Twitter Web Crawler Report

Martin Nolan (2313202N)

Source Code: https://github.com/mnolan99/2313202nWebScienceAssignment

CSV Data:
https://github.com/mnolan99/2313202nWebScienceAssignment/blob/master/Football%20Database.csv

Introduction

As part of this assessed exercise, I was tasked with the challenge of developing a data crawler to complete network based social media analytics. This network crawler makes use of the Twitter Streaming API to collect data from tweets in real time. Once there was enough data collected, I stored it in a database and clustered together similar information such as hashtags, usernames and keywords. I then used this information to conduct network analytics across all the information stored in the database.

To start with, this involved me creating a Twitter Developer account so that I could have access to the Twitter API. Once I was granted developer privileges, I was given access keys and consumer API tokens so that I could develop my app locally on my laptop and access the information which was being tweeted at that time. Before I began any development, I had to install the mongoDB Community Server onto my computer and install pymongo within my development environment.

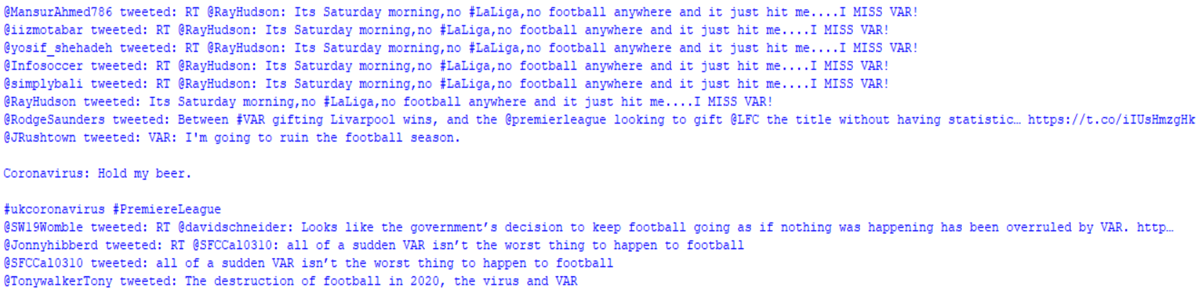
With this being the first time that I had created an app as a Twitter developer, I researched online using Stackoverflow, Github and read through the Twitter API [1] documentation to fully understand what the Twitter Streaming API is used for.

I used Visual Studio Code created a python script which makes use of the Tweepy [2] library to develop my solution to this assessment. The Tweepy API allowed me to access all of Twitter's RESTful API methods which meant that I could build a very efficient Twitter crawler. Firstly, I looked at 'hot' trending topics on Twitter to see what issues I could gather the most information about. I decided to pull tweets containing the keywords 'football', '#Football', 'VAR' and '#VAR' as it is a controversial topic that I am very interested in within the footballing environment.

Data Crawling

The data collected using the streaming API was uploaded to Twitter on 02/03/2020 between 21:13 and 21:48. This data was then stored within the football collection in my database using MongoDB. As we were to collect around 1% of tweets, I decided to only run my data crawler for around 35 minutes whilst many football games were being shown on television; this allowed me to collect information from 7007 tweets (including retweets).

There were many different APIs I used to collect data from the live tweets. **Firstly,** I used Twitter's standard search API [3] and the TwitterSearch library [4] which returns a collection of relevant tweets which match a pre-specified query (the keywords of 'Football' and 'VAR' in my case). I used this API to start with as it allowed me to very quickly and easily collect many tweets with basic information such as the author's username and the text of their tweet; this was then printed out onto the console.

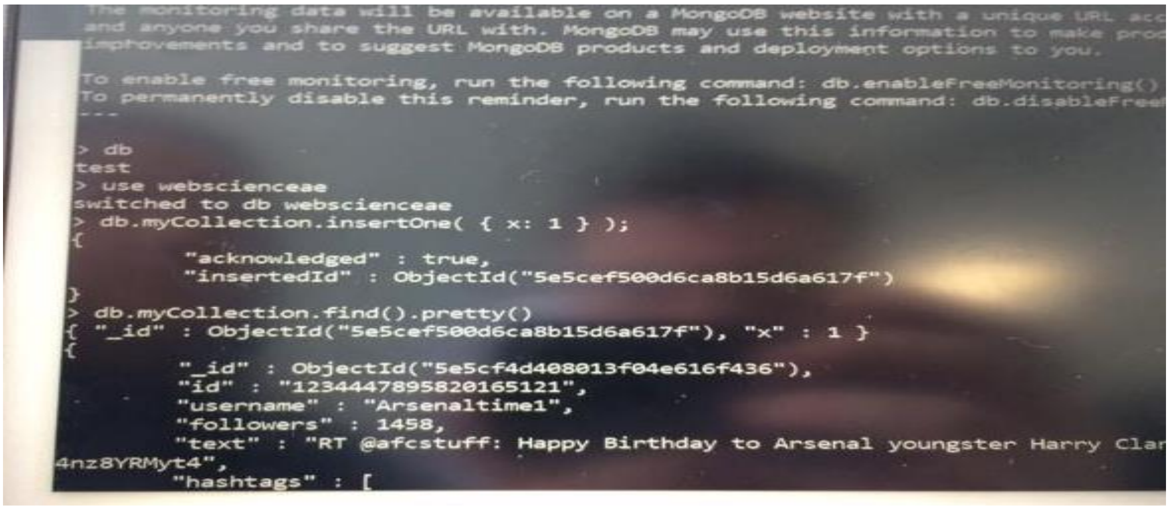


```
@MansurAhmed786 tweeted: RT @RayHudson: Its Saturday morning,no #LaLiga,no football anywhere and it just hit me....I MISS VAR!
@iizmotabar tweeted: RT @RayHudson: Its Saturday morning,no #LaLiga,no football anywhere and it just hit me....I MISS VAR!
@yosif_shehadeh tweeted: RT @RayHudson: Its Saturday morning,no #LaLiga,no football anywhere and it just hit me....I MISS VAR!
@Infosoccer tweeted: RT @RayHudson: Its Saturday morning,no #LaLiga,no football anywhere and it just hit me....I MISS VAR!
@simplybali tweeted: RT @RayHudson: Its Saturday morning,no #LaLiga,no football anywhere and it just hit me....I MISS VAR!
@RayHudson tweeted: Its Saturday morning,no #LaLiga,no football anywhere and it just hit me....I MISS VAR!
@RodgeSaunders tweeted: Between #VAR gifting Livarpool wins, and the @premierleague looking to gift @LFC the title without having statistic… https://t.co/iIUsHmzgHk
@JRushtown tweeted: VAR: I'm going to ruin the football season.

Coronavirus: Hold my beer.

#ukcoronavirus #PremiereLeague
@SW19Womble tweeted: RT @davidschneider: Looks like the government’s decision to keep football going as if nothing was happening has been overruled by VAR. http…
@Jonnyhibberd tweeted: RT @SFCCal0310: all of a sudden VAR isn’t the worst thing to happen to football
@SFCCal0310 tweeted: all of a sudden VAR isn’t the worst thing to happen to football
@TonywalkerTony tweeted: The destruction of football in 2020, the virus and VAR
```

Although the search API was very efficient at crawling twitter to find the tweets, many of the information provided was not valuable as there was a lot of duplicate tweets due to many users retweeting the same tweet. This caused me to me enhance my data crawling method by making use of the hybrid architecture of Twitter’s streaming and REST APIs. For this, I read through a Twitter crawler project on GitHub [5] and used the Tweepy library to help collect the data streams and filter the data whilst continuing to have integrity with OAuthentication.

Firstly, I had to set up my connection to my “WebScienceAE” database using MongoDB through the localhost. I then created a collection and decided to crawl Twitter for the specific keywords relating to football and filtered it so that it would only search for tweets that were written in English. Each tweet that was found was then added into my database collection with the relevant metadata in JSON format.

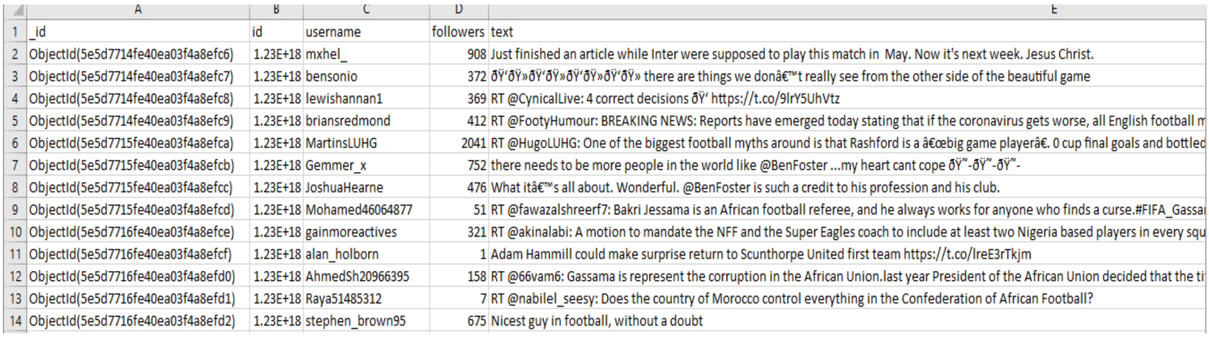
The information stored in the database included the unique tweet ID; the author’s username; the number of followers the author had; the timestamp of the tweet; the language the tweet was written in and finally the content of the tweet, including hashtags. This information was printed to the console and stored in the database as a single variable. I believed that using the Tweepy library to enhance my program would be best as it makes use of Twitter’s streaming API to collect a vast number of tweets. This is done very efficiently as Tweepy handles the authentication of users whilst connecting to Twitter’s database to read and collect the metadata from the tweets in real time.



```
The monitoring data will be available on a MongoDB website with a unique URL acc
and anyone you share the URL with. MongoDB may use this information to make prod
improvements and to suggest MongoDB products and deployment options to you.

To enable free monitoring, run the following command: db.enableFreeMonitoring()
To permanently disable this reminder, run the following command: db.disableFree
---

> db
test
> use webscienceae
switched to db webscienceae
> db.myCollection.insertOne( { x: 1 } );
{
        "acknowledged" : true,
        "insertedId" : ObjectId("5e5cef500d6ca8b15d6a617f")
}
> db.myCollection.find().pretty()
{ "_id" : ObjectId("5e5cef500d6ca8b15d6a617f"), "x" : 1 }
{
        "_id" : ObjectId("5e5cf4d408013f04e616f436"),
        "id" : "1234447895820165121",
        "username" : "Arsenaltime1",
        "followers" : 1458,
        "text" : "RT @afcstuff: Happy Birthday to Arsenal youngster Harry Clar
4nz8YRMyt4",
        "hashtags" : [
```

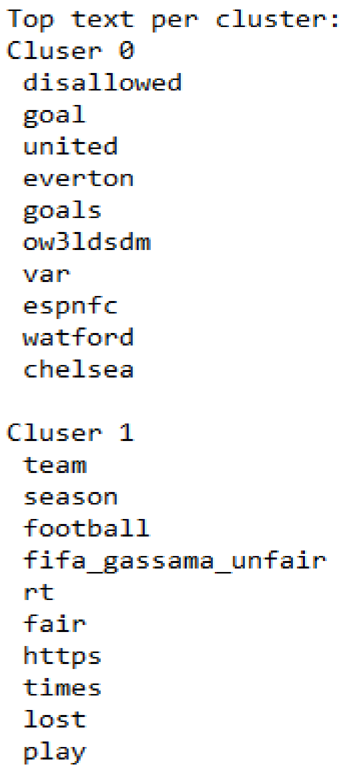
Context Analysis & Tweet Groupings

Once I had collected all the tweet metadata into my database, I exported all the information to an excel spreadsheet so that I could see the data in tabular form and group similar information together. A screenshot of some of this information can be found below.

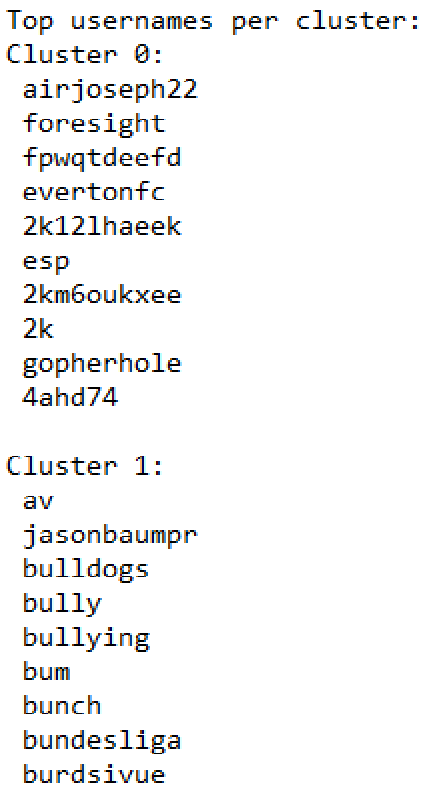


	A	B	C	D	E
1	_id	id	username	followers	text
2	ObjectId(5e5d7714fe40ea03f4a8efc6)	1.23E+18	mxhel_	908	Just finished an article while Inter were supposed to play this match in May. Now it's next week. Jesus Christ.
3	ObjectId(5e5d7714fe40ea03f4a8efc7)	1.23E+18	bensonio	372	ðŸ'ðŸ»ðŸ'ðŸ»ðŸ'ðŸ»ðŸ'ðŸ» there are things we donâ€™t really see from the other side of the beautiful game
4	ObjectId(5e5d7714fe40ea03f4a8efc8)	1.23E+18	lewishannan1	369	RT @CynicalLive: 4 correct decisions ðŸ' https://t.co/9lrY5UhVtz
5	ObjectId(5e5d7714fe40ea03f4a8efc9)	1.23E+18	briansredmond	412	RT @FootyHumour: BREAKING NEWS: Reports have emerged today stating that if the coronavirus gets worse, all English football m
6	ObjectId(5e5d7715fe40ea03f4a8efca)	1.23E+18	MartinsLUHG	2041	RT @HugoLUHG: One of the biggest football myths around is that Rashford is a â€œbig game playerâ€. 0 cup final goals and bottled
7	ObjectId(5e5d7715fe40ea03f4a8efcb)	1.23E+18	Gemmer_x	752	there needs to be more people in the world like @BenFoster ...my heart cant cope ðŸ˜-ðŸ˜-ðŸ˜-
8	ObjectId(5e5d7715fe40ea03f4a8efcc)	1.23E+18	JoshuaHearne	476	What itâ€™s all about. Wonderful. @BenFoster is such a credit to his profession and his club.
9	ObjectId(5e5d7715fe40ea03f4a8efcd)	1.23E+18	Mohamed46064877	51	RT @fawazalshreerf7: Bakri Jessama is an African football referee, and he always works for anyone who finds a curse.#FIFA_Gassa
10	ObjectId(5e5d7716fe40ea03f4a8efce)	1.23E+18	gainmoreactives	321	RT @akinalabi: A motion to mandate the NFF and the Super Eagles coach to include at least two Nigeria based players in every squ
11	ObjectId(5e5d7716fe40ea03f4a8efcf)	1.23E+18	alan_holborn	1	Adam Hammill could make surprise return to Scunthorpe United first team https://t.co/lreE3rTkjm
12	ObjectId(5e5d7716fe40ea03f4a8efd0)	1.23E+18	AhmedSh20966395	158	RT @66vam6: Gassama is represent the corruption in the African Union.last year President of the African Union decided that the ti
13	ObjectId(5e5d7716fe40ea03f4a8efd1)	1.23E+18	Raya51485312	7	RT @nabilel_seesy: Does the country of Morocco control everything in the Confederation of African Football?
14	ObjectId(5e5d7716fe40ea03f4a8efd2)	1.23E+18	stephen_brown95	675	Nicest guy in football, without a doubt

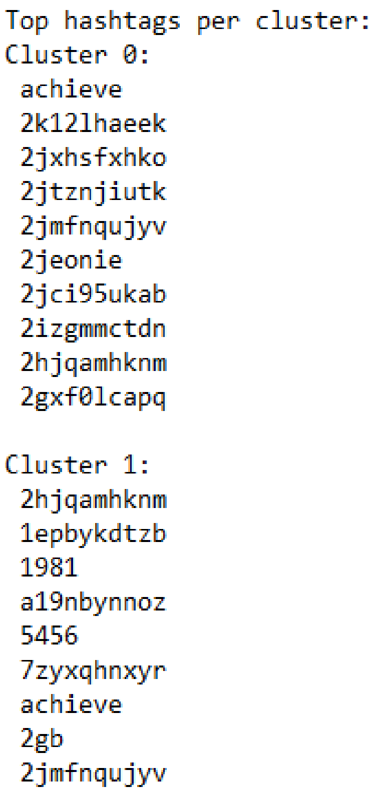
As suggested in the specification, I tried to use off-the shelf software to cluster the data but there was no software available which could import the data from a CSV file so I decided to create a python script to do this. This program allowed me to read in my CSV file and extract the author's username, the content of the tweet and any hashtag that was included in the tweet. Once all data had been imported, I made use of the sklearn library [6] to perform KMeans clustering on the dataset. This vector quantization allowed me to produce numerous groups of hashtags, usernames and text so that I could analyse it. I decided to write the findings out to a file which can be seen below.



```
Top text per cluster:
Cluser 0
 disallowed
 goal
 united
 everton
 goals
 ow3ldsdm
 var
 espnfc
 watford
 chelsea

Cluser 1
 team
 season
 football
 fifa_gassama_unfair
 rt
 fair
 https
 times
 lost
 play
```

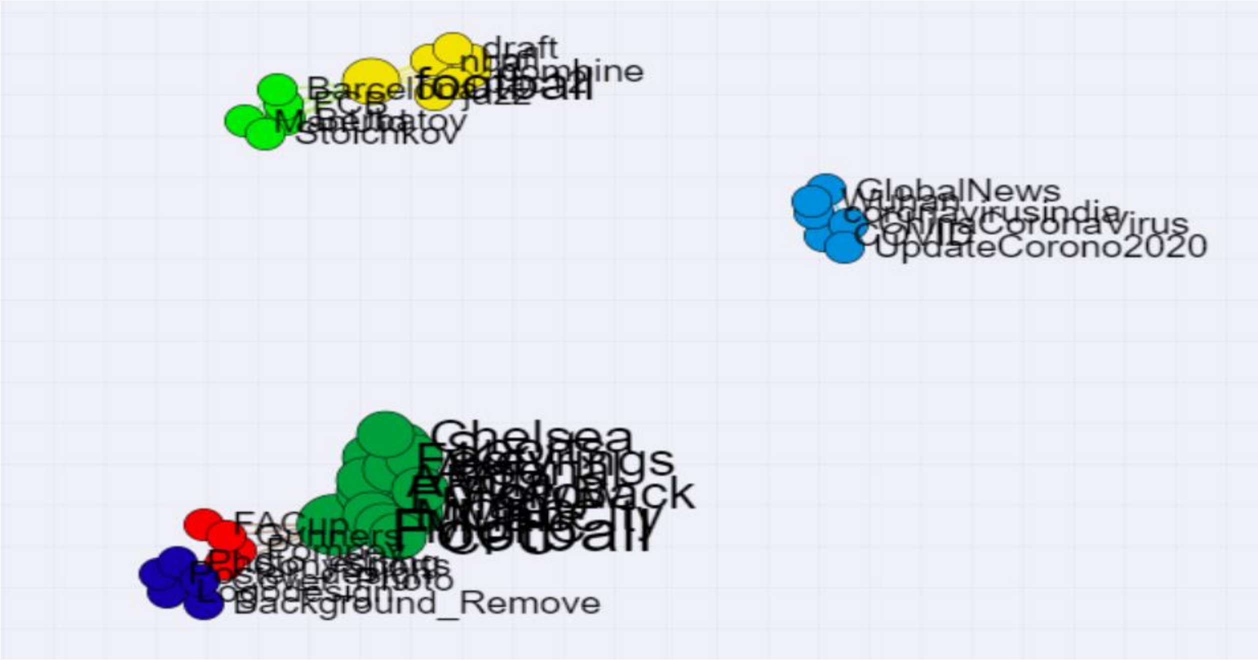


```
Top usernames per cluster:
Cluster 0:
 airjoseph22
 foresight
 fpwqtdeefd
 evertonfc
 2k12lhaeek
 esp
 2km6oukxee
 2k
 gopherhole
 4ahd74

Cluster 1:
 av
 jasonbaumpr
 bulldogs
 bully
 bullying
 bum
 bunch
 bundesliga
 burdsivue
```

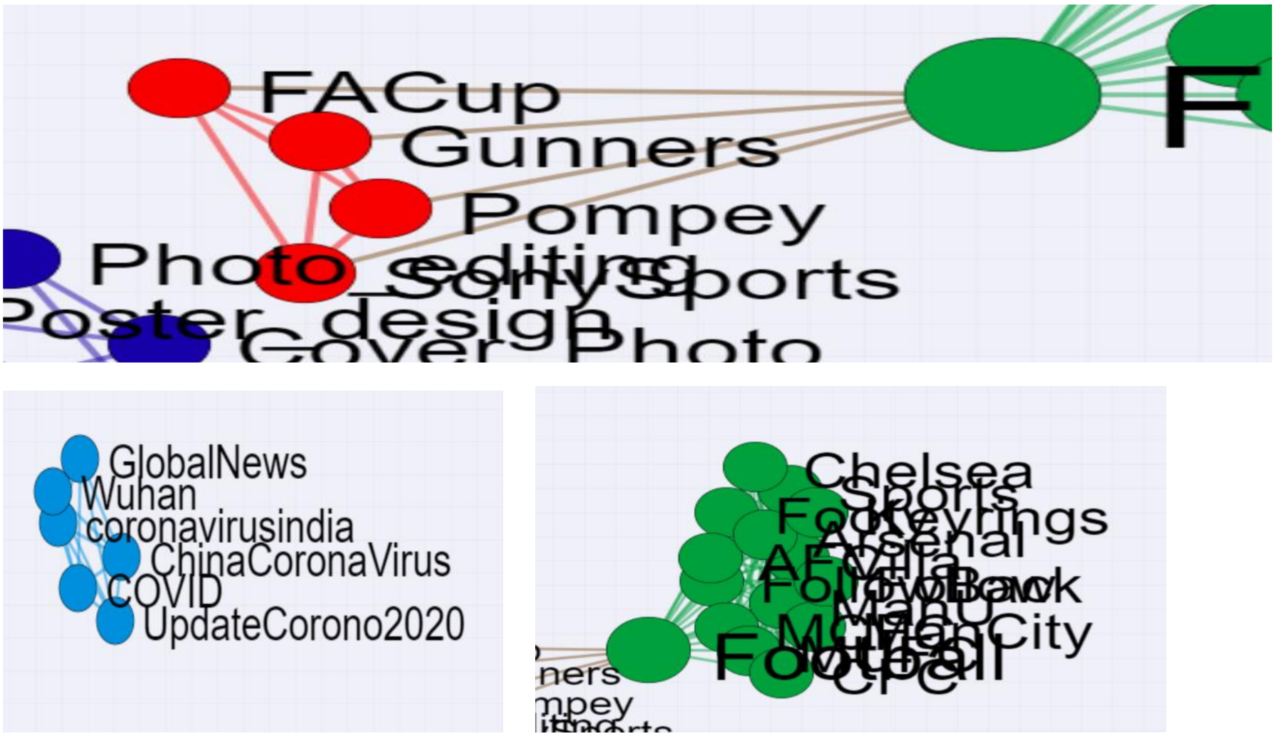


```
Top hashtags per cluster:
Cluster 0:
 achieve
 2k12lhaeek
 2jxhsfxhko
 2jtznjiutk
 2jmfnqujyv
 2jeonie
 2jci95ukab
 2izgmmctdn
 2hjqamhknm
 2gxf0lcapq

Cluster 1:
 2hjqamhknm
 1epbykdtzb
 1981
 a19nbynnoz
 5456
 7zyxqhnxyr
 achieve
 2gb
 2jmfnqujyv
```

In order to visualise the cluster information, I used a cortext manager [7] to create a new corpus which allowed me to analyse the network map of user's hashtags. I added my CSV file into the program, ensuring that tweets which had multiple hashtags were separated by commas. After the program had processed the file, it produced the interactive network maps which can be seen below.



The overall view of the cluster's network map.

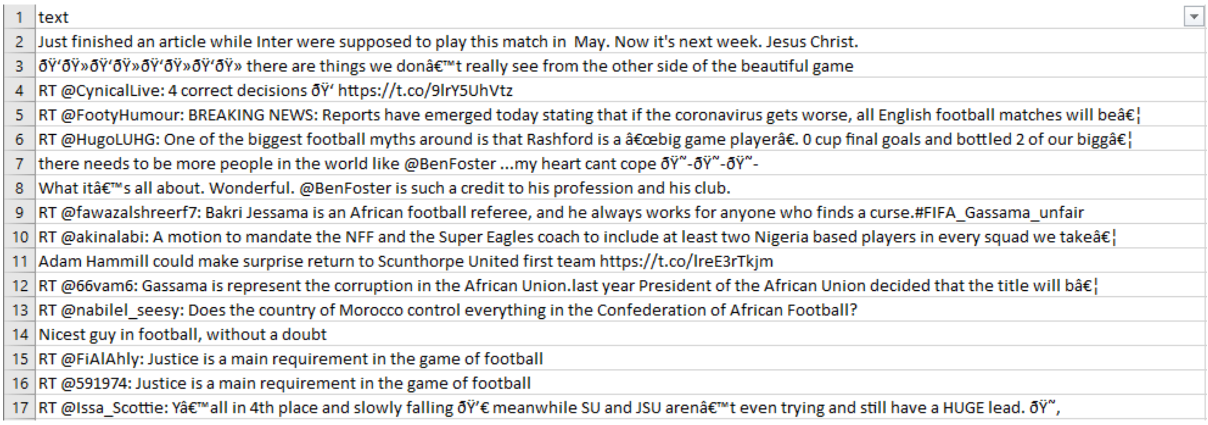


A magnified view of some of the clusters.

These network graphs are of vast importance in enabling me to analyse the hashtag data as they show co-occurring hashtag information. Within each cluster, there were a range of different hashtags, with the most used hashtags appearing as larger circles. The largest of clusters including hashtags of “FACup”, “Gunners”, “Football”, “COVID” and finally “UpdateCorono2020”. At the time when I was crawling Twitter for data, the English FA Cup (Round of 16) was being played between Portsmouth FC and Arsenal FC; hence, the Gunners and FACup hashtags. This was of key importance as it allowed me to collect more tweets in relation to my chosen field. Furthermore, with the outbreak of the worldwide virus ‘Coronavirus’, many football fans were worried about the impact that this virus could have on the footballing world. At the time of crawling the tweets, the virus had only begun to surface and there were rumours of FIFA (the international governing body of football) cancelling the European Football Championship which was due to take place in Summer. Hence, the “COVID” hashtag. Again, this was of great importance as allows me to very easily visualise the footballing communities’ current concerns.

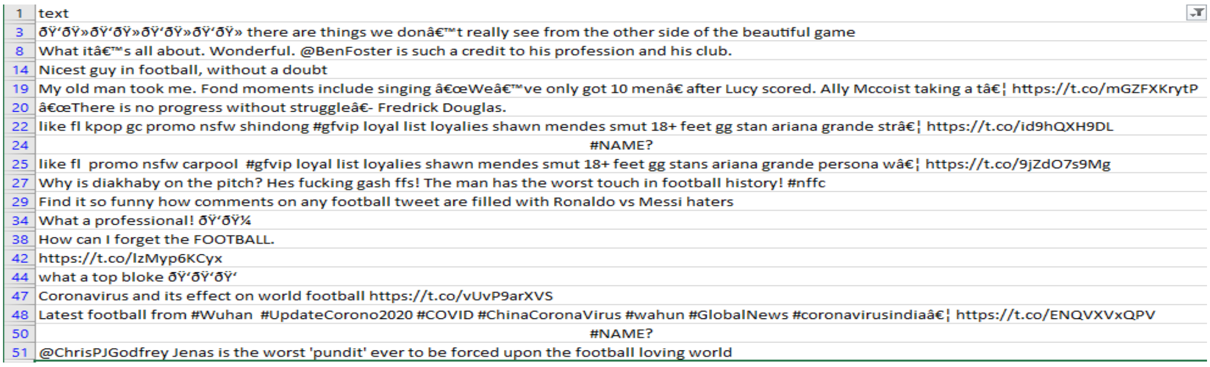
Capturing & Organising User Data

In order to organise all of the tweet data that I had collected in my CSV file, I decided to use the inbuilt filter function on Microsoft Excel to display tweets of specific categories so that I could analyse each category individually. These included all original tweets I had collected; tweets which were not retweets and finally tweets which were not retweets and mentioned another user. Screenshots of these different files can be found below.



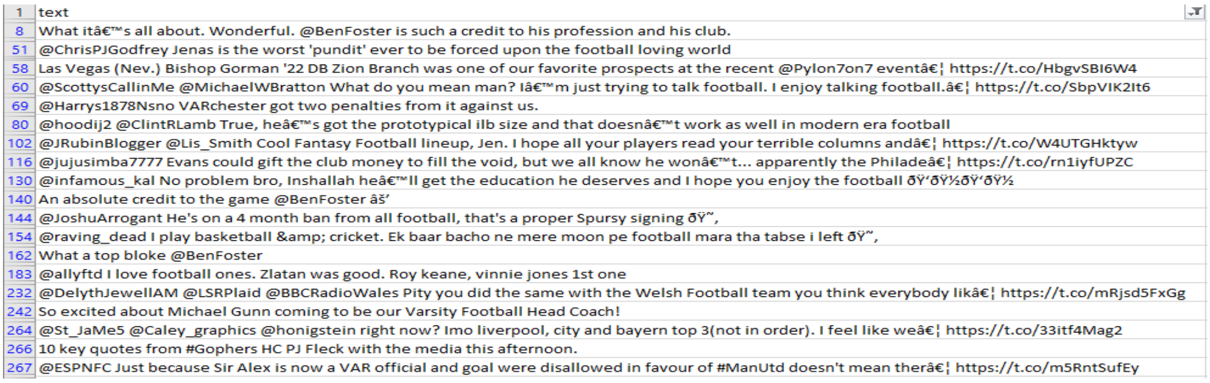
1	text
2	Just finished an article while Inter were supposed to play this match in May. Now it's next week. Jesus Christ.
3	ðŸ‘ðŸ»ðŸ‘ðŸ»ðŸ‘ðŸ»ðŸ‘ðŸ» there are things we donâ€™t really see from the other side of the beautiful game
4	RT @CynicalLive: 4 correct decisions ðŸ‘ https://t.co/9lrY5UhVtz
5	RT @FootyHumour: BREAKING NEWS: Reports have emerged today stating that if the coronavirus gets worse, all English football matches will beâ€¦
6	RT @HugoLUHG: One of the biggest football myths around is that Rashford is a â€œbig game playerâ€. 0 cup final goals and bottled 2 of our biggâ€¦
7	there needs to be more people in the world like @BenFoster ...my heart cant cope ðŸ˜ðŸ˜ðŸ˜
8	What itâ€™s all about. Wonderful. @BenFoster is such a credit to his profession and his club.
9	RT @fawazalshreerf7: Bakri Jessama is an African football referee, and he always works for anyone who finds a curse.#FIFA_Gassama_unfair
10	RT @akinalabi: A motion to mandate the NFF and the Super Eagles coach to include at least two Nigeria based players in every squad we takeâ€¦
11	Adam Hammill could make surprise return to Scunthorpe United first team https://t.co/lreE3rTkjm
12	RT @66vam6: Gassama is represent the corruption in the African Union.last year President of the African Union decided that the title will bâ€¦
13	RT @nabilel_seesy: Does the country of Morocco control everything in the Confederation of African Football?
14	Nicest guy in football, without a doubt
15	RT @FiAlAhly: Justice is a main requirement in the game of football
16	RT @591974: Justice is a main requirement in the game of football
17	RT @Issa_Scottie: Yâ€™all in 4th place and slowly falling ðŸ’€ meanwhile SU and JSU arenâ€™t even trying and still have a HUGE lead. ðŸ˜‚

Extract of data showing non-filtered CSV file.



1	text
3	ðŸ'ðŸ»ðŸ'ðŸ»ðŸ'ðŸ»ðŸ'ðŸ» there are things we donâ€™t really see from the other side of the beautiful game
8	What itâ€™s all about. Wonderful. @BenFoster is such a credit to his profession and his club.
14	Nicest guy in football, without a doubt
19	My old man took me. Fond moments include singing â€œWeâ€™ve only got 10 menâ€ after Lucy scored. Ally Mccoist taking a tâ€¦ https://t.co/mGZFXKrytP
20	â€œThere is no progress without struggleâ€- Fredrick Douglas.
22	like fl kpop gc promo nsfw shindong #gfvip loyal list loyalies shawn mendes smut 18+ feet gg stan ariana grande strâ€¦ https://t.co/id9hQXH9DL
24	#NAME?
25	like fl promo nsfw carpool #gfvip loyal list loyalies shawn mendes smut 18+ feet gg stans ariana grande persona wâ€¦ https://t.co/9jZdO7s9Mg
27	Why is diakhaby on the pitch? Hes fucking gash ffs! The man has the worst touch in football history! #nffc
29	Find it so funny how comments on any football tweet are filled with Ronaldo vs Messi haters
34	What a professional! ðŸ'ðŸ¼
38	How can I forget the FOOTBALL.
42	https://t.co/IzMyp6KCyx
44	what a top bloke ðŸ'ðŸ'ðŸ'
47	Coronavirus and its effect on world football https://t.co/vUvP9arXVS
48	Latest football from #Wuhan #UpdateCorono2020 #COVID #ChinaCoronaVirus #wahun #GlobalNews #coronavirusindiaâ€¦ https://t.co/ENQVXVxQPV
50	#NAME?
51	@ChrisPJGodfrey Jenas is the worst 'pundit' ever to be forced upon the football loving world

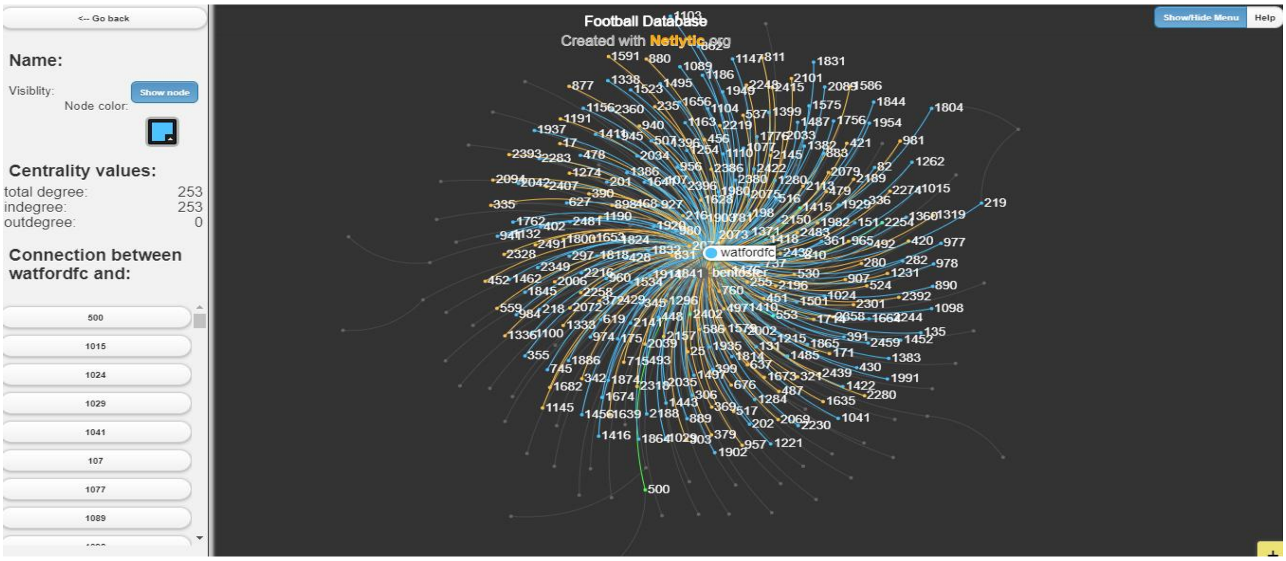
Extract of data showing data filtered by only showing tweets which are not retweets.



1	text
8	What itâ€™s all about. Wonderful. @BenFoster is such a credit to his profession and his club.
51	@ChrisPJGodfrey Jenas is the worst 'pundit' ever to be forced upon the football loving world
58	Las Vegas (Nev.) Bishop Gorman '22 DB Zion Branch was one of our favorite prospects at the recent @Pylon7on7 eventâ€¦ https://t.co/HbgvSBI6W4
60	@ScottysCallinMe @MichaelWBratton What do you mean man? Iâ€™m just trying to talk football. I enjoy talking football.â€¦ https://t.co/SbpVIK2It6
69	@Harrys1878Nsno VARchester got two penalties from it against us.
80	@hoodij2 @ClintRLamb True, heâ€™s got the prototypical ilb size and that doesnâ€™t work as well in modern era football
102	@JRubinBlogger @Lis_Smith Cool Fantasy Football lineup, Jen. I hope all your players read your terrible columns andâ€¦ https://t.co/W4UTGHktyw
116	@jujusimba7777 Evans could gift the club money to fill the void, but we all know he wonâ€™t... apparently the Philadeâ€¦ https://t.co/rn1iyfUPZC
130	@infamous_kal No problem bro, Inshallah heâ€™ll get the education he deserves and I hope you enjoy the football ðŸ'ðŸ½ðŸ'ðŸ½
140	An absolute credit to the game @BenFoster âš'
144	@JoshuArrogant He's on a 4 month ban from all football, that's a proper Spursy signing ðŸ˜‚
154	@raving_dead I play basketball &amp; cricket. Ek baar bacho ne mere moon pe football mara tha tabse i left ðŸ˜‚
162	What a top bloke @BenFoster
183	@allyftd I love football ones. Zlatan was good. Roy keane, vinnie jones 1st one
232	@DelythJewellAM @LSRPlaid @BBCRadioWales Pity you did the same with the Welsh Football team you think everybody likâ€¦ https://t.co/mRjsd5FxGg
242	So excited about Michael Gunn coming to be our Varsity Football Head Coach!
264	@St_JaMe5 @Caley_graphics @honigstein right now? Imo liverpool, city and bayern top 3(not in order). I feel like weâ€¦ https://t.co/33itf4Mag2
266	10 key quotes from #Gophers HC PJ Fleck with the media this afternoon.
267	@ESPNFC Just because Sir Alex is now a VAR official and goal were disallowed in favour of #ManUtd doesn't mean therâ€¦ https://t.co/m5RntSufEy

Extract of data showing data filtered by only showing tweets which are not retweets and contains a mention.

Once I had these files, I used a cortext manager to create a new corpus so that online text and social network analyser to complete network analysis on my datafiles [8]. This allowed me to discover, summarise and visualise the most important aspects of user interaction between the tweet's authors and other users who were either retweeting them or mentioning other Twitter users. These network graphs can be found below.



Name:

Visiblity: Show node

Node color:

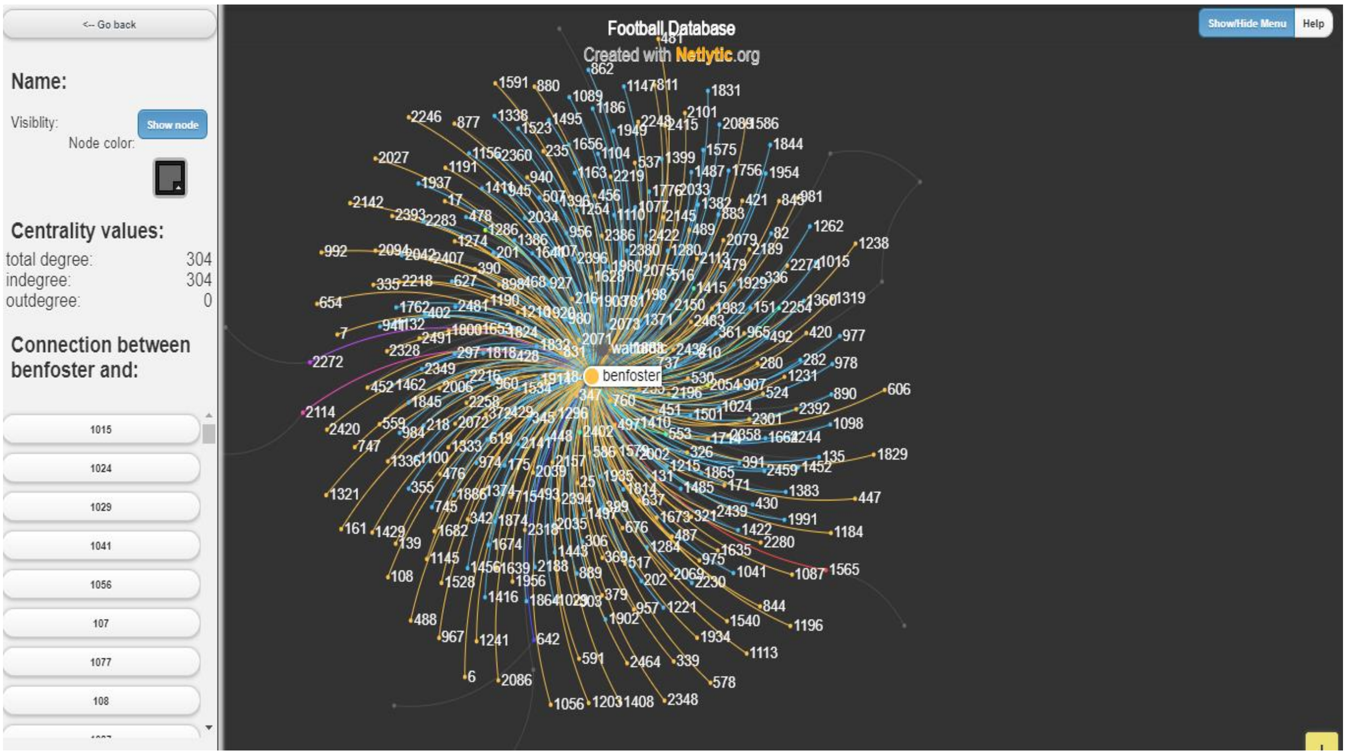
Centrality values:

total degree:	253
indegree:	253
outdegree:	0

Connection between watfordfc and:

500
1015
1024
1029
1041
107
1077
1089

Original dataset with all users who had tweeted about either football or VAR.



Name:

Visiblity:

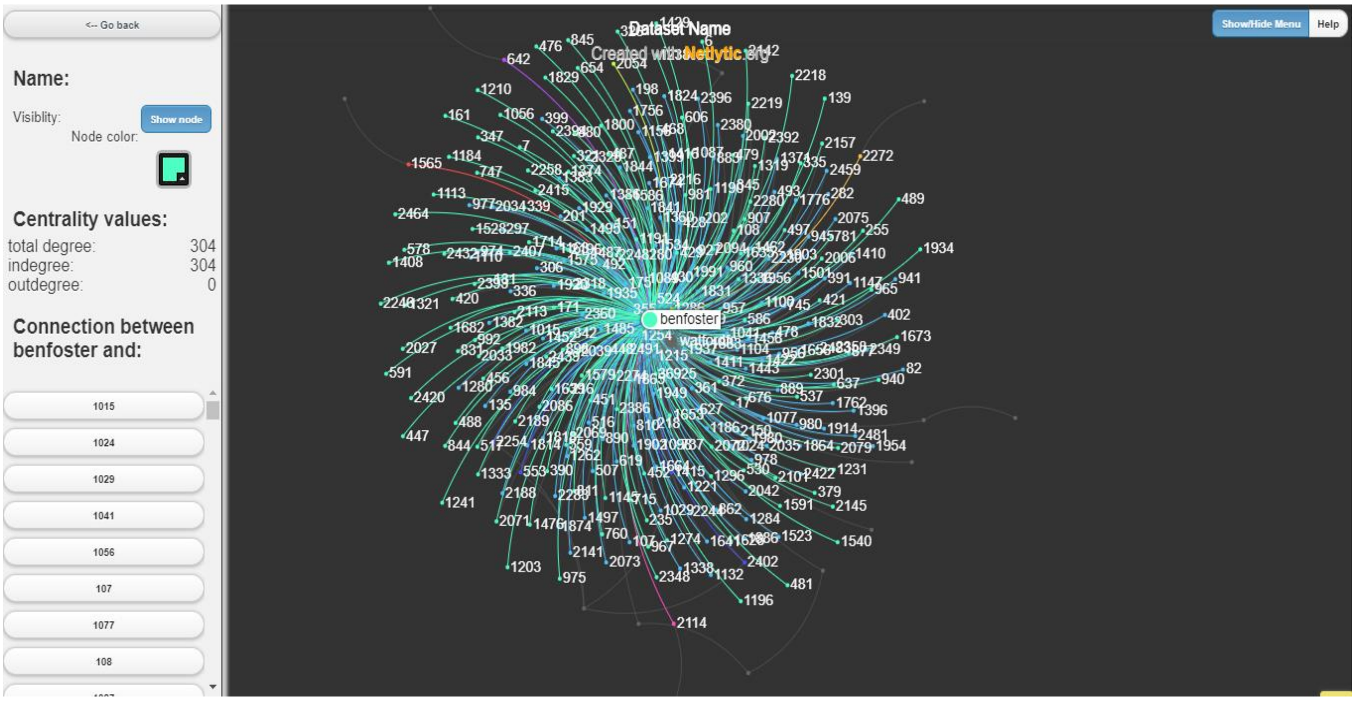
Node color:

Centrality values:

total degree:	304
indegree:	304
outdegree:	0

Connection between benfoster and:

1015
1024
1029
1041
1056
107
1077
108



Name:

Visiblity:

Node color:

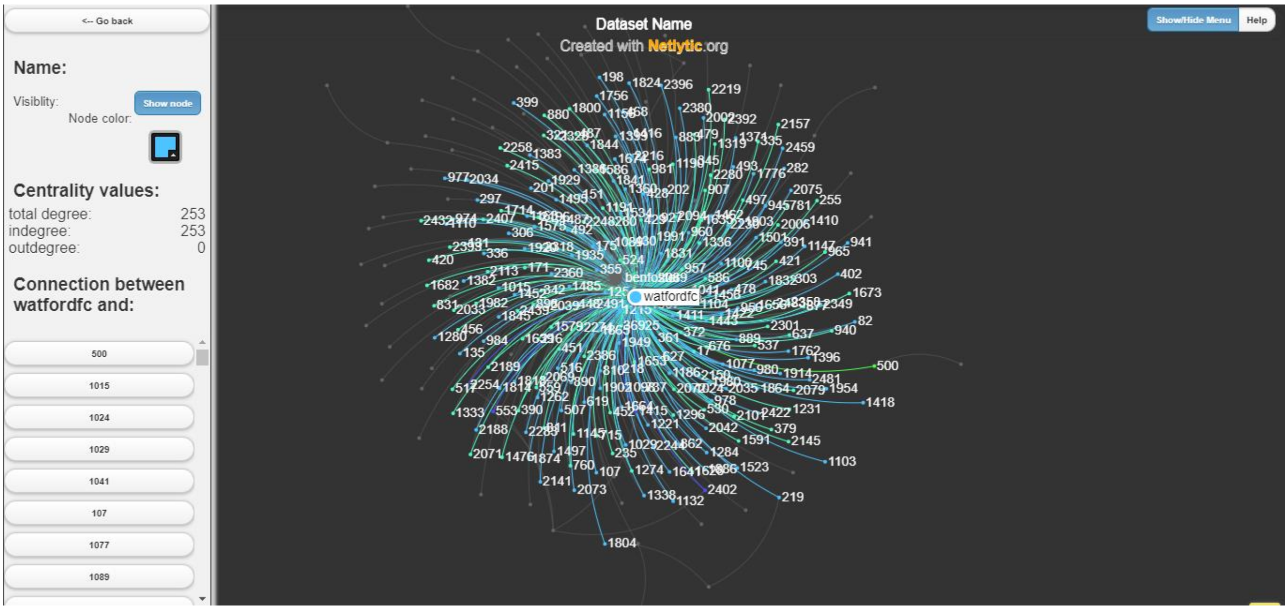
Centrality values:

total degree:	304
indegree:	304
outdegree:	0

Connection between benfoster and:

1015
1024
1029
1041
1056
107
1077
108

Network graph showing user's tweets which were not retweets.



Name:

Visiblity: Show node

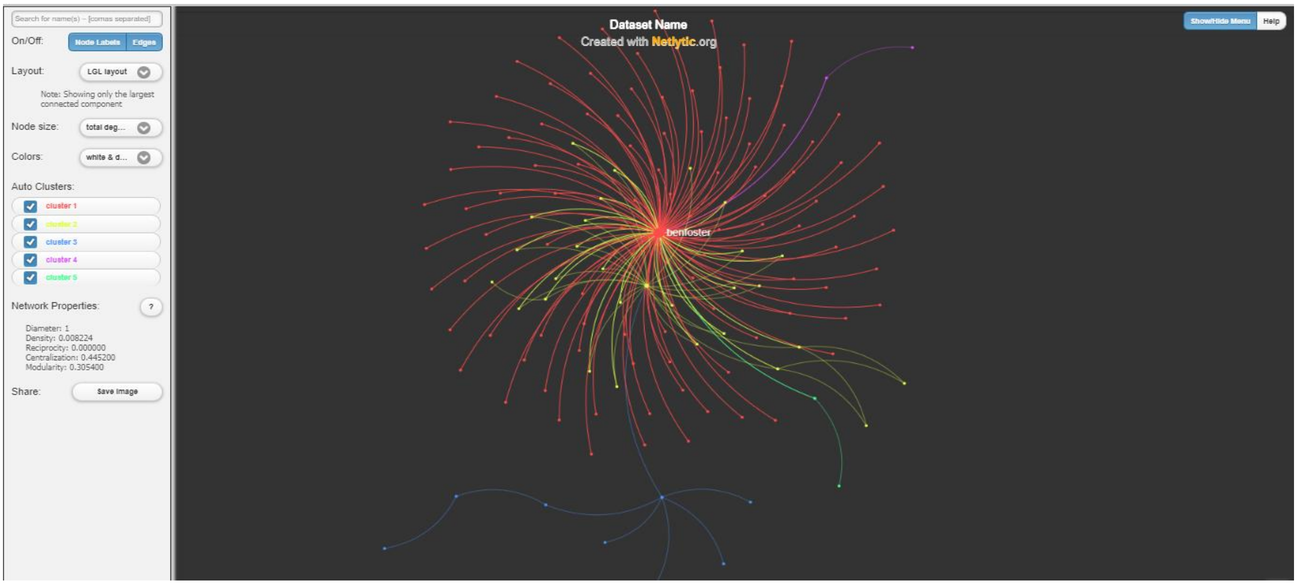
Node color:

Centrality values:

total degree:	253
indegree:	253
outdegree:	0

Connection between watfordfc and:

500
1015
1024
1029
1041
107
1077
1089



Search for name(s) – [comas separated]
Show/Hide Menu
Help
Dataset Name
Created with Netlytic.org
benfoster

On/Off: Node Labels | Edges

Layout: LGL layout

Note: Showing only the largest connected component

Node size: total deg...

Colors: white & d...

Auto Clusters:

☑ cluster 1
☑ cluster 2
☑ cluster 3
☑ cluster 4
☑ cluster 5

Network Properties: ?

Diameter: 1
Density: 0.008224
Reciprocity: 0.000000
Centralization: 0.445200
Modularity: 0.305400

Share: Save Image

Network graph showing user's tweets which mentioned other users and were not retweets.

Each of these network graphs shows a vast amount of data about the type of tweets that users were creating. From this, we can see the interconnections between the sets of entities (the tweets) and the context which has the highest in/out degree - in this case, it is clearly Watford FC and Ben Foster. This shows us that these topics are a large part of the overall football discussion on Twitter as there are many people who had mentioned Ben Foster or Watford FC in their tweet, creating an interconnected web between users. Overall, it is obvious to see that grouping and clustering of data is of significance when it comes to context analysis as it highlights non-trivial data discrepancies that might have been overlooked.

Bibliography

1. https://developer.twitter.com/en/docs
2. http://docs.tweepy.org/en/latest/
3. https://developer.twitter.com/en/docs/tweets/search/overview/standard
4. https://github.com/ckoepp/TwitterSearch
5. https://github.com/SamDelgado/twitter-to-mongo
6. https://stackoverflow.com/questions/27889873/clustering-text-documents-using-scikit-learn-kmeans-in-python?fbclid=IwAR13agTGUdH3e7Xdpt2x6ee6R8vrzjWCuguWgCgTklOcmcYBwVdO6ak8c3
7. https://managerv2.cortext.net/project/7996.

8. https://netlytic.org/