Chapter 1

Introduction

1.1 Aim:

Football is the most popular sport in the world, and the English Premier League is the most watched league in the world. It consists of 20 teams. A huge number of users would then naturally talk about it, spread news about it or even engage in a discussion about it; and what better place to do so than a Social networking site, precisely – Twitter! Twitter is probably the most popular social networking site present in the world consisting of several millions of users. It is a website which is tuned for fast communication, which is perfect for modern-day life which never sleeps. Every second, on average, around 6000 tweets are tweeted on twitter, which corresponds to 3.5 lakh tweets per minute and around 500 million tweets per day. Twitter contains over 300 million active users and such a huge number of people share so much knowledge here that twitter's popularity as an information source has led to the development of applications and research in many domains including football. In fact it is proven to be a valuable source of information, simply because it can provide, in some cases, more informative data than the ones found in other websites or statistical or historical sources. A few examples are players' injuries, sacked coaches, and the average sentiment amongst the fan base of each team. With such a huge repository of information on Twitter, it makes perfect sense to aggregate this data and represent the most useful and informative and trending tweets. This is what we are trying to achieve.

Analysis of the data at hand will help you make the best decision. Analysis of data uncovers individual characteristics of the subject and these trends and characteristics can be used in different kinds of fields. Similarly, in football, we can find out which is the most popular team, most popular player, etc. All this information will be present on Twitter. We just need to search for the correct information and then sort out this data according to the users' needs. This information will also be displayed and represented in the form of graphs to give that attractive visual appeal. Even the most trending tweets related to the English Premier League will be displayed on our dashboard. Weekly analysis, monthly analysis and end-of-season analysis would be given so as to rank each team based on different aspects.

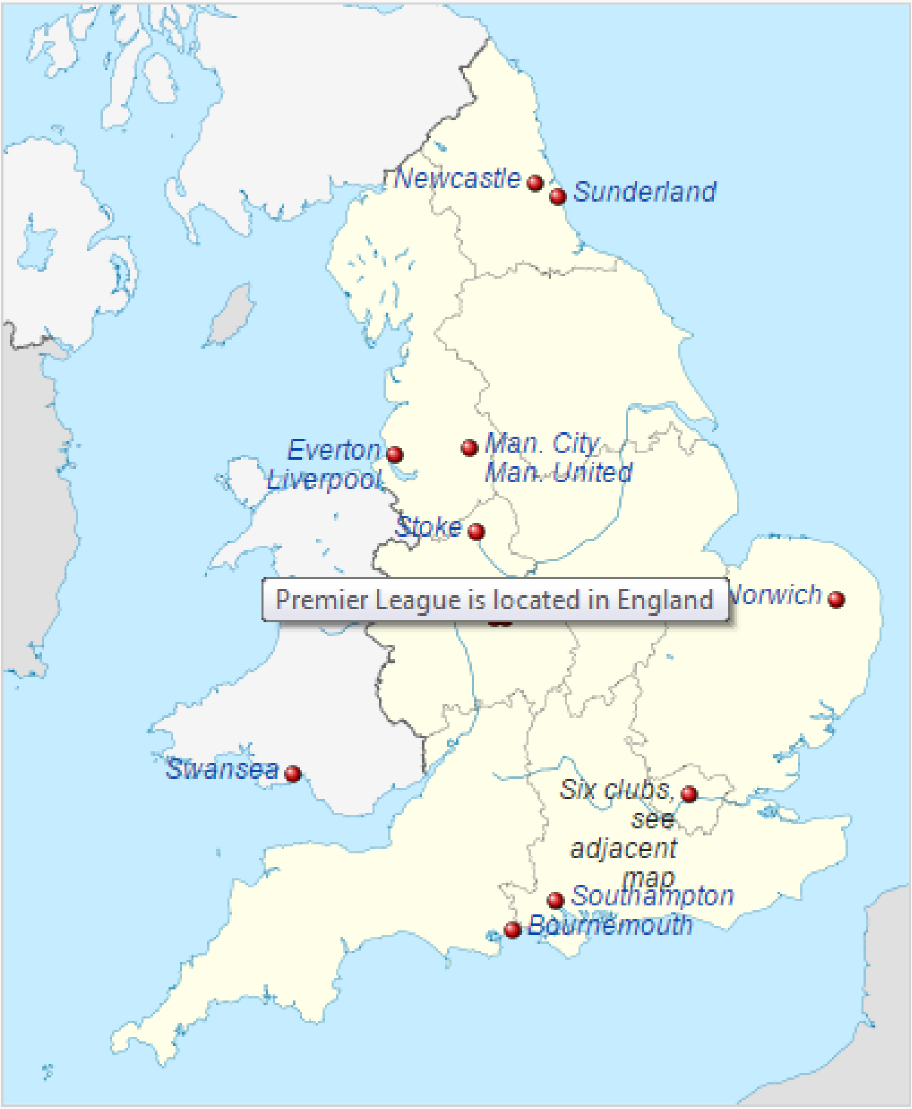
The English Premier League

[Fig. 1.1] Chelsea vs. Norwich (2016). This match was particularly popular because of Eden hazard (above with the ball) whose performance was ridiculed by many on twitter.

The English Premier league is the biggest, most popular and most viewed football league in the world. It is also the most profitable league in the world. It is contested by 20 teams based in England and wales, operating on a system based on promotion of teams from the lower leagues and relegation of teams from the higher leagues. The English Premier league is the highest league in the football league and the winner gets a premier league trophy along with huge financial incentives, and qualification to the UEFA Champions league. The league generates 2.2 Million EU per year in domestic and international rights. The league is broadcast in over 200 countries to a potential TV audience of 4.7 billion people. Games are played every week over the football year, August to May of the successive year. Teams play each other twice over the course of the season, one in their own stadium and one in their opponent’s stadium. A total of 38 games are played and games either happen on a weekend or a Monday night. A win grants the winner 3 points, a draw grants both the teams a single point and the loser isn’t awarded any points. In the end, the team with the highest number of points is awarded champions. In a situation of multiple teams ending with the same number of points, the goal difference (Goals scored minus goals given away) is used to separate the winner from the runner-up.

The Premier League is particularly popular in Asia, where it is the most widely distributed sports programme. In Australia, Fox Sports broadcasts almost all of the season's 380 matches live (38 Gameweeks x 10 Games per week), and Foxtel gives subscribers the option of selecting which Saturday 3pm match to watch. In India, the matches are broadcast live on STAR Sports. In China, the broadcast rights were awarded to Super Sports in a six-year agreement that began in the 2013–14 season.

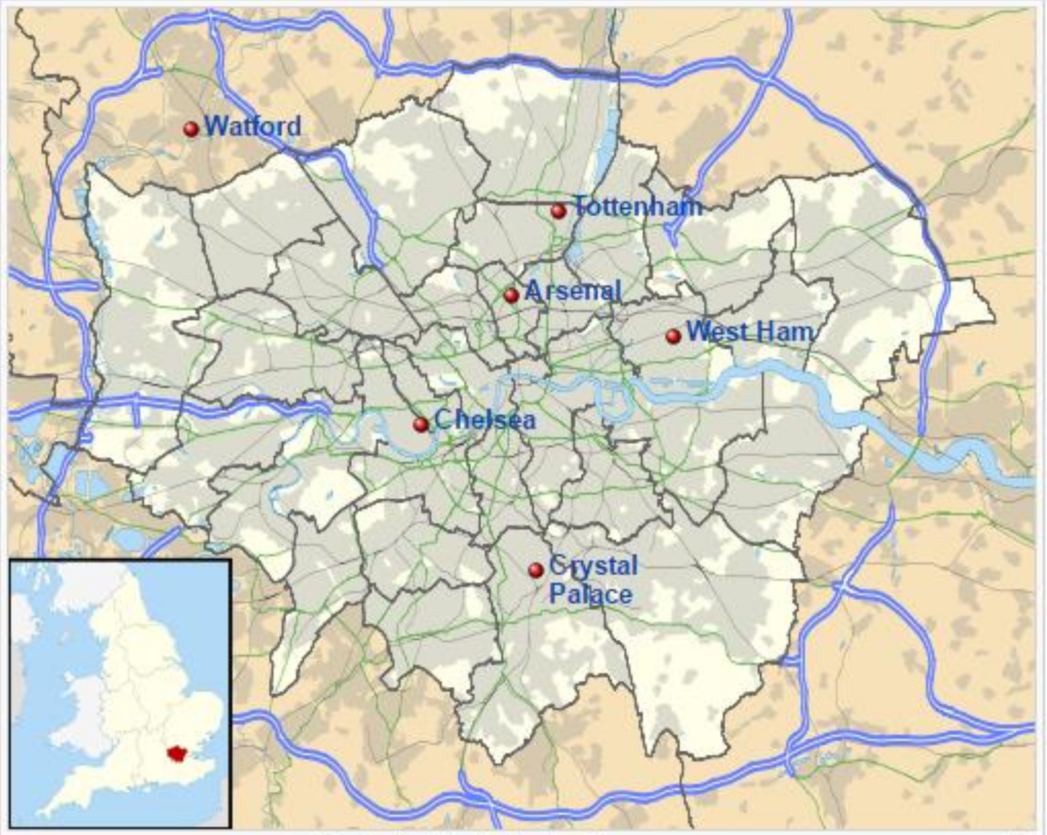
Each team has a head coach and a bunch of identity players. The head coach lends his own identity or philosophy to his team and the fans may like or dislike the philosophy. The head coach is often a personality cult and fans either ridicule or praise head coaches on a regular basis on social media.



Location of clubs for the 2015–16 Premier League season

Fig.1.2: Location of Premier League Clubs

Fig.1.3: Location of Premier League Clubs in London itself



Football clubs in the London area

The 2015-2016 season has proven to be one of the better seasons of football because it has thrown up a variety of surprises. Leicester City, a team who was bottom of the league in April 2015 is now in pole position; and Chelsea, who were premier league champions last year are languishing in 11th. Hence unsurprisingly, a lot of talk on social media has been about these two teams and their stories of hard work, grit and determination, and the lack of it.

1.2 Review of Literature:

Data, in the hands of the right people, is the most important asset an organization has. But the more data it has to deal with, the bigger the scale of the challenge. We are flooded by all kinds of data – scientific data, medical data, financial data, and behaviour data to name a few. The rise in cloud, social computing means that organizations are faced with ever increasing volumes of new types of data. Data is just getting bigger and bigger – in fact, between 2005 and 2010 digital data grew from 130 to 1227 exabytes. In the future data is expected to grow over 75 times in volume. People have no time to look at all this data. Human attention has become a precious resource. So, we must find ways to automatically analyze the data, to automatically classify it, to automatically summarize it, to automatically discover and characterize trends in it, and to automatically flag anomalies. This is one of the most active and exciting areas of big data analysis, and this is why big data analysis has become a fundamental part of corporate society, including football and sports. Let us take a look at the evolution of big data.

1.2.1: Big Media Data: Understanding, Search, and Mining (IEEE TRANSACTIONS ON BIG DATA)

Evolution Of Big Data:

Our capabilities of both generating and collecting data have been increasing rapidly. Contributing factors include the computerization of business, scientific, and government transactions; the widespread use of digital cameras, publication tools, and bar codes for most commercial products; and advances in data collection tools ranging from scanned text and image platforms to satellite remote sensing systems. In addition, popular use of the World Wide Web as a global information system has flooded us with a tremendous amount of data and information. The emergence of social media sites have provided the common man to express his/her feelings. The data in these social media sites is filled with information but is enormous and filled with large data sets, and that makes it difficult to process using data processing techniques. The challenges include analysis, pattern recognition, and visualization.

Big data analytics does refer to the process of collecting, organizing and analyzing large sets of data to discover patterns and other useful information, but the main focus is not just storage and analysis. The product of the analysis will help you to understand the information contained within the data, and it will also help identify the data that is most important to the business and future business decisions. Big data analysts basically want the knowledge that comes from analyzing the data, and major stakeholders want to use this knowledge to make important decisions involving a lot of time, effort, money, and people. However, due to large sample size, Big Data give rise to two additional goals: to understand heterogeneity and commonality across different subpopulations.

Big Data give promises for: (i) exploring the hidden structures of each subpopulation of the data, which is traditionally not feasible and might even be treated as ‘outliers’ when the sample size is small; (ii) extracting important common features across many subpopulations even when there are large individual variations. For our project, we can understand the similarities different fans of the same team have, and also the similarities that different teams have with each other. We can also understand the differences in a similar way, and these points of difference will add interest.

Big Data have numerous other applications. Taking social network data analysis for example, massive amount of social network data are being produced by Twitter, Facebook, LinkedIn and YouTube. These data reveal numerous individual’s characteristics and have been exploited in various fields.

In addition, the social media and Internet contain massive amount of information on the consumer preferences and confidences, leading economic indicators, business cycles, political attitudes, and the economic and social states of a society. It is anticipated that the social network data will continue to explode and be exploited for many new applications.

Several other new applications that are becoming possible in the Big Data era include:

A. Personalized services. With more personal data collected, commercial enterprises are able to provide personalized services adapt to individual preferences. For example, Target (a retailing company in the United States) is able to predict a customer’s need by analyzing the collected transaction records.

B. Internet security. When a network-based attack takes place, historical data on network traffic may allow us to efficiently identify the source and targets of the attack.

C. Personalized medicine. More and more health related metrics such as individual’s molecular characteristics, human activities, human habits and environmental factors are now available. Using these pieces of information, it is possible to diagnose an individual’s disease and select individualized treatments.

D. Digital humanities. Nowadays many archives are being digitized. For example, Google has scanned millions of books and identified about every word in every one of those books. This produces massive amount of data and enables addressing topics in the humanities, such as mapping the transportation system in ancient Roman, visualizing the economic connections of ancient China, studying how natural languages evolve over time, or analysing historical events.

The potential of big data is humongous in a commercial aspect. In fact the big data market is estimated to be worth about $100 Billion. The huge amount of money can drive major corporations to study and derive new information. Money certainly does talk!

1.2.2: Efficient Analysis of Big Data Using Map Reduce Framework (IJRDET)

Advantages of Big Data:

1. Dialogue with customers: Today's consumers are a tough nut to crack. They look around a lot before they buy, talk to their entire social network about their purchases, demand to be treated as unique and want to be sincerely thanked for buying your products. Big Data allows you to profile these increasingly vocal and fickle people in an easy manner so that you can engage in an almost one-on-one, real-time conversation with them. This is not actually a luxury. If you don't treat them like they want to, they will leave you in the blink of an eye. This is a great advantage of data mining and social media analysis in particular.
2. Perform risk analysis: Success not only depends on how you run your company. Social and economic factors are crucial for your accomplishments as well. Predictive analytics, fueled by Big Data allows you to scan and analyze newspaper reports or social media feeds so that you permanently keep up to speed on the latest developments in your industry and its environment. Detailed health-tests on your suppliers and customers or understanding the opinion of fans after a football club makes a major decision are another goodie that comes with Big Data. This will allow you to take action when one of them is in risk of defaulting.
3. Keeping data safe: You can map the entire data landscape across your company with Big Data tools, thus allowing you to analyze the threats that you face internally. You will be able to detect potentially sensitive information that is not protected in an appropriate manner and make sure it is stored according to regulatory requirements. With real-time Big Data analytics you can, for example, flag up any situation where 16 digit numbers – potentially credit card data - are stored or emailed out and investigate accordingly. Tools like MongoDB can store tweets in a very safe manner and also doesn't occupy much space.
4. Create new revenue streams: The insights that you gain from analyzing your market and its consumers with Big Data are not just valuable to you. You could sell them as non-personalized trend data to large industry players operating in the same segment as

you and create a whole new revenue stream. You can sell the data mined from football fans to football clubs who would like to know what they talk about and how popular their club or their rival clubs are.

5. Customize your website in real time: Big Data analytics allows you to personalize the content or look and feel of your website in real time to suit each consumer entering your website, depending on, for instance, their sex, nationality or from where they ended up on your site. The best-known example is probably offering tailored recommendations: Amazon’s use of real-time, item-based, collaborative filtering (IBCF) to fuel its ‘Frequently bought together’ and ‘Customers who bought this item also bought’ features or LinkedIn suggesting ‘People you may know’ or ‘Companies you may want to follow’. And the approach works: Amazon generates about 20% more revenue via this method. We can also stream live tweets mined from twitter for live customization of the website.
6. Reducing maintenance costs: Traditionally, factories estimate that a certain type of equipment is likely to wear out after so many years. Consequently, they replace every piece of that technology within that many years, even devices that have much more useful life left in them. Big Data tools do away with such unpractical and costly averages. The massive amounts of data that they access and use and their unequalled speed can spot failing grid devices and predict when they will give out. The result: a much more cost-effective replacement strategy for the utility and less downtime, as faulty devices are tracked a lot faster. A project focusing solely on the web and big data analysis would have extremely minimal tasks excluding labor costs, machinery costs, land costs, etc.

1.2.3: Data Mining in Sports: (MIS Masters Project)

The sports world is known for the vast amounts of statistics that are collected for each player, team, game, and season. There are also many types of statistics that are gathered for each – a football player will have data for goals, shots, passes, assists, tackles, blocks, aerial success, etc. for each game. Similarly a football team will have many types of statistics like number of followers, number of active followers, number of tweets per day, points, market share, income, valuation, number of fans, etc. This can result in information overload for those trying to derive meaning from the statistics. Hence, sports are ideal for data mining tools and techniques.

Sports organizations, due to the extremely competitive environment in which they operate, need to seek any edge that will give them an advantage over others. Traditionally knowledge of sports has been believed to be contained in the minds of its experts – the scouts, coaches, and managers. Only recently have sports organizations begun to realize that there is also a potential of important information contained in their computer data. Currently, most team sports organizations employ in-house statisticians and analysts to retrieve meaning and insight for the scouts who evaluate future prospects and talent, the coaches who are in charge of the team on the playing surface, as well as the general managers who are in charge of drafting or signing players. Making use of data mining analysts as well, is becoming very important nowadays.

The other main issue is scouting. Scouting is an integral part of every sport including football. There are two primary types of scouting efforts that are used by sports organizations. The first, is the scouting of potential talent. To do this, scouts travel across the nation, and also across the globe, to evaluate future players. These scouts compile reports and evaluations detailing each player's abilities, strengths, and weaknesses. The second form of scouting, called 'advance scouting', is the evaluation of upcoming competitors or teams with which they'll play against or compare with. These scouts travel to watch other teams and compile reports that are used to help make strategies and approaches when playing against them.

Traditionally, advance scouts in football were sent to games to collect data, count the different things and create reports important to team and player abilities.

With the availability of statistical analysis tools and video, scouting has changed a lot. Scouting has gone beyond the strengths and weaknesses of the opponent to analysis of typical player, coach, and team strategies and behaviors in certain situations. The usage of online data has become very important for scouting, like the evaluation of data in social media sites which has become very important for different types of analysis including sentiment analysis.

Data mining can be used by sports organizations in the form of statistical analysis, pattern discovery, as well as outcome prediction. Patterns in the data are often helpful in the forecast of future events. A program begun in 2002 by football club AC Milan (Italy) uses software to help predict player injuries by collecting data from workouts over a period of time (Flinders, 2002). The tool created by Computer Associates makes predictions from the medical statistics amassed for each player. Since athletes are their biggest investments, teams are hoping that prediction of injury will help save them millions of dollars.

Similarly, data mining can be used on physical aptitude test data in order to predict future physical performance. Data mining software was used to link test data of cadets at the United States Military Academy and their actual performance in a required fitness class. This type of analysis would have significant implications to sports. Among the included physical tests are the 40 yard dash, vertical and broad jump, as well as physical measurements. Throughout the years, football teams and experts have developed common standards on what are considered poor, good, and excellent results in the test based on the performance of the athletes throughout the previous years.

However, as most coaches would agree, statistics themselves can be very misleading. Certain players are able to build impressive stats but have little effect on a game. On the other hand, there are players who make a significant impact on the game without having impressive statistics. In football, take for instance the comparison of the wing back that is more prone to taking risks since he has to run forward to one who plays center back who has to be stronger. The center back may make more interceptions in a game than the wing back– a statistic often used to indicate a defender's value – but will also allow the opponent greater success when he misses an interception. So this has to be calculated.

Data mining is not meant to take the place of general managers, coaches, and scouts. Rather, it is a tool that can be used to aid in the decision-making processes that they make. In today's business world, a CEO or executive would not make any important decision without hard numbers and figures to back it up and so football clubs must be run similarly to those organizations in other industries. It is extensively used in the English Premier League.

1.3 Scope:

This project intended to provide a database for visualization, based on text analysis of each text record received from social networking database (twitter).

It supplies formal data after get processed based on the specification, to front-end visualization applications. So it needs to be platform independent, user friendly and easy maintenance. To satisfy usability of final outcome of this project, JAVA serves as object oriented programming language with assured platform independence, RStudio as an open source statistical programming language, MongoDB as open source database it provides free accessibility and cost free with optimum performance. All the API's and libraries are of open source with easy access.

IntelliJ is also a free software which we can use to do all sort of work in one place instead of shuffling over different applications through different coding language software. After obtaining permission from Twitter to extract the tweets, java code is used to extract the tweets which is stored in MongoDB and R Code is used to count the tweets, to make plots and to basically stastically analyse the information in different ways.

The biggest advantage of this project is that there is no need for any purchase of software, compatibility issues, hardware problems and complications related to maintenance. It would just be a dashboard, which would be present on a website, and such information will be easily accessible to those who have an internet connection. This information can also be used by certain organizations and even teams who would be astounded by the detail and can find it useful for their tasks. We want to emulate the effect done by certain organizations, for example, OptaStats, which displays statistical information related to football, and this information can be provided to football clubs who want better results

1.4 Motivation:

The motivation for doing this project was mainly an interest in undertaking a challenging project in an interesting area of big data analysis. The opportunity to learn about a new area of computing not covered in lectures was appealing to us. It would be very interesting to gather and then aggregate the social networking data so as to extract interesting patterns and recent trends from it.

We also wanted to monitor and chart the behavior and characteristics of Premier League supporters and to gauge their attitudes towards key performance areas for clubs, including the match day experience, stadium safety, ticketing facilities or mainly, the performance of the teams, which would be mentioned in the tweets.

Since RStudio is a fantastic tool for data mining and for processing data (since it's classified as a statistical programming language), we were excited to work with R Language. We will also be introduced to Mongo DB, which is a non-SQL database we have used to store tweets and collections of data. We were also exposed to the Twitter API and knowing how twitter provides permission and ensures authenticity of their tweets was valuable for us. We also hope that our findings, our research and our product can at least slightly contribute to mankind. In fact we feel that this project can make football more interesting and more appealing for fans to watch.

1.5 Problem Statement:

Traditional and popular websites like Skysports, goal.com, etc. displays all kinds of news from different teams including the transfer of players from one team to another and their rumours, live scores, playing statistics, fixtures and results, points and even and show news articles in an unbiased manner. However, there isn’t a single website which gives importance to the opinions of fans. Hence, we wanted to host a website which could at least partially value their opinions and show news reported by the fans themselves. Our project, Twitter Data Analysis of EPL will contain a website which contains statistical analysis of these tweets, which will be represented in an attractive and engaging way. It will be a website which is for the public and by the public.

Chapter 2

Report on the Present Investigation

2.1 Software Requirements:

Following are the software’s / technologies used in our project

1. **Twitter API:**

The social media site we chose to extract content from is twitter because of its shorter tweet size making things easier and primarily because of its hashtag concept making it easy for us to categorize tweets. Twitter is an online social networking service that enables users to send and read short 140-character messages called "tweets". Twitter allows you to interact with its data i.e. tweets & several attributes about tweets using Twitter APIs. You'd need to know a server side scripting language like php, python or ruby to make requests to twitter API and results would be in JSON format that can be easily read by your program. Registered users can read and post tweets, but unregistered users can only read them. Users access Twitter through the website interface, SMS, or mobile device app. Twitter Inc. is based in San Francisco and has more than 25 offices around the world. Users on Twitter generate over 400 million tweets everyday.

After making the twitter’s developers account, we got familiar with how they provide authentication to their tweets. The procedure is called OAuth. Each developer’s account had a unique consumer key, an access token, and a consumer secret key. We had to verify these keys on RStudio and only after correct verification did we get permission to extract tweets. The OAuth workflow is depicted above.

From Twitter we can extract the following types of information from Twitter: • Information about a user, • A user’s network consisting of his connections, • Tweets published by a user, and • Search results on Twitter. Twitter allows us to extract tweets in 2 different ways namely through the Rest API and streaming API. The Rest API is based on the REST architecture now popularly used for designing web APIs. To collect information a user must explicitly request it. It allows us to extract tweets containing one or more hashtags from a time period in the past. The only limitation of the Rest API is that it can only extract a limited number of tweets.

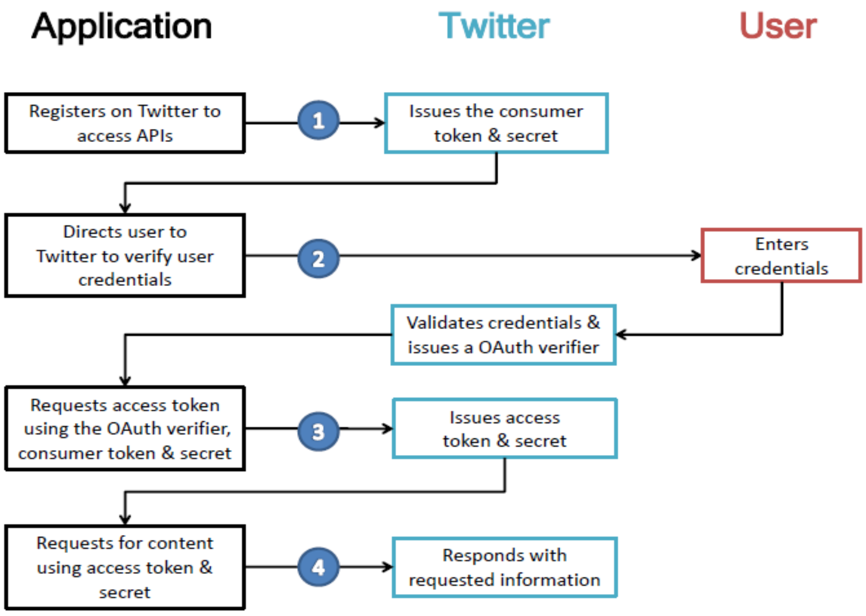

Application
Twitter
User
Registers on Twitter to access APIs
1
Issues the consumer token & secret
Directs user to Twitter to verify user credentials
2
Enters credentials
Validates credentials & issues a OAuth verifier
Requests access token using the OAuth verifier, consumer token & secret
3
Issues access token & secret
Requests for content using access token & secret
4
Responds with requested information

Fig. 2.1: OAuth workflow

The Streaming API however provides a continuous stream of public information from Twitter. Once a request for information is made, the Streaming APIs provide a continuous stream of updates with no further input from the user. The Streaming API however allowed us to extract live tweets. We used this on peak hours (match timings) to extract tweets pertaining to very popular hashtags (like #afc, Arsenal FC's official hashtag, which more than 500 mentions per hour as this exceeded limit). The Rest API on the other hand was used for less popular hashtags (#Kane referring to popular Spurs' player Harry Kane) The Rest API and Streaming API both were used for different comparisons and different purposes in our project and provided an equal method of comparisons for different players or teams.

2. RStudio:

Data collected over a period of time is processed by RStudio and some java code. R is a programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years. The source code for the R software environment is written primarily in C, FORTRAN, and R. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems. R uses a command line interface; there are also several graphical front-ends for it. RStudio is a software which allows users to use R Language.

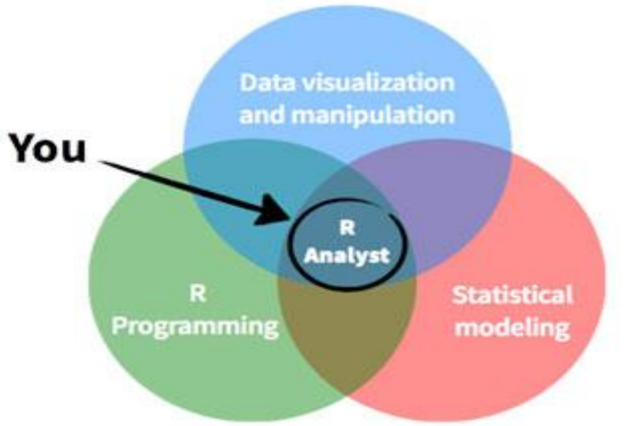


Fig.2.2: Statistical Analysis with R

We have used RStudio for data processing as well because it has different packages which can be used to analyze the data. R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made. Advanced users can write C, C++, Java, .NET or Python code to manipulate R objects directly. R is an interpreted language; users typically access it through a command-line interpreter.

It is R itself that allows extraction of tweets because of the “twitteR” package and “Rcurl” package. Using “plot” package, we can represent many forms of data into graphs. Some examples of the things we can do is make a wordcloud, bar plot, box plot, line graph, motion graph, scatter plot, text plot, find out the distribution and number of followers of a certain team across the world, and others.

Gameweek 32 Match Popularity Graph

- #AVLCHE: Hashtag used by people to mention the game: Aston Villa vs Chelsea
- #ARSWAT: Hashtag used by people to mention the game: Arsenal vs Watford
- #BOUMCI: Hashtag used by people to mention the game: Bournemouth vs Man City
- #NORNEW: Hashtag used by people to mention the game: Norwich vs Newcastle Utd
- #STOSWA: Hashtag used by people to mention the game: Stoke vs Swansea City
- #SUNWBA: Hashtag used by people to mention the game: Sunderland vs West Brom
- #WHUCRY: Hashtag used by people to mention the game: West Ham vs Crystal Palace
- #LIVTOT: Hashtag used by people to mention the game: Liverpool vs Tottenham
- #MUNEVE: Hashtag used by people to mention the game: Manchester Utd vs Everton
- #LEISOU: Hashtag used by people to mention the game: Leicester City v Southampton

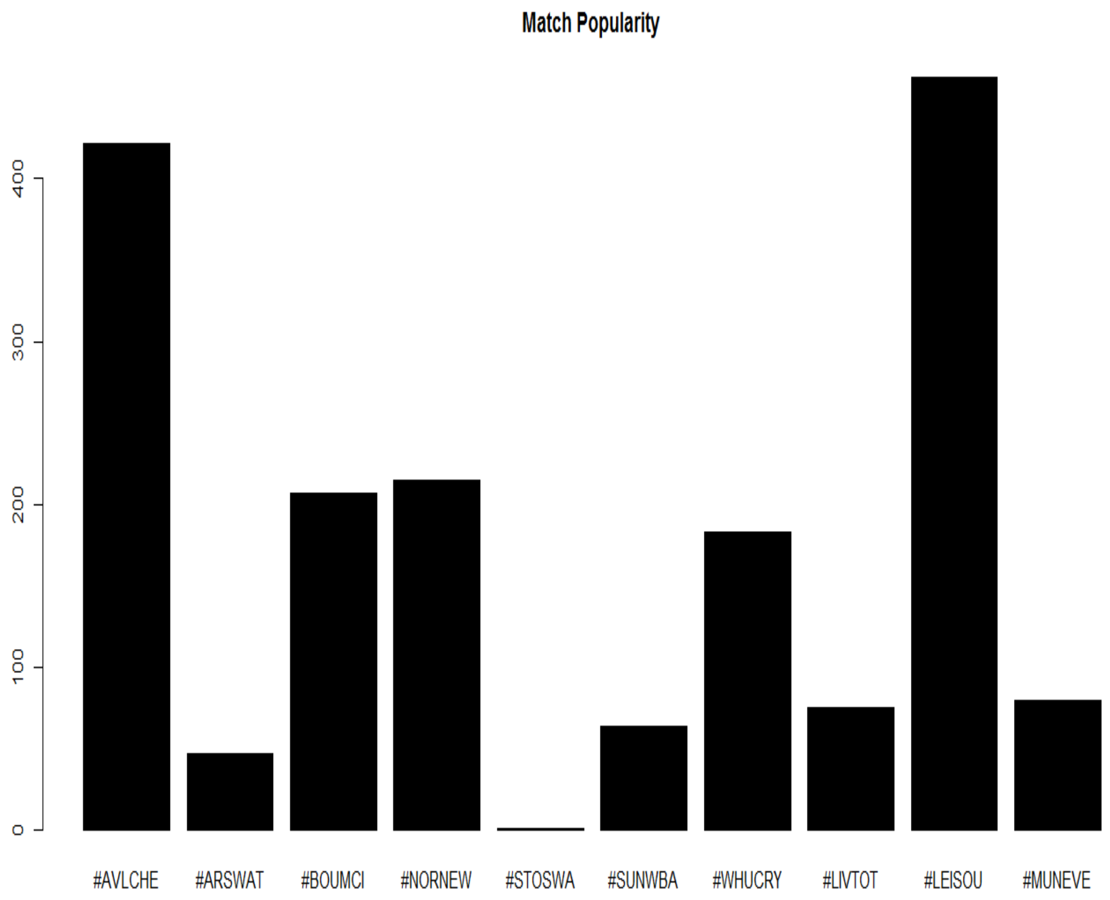


Fig 2.3: Match Popularity Graph created by RStudio

3. MongoDB:

Data is stored in a certain format in MongoDB. MongoDB is a NOSQL Database. It stores data in a very understandable manner. Tweets are stored along with their tweetID, Retweet count, Source and the tweet text. MongoDaemon starts the database from the command prompt and then we can view all the data in a command line interface. It stores each file on a collection and each collection is stored in a Database within MongoDB. We choose MongoDB for its adherence to the following principles:

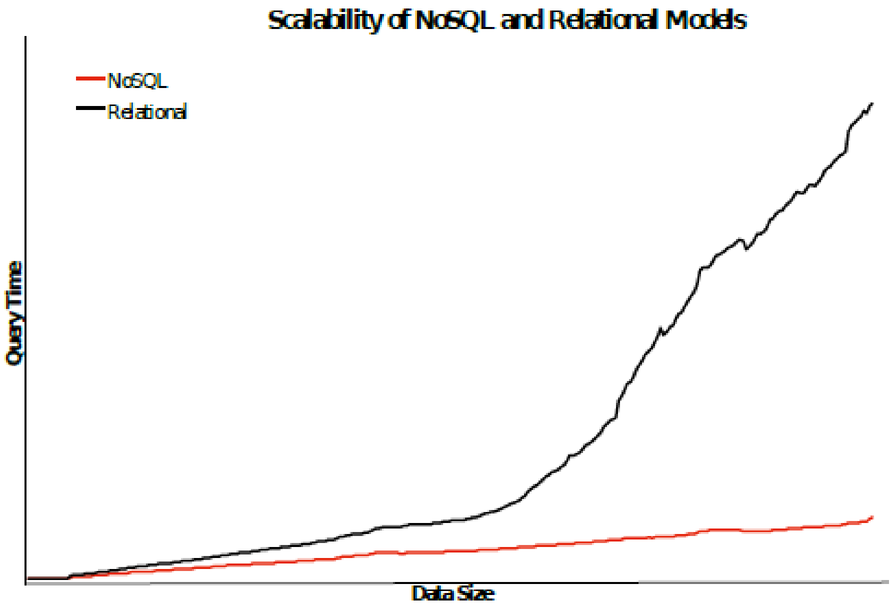
- Document-Oriented Storage. MongoDB stores its data in JSON-style objects. This makes it very easy to store raw documents from Twitter's APIs.

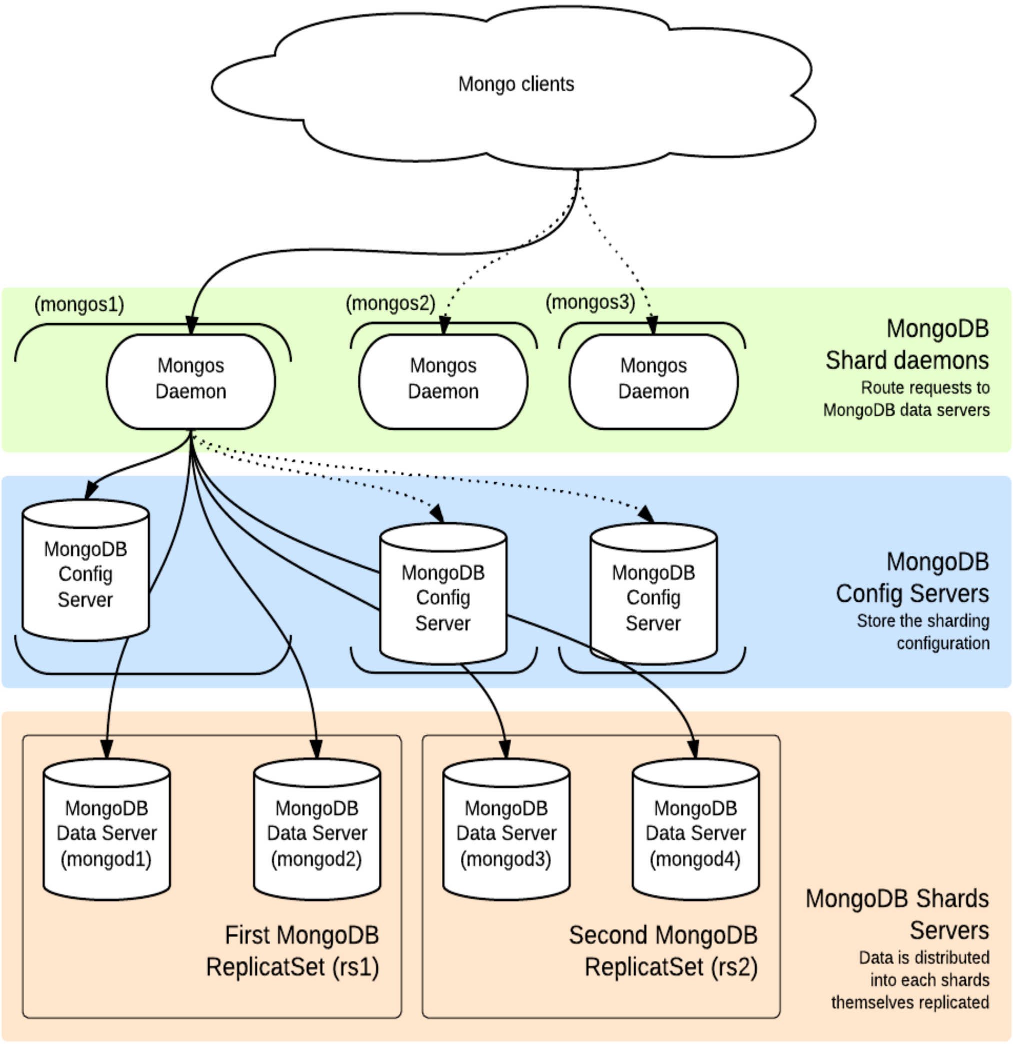
- Index Support. MongoDB allows for indexes on any field, which makes it easy to create indexes optimized for your application.

- Straightforward Queries. MongoDB's queries, while syntactically much different from SQL, are semantically very similar. In addition, MongoDB supports MapReduce, which allows for easy lookups in the data. Speed. Figure below shows a brief comparison of query speed between the relational model and MongoDB.

Fig.2.4: Storage space occupied by a traditional SQL Database compared to Non-SQL Database, i.e. MongoDB (below)

Fig.2.5: MongoDB Architecture (far below)


Scalability of NoSQL and Relational Models
NoSQL
Relational
Query Time
Data Size


Mongo clients
(mongos1)
(mongos2)
(mongos3)
Mongos Daemon
Mongos Daemon
Mongos Daemon
MongoDB Shard daemons
Route requests to MongoDB data servers
MongoDB Config Server
MongoDB Config Server
MongoDB Config Server
MongoDB Config Servers
Store the sharding configuration
MongoDB Data Server (mongod1)
MongoDB Data Server (mongod2)
MongoDB Data Server (mongod3)
MongoDB Data Server (mongod4)
First MongoDB ReplicatSet (rs1)
Second MongoDB ReplicatSet (rs2)
MongoDB Shards Servers
Data is distributed into each shards themselves replicated

4. HTML / CSS / PHP:

Various ongoing trends on social networking sites are aesthetically represented using plotting libraries in R and then added onto the website. The website called Footweets will contain several graphs comparing the popularity of different teams, comparing the popularity of different players, the follower's distribution map of different teams, the premier league table, the list of fixtures and results, list of popular hashtags fans used to mention their team, discussion forum and general information about the next fixtures, about the teams and the logo of the website. Data is added onto the website by connecting MongoDB to html.

2.2 Non Functional Requirements

1. Register the application with twitter and get the access keys.

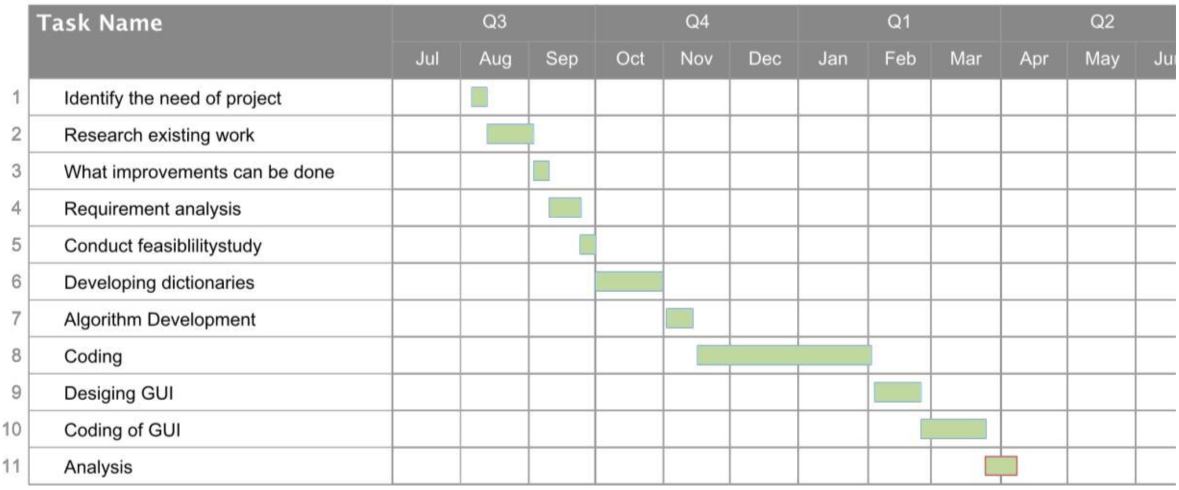
2. As a client application to the twitter, we need to provide 'consumer key', 'consumer secret' and 'access tokens'. Update the twitter4j.properties file with consumer key, consumer secret and access tokens, to access twitter though twitter4j.

3. As server side system it needs high performance CPU configuration, requires minimum 2.4 GHz processing speed with a physical memory of 3 GB.

4. Java class path set to external library jars.

5. Study the specifications and configuration setting of external libraries and API's, while integrating with user application.

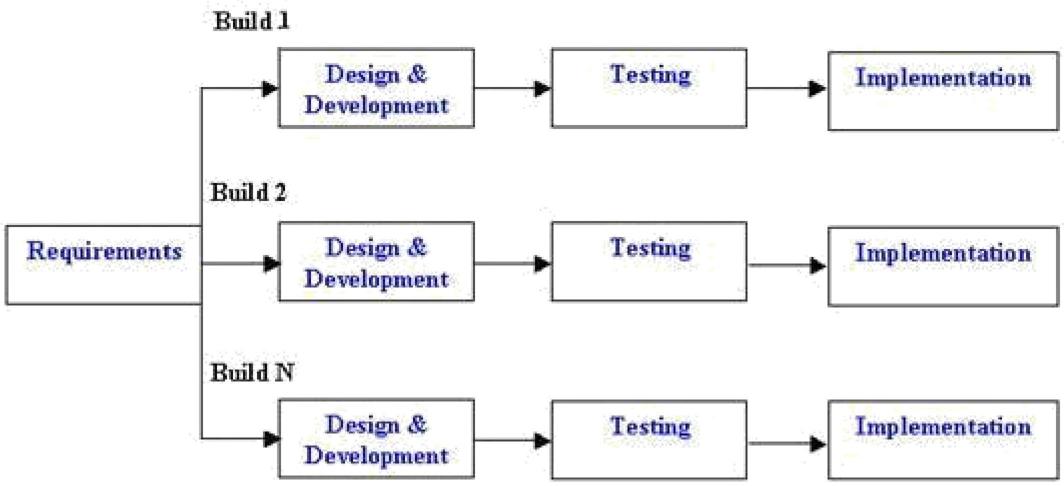
2.3 Timeline Chart:



	Task Name	Q3			Q4			Q1			Q2		
		Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Ju
1	Identify the need of project		🟩										
2	Research existing work		🟩	🟩									
3	What improvements can be done			🟩									
4	Requirement analysis			🟩									
5	Conduct feasiblilitystudy			🟩									
6	Developing dictionaries				🟩								
7	Algorithm Development					🟩							
8	Coding					🟩	🟩	🟩					
9	Desiging GUI								🟩				
10	Coding of GUI								🟩	🟩			
11	Analysis									🟥	🟥		

Fig.2.6: Timeline Chart

Fig.2.7: Incremental Model Procedure



2.4 Process Model:

Incremental Model:

In incremental model the whole requirement is divided into various builds. Multiple development cycles take place here, making the life cycle a “multi-waterfall” cycle. Cycles are divided up into smaller, more easily managed modules. Each module passes through the requirements, design, implementation and testing phases. A working version of software is produced during the first module, so we have working software early on during the software life cycle. Each subsequent release of the module adds function to the previous release. The process continues till the complete system is achieved.

At regular intervals, we will analyze the deficiencies of the project and formulate ways to improve the website. Suggestions from visitors will be accepted and sometimes incorporated into our development plan.

In our first version, we have implemented a website devoted to the evaluation and analysis of the twitter data mined in the last few months. It contains graphs showing popularity of different teams in different time periods based on tweet count mentioning the teams by their hashtags, graphs showing player popularity by tweet analysis, data about the next fixtures, the premier league table in full along with past results and details (goals scored, players who scored and assisted, red cards, etc.), followers distribution map, description of each team, discussion forum and a dashboard showing the live tweets.

In second increment, we will add sentiment analysis as one of our major features. For this, each tweet will be broken down into words and each word will be categorized as either a positive word, negative word or a neutral word. This will show what people think about the team through tweets. The next increment will contain player popularity and team popularity progression graphs. To make these graphs, we will have to analyze the whole season from start to finish, which is something we couldn’t do in the first increment. Simple line graphs depicting how the count of tweets has changed from one week to the next or one month to the next. Hence incremental process model is best suited for our project as with each increment we can add new features to the system. We can start off with a working model of the system and keep on improving it when demanded. Users can review the system and can state additional features or problems in the system, which can be added later on. The risk of changing environments is greatly reduced due to the implementation of the incremental model.

Chapter 3

Design

3.1 Data Flow Diagram

Twitter contains all the data we need. So we want to collect hashtag-specific data from twitter and display it into our website. Alas, it isn't that straightforward. Data flows in this way

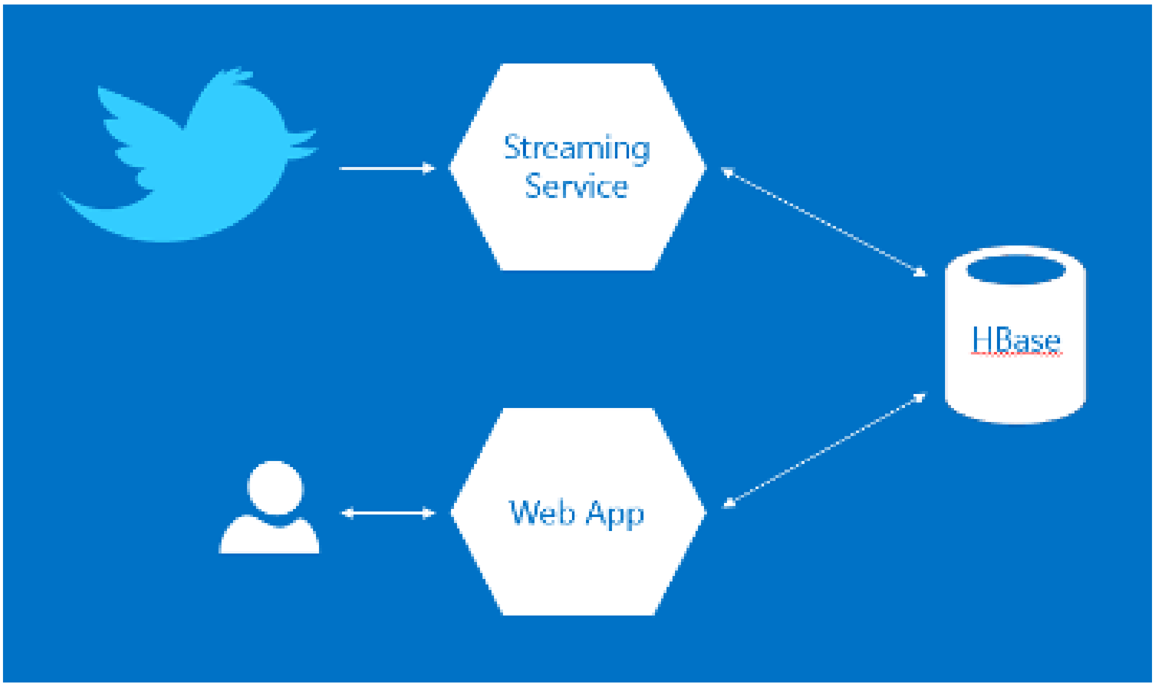
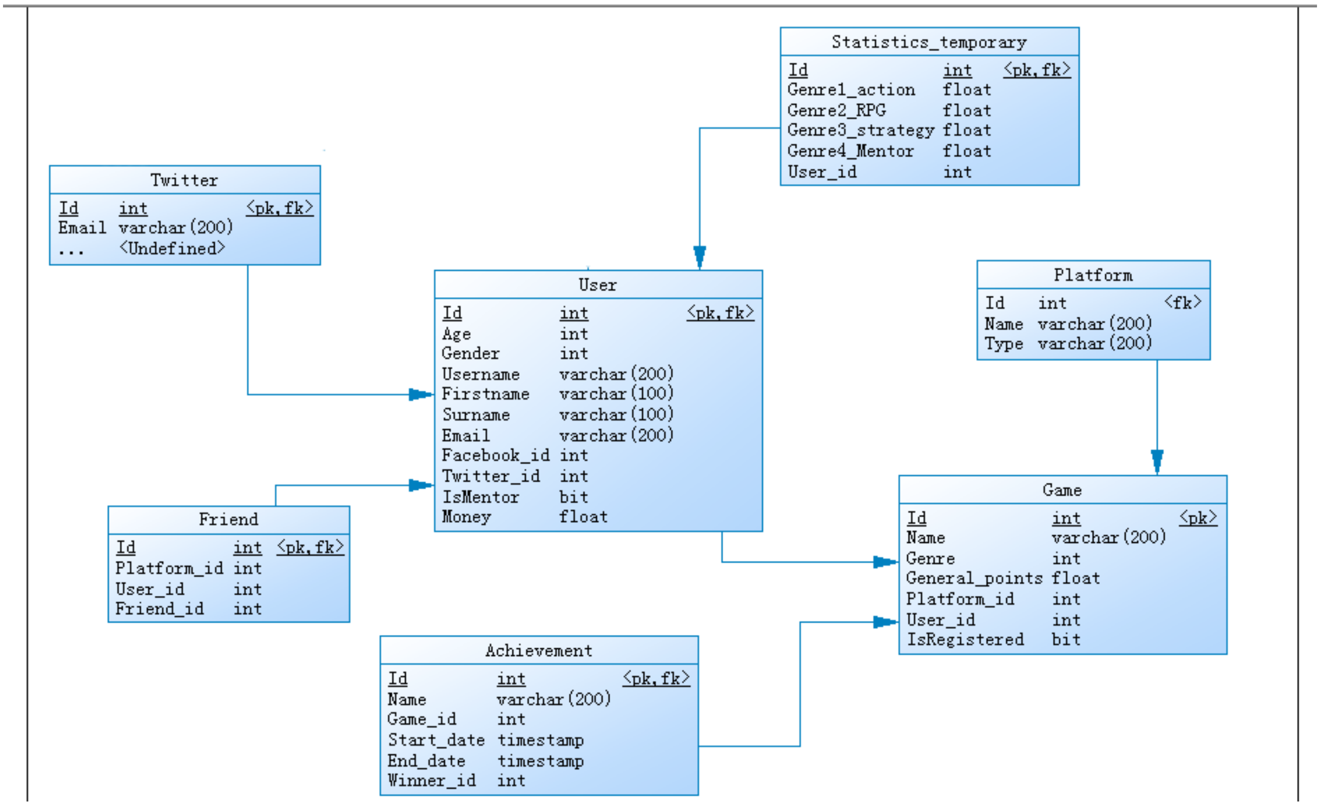


Fig.3.1: Where does data flow towards?

Through twitter's rest API or streaming API, we can extract the tweets we need by applying certain constraints on the code. RStudio extracts the tweets. We also used some java code to make the process quick and automatic. We ran the code on IntelliJ, which is a cross-code platform enabling us to integrate different softwares and coding languages. Then we stored the tweets on a database. Instead of HBase, we used MongoDB. MongoDB according to our observations occupied less space. We could also start the database easily. Then we added all of this data to our web application. Hence the main objectives of our project is **Content Retrieval, Data Processing, Storage, Data Analysis and Visualization**.

3.2: E R Diagram



Statistics_temporary		
Id	int	<pk,fk>
Genre1_action	float	
Genre2_RPG	float	
Genre3_strategy	float	
Genre4_Mentor	float	
User_id	int	

Twitter		
Id	int	<pk,fk>
Email	varchar(200)	
...	<Undefined>	

User		
Id	int	<pk,fk>
Age	int	
Gender	int	
Username	varchar(200)	
Firstname	varchar(100)	
Surname	varchar(100)	
Email	varchar(200)	
Facebook_id	int	
Twitter_id	int	
IsMentor	bit	
Money	float	

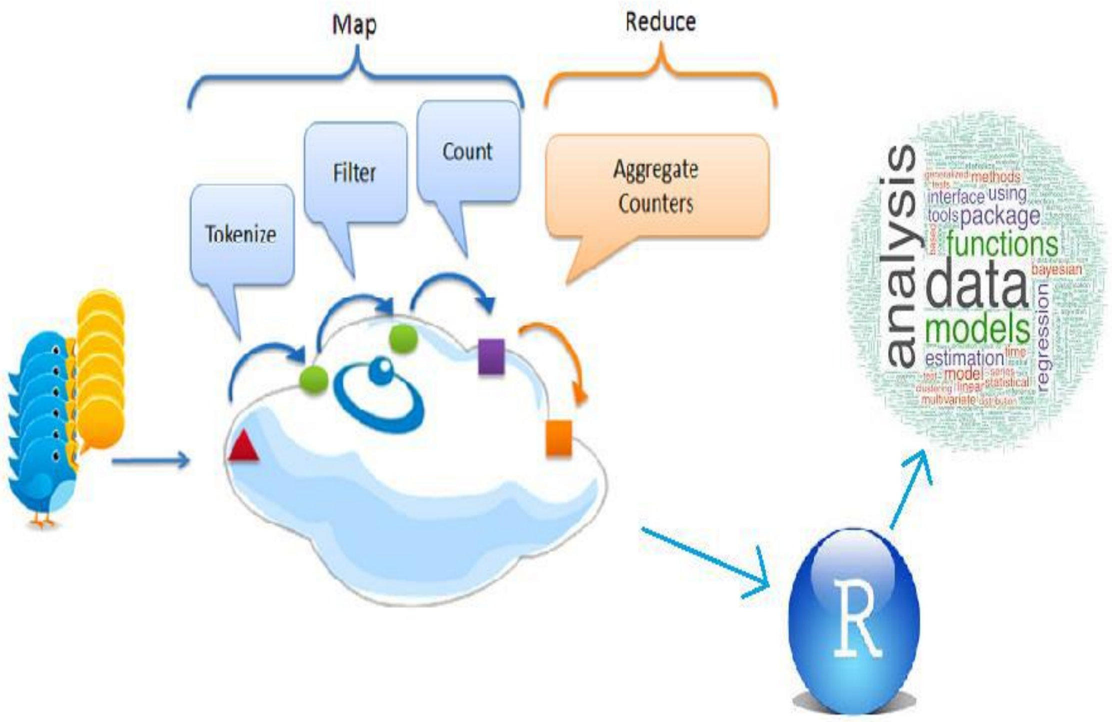
Platform		
Id	int	<fk>
Name	varchar(200)	
Type	varchar(200)	

Friend		
Id	int	<pk,fk>
Platform_id	int	
User_id	int	
Friend_id	int	

Game		
Id	int	<pk>
Name	varchar(200)	
Genre	int	
General_points	float	
Platform_id	int	
User_id	int	
IsRegistered	bit	

Achievement		
Id	int	<pk,fk>
Name	varchar(200)	
Game_id	int	
Start_date	timestamp	
End_date	timestamp	
Winner_id	int	

3.3: System Flowchart:



The system's architecture doesn't show the front-end, i.e. the content on the website. The database used for the storage of tweets is MongoDB. The MongoDB file classification system can be shown like this:

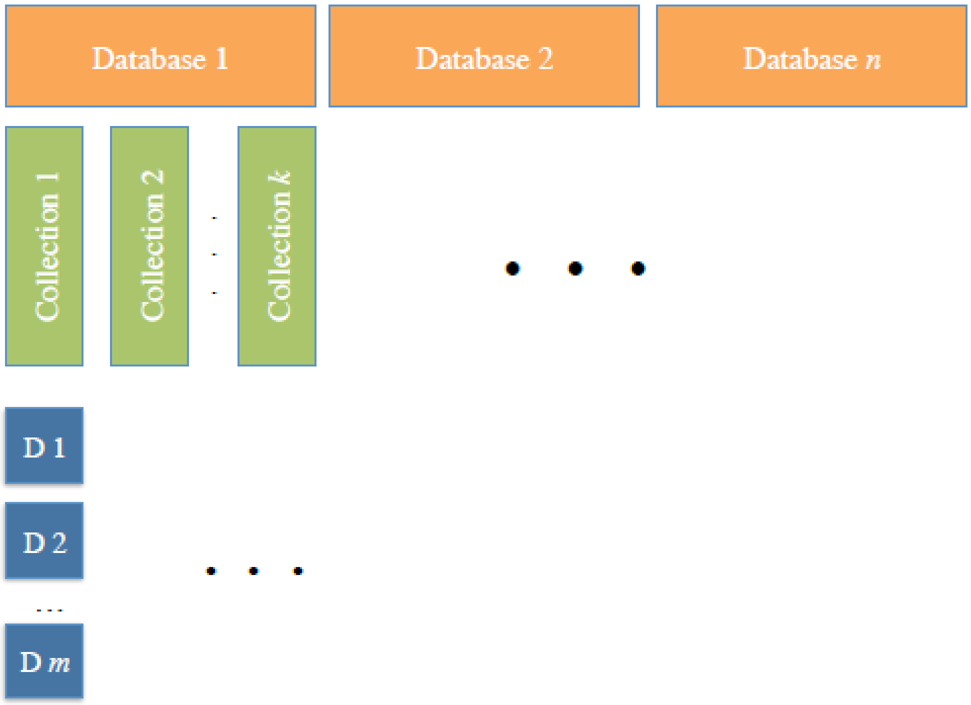


Fig. 3.2: Organization of MongoDB data.

MongoDB stores its files as collections of data. A sum of data streams make up a collection. We used each data stream to store the set of tweets for each team.

3.4: Feasibility Analysis:

A web application on data mining surely is a feasible option for a variety of reasons. Through this project, we have identified different patterns of data from social media, which is the primary objective of data mining and provided a platform to showcase the analysis and findings.

Different graphs and interesting news reports, and providing the premier league points table on the website for easy viewing means that the website is a good option for football fans to visit, even on a daily basis. What makes us different is that the website is more conceptual and unique since no other football website showcases data mined from twitter on their website. This will attract scholars and stakeholders of football clubs or sponsors who want to see what the fans are talking about, which could be extremely important from a marketing perspective. Over a period of time, if fans engage in a healthy discussion on their website, it will only bring in more fame and attraction.

If there is a bigger and faster need for processed and analysed data, regular maintenance is required. However, a website and system like this is extremely easy to maintain. IntelliJ, mongodb, and RStudio can be used multiple number of times a day. Costs involved are also minimal as all these softwares are open source softwares.

Regular updates to the website will be done to make it interesting for the user. Fans would like to see changes sometimes and we need changes to adapt to the needs of modern life. Considering that the website becomes popular, ads will start creeping into the sides of the website, and that could give us some extra income as well.

3.5: Technical Analysis

Some of the major limitations with the project is that the system lacks a server, which could extract tweets pertaining to different conditions and store it into different data frames. With the current system, one full computer is required to extract the tweets of one team and store it into one data frame. This does reduce the effectiveness of the comparisons of different teams since we cannot compare all the teams at the same time. Hence we used 4 computers when their matches were going on and compared each team for 30 minutes. The whole process ended in 2 and a half hours. The process isn’t perfect but the results seemed good enough and expected.

Another major limitation is the problem with Rest API’s limit. This means we cannot compare big hashtags like the teams after the time has passed. Hence we used this only for the less popular hashtags. This major limitation doesn’t allow us to compare the teams on a season wide basis. However, if there are multiple computers present at the time of evaluation, then it’s possible to evaluate tweets every day or every week and find out the results or the change in popularity. This is where a very powerful computer like a server would come handy.

What initially posed to be a problem, mongo dB silenced us and worked perfectly. The collections don’t take up much space (1 hour tweet analysis of 20 teams occupied only 0.01 GB), and starting mongo dB was very fast as well. It also worked perfectly in sync with html and php and hence we were able to show our findings on to the website easily.

RStudio, expectedly was the most important software for this project. We stored the count of the number of tweets referring each and every team, or the popular players and performed comparisons. RStudio has inbuilt packages, which allowed us to make graphs and scatterplots in png format so we could easily show those to the website as well. We were also able to create followers maps which proved to be very difficult but highly rewarding as the findings were extremely interesting.

Overall, the technical aspects of the project seems alright and we can conclude that is feasible to continue this project on the long run with the same or similar technical system.

3.6: Algorithm

An algorithm in short depicting the most important aspect of our project: Tweet mining.

1. Connect to twitter and fetch tweets from a geographic location.
    - a. Do not duplicate the fetched tweets.
    - b. Retrieve the metadata of each tweet along with text content.
        1. Tweet Id.
        2. Sender Id.
        3. Receiver Id.
        4. Sender Name.
        5. Receiver Name.
        6. Date and time of tweet creation.
        7. Geo Coordinates (latitude and longitude).
        8. Sending source.
        9. Sender’s place.
        10. Actual text of the tweet.
    - c. Send http request to twitter for every 1 minute interval to fetch the tweets.

2. Make followers map of the tweets which are geo tagged

a. Using the geo coordinates, show each team’s followers with arrows pointing to the location of the tweet

3. Perform analysis on the count of tweets extracted for different topics in R and find out

a. Most popular team

b. Most popular player

c. Match popularity

d. Manager popularity.

Chapter 4

Implementation

4.1 Features:

Three main features have been implemented in our project are as follows:

1. Comparisons through different graphs, maps and word clouds

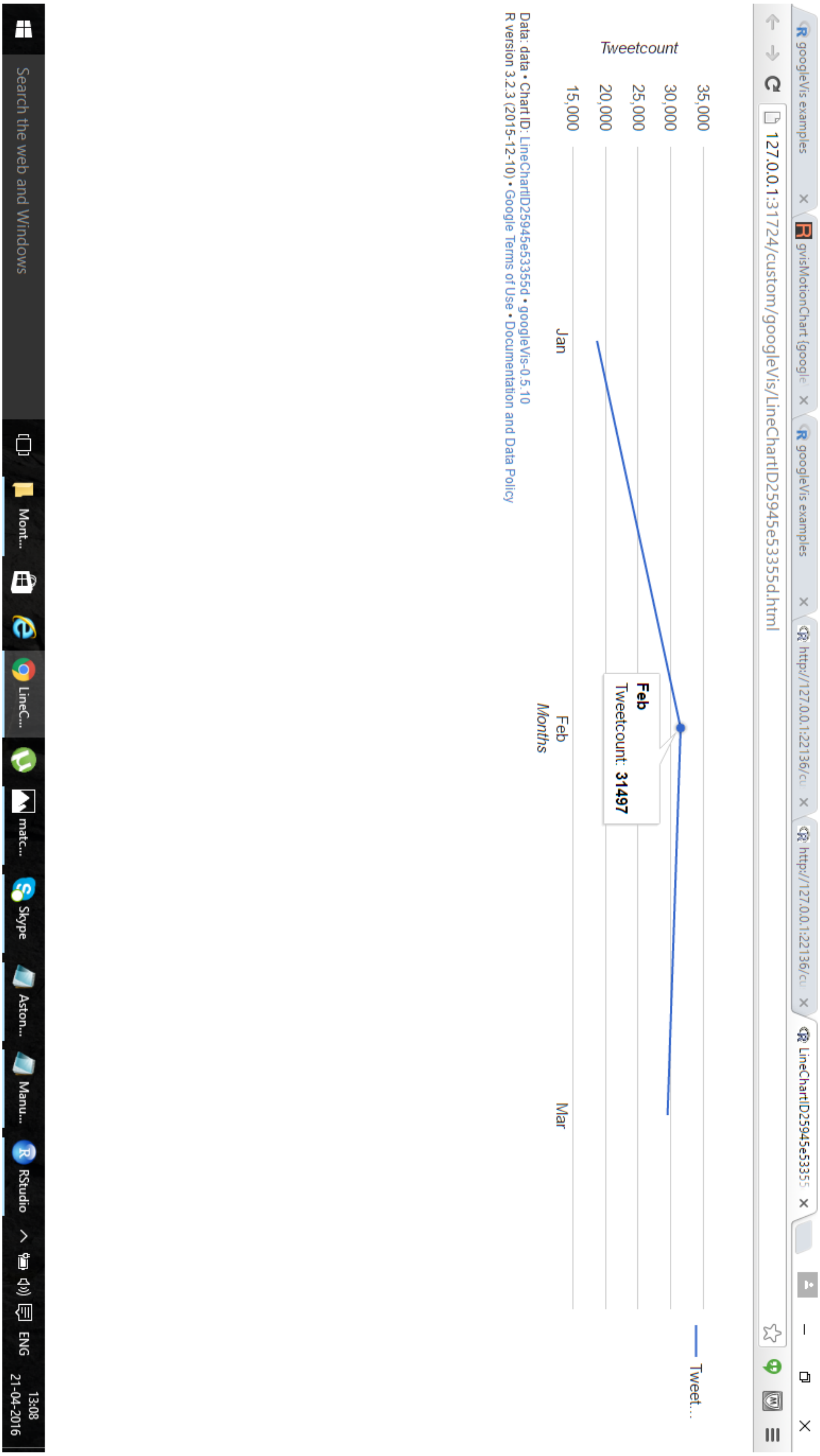
The main feature and aim of the project was to provide comparisons of different teams on based on different characteristics. These characteristics were analyzed by counting the number of tweets for various teams, players, users, etc. Graphs were then generated by RStudio because of its ggplot2 and googletextplot package. These packages allowed us to show the results of comparisons through interesting ways like bar graphs, text plots, word clouds, followers' maps, donut charts, and so on.

Things that can be shown through statistical diagrams are:

- Most popular team
- Most popular player
- Most popular match in a game week
- Most number of followers in a given area of the world
- Popular hashtags and their relative frequencies
- Most popular head coaches / managers.

Manchester United Tweet count line-graph (Monthwise)

Fig.4.1: Manchester United Tweet count graph



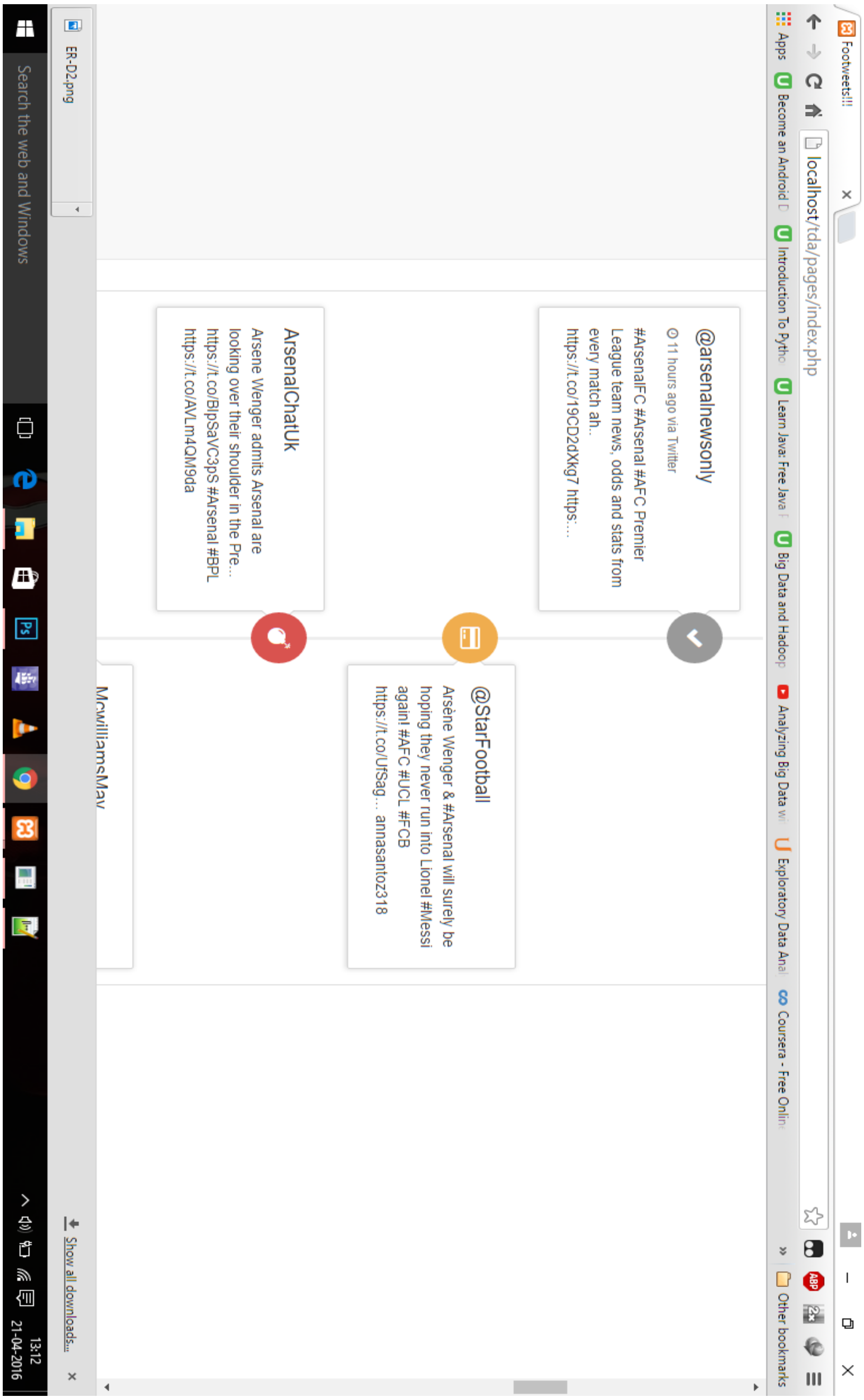
127.0.0.1:31724/custom/googleVis/LineChartID25945e53355d.html

Feb
Tweetcount: 31497

Data: data • Chart ID: LineChartID25945e53355d • googleVis-0.5.10
R version 3.2.3 (2015-12-10) • Google Terms of Use • Documentation and Data Policy

2. Display of popular tweets

Since we are mining tweets from twitter, we have chosen to show these tweets in a side bar so that people can read them. Tweets are connected to the website through a mongodb-php driver and situated on the side. Tweets with the hashtag #epl will be shown on this sidebar. This is actually a source of news as well, so we are sure that this will keep users interested.



@arsenalnewsonly

11 hours ago via Twitter

#ArsenalFC #Arsenal #AFC Premier League team news, odds and stats from every match ah. https://t.co/19CD2dXkg7 https:...

ArsenalChatUk

Arsene Wenger admits Arsenal are looking over their shoulder in the Pre... https://t.co/BIpSaVC3pS #Arsenal #BPL https://t.co/AVLm4QM9da

@StarFootball

Arsène Wenger & #Arsenal will surely be hoping they never run into Lionel #Messi again! #AFC #UCL #FCB https://t.co/UfSag... annasantoz318

McwilliamsMav

Fig.4.2: Display of tweets on the website

3. Display of other important football information:

Information which cannot be derived or extracted from tweets, but which are important to the fans nonetheless, are displayed in various different pages on the website. Information such as upcoming fixtures in the next game week, the premier league table, results and details of past matches, general information about the teams are all included in the website. Due to this, visitors to the website will feel that they are encountering one of the most important aspects of a traditional football related website.

4.2 Datasets:

Datasets included in our projects were the tweets collected from twitter through a twitter developers' account and the premier league table, fixtures and results list collected from www.footstats.co.uk & www.fantasypremierleague.com. Tweets are stored in mongodb and other important information was either noted down or saved in a .jpg, .png, or .pdf format.

1. **MongoDB:**

MongoDB is started through the cmd. MongoDaemon starts the Mongo Process, making it fit for use. After this, we make databases. Inside these databases we make collections and each collection stores the tweet collection of a team.

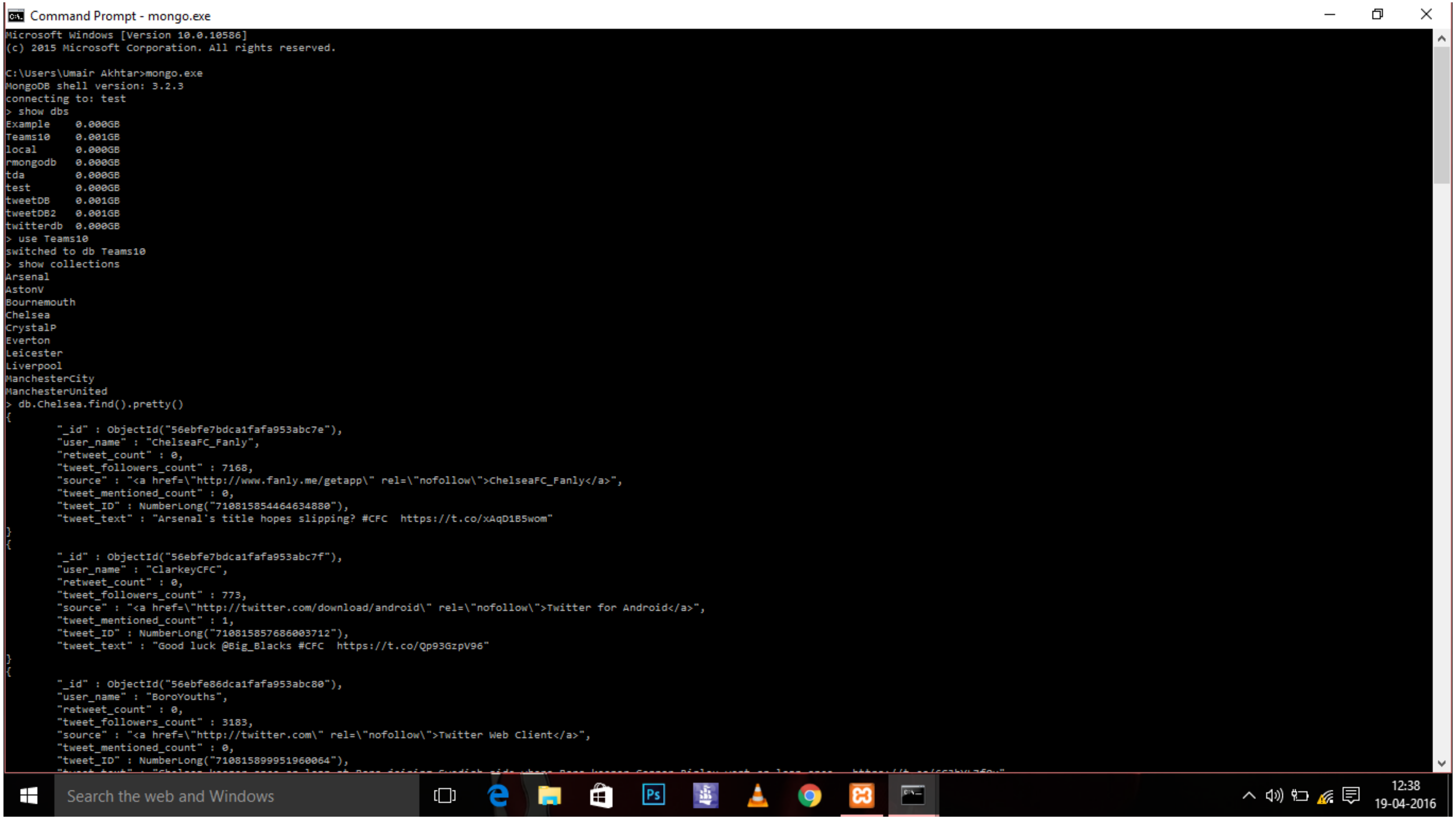
➘ Image Below

2. **IntelliJ Java Extraction**

IntelliJ IDEA software is linked up with jdk and twitterJ to extract tweets and a library for mongodb is installed with which ones the tweets are extracted they get directly stored in the collection specified on mongo database.

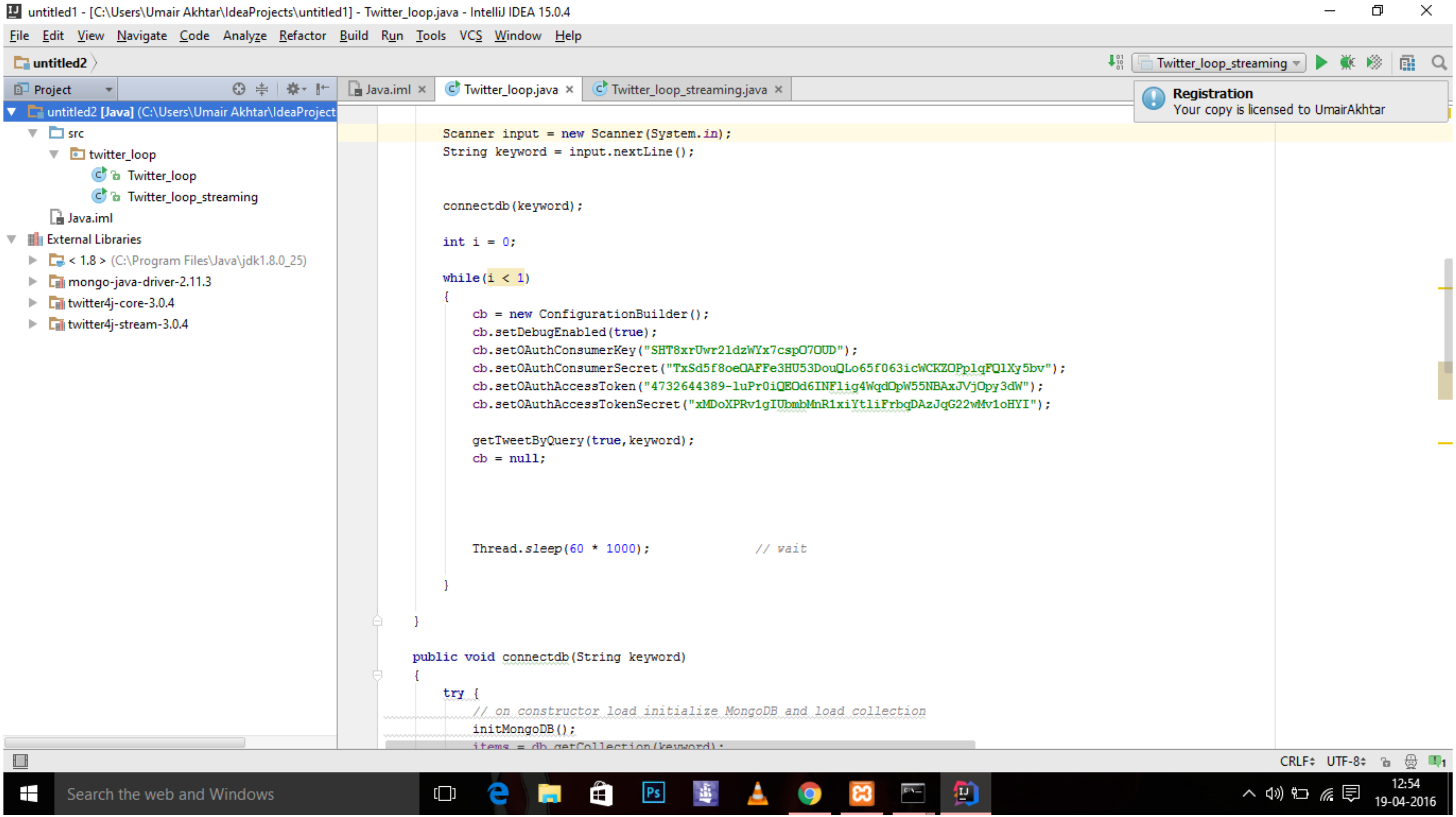
➘ Image Below

Fig.4.3: MongoDB: Showing Collections in which tweets are stored.



```
Microsoft Windows [Version 10.0.10586]
(c) 2015 Microsoft Corporation. All rights reserved.

C:\Users\Umair Akhtar>mongo.exe
MongoDB shell version: 3.2.3
connecting to: test
> show dbs
Example    0.000GB
Teams10    0.001GB
local      0.000GB
rmongodb   0.000GB
tda        0.000GB
test       0.000GB
tweetDB    0.001GB
tweetDB2   0.001GB
twitterdb  0.000GB
> use Teams10
switched to db Teams10
> show collections
Arsenal
AstonV
Bournemouth
Chelsea
CrystalP
Everton
Leicester
Liverpool
ManchesterCity
ManchesterUnited
> db.Chelsea.find().pretty()
{
        "_id" : ObjectId("56ebfe7bdca1fafa953abc7e"),
        "user_name" : "ChelseaFC_Fanly",
        "retweet_count" : 0,
        "tweet_followers_count" : 7168,
        "source" : "<a href=\"http://www.fanly.me/getapp\" rel=\"nofollow\">ChelseaFC_Fanly</a>",
        "tweet_mentioned_count" : 0,
        "tweet_ID" : NumberLong("710815854464634880"),
        "tweet_text" : "Arsenal's title hopes slipping? #CFC  https://t.co/xAqD1B5wom"
}
{
        "_id" : ObjectId("56ebfe7bdca1fafa953abc7f"),
        "user_name" : "ClarkeyCFC",
        "retweet_count" : 0,
        "tweet_followers_count" : 773,
        "source" : "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",
        "tweet_mentioned_count" : 1,
        "tweet_ID" : NumberLong("710815857686003712"),
        "tweet_text" : "Good luck @Big_Blacks #CFC  https://t.co/Qp93GzpV96"
}
{
        "_id" : ObjectId("56ebfe86dca1fafa953abc80"),
        "user_name" : "BoroYouths",
        "retweet_count" : 0,
        "tweet_followers_count" : 3183,
        "source" : "<a href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>",
        "tweet_mentioned_count" : 0,
        "tweet_ID" : NumberLong("710815899951960064"),
```

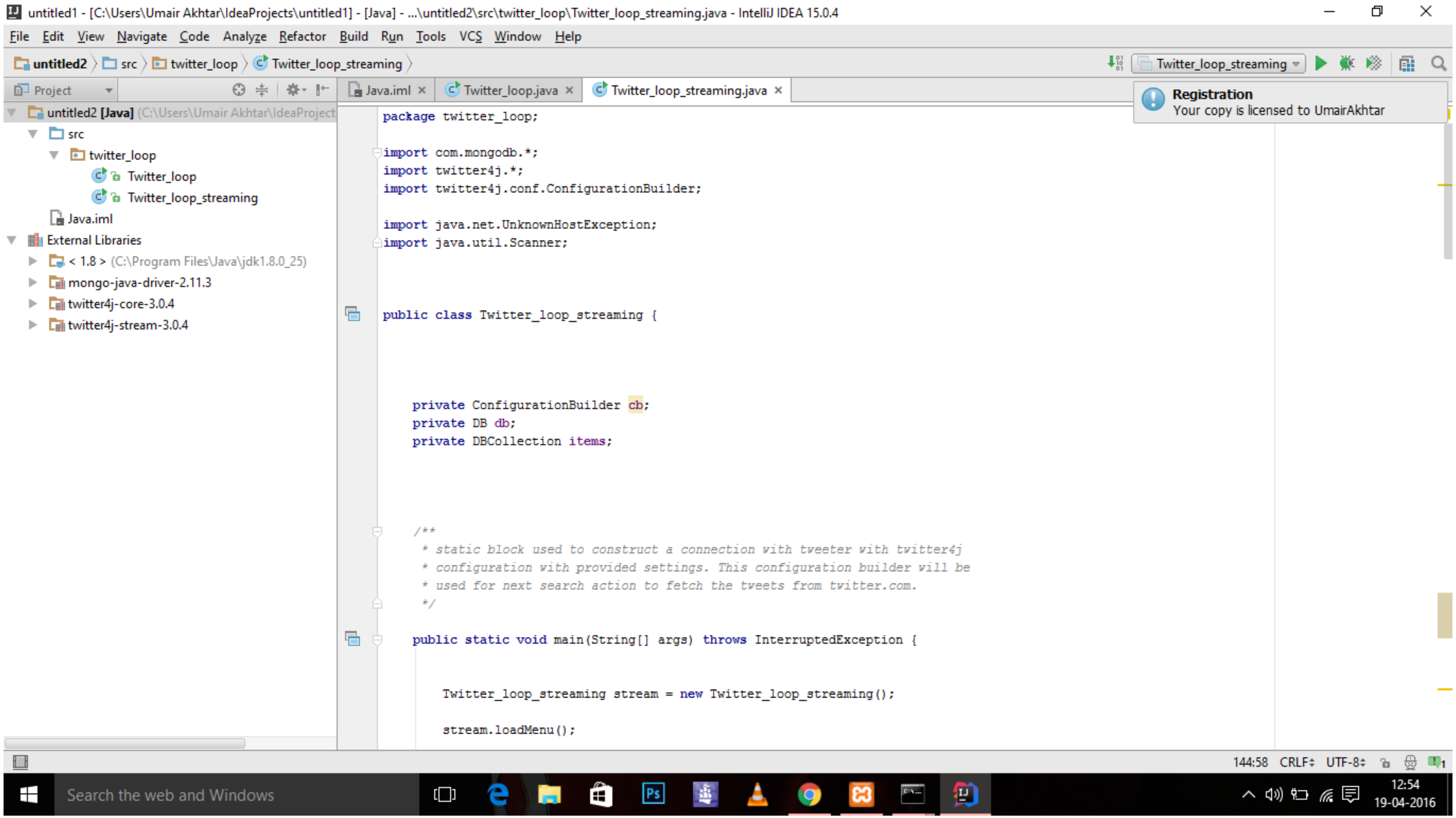
Fig.4.4: IntelliJ: Showing Java code synced with MongoDB database. Contains code to implement both Rest and Streaming API



untitled1 - [C:\Users\Umair Akhtar\IdeaProjects\untitled1] - Twitter_loop.java - IntelliJ IDEA 15.0.4
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
untitled2
Twitter_loop_streaming
Project
Java.iml
Twitter_loop.java
Twitter_loop_streaming.java
untitled2 [Java] (C:\Users\Umair Akhtar\IdeaProject
src
twitter_loop
Twitter_loop
Twitter_loop_streaming
Java.iml
External Libraries
< 1.8 > (C:\Program Files\Java\jdk1.8.0_25)
mongo-java-driver-2.11.3
twitter4j-core-3.0.4
twitter4j-stream-3.0.4
Registration
Your copy is licensed to UmairAkhtar
CRLF÷ UTF-8÷
Search the web and Windows
12:54
19-04-2016

```
        Scanner input = new Scanner(System.in);
        String keyword = input.nextLine();


        connectdb(keyword);

        int i = 0;

        while(i < 1)
        {
            cb = new ConfigurationBuilder();
            cb.setDebugEnabled(true);
            cb.setOAuthConsumerKey("SHT8xrUwr2ldzWYx7cspO7OUD");
            cb.setOAuthConsumerSecret("TxSd5f8oeOAFFe3HU53DouQLo65f063icWCKZOPplqFQlXy5bv");
            cb.setOAuthAccessToken("4732644389-luPr0iQEOd6INFlig4WqdOpW55NBAxJVjOpy3dW");
            cb.setOAuthAccessTokenSecret("xMDoXPRv1gIUbmbMnR1xiYtliFrbqDAzJqG22wMv1oHYI");

            getTweetByQuery(true,keyword);
            cb = null;



            Thread.sleep(60 * 1000);                // wait

        }

    }

    public void connectdb(String keyword)
    {
        try {
            // on constructor load initialize MongoDB and load collection
            initMongoDB();
            items = db.getCollection(keyword);
```

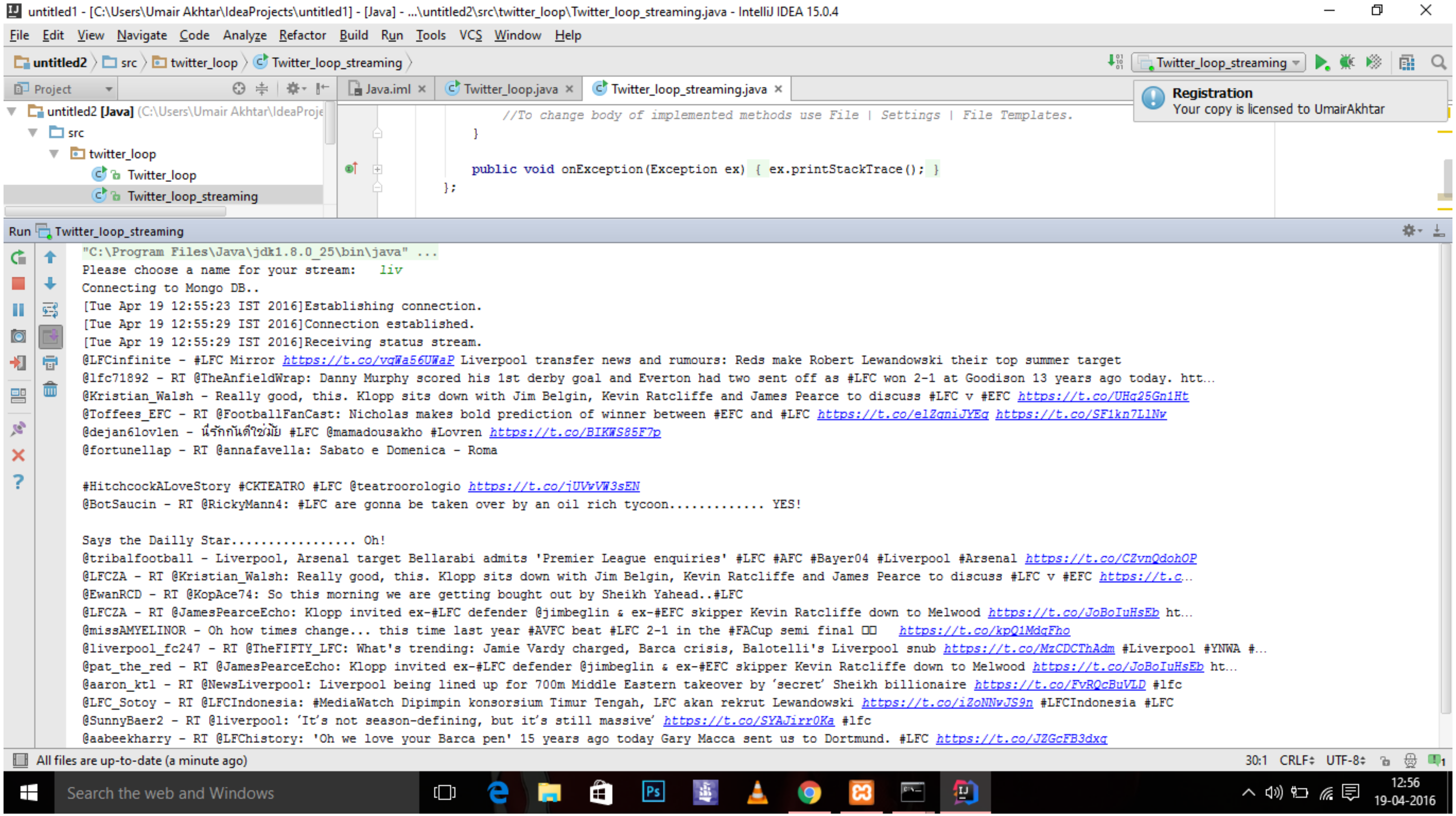
Fig.4.5: IntelliJ: Contains Twitter API Codes


untitled1 - [C:\Users\Umair Akhtar\IdeaProjects\untitled1] - [Java] - ...\untitled2\src\twitter_loop\Twitter_loop_streaming.java - IntelliJ IDEA 15.0.4
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
untitled2 〉 src 〉 twitter_loop 〉 Twitter_loop_streaming 〉
Twitter_loop_streaming
Project
Java.iml ×
Twitter_loop.java ×
Twitter_loop_streaming.java ×
untitled2 [Java] (C:\Users\Umair Akhtar\IdeaProject
src
twitter_loop
Twitter_loop
Twitter_loop_streaming
Java.iml
External Libraries
< 1.8 > (C:\Program Files\Java\jdk1.8.0_25)
mongo-java-driver-2.11.3
twitter4j-core-3.0.4
twitter4j-stream-3.0.4
Search the web and Windows
144:58 CRLF÷ UTF-8÷
12:54
19-04-2016

Registration
Your copy is licensed to UmairAkhtar

```
package twitter_loop;

import com.mongodb.*;
import twitter4j.*;
import twitter4j.conf.ConfigurationBuilder;

import java.net.UnknownHostException;
import java.util.Scanner;


public class Twitter_loop_streaming {



    private ConfigurationBuilder cb;
    private DB db;
    private DBCollection items;



    /**
     * static block used to construct a connection with tweeter with twitter4j
     * configuration with provided settings. This configuration builder will be
     * used for next search action to fetch the tweets from twitter.com.
     */
    public static void main(String[] args) throws InterruptedException {


        Twitter_loop_streaming stream = new Twitter_loop_streaming();

        stream.loadMenu();
```

Fig.4.6: IntelliJ: Tweets are extracted and shown below..



untitled1 - [C:\Users\Umair Akhtar\IdeaProjects\untitled1] - [Java] - ...\untitled2\src\twitter_loop\Twitter_loop_streaming.java - IntelliJ IDEA 15.0.4
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
untitled2 src twitter_loop Twitter_loop_streaming
Twitter_loop_streaming
Project
Java.iml
Twitter_loop.java
Twitter_loop_streaming.java
untitled2 [Java] (C:\Users\Umair Akhtar\IdeaProje
src
twitter_loop
Twitter_loop
Twitter_loop_streaming
Registration
Your copy is licensed to UmairAkhtar
Run Twitter_loop_streaming
All files are up-to-date (a minute ago)
30:1 CRLF UTF-8
Search the web and Windows
12:56
19-04-2016

```
        //To change body of implemented methods use File | Settings | File Templates.
    }

    public void onException(Exception ex) { ex.printStackTrace(); }
};
```

```
"C:\Program Files\Java\jdk1.8.0_25\bin\java" ...
Please choose a name for your stream:   liv
Connecting to Mongo DB..
[Tue Apr 19 12:55:23 IST 2016]Establishing connection.
[Tue Apr 19 12:55:29 IST 2016]Connection established.
[Tue Apr 19 12:55:29 IST 2016]Receiving status stream.
@LFCinfinite - #LFC Mirror https://t.co/vqWa56UWaP Liverpool transfer news and rumours: Reds make Robert Lewandowski their top summer target
@lfc71892 - RT @TheAnfieldWrap: Danny Murphy scored his 1st derby goal and Everton had two sent off as #LFC won 2-1 at Goodison 13 years ago today. htt…
@Kristian_Walsh - Really good, this. Klopp sits down with Jim Belgin, Kevin Ratcliffe and James Pearce to discuss #LFC v #EFC https://t.co/UHq25Gn1Ht
@Toffees_EFC - RT @FootballFanCast: Nicholas makes bold prediction of winner between #EFC and #LFC https://t.co/elZqniJYEq https://t.co/SF1kn7L1Nw
@dejan6lovlen - นี่รักกันดีใช่มั้ย #LFC @mamadousakho #Lovren https://t.co/BIKWS85F7p
@fortunellap - RT @annafavella: Sabato e Domenica - Roma

#HitchcockALoveStory #CKTEATRO #LFC @teatroorologio https://t.co/jUVvVW3sEN
@BotSaucin - RT @RickyMann4: #LFC are gonna be taken over by an oil rich tycoon............. YES!

Says the Dailly Star................. Oh!
@tribalfootball - Liverpool, Arsenal target Bellarabi admits 'Premier League enquiries' #LFC #AFC #Bayer04 #Liverpool #Arsenal https://t.co/CZvnQdohOP
@LFCZA - RT @Kristian_Walsh: Really good, this. Klopp sits down with Jim Belgin, Kevin Ratcliffe and James Pearce to discuss #LFC v #EFC https://t.c…
@EwanRCD - RT @KopAce74: So this morning we are getting bought out by Sheikh Yahead..#LFC
@LFCZA - RT @JamesPearceEcho: Klopp invited ex-#LFC defender @jimbeglin & ex-#EFC skipper Kevin Ratcliffe down to Melwood https://t.co/JoBoIuHsEb ht…
@missAMYELINOR - Oh how times change... this time last year #AVFC beat #LFC 2-1 in the #FACup semi final □□  https://t.co/kpQ1MdqFho
@liverpool_fc247 - RT @TheFIFTY_LFC: What's trending: Jamie Vardy charged, Barca crisis, Balotelli's Liverpool snub https://t.co/MzCDCThAdm #Liverpool #YNWA #…
@pat_the_red - RT @JamesPearceEcho: Klopp invited ex-#LFC defender @jimbeglin & ex-#EFC skipper Kevin Ratcliffe down to Melwood https://t.co/JoBoIuHsEb ht…
@aaron_ktl - RT @NewsLiverpool: Liverpool being lined up for 700m Middle Eastern takeover by 'secret' Sheikh billionaire https://t.co/FvRQcBuVLD #lfc
@LFC_Sotoy - RT @LFCIndonesia: #MediaWatch Dipimpin konsorsium Timur Tengah, LFC akan rekrut Lewandowski https://t.co/iZoNNvJS9n #LFCIndonesia #LFC
@SunnyBaer2 - RT @liverpool: 'It's not season-defining, but it's still massive' https://t.co/SYAJirr0Ka #lfc
@aabeekharry - RT @LFChistory: 'Oh we love your Barca pen' 15 years ago today Gary Macca sent us to Dortmund. #LFC https://t.co/JZGcFB3dxq
```

3. Websites:

The official premier league website contains the premier league points table, which shows who is on course to win the premier league title and which 3 teams are expected to be relegated or evicted from the league. We have shown this table on our website. We have also shown each team's results throughout the season in a grid format, so that it's easy to understand and easy to view (fits in one page). On the notification bar, we have shown upcoming fixtures. In the tables section we have shown the premier league table and in the sidebar we have shown the results grid

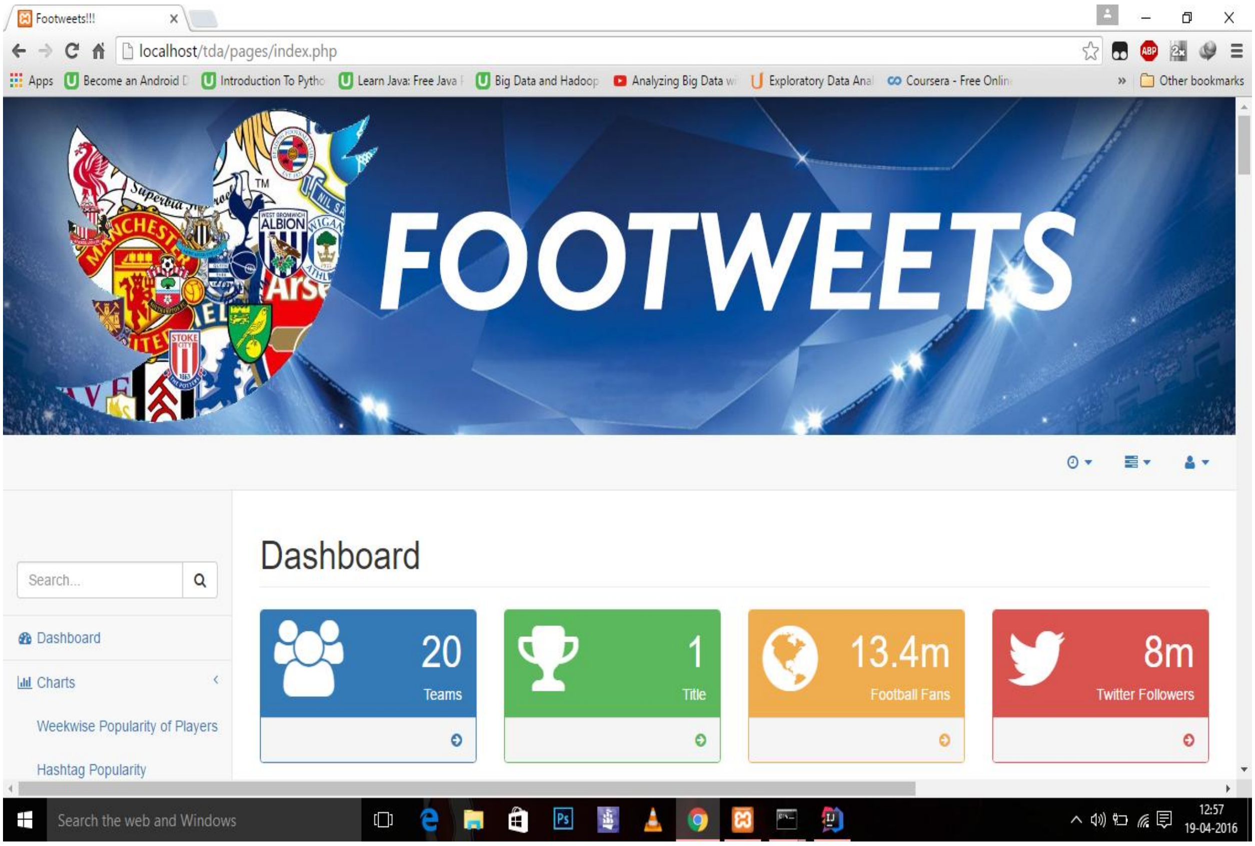
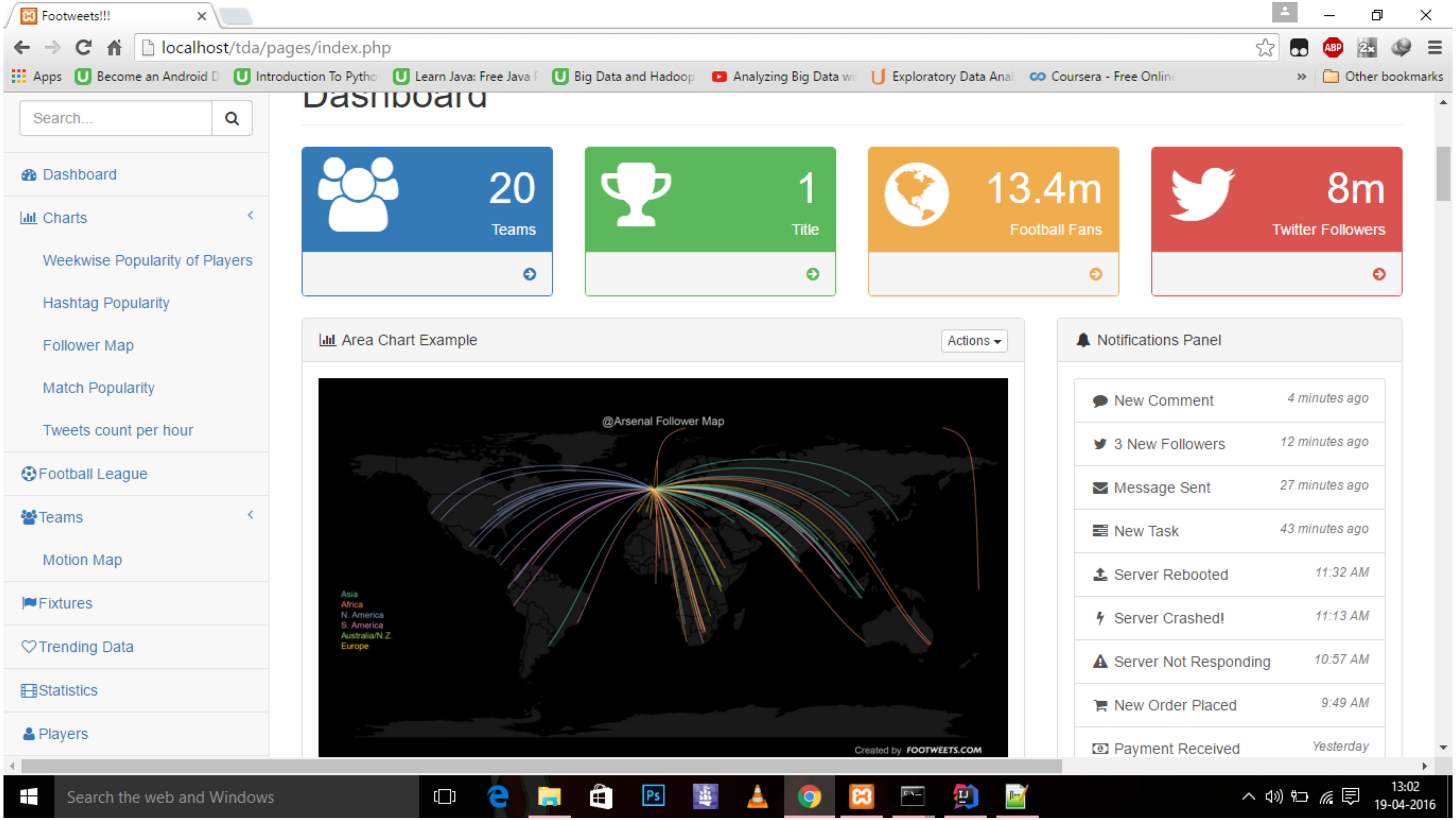


Fig.4.7: Dashboard of the Website

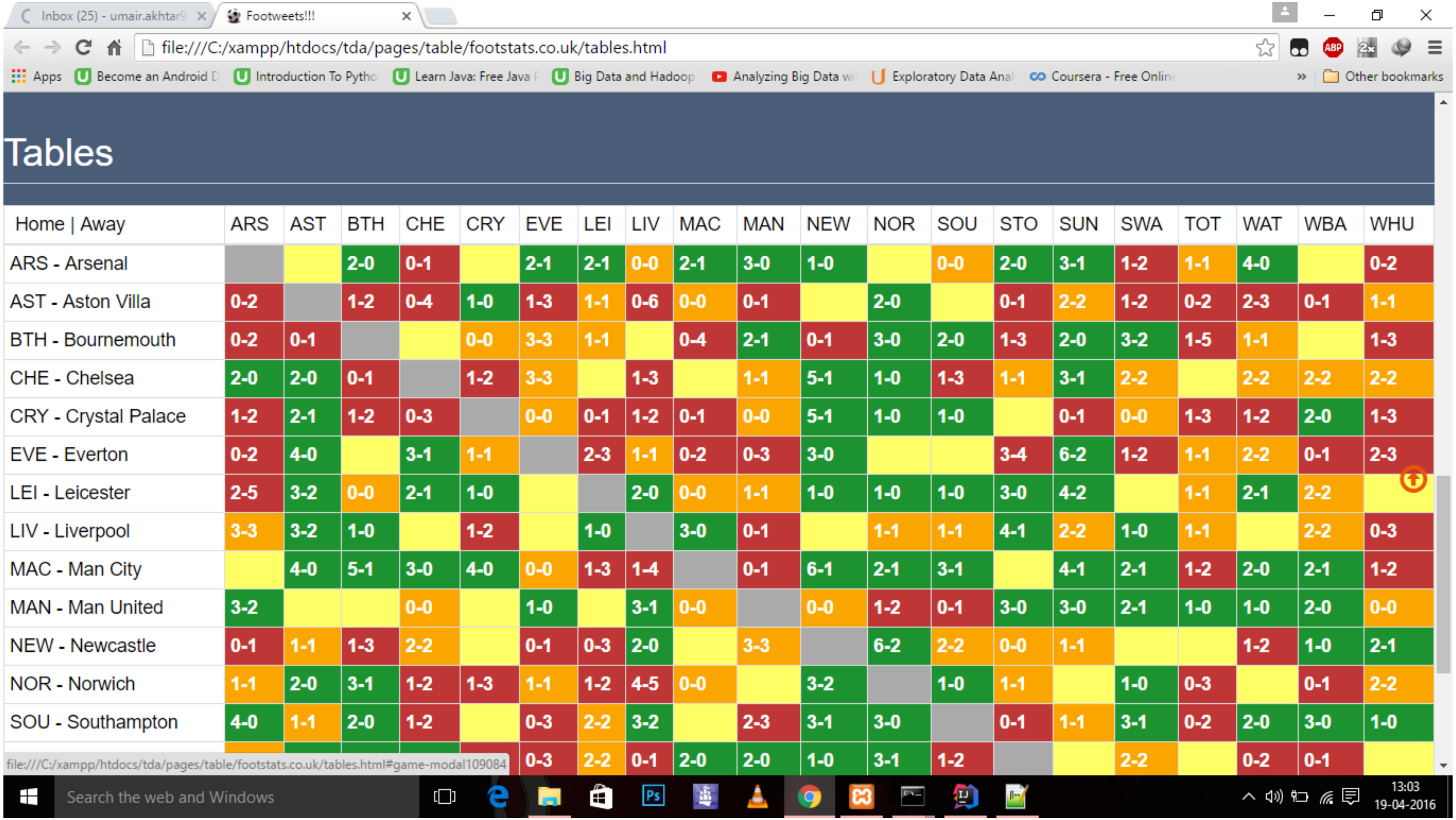
This the main page of the website. We have created the logo and the background wallpaper. The clock, when clicked, will show the next fixtures in the gameweek. The dashboard shows that there are 20 teams in the premier league, and all the 20 teams are fighting for one trophy. @Premierleague (Premier League's official twitter handle) has 8 million followers, and 13.4 million fans. The sidebar contains various options and pages showing different charts, tables, tweets, information about the teams and analysis.

Fig.4.8: Main Page. The Main Page contains a slider window: These images sow the popularity of EPL in general



New Comment 4 minutes ago
3 New Followers 12 minutes ago
Message Sent 27 minutes ago
New Task 43 minutes ago
Server Rebooted 11:32 AM
Server Crashed! 11:13 AM
Server Not Responding 10:57 AM
New Order Placed 9:49 AM
Payment Received Yesterday

Fig.4.9: The Premier League Fixtures Grid



Tables

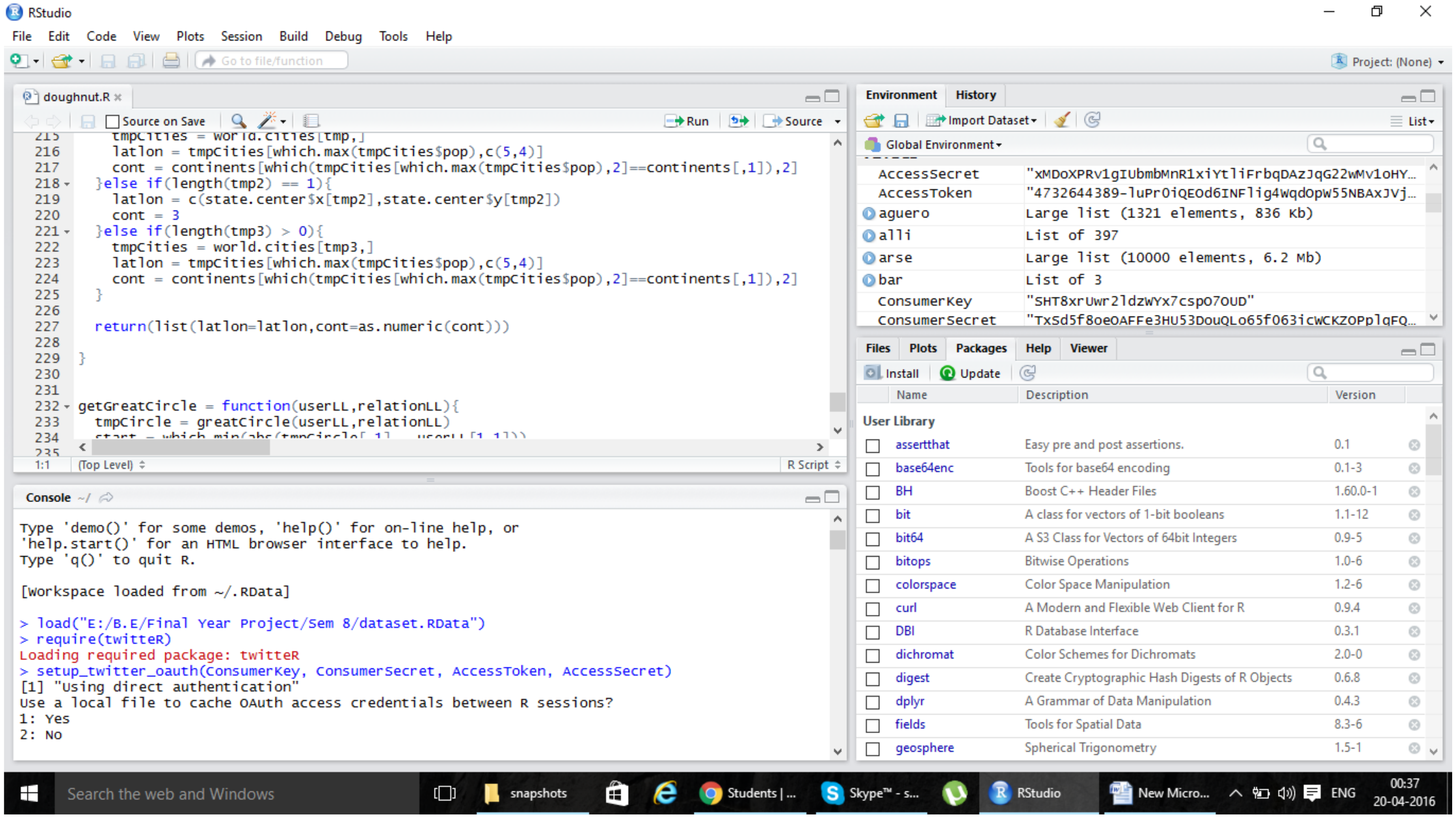
Home | Away	ARS	AST	BTH	CHE	CRY	EVE	LEI	LIV	MAC	MAN	NEW	NOR	SOU	STO	SUN	SWA	TOT	WAT	WBA	WHU
ARS - Arsenal			2-0	0-1		2-1	2-1	0-0	2-1	3-0	1-0		0-0	2-0	3-1	1-2	1-1	4-0		0-2
AST - Aston Villa	0-2		1-2	0-4	1-0	1-3	1-1	0-6	0-0	0-1		2-0		0-1	2-2	1-2	0-2	2-3	0-1	1-1
BTH - Bournemouth	0-2	0-1			0-0	3-3	1-1		0-4	2-1	0-1	3-0	2-0	1-3	2-0	3-2	1-5	1-1		1-3
CHE - Chelsea	2-0	2-0	0-1		1-2	3-3		1-3		1-1	5-1	1-0	1-3	1-1	3-1	2-2		2-2	2-2	2-2
CRY - Crystal Palace	1-2	2-1	1-2	0-3		0-0	0-1	1-2	0-1	0-0	5-1	1-0	1-0		0-1	0-0	1-3	1-2	2-0	1-3
EVE - Everton	0-2	4-0		3-1	1-1		2-3	1-1	0-2	0-3	3-0			3-4	6-2	1-2	1-1	2-2	0-1	2-3
LEI - Leicester	2-5	3-2	0-0	2-1	1-0			2-0	0-0	1-1	1-0	1-0	1-0	3-0	4-2		1-1	2-1	2-2	
LIV - Liverpool	3-3	3-2	1-0		1-2		1-0		3-0	0-1		1-1	1-1	4-1	2-2	1-0	1-1		2-2	0-3
MAC - Man City		4-0	5-1	3-0	4-0	0-0	1-3	1-4		0-1	6-1	2-1	3-1		4-1	2-1	1-2	2-0	2-1	1-2
MAN - Man United	3-2			0-0		1-0		3-1	0-0		0-0	1-2	0-1	3-0	3-0	2-1	1-0	1-0	2-0	0-0
NEW - Newcastle	0-1	1-1	1-3	2-2		0-1	0-3	2-0		3-3		6-2	2-2	0-0	1-1			1-2	1-0	2-1
NOR - Norwich	1-1	2-0	3-1	1-2	1-3	1-1	1-2	4-5	0-0		3-2		1-0	1-1		1-0	0-3		0-1	2-2
SOU - Southampton	4-0	1-1	2-0	1-2		0-3	2-2	3-2		2-3	3-1	3-0		0-1	1-1	3-1	0-2	2-0	3-0	1-0
						0-3	2-2	0-1	2-0	2-0	1-0	3-1	1-2			2-2		0-2	0-1	

file:///C:/xampp/htdocs/tda/pages/table/footstats.co.uk/tables.html#game-modal109084

4.3 Screenshots

4.3.1 Setting up OAuth using the provided Consumer key and Secret and Access token and Secret in RStudio

Fig.4.9: RStudio, OAuth Procedure



```
    tmpCities = world.cities[tmp,]
    latlon = tmpCities[which.max(tmpCities$pop),c(5,4)]
    cont = continents[which(tmpCities[which.max(tmpCities$pop),2]==continents[,1]),2]
  }else if(length(tmp2) == 1){
    latlon = c(state.center$x[tmp2],state.center$y[tmp2])
    cont = 3
  }else if(length(tmp3) > 0){
    tmpCities = world.cities[tmp3,]
    latlon = tmpCities[which.max(tmpCities$pop),c(5,4)]
    cont = continents[which(tmpCities[which.max(tmpCities$pop),2]==continents[,1]),2]
  }

  return(list(latlon=latlon,cont=as.numeric(cont)))
}


getGreatCircle = function(userLL,relationLL){
  tmpCircle = greatCircle(userLL,relationLL)
  start = which.min(abs(tmpCircle[,1] - userLL[1,1]))
```

Console ~/

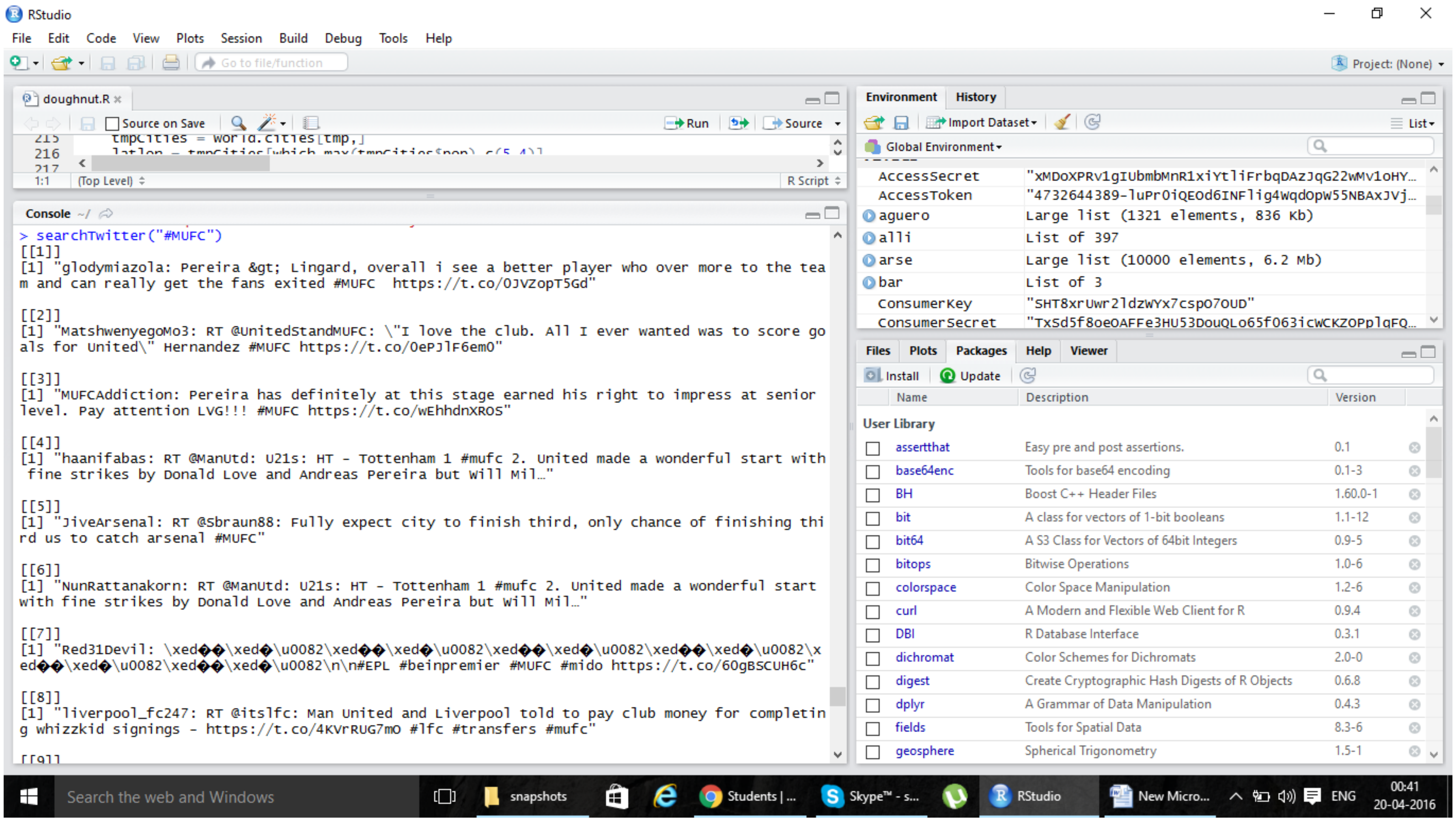
```
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> load("E:/B.E/Final Year Project/Sem 8/dataset.RData")
> require(twitteR)
Loading required package: twitteR
> setup_twitter_oauth(ConsumerKey, ConsumerSecret, AccessToken, AccessSecret)
[1] "Using direct authentication"
Use a local file to cache OAuth access credentials between R sessions?
1: Yes
2: No
```

AccessSecret	"xMDoXPRv1gIUbmbMnR1xiYtliFrbqDAzJqG22wMv1oHY…
AccessToken	"4732644389-luPr0iQEOd6INFlig4WqdOpW55NBAxJVj…
aguero	Large list (1321 elements, 836 Kb)
alli	List of 397
arse	Large list (10000 elements, 6.2 Mb)
bar	List of 3
ConsumerKey	"SHT8xrUwr2ldzWYx7csp07OUD"
ConsumerSecret	"TxSd5f8oeOAFFe3HU53DouQLo65f063icWCKZOPplqFQ…

	Name	Description	Version
	User Library		
□	assertthat	Easy pre and post assertions.	0.1
□	base64enc	Tools for base64 encoding	0.1-3
□	BH	Boost C++ Header Files	1.60.0-1
□	bit	A class for vectors of 1-bit booleans	1.1-12
□	bit64	A S3 Class for Vectors of 64bit Integers	0.9-5
□	bitops	Bitwise Operations	1.0-6
□	colorspace	Color Space Manipulation	1.2-6
□	curl	A Modern and Flexible Web Client for R	0.9.4
□	DBI	R Database Interface	0.3.1
□	dichromat	Color Schemes for Dichromats	2.0-0
□	digest	Create Cryptographic Hash Digests of R Objects	0.6.8
□	dplyr	A Grammar of Data Manipulation	0.4.3
□	fields	Tools for Spatial Data	8.3-6
□	geosphere	Spherical Trigonometry	1.5-1

4.3.2 Extracting tweets using R

Fig.4.10: RStudio, Tweet Extraction



```
215    tmpCities = world.cities[tmp,]
216    latlon = tmpCities[which.max(tmpCities$pop), c(5,4)]
1:1
```

```
> searchTwitter("#MUFC")
[[1]]
[1] "glodymiazola: Pereira &gt; Lingard, overall i see a better player who over more to the tea
m and can really get the fans exited #MUFC  https://t.co/OJVZopT5Gd"

[[2]]
[1] "MatshwenyegoMo3: RT @UnitedStandMUFC: \"I love the club. All I ever wanted was to score go
als for United\" Hernandez #MUFC https://t.co/0ePJlF6em0"

[[3]]
[1] "MUFCAddiction: Pereira has definitely at this stage earned his right to impress at senior
level. Pay attention LVG!!! #MUFC https://t.co/wEhhdnXROS"

[[4]]
[1] "haanifabas: RT @ManUtd: U21s: HT - Tottenham 1 #mufc 2. United made a wonderful start with
 fine strikes by Donald Love and Andreas Pereira but will Mil…"

[[5]]
[1] "JiveArsenal: RT @Sbraun88: Fully expect city to finish third, only chance of finishing thi
rd us to catch arsenal #MUFC"

[[6]]
[1] "NunRattanakorn: RT @ManUtd: U21s: HT - Tottenham 1 #mufc 2. United made a wonderful start
with fine strikes by Donald Love and Andreas Pereira but will Mil…"

[[7]]
[1] "Red31Devil: \xed��\xed�\u0082\xed��\xed�\u0082\xed��\xed�\u0082\xed��\xed�\u0082\x
ed��\xed�\u0082\xed��\xed�\u0082\n\n#EPL #beinpremier #MUFC #mido https://t.co/60gBSCUH6c"

[[8]]
[1] "liverpool_fc247: RT @itslfc: Man United and Liverpool told to pay club money for completin
g whizzkid signings - https://t.co/4KVrRUG7mO #lfc #transfers #mufc"

[[9]]
```

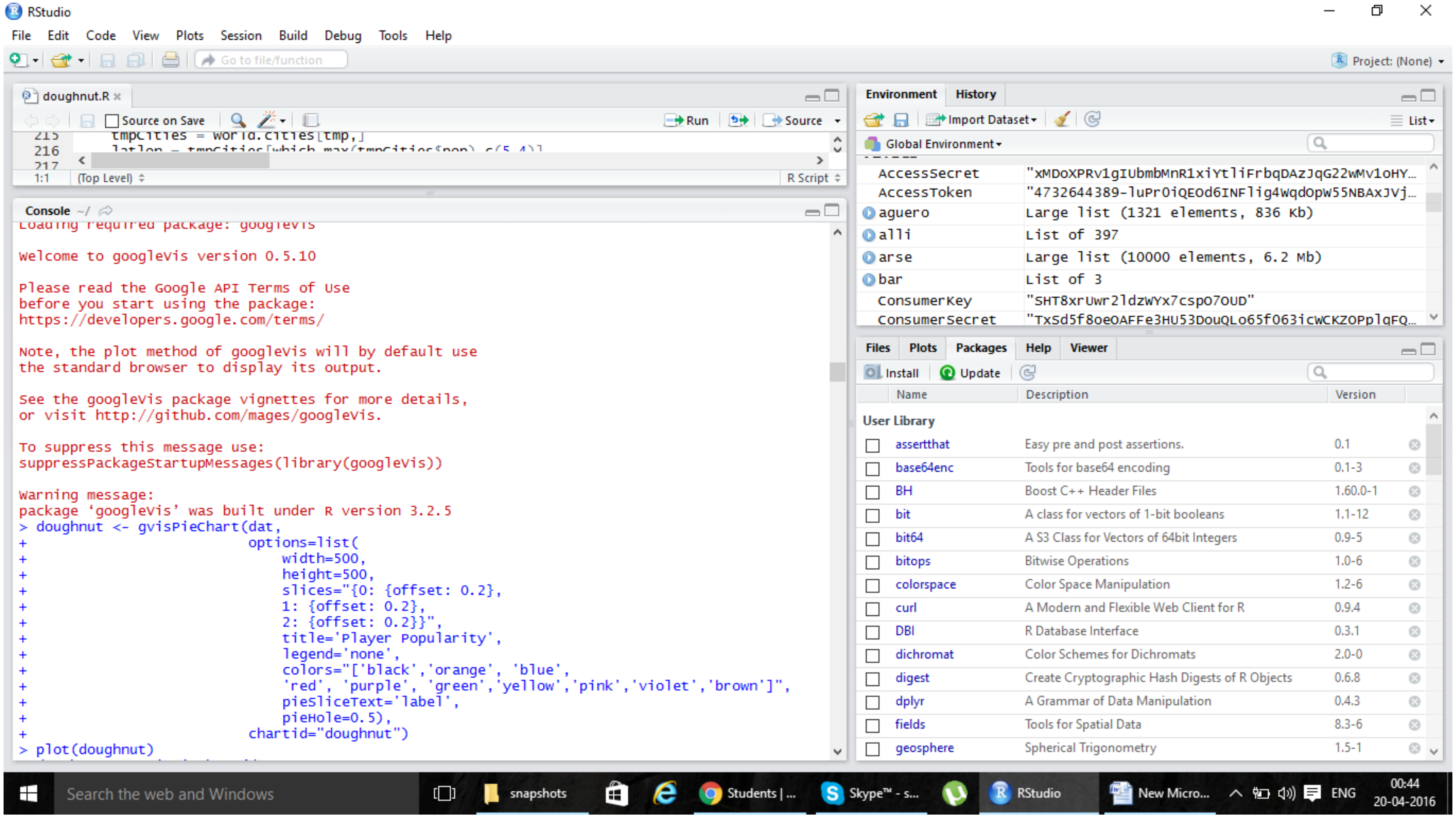
```
AccessSecret     "xMDoXPRv1gIUbmbMnR1xiYtliFrbqDAzJqG22wMv1oHY…
AccessToken      "4732644389-luPr0iQEOd6INFlig4WqdOpW55NBAxJVj…
aguero           Large list (1321 elements, 836 Kb)
alli             List of 397
arse             Large list (10000 elements, 6.2 Mb)
bar              List of 3
ConsumerKey      "SHT8xrUwr2ldzWYx7cspO7OUD"
ConsumerSecret   "TxSd5f8oeOAFFe3HU53DouQLo65f063icWCKZOPplqFQ…
```

User Library

	Name	Description	Version
□	assertthat	Easy pre and post assertions.	0.1
□	base64enc	Tools for base64 encoding	0.1-3
□	BH	Boost C++ Header Files	1.60.0-1
□	bit	A class for vectors of 1-bit booleans	1.1-12
□	bit64	A S3 Class for Vectors of 64bit Integers	0.9-5
□	bitops	Bitwise Operations	1.0-6
□	colorspace	Color Space Manipulation	1.2-6
□	curl	A Modern and Flexible Web Client for R	0.9.4
□	DBI	R Database Interface	0.3.1
□	dichromat	Color Schemes for Dichromats	2.0-0
□	digest	Create Cryptographic Hash Digests of R Objects	0.6.8
□	dplyr	A Grammar of Data Manipulation	0.4.3
□	fields	Tools for Spatial Data	8.3-6
□	geosphere	Spherical Trigonometry	1.5-1

4.3.3 Preparing a doughnut chart from the obtained data

Fig.4.10: RStudio, Creating Donut Chart



```
tmpCities = world.cities[tmp,]
latlon = tmpCities[which.max(tmpCities$pop),c(5,4)]
```

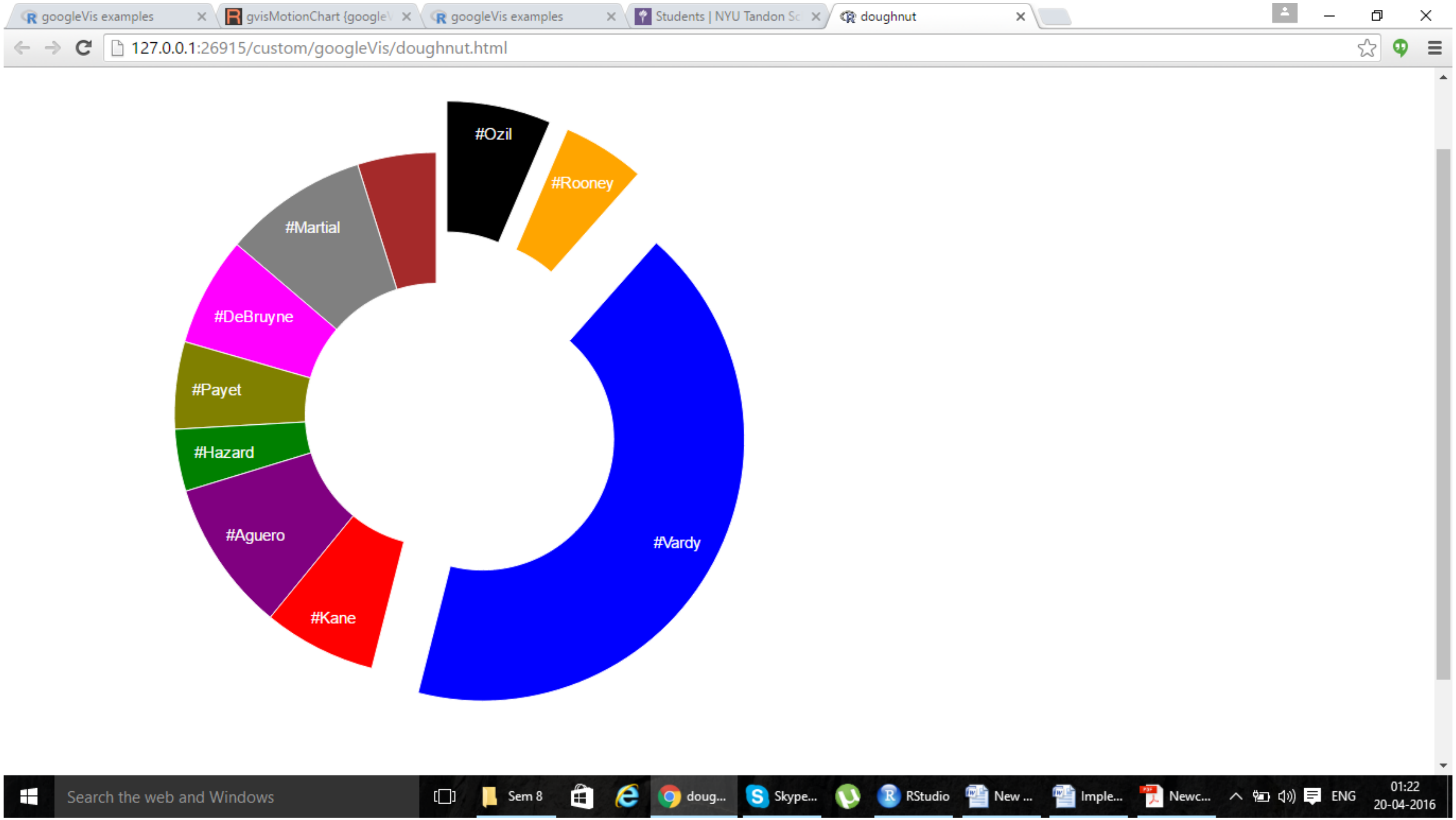
```
Loading required package: googleVis

Welcome to googleVis version 0.5.10

Please read the Google API Terms of Use
before you start using the package:
https://developers.google.com/terms/

Note, the plot method of googleVis will by default use
the standard browser to display its output.

See the googleVis package vignettes for more details,
or visit http://github.com/mages/googleVis.

To suppress this message use:
suppressPackageStartupMessages(library(googleVis))

Warning message:
package ‘googleVis’ was built under R version 3.2.5
> doughnut <- gvisPieChart(dat,
+                          options=list(
+                              width=500,
+                              height=500,
+                              slices="{0: {offset: 0.2},
+                              1: {offset: 0.2},
+                              2: {offset: 0.2}}",
+                              title='Player Popularity',
+                              legend='none',
+                              colors="['black','orange', 'blue',
+                              'red', 'purple', 'green','yellow','pink','violet','brown']",
+                              pieSliceText='label',
+                              pieHole=0.5),
+                          chartid="doughnut")
> plot(doughnut)
```

```
AccessSecret     "xMDoXPRv1gIUbmbMnR1xiYtliFrbqDAzJqG22wMv1oHY…
AccessToken      "4732644389-luPr0iQEOd6INFlig4wqdOpW55NBAxJVj…
aguero           Large list (1321 elements, 836 Kb)
alli             List of 397
arse             Large list (10000 elements, 6.2 Mb)
bar              List of 3
ConsumerKey      "SHT8xrUwr2ldzWYx7cspO7OUD"
ConsumerSecret   "TxSd5f8oeOAFFe3HU53DouQLo65f063icWCKZOPplqFQ…
```

User Library

	Name	Description	Version
□	assertthat	Easy pre and post assertions.	0.1
□	base64enc	Tools for base64 encoding	0.1-3
□	BH	Boost C++ Header Files	1.60.0-1
□	bit	A class for vectors of 1-bit booleans	1.1-12
□	bit64	A S3 Class for Vectors of 64bit Integers	0.9-5
□	bitops	Bitwise Operations	1.0-6
□	colorspace	Color Space Manipulation	1.2-6
□	curl	A Modern and Flexible Web Client for R	0.9.4
□	DBI	R Database Interface	0.3.1
□	dichromat	Color Schemes for Dichromats	2.0-0
□	digest	Create Cryptographic Hash Digests of R Objects	0.6.8
□	dplyr	A Grammar of Data Manipulation	0.4.3
□	fields	Tools for Spatial Data	8.3-6
□	geosphere	Spherical Trigonometry	1.5-1

Fig.4.11: Donut Chart



4.3.4 Creating wordclouds

STEP 1: Identifying Hashtags

Identify the hashtags used for all the teams. Only the official ones and the most popular ones are used. We’ve also avoided hashtags which could have signified something else (#Spurs is one hashtag used by Tottenham hotspur fans. However #Spurs also stands for San Antonio Spurs, which is a basketball team. We’ve avoided this hashtag as it creates a confusion and also because we noticed that not a lot of football fans use the hashtag #Spurs.

Popular Team Hashtags:

Team	Predefined Hashtags	Team	Predefined Hashtags
Arsenal	#arsenal #afc #coyg #gunners #arsenalfc #gooner #goonerfamily #onearsenal	Aston Villa	#avfc #villa #astonvilla #vtid
Bournemouth	#afcb #cherries #utciad #thecherries #bournemouthfc #afcbournemouth	Chelsea	#cfc #chelsea #cfcfamily #cfclive #chelseafc #weareblues
Crystal Palace	#cpfc #crystalpalace	Everton	#efc #everton #coyb
Leicester	#lcfc #leicester	Liverpool	#lfc #ynwa #liverpool #liverpoolfc #thekop #lfcfamily #wegoagain #thereds
Manchester City	#mcfc #mancity #bluemoon #manchestercity	Manchester United	#mufc #comeonyoureds #mufcofficial #ggmu #mufcfanpics #manu #manchesterunited #thereddevils #manutd #manunited #united #coyr
Norwich City	#ncfc #otbc #canaries	Newcastle United	#nufc #newcastle
Southampton	#SaintsFC #safc #southampton	Stoke City	#scfc #stoke #stokecity #potters
Sunderland	#safc #sunderland #coyi	Swansea City	#Swans #swanseacity #swansea
Watford	#WatfordFC #watford	WestBrom Albion	#WBA #coyb #westbrom

Hashtags used for every game:

Gameweek 31: 19 & 20th March

Premier League Hashtags
#ARS
#AVL
#BOU
#CHE
#CRY
#EVE
#LEI
#LIV
#MCI
#MUN
#NEW
#NOR
#SOU
#SUN
#STK
#SWA
#TOT
#WAT
#WBA
#WHU

SAT 19TH MARCH:

#EVEARS
#WBANOR
#WATSTK
#CRYLEI
#SWAAVL
#CHEWHU

SUN 20TH MARCH:

#NEWSUN
#SOULIV
#MCIMUN
#TOTBOU

STEP 2: Extracting tweets (date: from 24-02-2016 to 25-02-2016)

```
epl <- searchTwitter("EPL", n=100, lang="en", since = "2016-02-24", until = "2016-02-25")
> mufc <- searchTwitter("Manutd", n=100, lang="en", since = "2016-02-24", until = "2016-0
2-25")
> cfc <- searchTwitter("Chelsea", n=100, lang="en", since = "2016-02-24", until = "2016-02-
25")
```

Sample Output:

```
"PremierLeagueX: Italy keep quiet on future of Chelsea target Conte https://t.co/qDbKIT2eH
7 #Soccer #EPL"

[[2]]
[1] "vofnzambia: RT @Zambia: Multichoice Offers EPL, La Liga On Compact https://t.co/M
AWc0qQLK5"

[[3]]
[1] "csgo_guru: #EPL continues with another Tier-1 match: @astralisgg VS @TeamVirtuspro
!\n\nOVERPASS: https://t.co/cYNuCCqwWE\nCACHE: https://t.co/a7J8WuZ90c"

[[4]]
[1] "OsoBipolar7: RT @alshmlaini: @tjpothuraju @TrollFootball probably if Granada was pl
aying in Epl they will be level with Leicester city"
```

STEP 3: Creating data frames and extracting only the Hashtags

```
epl_df <- twListToDF(epl)
> mufc_df <- twListToDF(mufc)
> cfc_df <- twListToDF(cfc)
>
> epl_hash <- str_extract_all(epl_df$text, "#\\w+")
> mufc_hash <- str_extract_all(mufc_df$text, "#\\w+")
> cfc_hash <- str_extract_all(cfc_df$text, "#\\w+")
> epl_hash
```

Sample Output:

```
[[31]]
[1] "#EPL"       "#football"  "#twitter92"

[[32]]
character(0)

[[33]]
[1] "#weekender" "#THFC"      "#Swans"     "#EPL"

[[34]]
character(0)

[[35]]
[1] "#WHUFC" "#SAFC"  "#EPL"
```

STEP 4: Calculating frequencies of the Hashtags

```
epl_hash <- unlist(epl_hash)
> mufc_hash <- unlist(mufc_hash)
> cfc_hash <- unlist(cfc_hash)
>
> epl_freq <- table(epl_hash)
> mufc_freq <- table(mufc_hash)
> cfc_freq <- table(cfc_hash)
> all_tags <- c(epl_freq, mufc_freq, cfc_freq)
```

Sample Output:

```
 > cfc_freq
 cfc_hash
      #Atletico         #Blues        #brunch           #cfc           #CFC
              3              3              1              1              5
       #chelsea       #Chelsea      #football    #GaryCahill         #Inter
              2              3              5              1              1
         #Italy        #KTBFFH           #LFC     #nightmare   #PFerreira19
              1              1              2              1              1
 #premierleague      #RadioCFC        #Saints        #Soccer   #Southampton
              3              1              3              3              3
   #spinclasses    #theoneshow #WinItWednesday           #YR
              1              1              4              1
```

OUTPUT:

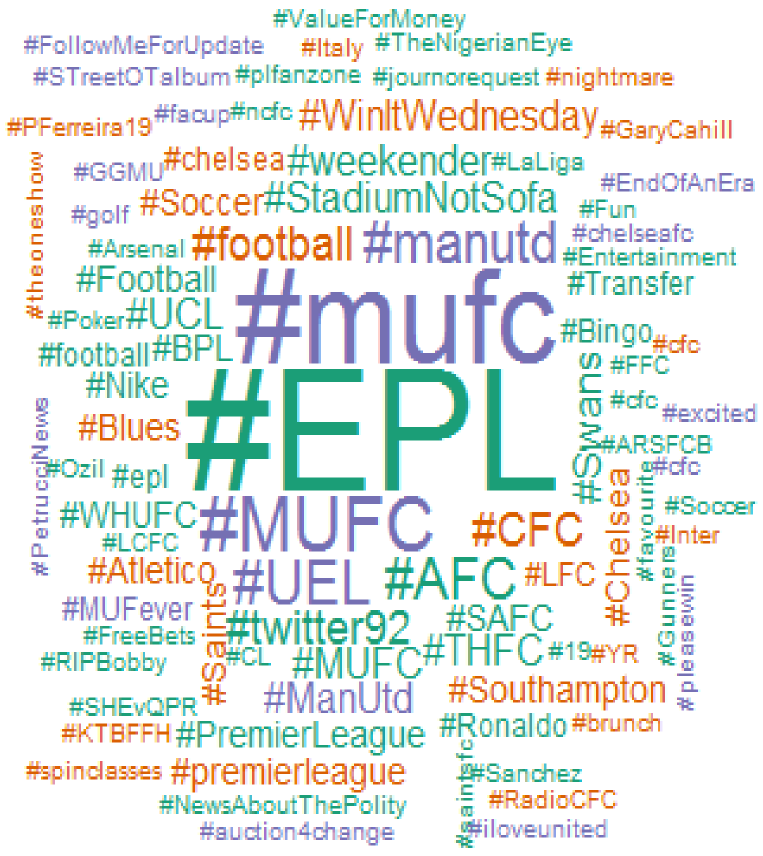
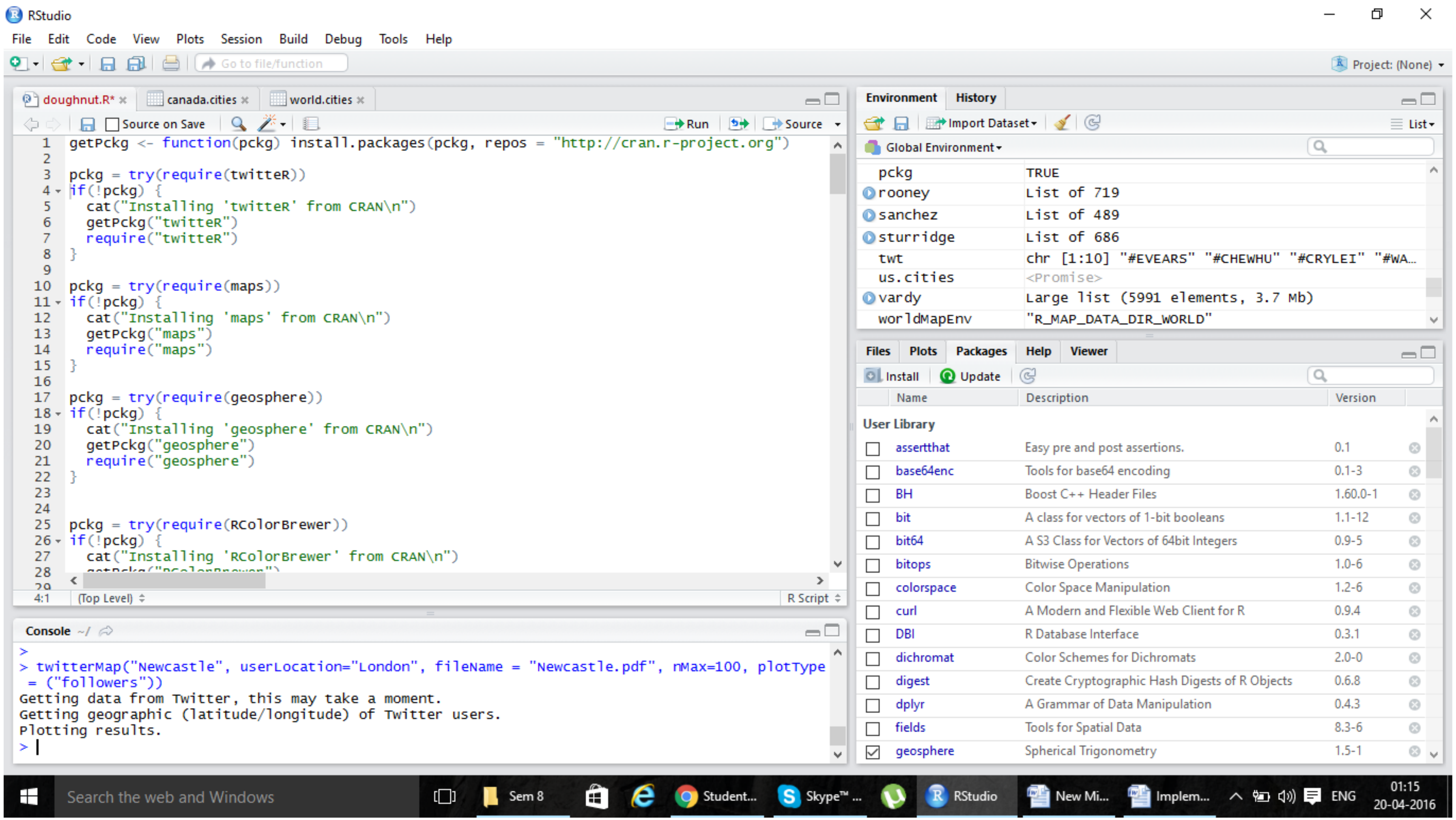


Fig.4.12: Wordcloud of #epl, Created by RStudio

4.3.5 To create followers map



```
getPckg <- function(pckg) install.packages(pckg, repos = "http://cran.r-project.org")

pckg = try(require(twitteR))
if(!pckg) {
  cat("Installing 'twitteR' from CRAN\n")
  getPckg("twitteR")
  require("twitteR")
}

pckg = try(require(maps))
if(!pckg) {
  cat("Installing 'maps' from CRAN\n")
  getPckg("maps")
  require("maps")
}

pckg = try(require(geosphere))
if(!pckg) {
  cat("Installing 'geosphere' from CRAN\n")
  getPckg("geosphere")
  require("geosphere")
}


pckg = try(require(RColorBrewer))
if(!pckg) {
  cat("Installing 'RColorBrewer' from CRAN\n")
```

Console ~/

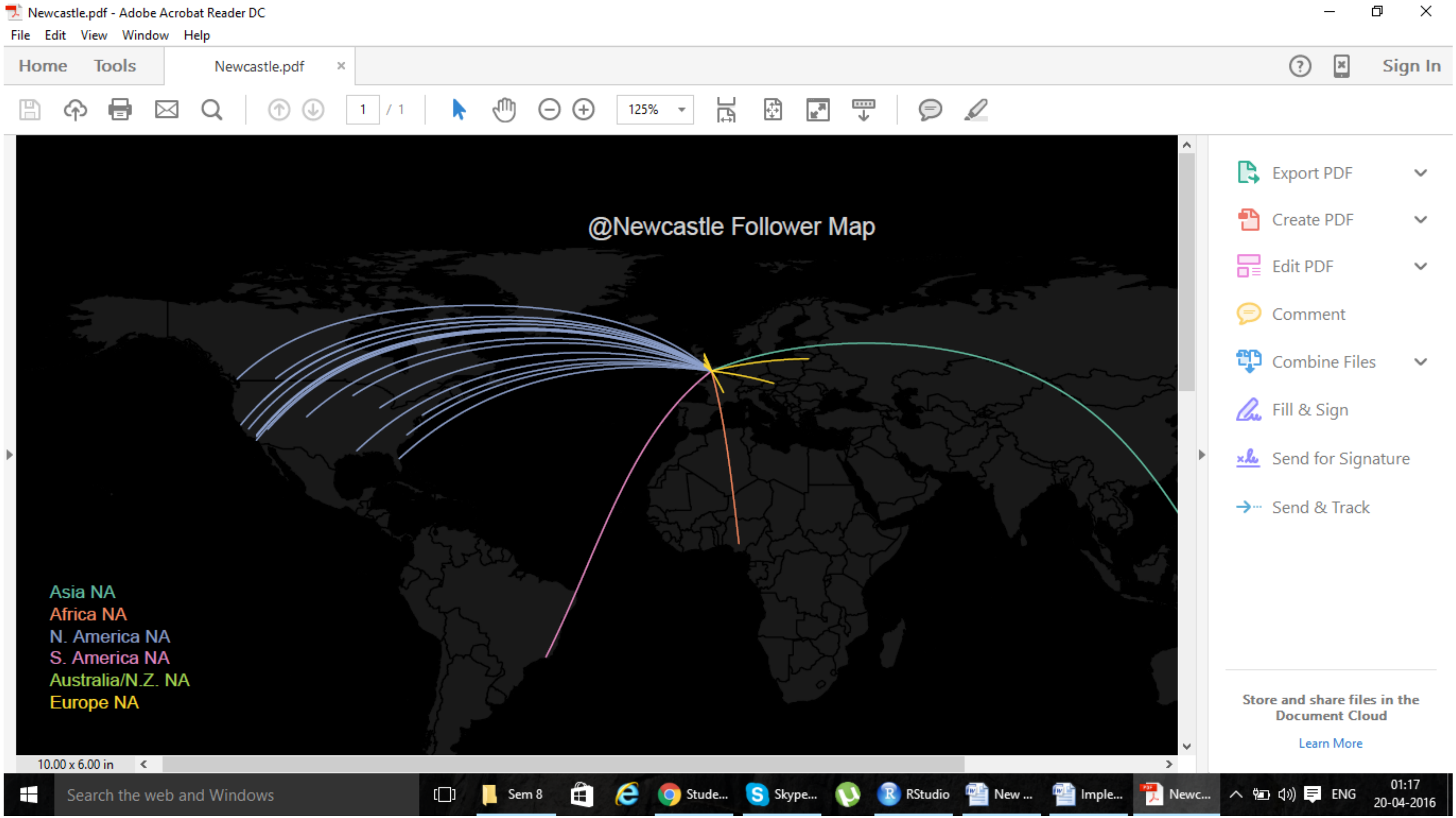
```
>
> twitterMap("Newcastle", userLocation="London", fileName = "Newcastle.pdf", nMax=100, plotType
 = ("followers"))
Getting data from Twitter, this may take a moment.
Getting geographic (latitude/longitude) of Twitter users.
Plotting results.
> |
```

```
pckg        TRUE
rooney      List of 719
sanchez     List of 489
sturridge   List of 686
twt         chr [1:10] "#EVEARS" "#CHEWHU" "#CRYLEI" "#WA…
us.cities   <Promise>
vardy       Large list (5991 elements, 3.7 Mb)
worldMapEnv "R_MAP_DATA_DIR_WORLD"
```

	Name	Description	Version
	User Library		
□	assertthat	Easy pre and post assertions.	0.1
□	base64enc	Tools for base64 encoding	0.1-3
□	BH	Boost C++ Header Files	1.60.0-1
□	bit	A class for vectors of 1-bit booleans	1.1-12
□	bit64	A S3 Class for Vectors of 64bit Integers	0.9-5
□	bitops	Bitwise Operations	1.0-6
□	colorspace	Color Space Manipulation	1.2-6
□	curl	A Modern and Flexible Web Client for R	0.9.4
□	DBI	R Database Interface	0.3.1
□	dichromat	Color Schemes for Dichromats	2.0-0
□	digest	Create Cryptographic Hash Digests of R Objects	0.6.8
□	dplyr	A Grammar of Data Manipulation	0.4.3
□	fields	Tools for Spatial Data	8.3-6
☑	geosphere	Spherical Trigonometry	1.5-1

Fig.4.13: RStudio, Creating Followers Map

Fig.4.14: Followers Map of Newcastle United



Chapter 5

Results and Discussions

5.1 Performance Measure

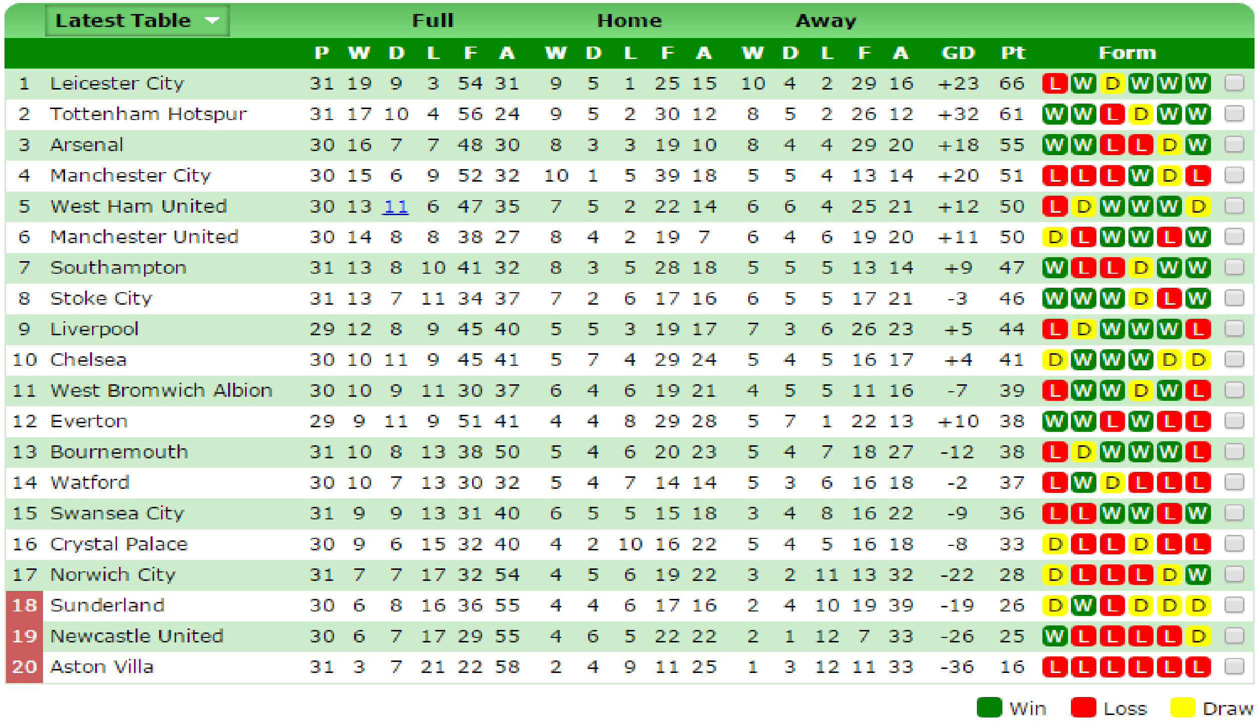
This project is based on Big Data Analysis where we have analyzed the tweets regarding the English Premier League. An experiment is conducted on the English Premier League using a static dataset which contains tweets that are extracted from Twitter over a 3 month period. Input is given to the system in the form of tweets after which the tweets are cleaned and are further analyzed to get the frequency of particular hashtags of Teams, Players, Matches, Managers etc. and then output is generated in the form of various graphs which are displayed on a dashboard. Popularity of a team is obtained by calculating the frequency of hashtags, number of tweets tweeted, number of followers of that particular team all around the world.

For each player tweets were extracted and the final results are displayed below:
According to the results, Jamie Vardy (#Vardy; tweet count in hundreds: 5991) is the most popular player in the English premier league as of now.

For the followers' map, the geo-location of the followers who are following their respective teams were analyzed. We saw that for some teams, there are very few followers but for the others, there is a regular distribution of followers all over the world. In India, the most popular team is Liverpool FC, who are ninth in the table. Manchester United and arsenal are other popular teams as well.

The Premier league table is the most efficient method of analysis based on performance of their team on the football pitch. Each team is awarded 3 points for a win and a point for a draw. At the end of the season, the team with the most points wins the trophy and the three teams with the least amount of points get demoted to the second tier of English football. Often it's these points which attract fans since fans are interested in the success of teams and they like to be associated with a trophy-laden team rather than a team which cannot win regularly.

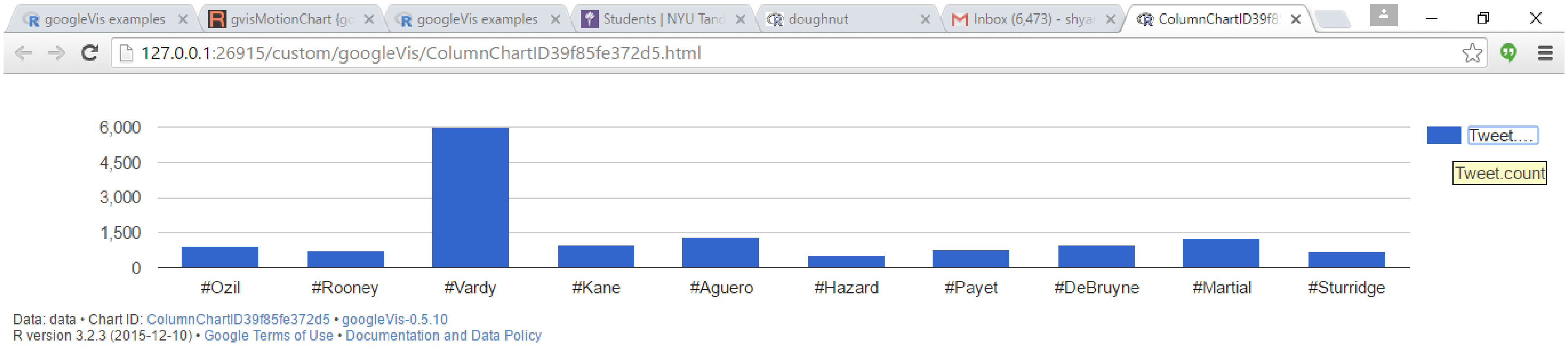
Fig.5.1: The Premier League Table →



Latest Table		Full						Home					Away							Form	
		P	W	D	L	F	A	W	D	L	F	A	W	D	L	F	A	GD	Pt	Form	
1	Leicester City	31	19	9	3	54	31	9	5	1	25	15	10	4	2	29	16	+23	66	L W D W W W	□
2	Tottenham Hotspur	31	17	10	4	56	24	9	5	2	30	12	8	5	2	26	12	+32	61	W W L D W W	□
3	Arsenal	30	16	7	7	48	30	8	3	3	19	10	8	4	4	29	20	+18	55	W W L L D W	□
4	Manchester City	30	15	6	9	52	32	10	1	5	39	18	5	5	4	13	14	+20	51	L L L W D L	□
5	West Ham United	30	13	11	6	47	35	7	5	2	22	14	6	6	4	25	21	+12	50	L D W W W D	□
6	Manchester United	30	14	8	8	38	27	8	4	2	19	7	6	4	6	19	20	+11	50	D L W W L W	□
7	Southampton	31	13	8	10	41	32	8	3	5	28	18	5	5	5	13	14	+9	47	W L L D W W	□
8	Stoke City	31	13	7	11	34	37	7	2	6	17	16	6	5	5	17	21	-3	46	W W W D L W	□
9	Liverpool	29	12	8	9	45	40	5	5	3	19	17	7	3	6	26	23	+5	44	L D W W W L	□
10	Chelsea	30	10	11	9	45	41	5	7	4	29	24	5	4	5	16	17	+4	41	D W W W D D	□
11	West Bromwich Albion	30	10	9	11	30	37	6	4	6	19	21	4	5	5	11	16	-7	39	L W W D W L	□
12	Everton	29	9	11	9	51	41	4	4	8	29	28	5	7	1	22	13	+10	38	W W L W L L	□
13	Bournemouth	31	10	8	13	38	50	5	4	6	20	23	5	4	7	18	27	-12	38	L D W W W L	□
14	Watford	30	10	7	13	30	32	5	4	7	14	14	5	3	6	16	18	-2	37	L W D L L L	□
15	Swansea City	31	9	9	13	31	40	6	5	5	15	18	3	4	8	16	22	-9	36	L L W W L W	□
16	Crystal Palace	30	9	6	15	32	40	4	2	10	16	22	5	4	5	16	18	-8	33	D L L D L L	□
17	Norwich City	31	7	7	17	32	54	4	5	6	19	22	3	2	11	13	32	-22	28	D L L L D W	□
18	Sunderland	30	6	8	16	36	55	4	4	6	17	16	2	4	10	19	39	-19	26	D W L D D D	□
19	Newcastle United	30	6	7	17	29	55	4	6	5	22	22	2	1	12	7	33	-26	25	W L L L L D	□
20	Aston Villa	31	3	7	21	22	58	2	4	9	11	25	1	3	12	11	33	-36	16	L L L L L L	□

🟩 Win 🟥 Loss 🟨 Draw

Fig.5.2: Players popularity graph



Data: data • Chart ID: ColumnChartID39f85fe372d5 • googleVis-0.5.10
R version 3.2.3 (2015-12-10) • Google Terms of Use • Documentation and Data Policy

5.2 Analysis:

Due to twitter API's limitations, we could only extract tweets for the last 3 months. However, from the analysis we have performed over this time, (from January to March), we've obtained some surprising results. Leicester City, who were bottom of the premier league table last year, are in the summit at the moment. This has led to an enormous increase in the number of tweets pertaining to Leicester city and its players. Everyone all over the world is praising Leicester City's character, determination, grit and their play style and they will surely gain a lot of new followers over the years. This rise in popularity has led to Vardy's (and to a similar extent Mahrez's) rise as well. He is head and shoulders above the other players when it comes to tweet count mentioning him including players like Rooney and Ozil who are expected to be popular every season (as they are, and have been very good players from the beginning). This also explains why Game week 32's most popular game was Leicester vs. Southampton, instead of Liverpool vs. Tottenham or Manchester United vs. Everton, which were big games involving big teams followed all over the world unlike Southampton or Leicester.

When we analyzed team popularity, we expected traditional power horses Manchester United, Chelsea, Liverpool and Manchester City to have the highest number of tweet mentions. The results were some both surprising and expected. Manchester United and Liverpool both had a staggering amount of mentions compared to the others, including Chelsea and Manchester city. The lack of tweets mentioning Chelsea could be explained by their poor standing in the table (10th) compared to last year when they were winners, boring performances on the pitch and lack of a charismatic manager, unlike the one they had last year. Manchester City's lack of tweets however, couldn't be explained, although it might have something to do with a number of injuries to key players throughout the second half of the season. This meant that players like Kompany and Silva did not participate in a number of matches, hence discouraging fans to tune into their television sets to watch the game or tweet about it. Tweets mentioning Liverpool F.C were the most frequent, followed somewhat closely by Manchester United. Both of these teams have eccentric managers. A huge number of fans rant about United's manager Louis Van Gaal. His habit of making surprising decisions and changes has confused fans throughout the season. Klopp, liverpool's manager, however is a very popular personality and fans praise about him on twitter. Additionally, entertaining Liverpool games with nail biting finishes, fans have a lot of reasons to talk about their favourite team.

The lack of tweets mentioning Bournemouth and crystal palace, which aren't popular teams wasn't surprising at all. Teams that were somewhat popular but also not un-popular had a moderate number of tweets, results which were expected.

From the follower's map, we found out that there are a lot of fans for teams such as West Brom and West Ham United, even though they aren't big successful teams or those who play attractive football. In fact in North America, Tottenham Hotspur is the most popular team ahead of Manchester United, Arsenal, Liverpool and Chelsea. Newcastle United, too, has a lot of fans all over the world, despite languishing in 19th place. This shows the reach of the premier league all over the world and it also shows that fans don't just support successful teams. According to the premier league fan survey, the reasons to follow a team are: Whether or not it's a local club, matchday atmosphere experience, Family / friends influence, The way the team plays, Good stadium facilities, success of the teams, To see a particular player and Area.

The most interesting aspect of the analysis is to find out the unpredictable and interesting results of it. The result from one week aren't necessarily the results in the next. For example, Newcastle United was the most popular team in the middle of March because they appointed a new and prestigious manager – Rafa Benitez. Thus, fans were interested to discuss about the impact he would have on the title but the hype soon faded out. What we can conclude is that football is always popular.

Chapter 6

Conclusion

As we have entered an era of Big Data, processing large volumes of data has never been greater. Analysis of Big Data guarantees faster advances in many scientific disciplines and improving the profitability and success of many enterprises. Our study shows that there exists enough data in twitter to perform thorough analysis which is complete. It also shows that if the data is collected and analysed appropriately, then it can be useful to many organizations and football fans.

In our results we can't neglect a number of limitations of this study. First and foremost, the experiment was not conducted on the whole English Premier League Season, but instead on a 3 month period (January to March). It would be interesting to perform the same study but on a complete season. Acquiring data from the past proved to be extremely difficult and time-consuming. From our queries, we noticed that tweets longer than 60 days old were not shown. This could be a major side effect and to solve this, we have to analyse and collect tweets every week or every month.

Nevertheless, this fact does not limit the conclusion of the study which is that Twitter contains information which is useful for fans and organizations alike. It would be very beneficial to football teams to analyse themselves, to see where they stand and to see what the people think about their teams. It would also be beneficial to certain organizations such as OptaStats, which can use this information. An increased fan following would also be an added benefit of the success of this project since we will represent all information in a visually appeasing manner through graphs and interesting statistics.

Now there is another reason to watch football. Now there is another reason to discuss about football on twitter

Chapter 7

References

Big Data:

[1] Big-data content retrieval, storage and analysis foundations of data-intensive computing Website: http://www.acsu.buffalo.edu/~mjalimin/

[2] Dr. Siddaraju, Sowmya C L, Rashmi K and Rahul M, "Efficient Analysis of Big Data Using Map Reduce Framework", International Journal of Recent Development in Engineering and Technology, (ISSN 2347-6435(Online) Volume 2, Issue 6, June 2014)

[3]Wang, J., Qi, G., Sebe, N., & Aggarwal, C. C. (2015). Guest Editorial: Big Media Data: Understanding, Search, and Mining. *IEEE Transactions on Big Data IEEE Trans. Big Data*, *1*(3), 82-83.

[4] Big Data and the Future of Business http://www.technologyreview.com/view/538916/big-data-and-the-future-of-business/

[5] Jianqing Fan1, Fang Han and Han Liu, Challenges of Big Data analysis, National Science Review Advance Access published February, 2014.

Twitter Data Analysis:

[1] Twitter Developers' Website: https://dev.twitter.com/

[2] Kumar, S., Morstatter, F., & Liu, H. (2013). Visualizing Twitter Data. *Twitter Data Analytics SpringerBriefs in Computer Science*, 49-69. doi:10.1007/978-1-4614-9372-3_5\

[3] Yanchang Zhao, “Text Mining with R – Twitter Data Analysis”, Presented at AusDM 2014 (QUT, Brisbane) in Nov 2014 and at UJAT (Mexico) in Sept 2014

[4] How-to: Analyze Twitter Data with MongoDB
http://blog.cloudera.com/blog/2012/09/analyzing-twitter-data-with-mongodb

[5] Adarsh M J and Pushpa Ravikumar, “Survey: Twitter data Analysis using Opinion Mining”, International Journal of Computer Applications (0975 – 8887), Volume 128 – No.5, October 2015

[6] David Ediger, Courtney Corley and William N. Reynolds, “Massive Social Network Analysis: Mining Twitter for Social Good”, 2010 39th International Conference on Parallel Processing, IEEE

[7] Manoj Kumar Danthala, “Tweet Analysis: Twitter data processing using apache hadoop”, International Journal Of Core Engineering & Management (IJCEM) Volume 1, Issue 11, February 2015

[8] https://www.credera.com/blog/business-intelligence/twitter-analytics-using-r-part
How-to: Analyze Twitter Data with Apache Hadoop

Data Mining (Sports)

[1] Osama K. Solieman: Data Mining in Sports: A Research Overview
MIS Masters Project

[2] International Journal of Recent Development in Engineering and Technology Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 2, Issue 6, June 2014)

Football Websites

[1] Scudamore, Richard. "National Fan Survey (EPL)." 12 May 2008.

[2] Footstats**:** the Premier Football Statistics and Analysis site http://www.footstats.co.uk