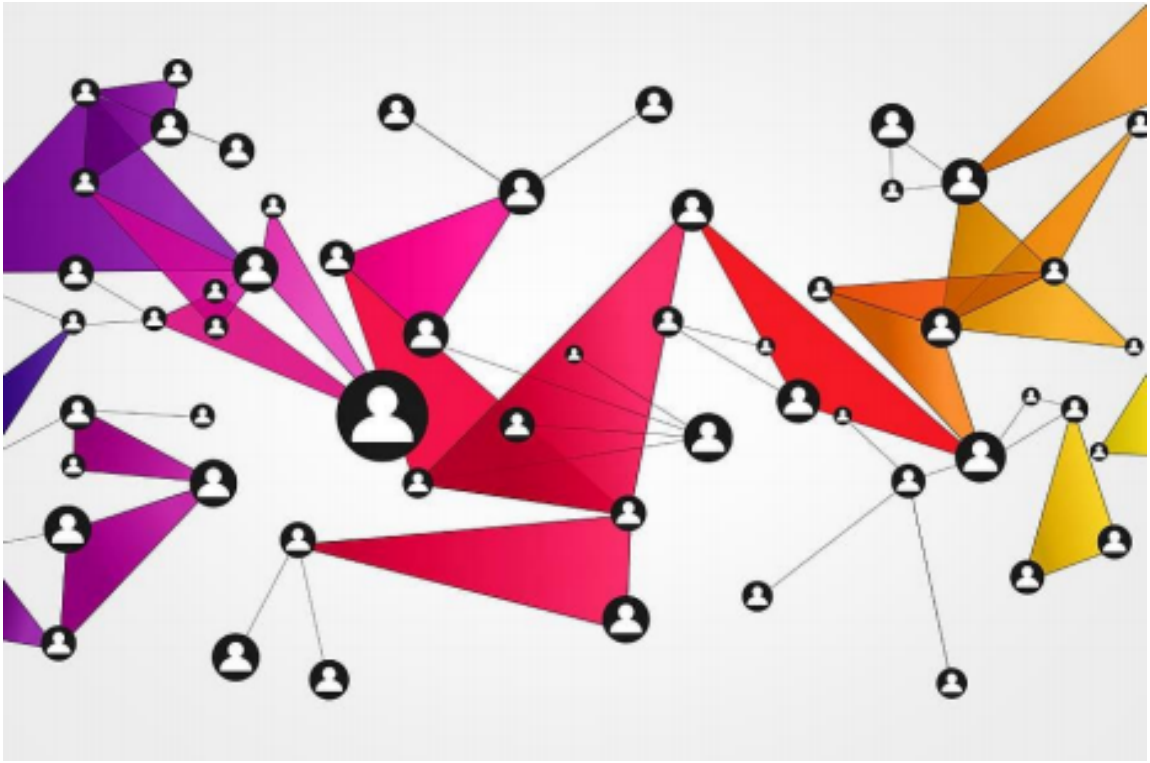
Group Project

19OH01 - Social and Economic Network Analysis



Visualise and Analyse Hashtags - IPL Auction 2021

1. Problem Statement

In this project, the tweets related to the most sensational IPL Auction that happened during February 2021 are gathered and the hashtags and accounts mentioned in the tweets are extracted. The most trending hashtags and mentioned accounts are filtered and visualised as graphs and word cloud using python libraries. Then the similar hashtags are clustered together and communities are detected.

- Identify the hashtags related to IPL Auction 2021 (Like #IPLAuction #CSK #Dhoni #MI #Gayle etc.)
- Scrap the tweets from twitter using twint
- Identify the counts of hashtags and mentioned accounts and display it as graph and word cloud using seaborn and wordcloud libraries
- Identify communities or clusters within hashtags using networkx and Gephi and interpret its meaningfulness.

2. Dataset description

a. The dataset has a total of 72617 tweets (using TWINT library)
b. IPL Auction 2021 was held on 18th February 2021. So, tweets were collected from *17th February to 19th February*
c. TWINT library didn't require any API credentials. It takes input search string and provides dataset with attributes TWEETID, DATE, TIME, TWEET_TEXT
d. Advanced search filters were used to filter tweets by time and to include multiple keywords(hashtag) in search.
e. Eg. SEARCH STRING of the form `(#CSI OR #MI OR #KKR OR #DC OR #RCB OR #KXIP OR #SRH OR #RR) until:2021-02-19 since:2021-02-17`
f. Syntax for scrapping (Clone the git repository)

```
import twint
# Configure
c = twint.Config()
c.Search = "(#CSI OR #MI OR #KKR OR #DC OR #RCB OR #KXIP OR #SRH OR #RR) until:2021-02-19 since:2021-02-17"
# Run
twint.run.Search(c)
```

g. Link for Dataset:
   i. Twitter data - Text files and json

3. Tools used

a. **TWINT** library was used.

   Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles *without using Twitter's API*.

b. Libraries **Seaborn, Wordcloud**

   Visualise the frequent hashtags as graph and wordcloud.

c. **Networkx and Gephi**

   Louvain algorithm in **Networkx** and visualise clusters in **Gephi.**

4. Challenges Faced

a. *Twitter API* was not able to provide data older than 7 days in detail. So, the TWINT tool was used.
b. The *TWINT tool* didn't install properly in windows through the *pip* command. So, the git repository for TWINT was cloned in the project directory in linux.
c. *Search string had an upper limit* of the number of characters. So, the available hashtags were split as 5 different search strings and later duplicates were removed using tweet id.
d. *Wordcloud* masks led to crowded and unclear images. Cricket related masks were not giving better images. So, India map image was used as MASK
e. *Louvain Algorithm* for clustering gave 5 clusters. But only 2 were interpretable, remaining 3 clusters had hashtags from many teams and many players without much relation between them.
f. *Louvain Algorithm* gave different clusters when executed again.

5. Contribution of Team Members

Roll No.	**Name**	**Contribution**
18Z312	Deepthishree GS	Helped scrap the tweets from text files to readable by Python, remove duplicates using Tweet ID and preprocessed it
18Z323	Iswarya GP	Used Twint tool to scrap the tweets in Linux and handled the errors which came with it. Also helped in documentation
18Z324	Janani R	Explored twitter API and found the advanced search settings to write down the search strings to be used with twint. Also helped in documentation
18Z326	Jeffrey Bianca	Generated bar graphs and word clouds for the processed tweet data. Explored possible masks for word cloud.
18Z360	Swetha M	Louvain algorithm for the graph generated using Networkx and visualising it using Gephi.

6. Annexure I: Code

GITHUB REPO with code, output and reports

7. Annexure II: Snapshots of the Output

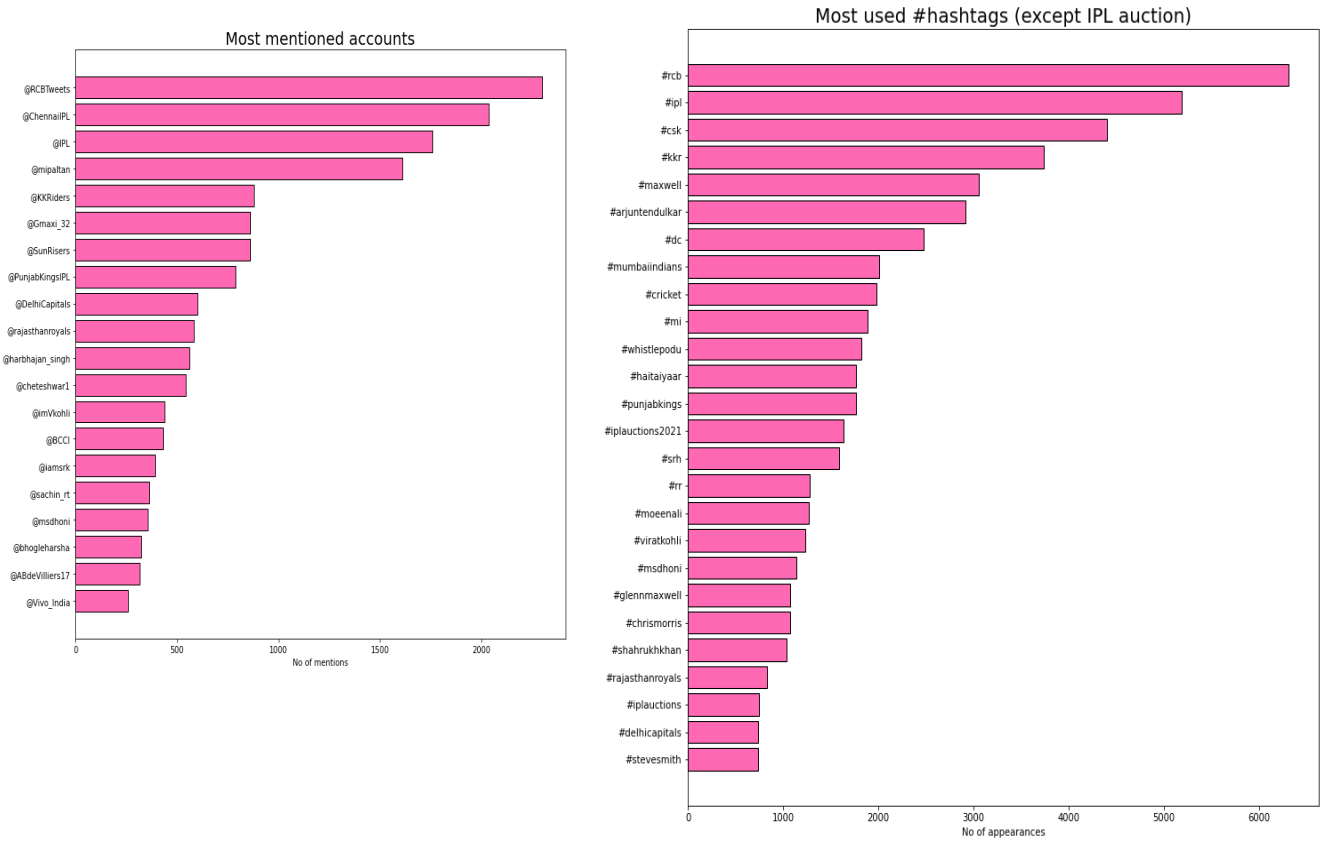


Fig. 1 Barcharts depicting the most used hashtags and mentions

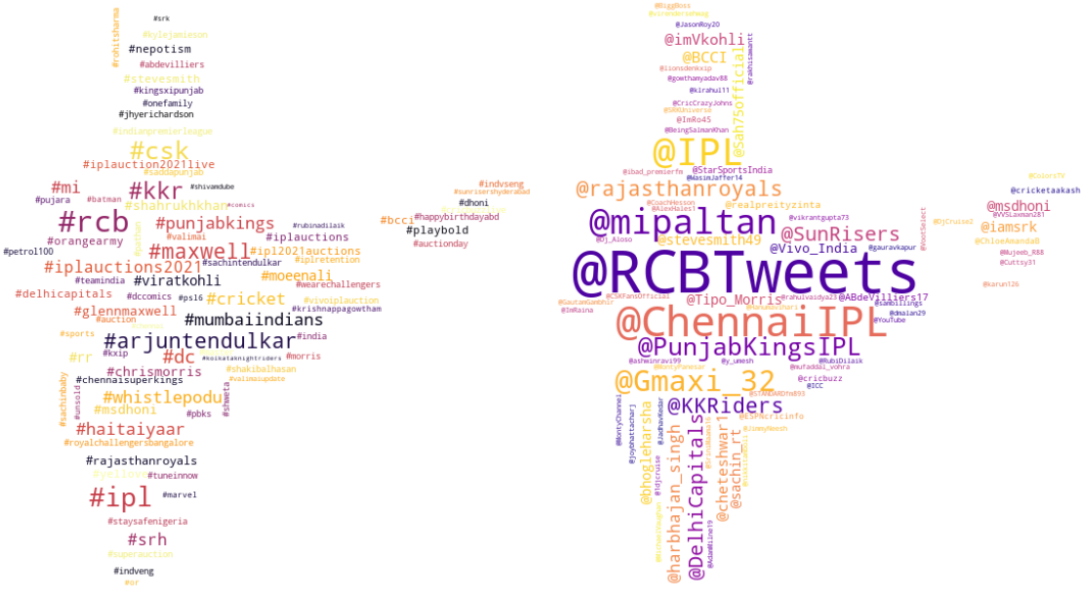


Fig. 2 Wordcloud representation of most used hashtags and mentions

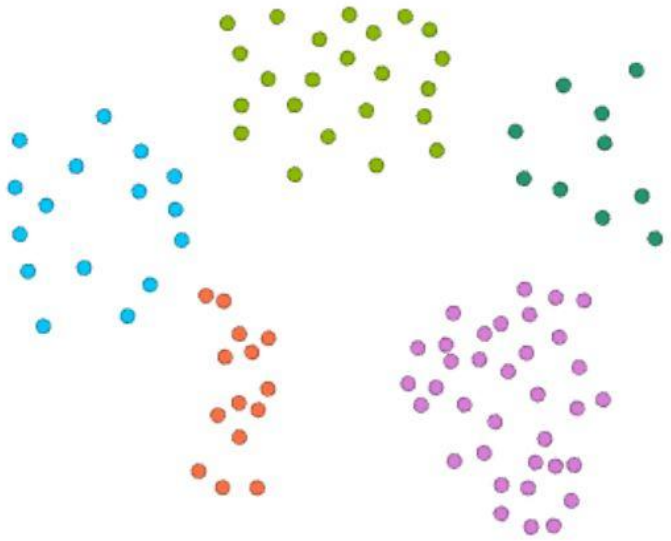


Fig. 3 Cluster of hashtags according to Louvain algorithm

Fig. 4 One cluster with CSK related tags (green)

Fig. 5 One cluster with MI related tags (blue)

8. References

[1] Community detection for NetworkX's documentation — Community detection for NetworkX 2 documentation

[2] Communities — NetworkX 2.6rc1.dev0 documentation

[3] Visualisation of Information from Raw Twitter Data — Part 1 | by Jaime Zornoza | Towards Data Science

[4] TWINT for extracting tweets

[5] (#IPLAuction2021 OR #IPL2021 OR #CSK OR #MI OR #IPLAuctions OR #RCB) until:2021-02-19 since:2021-02-17 - Twitter Search / Twitter

[6] Make a simple Wordcloud

[7] Generate wordcloud

[8] Python - Wordcloud official documentation

[9] Issue in date parameters - Stack Overflow

[10] AttributeError: module 'twint' has no attribute 'Config' #92