and Netflix series has 900+ retweets
- Talks about regional channels like Big Hit Entertainment

Results with number of likes are also similar. To get the influential tweets we need to look text of people with high number of followers. Before beginning we noticed that there is no significant correlation between *retweet_count* and *follower count*.

The next visualization between *retweet_count* and *follower_count* is highly affected by outliers. We can also notice that having high number of followers doesn't give high number of retweets or likes. Thus, analysing tweets with high following count will not give the exact trend. Data can be further analysed using log transformation.



Topic Modelling

Topic modelling is done with NLTK library packages. Latent Dirichlet Allocation method with 10 components and online method is applied on tweets where retweet_count is greater than one. Number of tweets reduced to 691 using these filters. This will make sure to get more accurate results. The next visualization shows the results from top 10 extracted topics.

Some of the key finding from topic Modelling

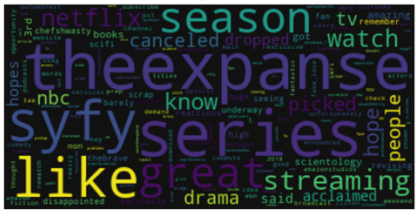
- TV series are most popular in online streaming. Netflix and Hulu are the biggest players
- Users tweets about next episode of the series a lot
- Few users (Techies) talk about technologies related to internet streaming. Like video adoption rate, download, upload, 5K televisions,
- Music Streaming, Youtube and radio are also part of streaming.
- Data is highly volatile and affected with events like grand prix, DC comics, doom superhero

Results from filtered topic modelling has given few insights about online streaming.

Topic 1:

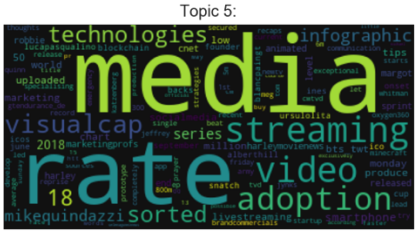
Topic 2:

Topic 3:



Topic 4:

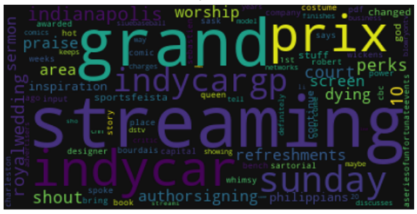
Topic 5:



Topic 6:

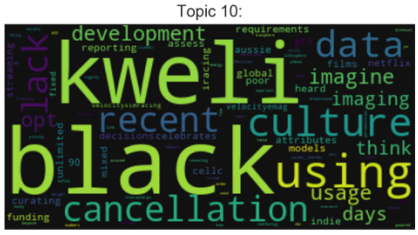
Topic 7:

Topic 8:

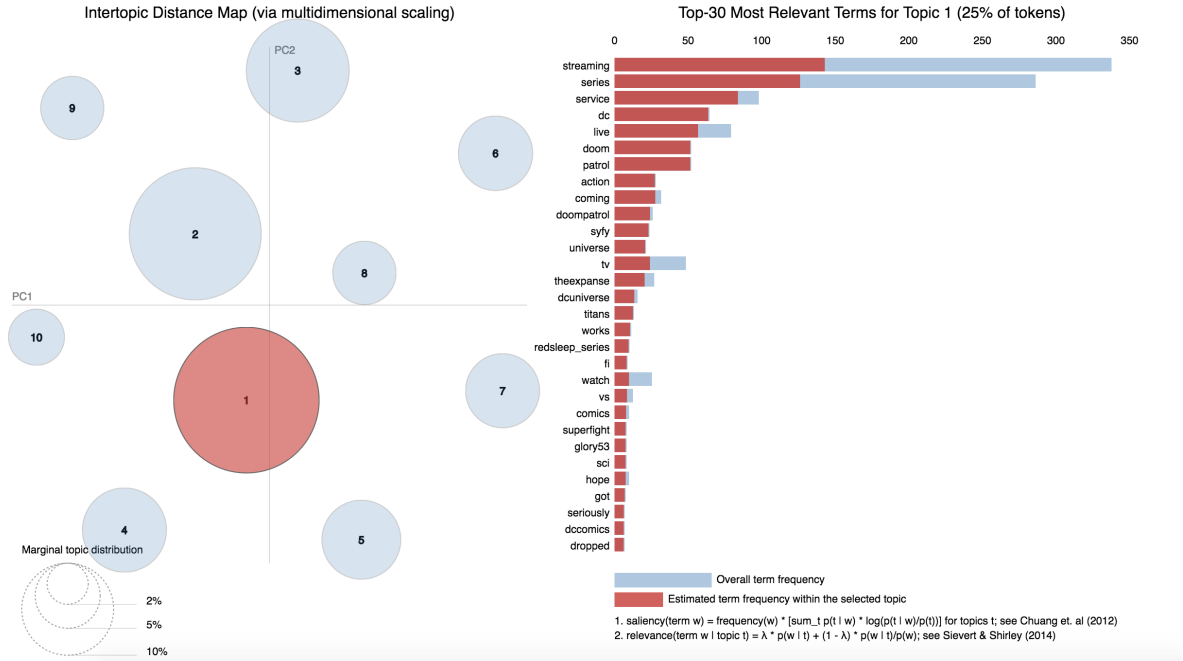


Topic 9:

Topic 10:



We also made LDA visualization on python web server. The key findings are almost similar to word cloud.



Network Model

From the results of sentiment analysis and topic modelling, we saw two major trends

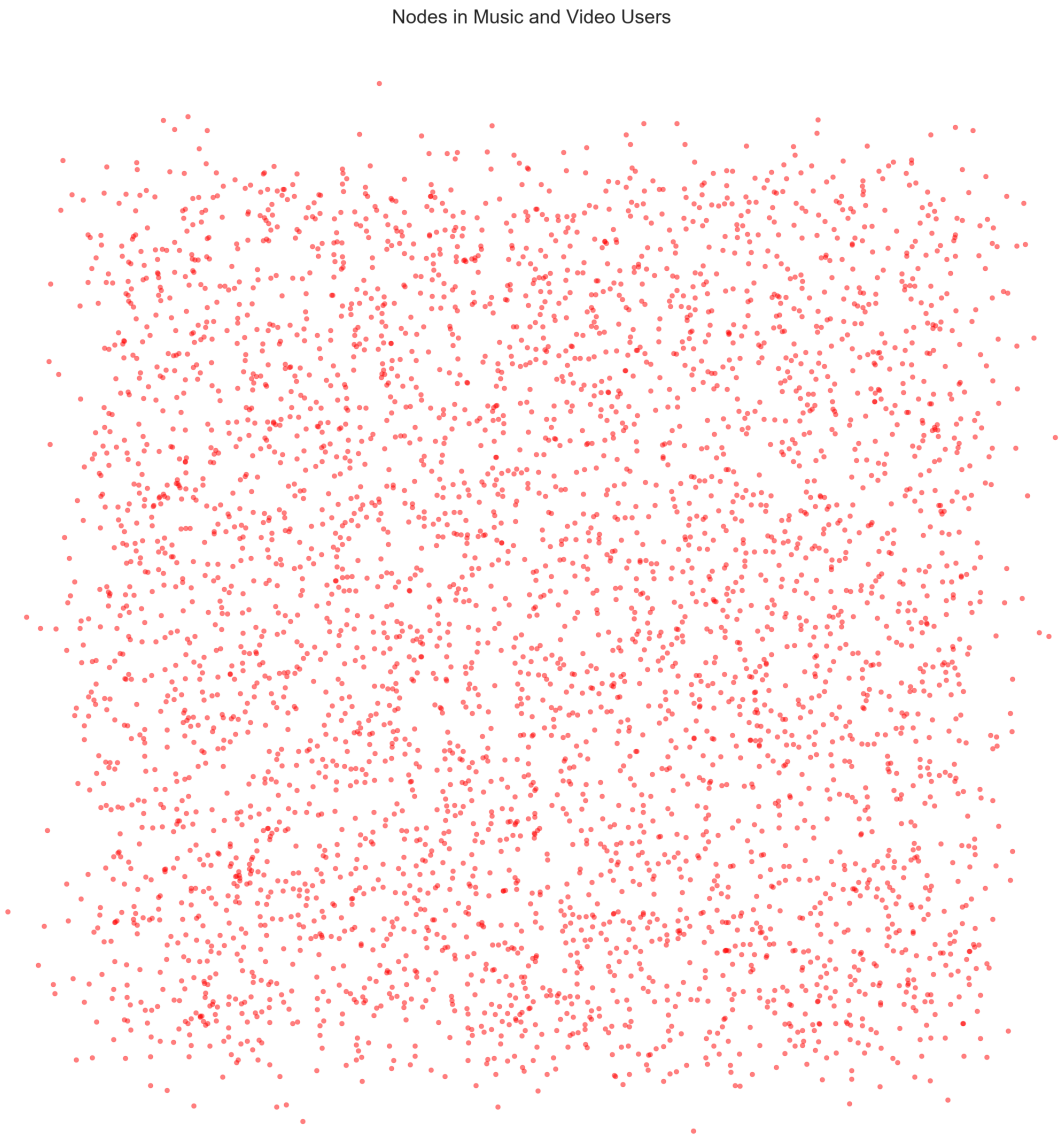
- Netflix and Hulu as video streaming
- Spotify and YouTube as music streaming

Next we extracted 2500 for both the streams and created network model of unique users

In video streaming, we found that 49 users are connected with each other. Below are the trends in video streaming

In Music streaming, we found that 10 users are connected with each other

Comparing both with each other to create community, we found only one user to be connected. Networkx package [2] is used to create graph



Limitation

- The data is collected from one source only. More unstructured data can be collected from other social networking sites to get more insights
- Results are biased as only 9 days of tweets are collected. This is highly affected with noise like viral marketing
- More analysis could be performed like ngram, network modelling which is not performed in this report
- Structured data can be used to support the results from sentiment analysis

Conclusion

Social media analysis is quite powerful to get trends. Validity of this study is quite small and accurate for short period of time. Here in online streaming, we found about, Netflix, Hulu, Spotify and Youtube are most popular ones. Users are mostly not connected from video and music streaming channels. Most of the topics are related with recent events.

Overall, sentiment analysis and topic modelling can give an overall glance of trends happening over social media. Network modelling helped to get the community information between multiple channels.

References

[1] *feature Extraction* http://scikit-learn.org/stable/modules/feature_extraction.html

[2] *Networkx* https://networkx.github.io/documentation/stable/