Kingdom of Saudi Arabia
Ministry of Education
King Abdulaziz University
Faculty of Computing and Information Technology
Department of Information Systems

المملكة العربية السعودية
وزارة التعليم
جامعة الملك عبد العزيز
كلية الحاسبات وتقنية
المعلومات
قسم نظم المعلومات

Analyzing Performance of Saudi Universities on Twitter

1. Amjad Dahlawi	1406448
2. Ali Almalki	1408243

Supervisor: Rabeeh Abbasi

Abstract

The issue of analyzing tweets related to universities is very significant subject. There is so much data on Twitter that needs to be understood by analyzing the data and get useful information's from it. Therefore, we intend in our project to enable the user to analyze tweets, searching for common topics or keywords and presenting relationships in graphs. Taking into concern 25 studies related to the subject of our project were summarized and relied upon.

The analysis has been conducted on data related to Saudi universities accounts on Twitter and their follower's number, number of retweets from each of the ten universities and the frequent tweeting times of the ten universities. Afterward, it was concluded from the analysis that the users of Twitter and especially users who are interested in educational purposes with their universities on Twitter.

The system is made up of several functional requirements and each functional requirement has an appointed diagram for it that illustrates its function in our system. There are many scenarios for the user activities through our system consequently it has been covered most of the scenarios. The unique thing about the system is that all users have same privileges.

Our project will focus on analyzing unstructured data on Twitter regarding Saudi universities and designing methodologies. Taken into consideration the issues that has been discussed in this report. From this project, we hope to build an effective analyzing system for educational purposes.

Acknowledgment

The senior project opportunity we had was a great chance for learning and professional development. Therefore, we consider ourselves as a very lucky individuals as we were provided with an opportunity to be a part of it. We also grateful for practicing and learning wonderful things and exploring new science which is social media analytics through this semester period.

We would like to convey our gratefulness to our supervisor Dr. Rabeeh Abbasi and the lecturer Mr. Abdullah Al Saleh for their generous support, coaching and companionship during the project period. The skills and knowledge which we have gained throughout our practical work we perceive it as very valuable component in our future career development. As you know this senior project period was part of our educational curricula and therefore we have to thank you for providing us this opportunity.

Our sincere thanks to Dr. Abdulrahman Al-Talhi Dean of the Faculty, for the continuous encouragement.

Thank you once again for your great support in the successful completion of our senior project.

Sincerely,

Ali Almalki & Amjad Dahlawi

Table of Contents

Table of Figures

1. Chapter 1| Project Outlines

1.1 Introduction:

The goal of this project is to analyze tweets related to universities. The analysis will include topics being discussed by Saudi universities, analysis of the tweets of the people interacting with Saudi universities through twitter, identification of important events, etc. The project will help the universities in analyzing their engagement with public. It will also help students in identifying topics discussed by universities and how they interact with public.

Twitter is widely used by universities all over the world to disseminate information and to interact with people. All the 37 Saudi universities (both public and private) have twitter accounts. Universities use these account to disseminate information about admissions, exams, events, new programs, conferences, achievements etc. People use twitter to ask questions and provide feedback to the universities.

1.2 Problem Statement:

The problem is we have so much data and we need to understand it by analyzing the data and get useful information's from it.

Twitter is one of the most popular applications of social media simply to send messages called "tweets" not more than 140 characters, but with the new update it has been extended the length of the tweet to 280 characters, You can attach a picture, video or link in the tweet, choose who you want to follow and others also can follow you back. Twitter is an important social media application for news, announcements, data transmission and communication between students and their universities, which is the subject of our project.

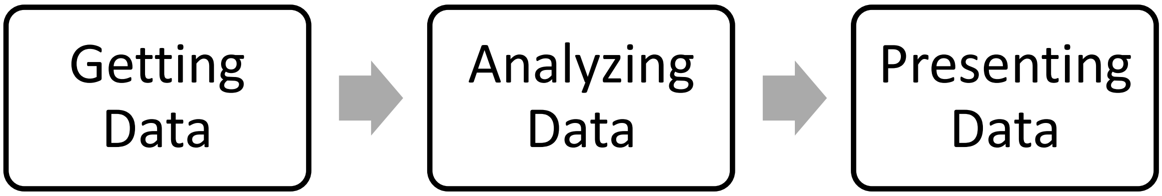
There is a difficulty in knowing the status of universities, academic level and universities ranking in different fields and because the great use of Twitter by universities leads to the importance of analysis of this use to extract useful information in this regard.

The project will analyze student's tweets with their universities and extract useful information from it, also to determine the ranking of the best Saudi universities and the most interactive universities with their Twitter accounts.

1.3 Objectives

- Develop analytical tool for academic improvements.
- Create a website that will help in analyzing universities information.
- Decrease effort of evaluating universities service performance.
- Present results in creative graphical representation.
- Improve the way of getting notified about universities upcoming events.
- Identify universities participating in international scientific competitions.
- Provide statistics and charts illuminate the most Saudi universities interacting in Twitter.
- Extract from tweets how people react to different topics.

1.4 Methodology:


Getting Data

Analyzing Data

Presenting Data

Getting Data:

Getting data phase will start with retrieving tweets from all Saudi students and Saudi universities. This will include the process of reading tweets throughout twitter API and storing it in a flexible way using MongoDB.

Analyzing Data:

After retrieving the data the program will start analyzing tweets using python libraries including pandas and NumPy. Pandas will provide data manipulation and analysis In particular; it offers data structures and operations for manipulating numerical tables and time series (1). While numpy will add support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays (2).

Presenting Data:

The next step is presenting data in a creative professional by converting Python code to a website using Flask library in python.

1.5 Project Plan:

Phases		**Deadline**
Major Phases	**Minor Phases**	
Project Proposal		Week #3
Literature Review		Week #4
Analysis Phase	Data Gathering	Week #5
	Data Analysis	Week #6
	System Requirements	Week #7
	Functional Requirements	
	Non- Functional Requirements	
Design Phase	Systems Architecture	Week #8
	Use Case Model	
	Data Flow Diagram	Week #9
	Class Diagram	
	Database Architecture	Week #10
	Database Design	
Interface/Prototype		Week #13
Final Report Preparation		Week #14

2. Chapter 2| Literature Review

2.1 Background and Overview of Related Work

2.1.1 Analytics

2.1.1.1 A framework for real-time Twitter data analysis

Twitter has a lot of data and in order to extract useful information from it and to identify relevant topics we need to do some analysis. Proposes a system for such kind of analysis, they call it "Twitter Live Detection Framework" (TLDF). It is used to analyze the data in the tweets and to define popular events. This analysis is an improvement of Soft Frequent Pattern Mining (SFPM) algorithm. Their goal in this analysis is to determine the user perspective on related events. TLDF is a framework to analyze the text documents in Twitter to identify topics, events and related global news. The analysis works after choosing some generic terms to query Twitter, and the stream of tweets divided into dynamics window which has been analyzed to identify relevant topics. In the analysis they have keywords indicting the nature of the topics or events being discussed in the tweets and it is updated from time to time to include new important terms or to delete the terms that have been unused anymore. This real-time Twitter data analysis algorithm has a better performance than SFPM in detecting relevant topics, but SFPM is good in offline detection scenarios they get the data using Steaming API (3).

2.1.1.2 A System for Querying, Analyzing, and Visualizing Geotagged Microblogs

This study shows an efficient and scalable way of querying, analyzing, and visualizing system for microblogs developed by *Taghreed*. It supports arbitrary queries on a very large number of microblogs. Dealing with microblogs is like dealing with big data so the main focus of her work is optimizing continuous queries and this system is made of four major components (4):

1- Indexer.
2- Query Engine.
3- Recovery Manager.
4- Visualizer.

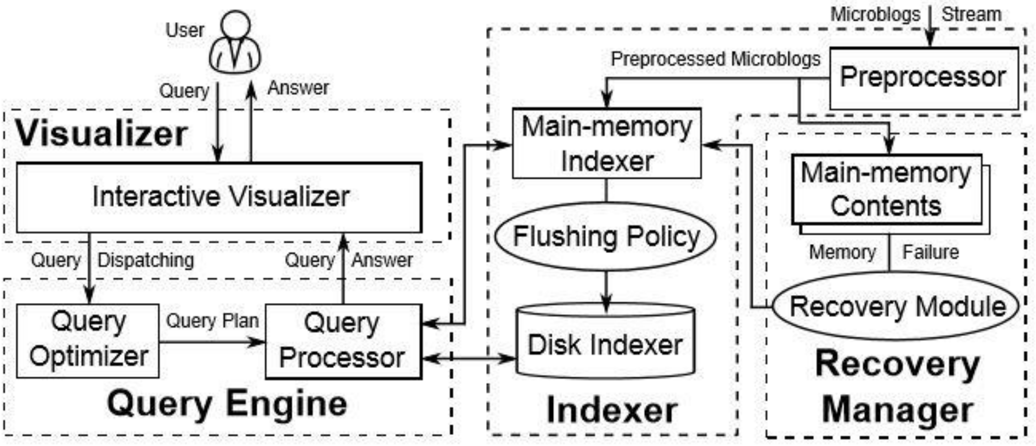


Figure 1: Taghreed's Major Components (4)

2.1.1.3 Web-Based Visual Analytics for Social Media

This study is about giving information about traffic and weather for people of Seattle by using SRS (Scalable Reasoning System). SRS is modular framework will lead to rapid prototyping of visual and analytic applications. Three layers have been used in this client framework:

1- Server interaction layer (model).
2- Visualization layer (view).
3- The transform layer (controller).

Combining visual analytics with a server side algorithm provides situational awareness and assist in the decision making process (5).

2.1.1.4 A Broad-Scale Study on Visual Social Media Analytics for Public Safety

Media analytics for crisis intelligence in social media has continuous usage in areas like data mining and visualization. This study use two phases. The first phase includes data and tasks, study setup (Domain experts: crisis response and critical infrastructures), and results (task performance, comments and suggestions). The second phase includes Event digest (Event detection, relevant media display, significance score), user study (study setup, performance comparison, questionnaire). ScatterBlogs gave the experts the complete idea of visualization and analysis techniques (6).

2.1.1.5 A Comparative Analysis on Weibo and Twitter

This study shows the comparison between Weibo and twitter. Weibo is similar to twitter but people especially in china use weibo more often. Weibo user network contains 222 million users. This study will show different comparison techniques including degree distributions, distance distributions and clustering coefficient. People interact in a different way in weibo. Also clustering coefficient is a lot smaller for most accounts. The real interaction between people is extraordinarily weaker in weibo than it is in twitter (7).

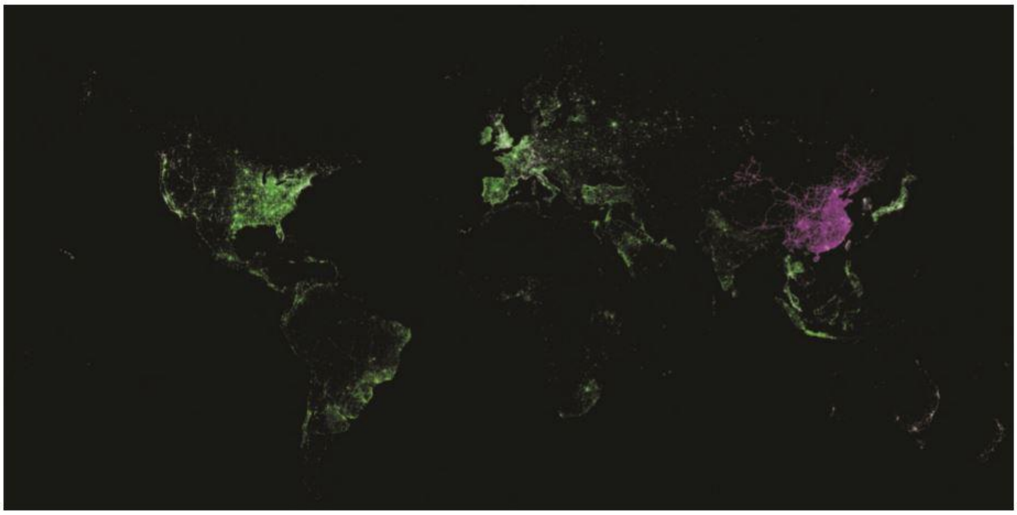


Figure 2:Map of weibo and twitter user location three years after launch Green: Twitter, Purple: Weibo White: Both (7)

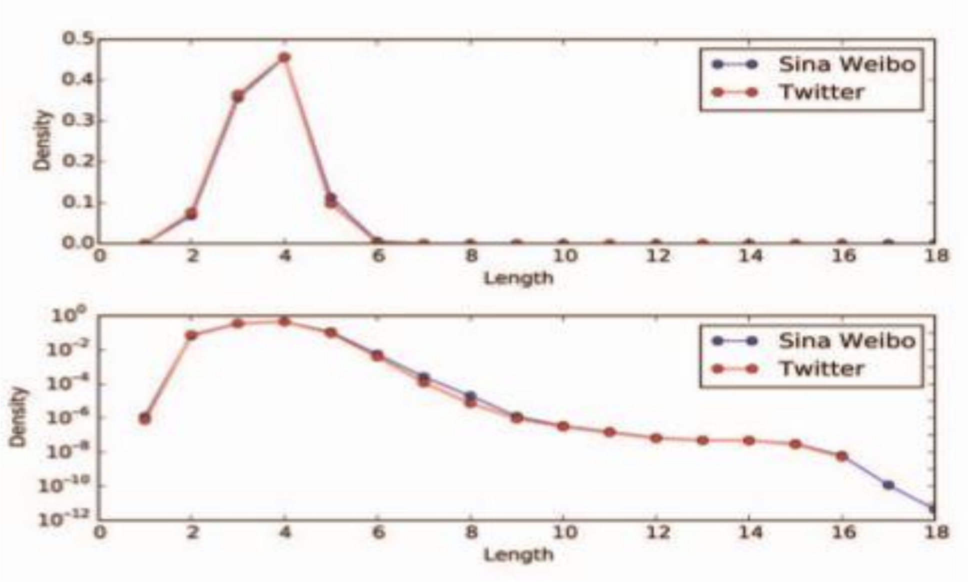


Figure 3: Distribution of distances from sampled seeds (7)

2.1.1.6 A Data Analytic Framework for Unstructured Text

This paper shows how to handle and manage big data, extract useful information from it and the ability to predict what's going to happen next based on the available data. The goal of the study is to give a clear view about the unstructured data and how to manage it and implement it. The process is done in Python to get the sentiment score. Big data is a large data set that is analyzed in a way to find some patterns, events, trends and associations between data.so there is an approach to analyze the big data on Twitter called unstructured data framework which analysis the twitter data based on the 3Vs of big data:

- variety
- velocity
- volume

And there are two relevant features which is the Value and complexity. The massive and rapid proliferation of big data is increasing nowadays. Therefore, it's so important to analyze to get a value from it. Data mining and OLAP were used in the first studies on Twitter. The data manager has two managers: metadata extraction manager and knowledge discovery manager all of them works together to get and retrieve all the metadata of unstructured data on the network and then store it in the storage (8).

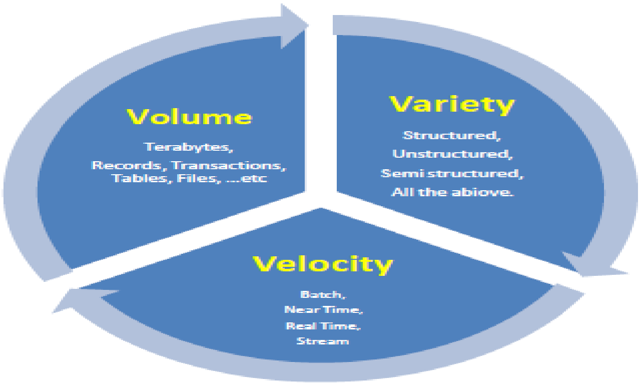


Figure 4: Big Data definition (3 Vs) (8)

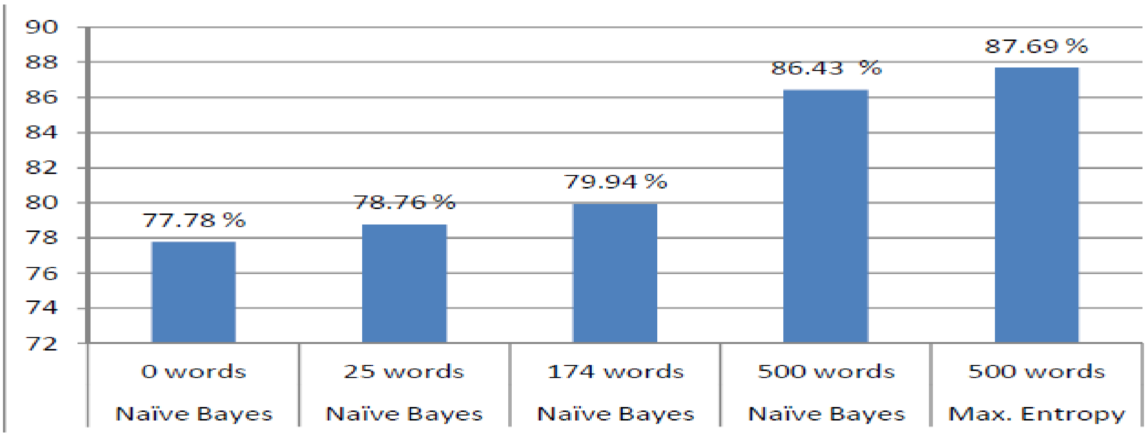


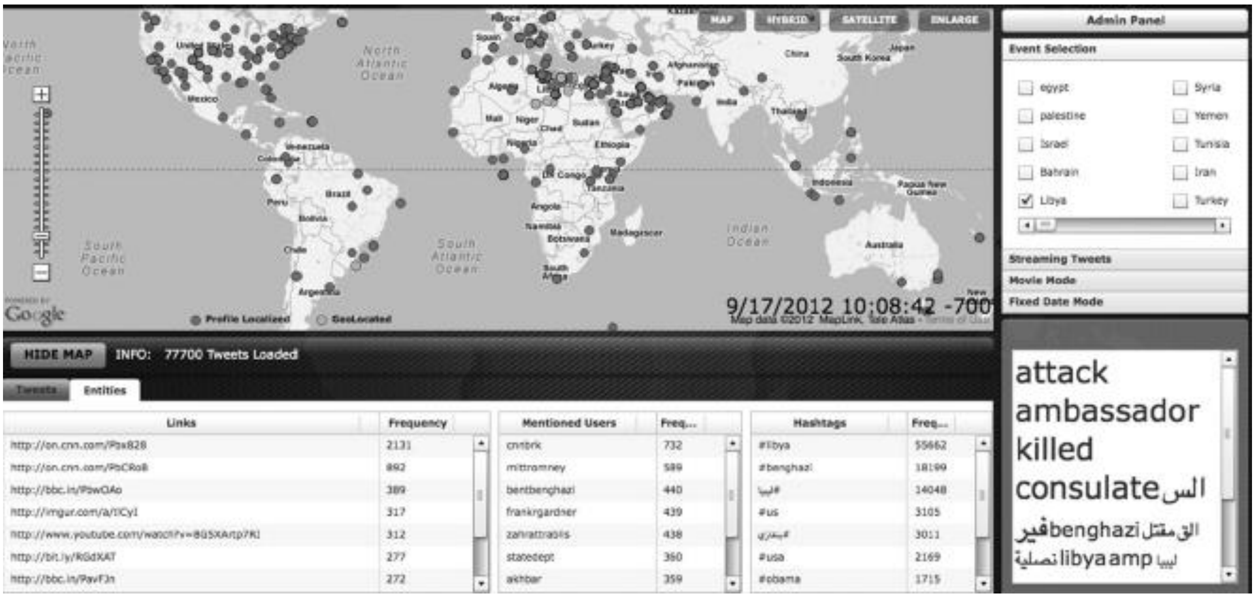
Figure 5: Accuracy evaluation of proposed classifiers (8)

2.1.1.7 Embassies burning: toward a near-real-time assessment of social media using geo-temporal dynamic network analytics

The phenomenon of crises causing a big problem that needs to be resolved and evaluated quickly by rapid assessment in order to develop plans to deal with these crises. Social media such as Twitter is very important for this assessment. This paper shows rapid ethnographic approach to extract useful information from Twitter and then doing the assessment by dynamic network analysis techniques. The text mining and visualization are involved in this analysis approach. This analysis approach is based on the data collection from the social media rather than collect it on the ground with the people. One of the tools that deal with the analysis related to crisis called TweetTracker which is a tool to analyze tweets in real time in form of parameters related to the event, they analyze it. It has three parameters:

- Keywords
- Geographical boundary boxes
- Analyst timelines

There are other tools like REA, and ORA-NetScenes. What make these analysis special is that they are done within hours of the event (9).



INFO: 77700 Tweets Loaded

Links	Frequency
http://on.cnn.com/Psx828	2131
http://on.cnn.com/PbCRoB	892
http://bbc.in/PswOAo	389
http://imgur.com/a/tICyI	317
http://www.youtube.com/watch?v=8G5XArtp7RI	312
http://bit.ly/RGdXAT	277
http://bbc.in/PavF3n	272

Mentioned Users	Freq...
cnnbrk	732
mittromney	589
bentbenghazi	440
frankrgardner	439
zahrattrablis	438
statedept	360
akhbar	359

Hashtags	Freq...
#libya	55662
#benghazi	18199
#ليبيا	14048
#us	3105
#بنغازي	3011
#usa	2169
#obama	1715

Event Selection:

□ egypt □ Syria
□ palestine □ Yemen
□ Israel □ Tunisia
□ Bahrain □ Iran
☑ Libya □ Turkey

attack
ambassador
killed
consulate السفير
benghazi القنصلية مقتل
libya amp ليبيا

Figure 6: TweetTracker user Interface (9)

2.1.1.8 Twitter Data Analytics

This book talks about the analysis of data on Twitter which has more than 140 million active users and they have posted an enormous number of tweets more than 400 million tweets every day and it's increasing. It's showing the basics of collecting, storing, and analyzing Twitter data with the API provided by Twitter. It's also talk about the visual analytics which is an analysis based on the intuitive visualizations. API tool it is used for accessing Twitter data and it has two types which is:

- REST APIs used for data retrieval by pull strategy and to collect information's from the user.
- Streaming APIs that gives a stream of public information's from Twitter.

There are three types of the streaming API on Twitter:

- Public streams
- User streams
- Site streams

Open Authentication (OAuth) it's a standard provided Twitter to authenticate users when accessing protected information. When you want to do a text visualizing you should use an important technique which is called Word Cloud. It highlights important words and the frequency to determine its importance (10).

Figure 7: Word cloud containing top 60 words (10)

2.1.1.9 Framework for Social Media Big Data Quality Analysis

This paper shows how to deal with the big data with the appropriate analysis to get a value from it. Big data is unstructured data and unorganized data that is heavy on social media sites and it's increasing so fast in social media nowadays. So, the paper discusses the importance to create an analysis with a good quality to extract useful information from the big data on social media sites. There are four properties of big data (4 Vs): Volume, Velocity, Varity and Veracity.

- Volume: The heavy amount of big data.
- Velocity: The speed of the big data spreading on social media sites.
- Varity: The big data is a collection of data with different styles so we need to combine it all together to do the analysis.
- Veracity: The accuracy and the truthful of the data.

Analyzing the big data is very important to the organization to make the right decisions. Also, the paper talked about the metadata that influences the big data analysis process. Metadata is data about data and how to collect it and store it for the analysis. This paper proposes a framework with a good quality analysis for the big data in social media sites especially twitter (11).

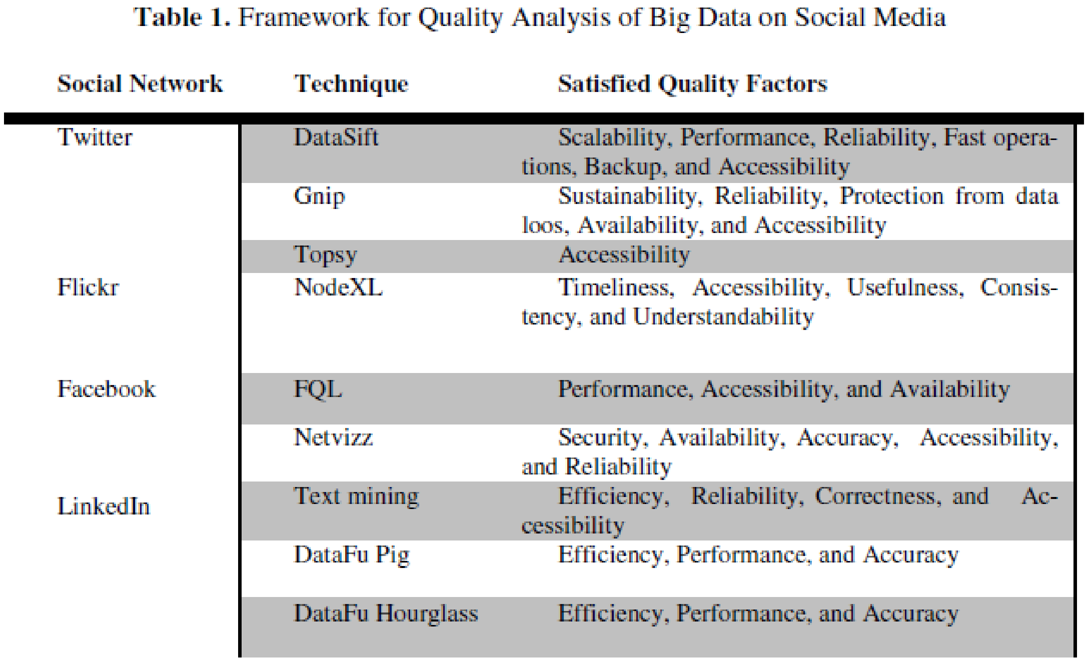


Table 1. Framework for Quality Analysis of Big Data on Social Media

Social Network	Technique	Satisfied Quality Factors
Twitter	DataSift	Scalability, Performance, Reliability, Fast operations, Backup, and Accessibility
	Gnip	Sustainability, Reliability, Protection from data loos, Availability, and Accessibility
	Topsy	Accessibility
Flickr	NodeXL	Timeliness, Accessibility, Usefulness, Consistency, and Understandability
Facebook	FQL	Performance, Accessibility, and Availability
	Netvizz	Security, Availability, Accuracy, Accessibility, and Reliability
LinkedIn	Text mining	Efficiency, Reliability, Correctness, and Accessibility
	DataFu Pig	Efficiency, Performance, and Accuracy
	DataFu Hourglass	Efficiency, Performance, and Accuracy

Figure 8: Framework for Quality Analysis of Big data on Social Media (11)

2.1.1.10 Visualizing and Estimating Happiness in Italian Cities from Geotagged Tweets

This research highlights the analysis of the geotagged tweets in Italian cities related to the human behavior. To do such kind of analysis, the research proposes a framework called Felicitt`a which is a visualization system estimating the happiness in a certain geographical area based on the geotagged tweets with different visualization techniques. This system contains sentiment analysis that detect the Italian tweets sentiment. The Felicitt `a framework contains:

- Geotagged information retrieved by APIs.
- Data Gathering
- Data analysis estimating the sentiment.
- Visualization of the analysis results.

Sentiment analysis have three major steps: pre-parsing, parsing and analysis (12).

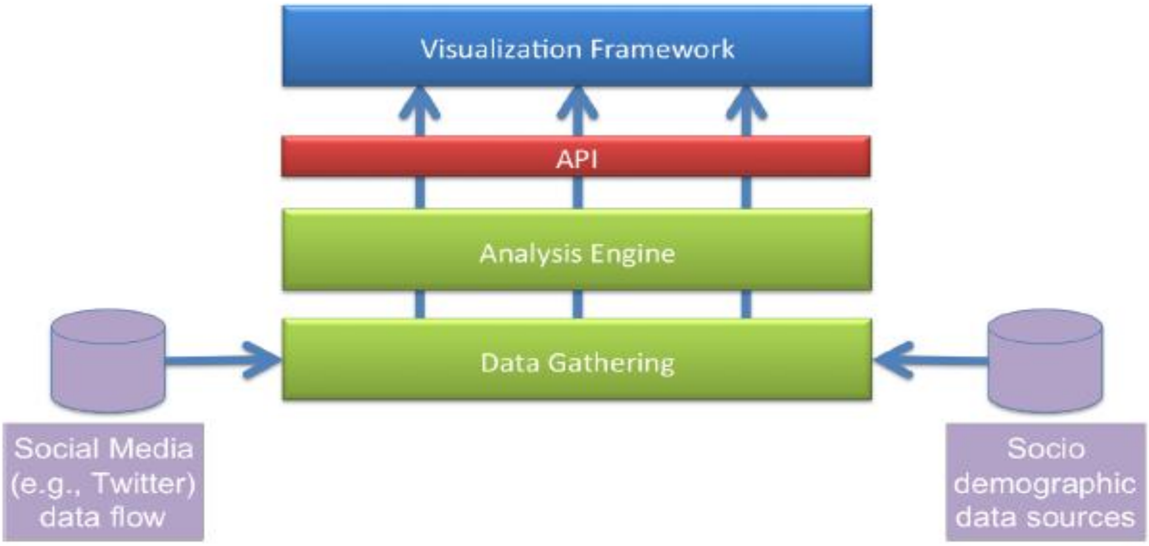

Visualization Framework
API
Analysis Engine
Data Gathering
Social Media (e.g., Twitter) data flow
Socio demographic data sources

Figure 9: The Felicitta framework (12)

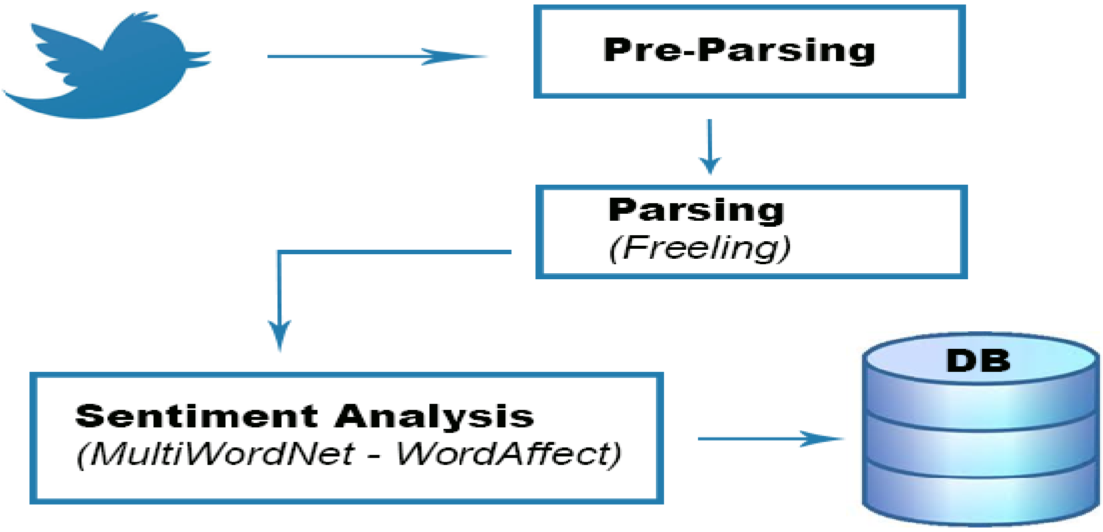


Figure 10: The Felicitta framework 2

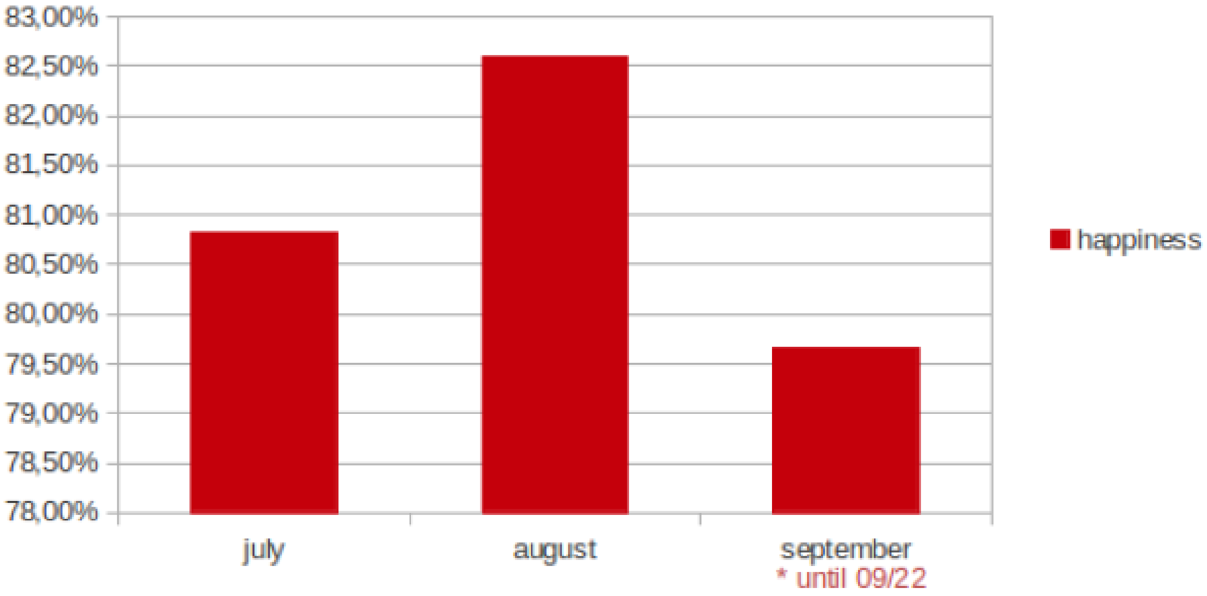
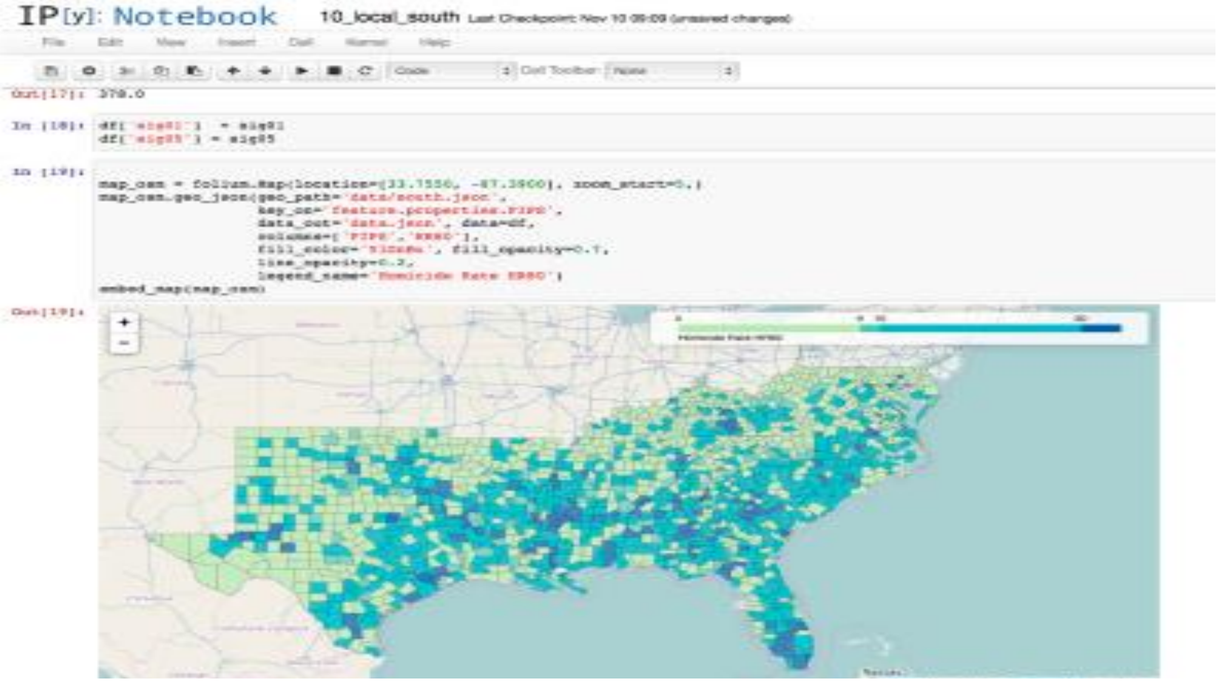


Figure 11: Variation of Happiness (12)

2.1.1.11 Open Geospatial Analytics with PySAL

This article shows the PySAL python library and its contribution in the spatial analysis and geospatial analysis. There is a need for open source library like PySAL to analyze the large amount of data and live data stream related to the spatial analysis. Geospatial analysis has some challenges like scalability challenge because the spatial statics in the prototype system was not designed for big data analysis. In this article, the efforts are focusing on these challenges. Format of PySAL library contains desktop programs, web application, and decision support system. PySAL have many platforms for the range of different use cases and it's designing to accommodate the spread of high-level applications data for the spatial analysis. PySAL contains modules helping the process of geospatial analysis like **weights** module. Weight module support three classes: contiguity based, distance based, hybrid. Also, it supports the reading and writing of 13 different formats for the spatial weights to make it easy to the spatial analysis packages. PySAL have another important module called **esda** module includes array of classification to be used with choropleth maps. Furthermore, **inequality** module that deals with statics to analyze inequality in spatial distributions (13).


IP[y]: Notebook
10_local_south Last Checkpoint: Nov 10 09:09 (unsaved changes)
File Edit View Insert Cell Kernel Help
Code
Cell Toolbar: None
+
−
Homicide Rate 1980

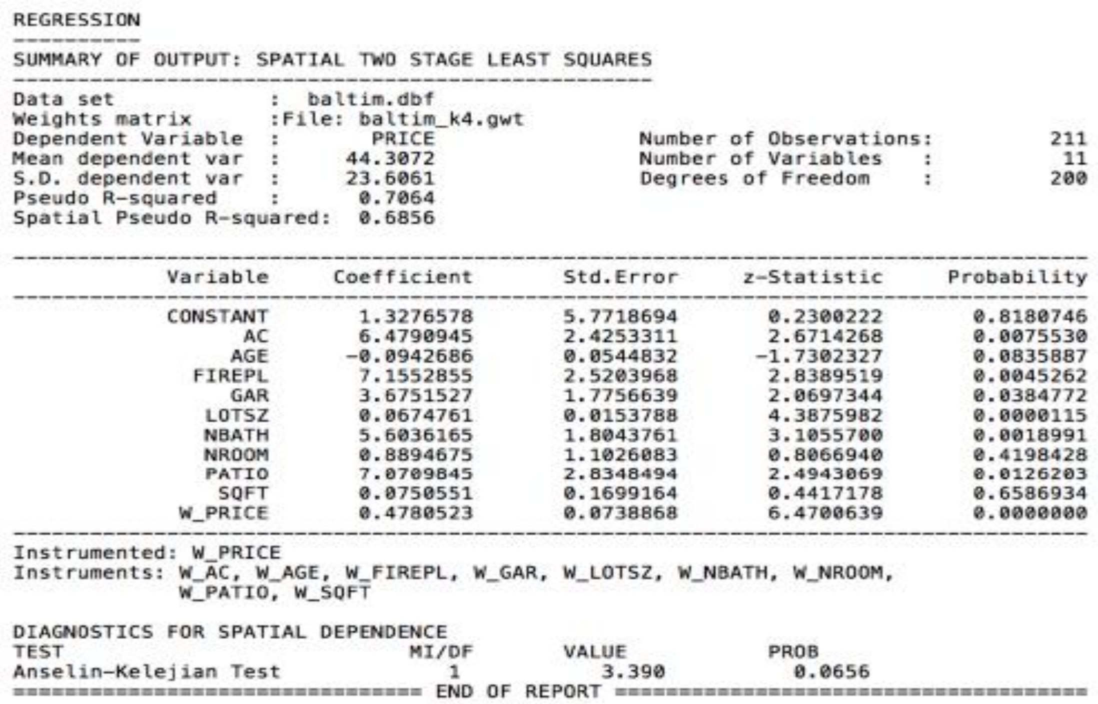
```
Out[17]: 378.0

In [18]: df['sig01'] = sig01
         df['sig05'] = sig05

In [19]:
         map_osm = folium.Map(location=[33.7550, -87.3900], zoom_start=5.)
         map_osm.geo_json(geo_path='data/south.json',
                          key_on='feature.properties.FIPS',
                          data_out='data.json', data=df,
                          columns=['FIPS','HR80'],
                          fill_color='YlGnBu', fill_opacity=0.7,
                          line_opacity=0.2,
                          legend_name='Homicide Rate 1980')
         embed_map(map_osm)

Out[19]:
```

Figure 12: PySal in IPython notebook (13)



```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
--------------------------------------------------
Data set            :  baltim.dbf
Weights matrix      :File: baltim_k4.gwt
Dependent Variable  :        PRICE               Number of Observations:          211
Mean dependent var  :      44.3072               Number of Variables    :          11
S.D. dependent var  :      23.6061               Degrees of Freedom     :         200
Pseudo R-squared    :       0.7064
Spatial Pseudo R-squared:   0.6856

-------------------------------------------------------------------------------------
             Variable     Coefficient       Std.Error     z-Statistic     Probability
-------------------------------------------------------------------------------------
             CONSTANT       1.3276578       5.7718694       0.2300222       0.8180746
                   AC       6.4790945       2.4253311       2.6714268       0.0075530
                  AGE      -0.0942686       0.0544832      -1.7302327       0.0835887
               FIREPL       7.1552855       2.5203968       2.8389519       0.0045262
                  GAR       3.6751527       1.7756639       2.0697344       0.0384772
                LOTSZ       0.0674761       0.0153788       4.3875982       0.0000115
                NBATH       5.6036165       1.8043761       3.1055700       0.0018991
                NROOM       0.8894675       1.1026083       0.8066940       0.4198428
                PATIO       7.0709845       2.8348494       2.4943069       0.0126203
                 SQFT       0.0750551       0.1699164       0.4417178       0.6586934
              W_PRICE       0.4780523       0.0738868       6.4700639       0.0000000
-------------------------------------------------------------------------------------
Instrumented: W_PRICE
Instruments: W_AC, W_AGE, W_FIREPL, W_GAR, W_LOTSZ, W_NBATH, W_NROOM,
             W_PATIO, W_SQFT

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                            MI/DF      VALUE           PROB
Anselin-Kelejian Test              1          3.390          0.0656
================================ END OF REPORT ======================================
```

Figure 13: GeoDaSpace for spatial econometrics functionality from PySal

2.1.1.12 Divergent discourse between protests and counter-protests

After a black teenager Michael Brown have been killed by a white officer Darren Wilson people created a hashtag on twitter #BlackLivesMatter. This hashtag was created for complaining against killing black people. Other twitter users created another hashtag #AllLivesMatter in response to the other hashtag. Author used three methods to reach to his final results. Data Collection, Entropy and diversity, Jensen-Shannon Divergence as a result of this study, its suggests that #BlackLivesMatter hashtag movement was able to grow, "exhibit diverse conversations, and avoid derailment on social media by making discussion of counter-protest opinions a central topic of #AllLivesMatter, rather than the movement itself" (14).

2.1.1.13 Geo-Twitter Analytics: Applications in Crisis Management

The study shows how to retrieve data from social media based on geographical way, "This strategy integrates computational methods for capturing, storing, and indexing tweets with visual query and analysis methods". Providing crisis analysis is helpful for detecting the problem in a fast way, but it hard to explain these disasters in twitter so the use of abbreviations is common, twitter used to locate disasters and crisis. SensePlace2 is a way of detecting tweets that have interesting topics and it used for finding disaster locations (15).

2.1.1.14 Twitter Analytics: Architecture, Tools and Analysis

This experiment used python as the programming part and MySQL as the representation and storage part, in order to collect the required data for this analysis, it is essential to use an API for two important reasons:

1- Low complexity of implementation,
2- Compatible with the python software.

This study implemented two Twitter APIs namely REST and Twython to enable data collection process, Twython API used to determine the location of the tweet and REST allow users to enter customized parameters to extract tweets (16).

2.1.1.15 Novel applications of social media analytics

This paper talks over social media and how it is so fast spreading everywhere with so much data with it. Not just the regular user uses the social media platforms like Facebook, Twitter, YouTube, Weibo, to share their information's in the real life but also the companies and government agencies contribute in this social media world. With the fast and large growth of social media usage caused huge user generated contents (UGCs). For that reason, comes the rule of social media analytics (SMA) to collect, summarize, analyze and visualize data to extract useful information's. Social media analytics have three major process: capture, understand, and present. So, they proposed novel solutions for social media analytics to understand the content and usage of social media. Also, to find some appropriate algorithms for scalable social media analytics. They find social media analytics framework that contains sentiment benchmarks to detect industry-specific marketing intelligence. Furthermore, An Empirical Analysis of Users to check users' privacy disclosure behavior via SMA. The result was that males and females are different from each other in privacy disclosure patterns (17).

2.1.1.16 COSMOS: Towards an integrated and scalable service for analyzing social media on demand

Data analysis and the analysis of social communication, develop the ability to reach the heart of the society, which is largely increased by social computing. The tools used to reach this type of analysis are often not shared between people and expensive. "The collaborative online social media observatory (COSMOS), an integrated social media analysis tool is presented, developed for open access within academia". COSMOS can analyze big data and can do it rapidly (18).

2.1.1.17 Tweetviz: Visualizing Tweets for Business Intelligence

Extracting data from social media can provide huge business opportunities, Tweetviz is another way to help business retrieve large raw data from twitter, Tweetviz can use tweet locations to a real benefit by knowing where businesses should consider to take action toward reaching a business goal (19).

2.1.2 Universities

2.1.2.1A case study of Israeli higher-education institutes sharing scholarly information with the community via social networks

This case study inspects the cases in Social Networking Sites (SNS) that have been used for scholarly purposes by higher education institutes in Israel. The research concerns with the data, patterns, interactivity with the two social networking sites Facebook and Twitter accounts of these institutes. So, the research has been conducted on 47 Facebook accounts and 26 Twitter accounts of Israeli universities or colleges within these institutes. All tweets within Twitter Official accounts of these institute have been analyzed and classified into categories based on their content. The world of Online Social Networks has become very important to share the academic knowledge between students and their universities, Also, for learning activities throughout these online social networks. Higher education institutes depend and adapting online education more their need of the traditional learning and there is so many trends about it like Open Courseware(OCW) that offer free courses online full of academic materials. The analysis of the data in the accounts of higher education institutes in Israel will be based upon the characteristics, patterns, and interactivity in the accounts. The methodology of the analysis as the following: Collect data from 47 Facebook accounts and 26 Twitter accounts, Facebook accounts of fan pages, Facebook and Twitter special features like first and last data of the post, total number of posts and tweets, number of likes and comments…etc. Furthermore, the wall in Facebook is the main feature to indicate the activities. The results of the analysis will be useful to get some values and extract useful information's regarding higher education institutes in Israel (20).

A. Forkosh-Baruch, A. Hershkovitz / Internet and Higher Education 15 (2012) 58–68

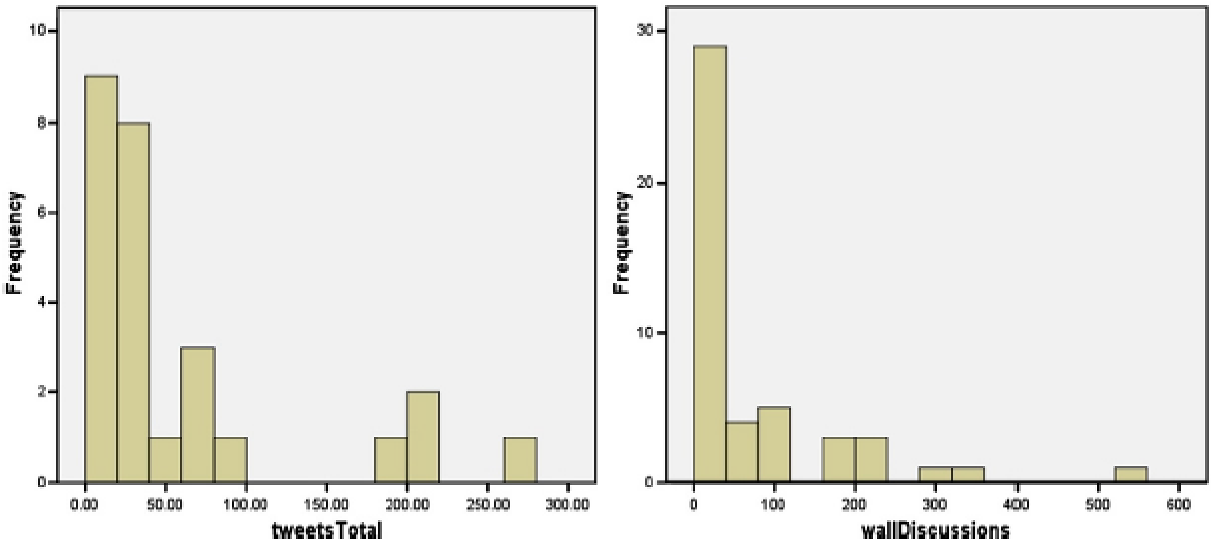


Figure 14: Distribution of overall number of tweets (left) and facebook wall messages (right) (20)

A. Forkosh-Baruch, A. Hershkovitz / Internet and Higher Education 15 (2012) 58–68

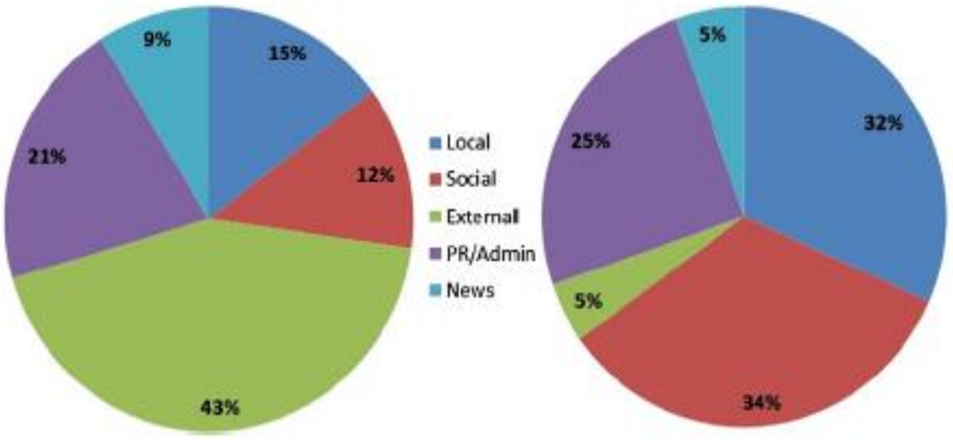


Figure 15: Distribution of universities tweets (left, N=699, 12 accounts) and colleges tweets (right, N=1095, 14 accounts)by content (20)

2.1.2.2 Twitter as a resource to evaluate the university teaching process

The purpose of the study is to evaluate the teaching process in Twitter through an experiment and questionnaire with open and closed questions delivered to the students to do it. The number of the students was 146 students voluntarily or anonymously divided into three groups. They evaluated each class in the categories through tweets. The analysis of the evaluation is done through many indicator systems. The study agrees with Kassens-Noor (2012) about that Twitter can be used as a tool for continues and formative teaching evaluation. The methodology of the analysis will be based on a qualitative design on the analysis of the data generated. Collection of the data was through Twitter to get the feedback from the students on the activities in the classroom, and questionnaire with three questions, two closed, one open for the student to write his opinion about the use of Twitter in the learning process. Also, to collect the data the teachers wrote a tweet in their accounts to encourage the students to participate in the hashtags to give their opinions freely and honestly about the courses and how to improve the teaching process. Qualitative data was gathered and organized for the analysis and processing using NVivo 10 software. The total results were 495 tweets from the students and based on the number of tweets it has been noticed that the most participation appears around conferences and practical classes. Finally, to determine the evaluation of the entire teaching process, the students tweets classified into three aspects: Positive, Negative and Reflections. They conclude at the end that Twitter as a tool for evaluating classes is broadly positive (21).

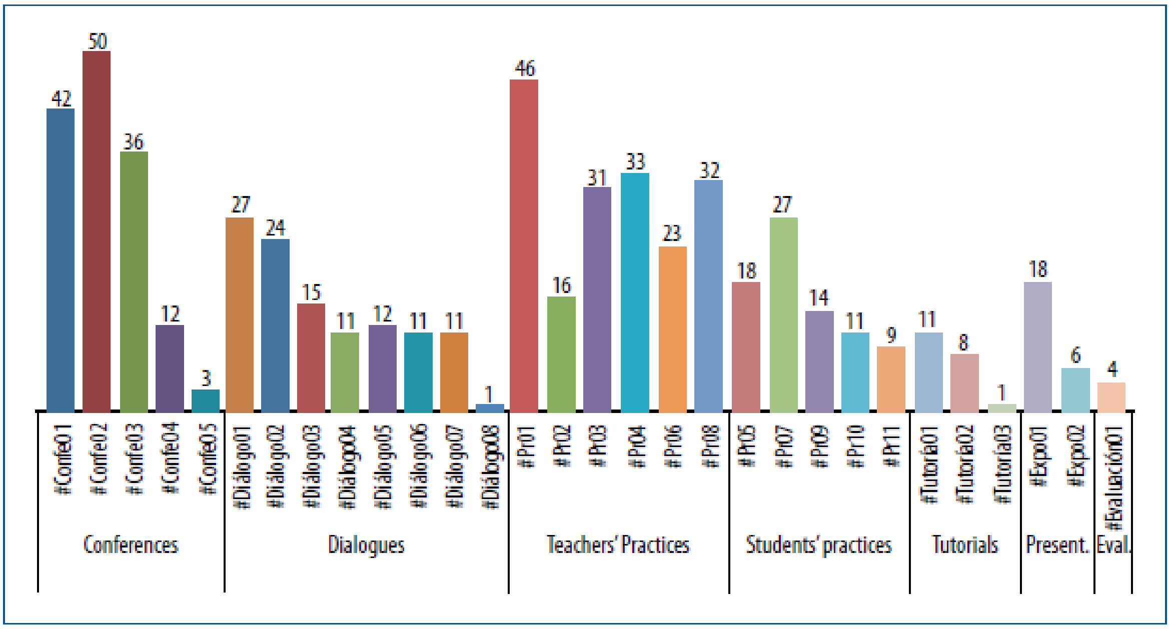


Figure 16: Number of tweets in each class organized by hashtags (21)

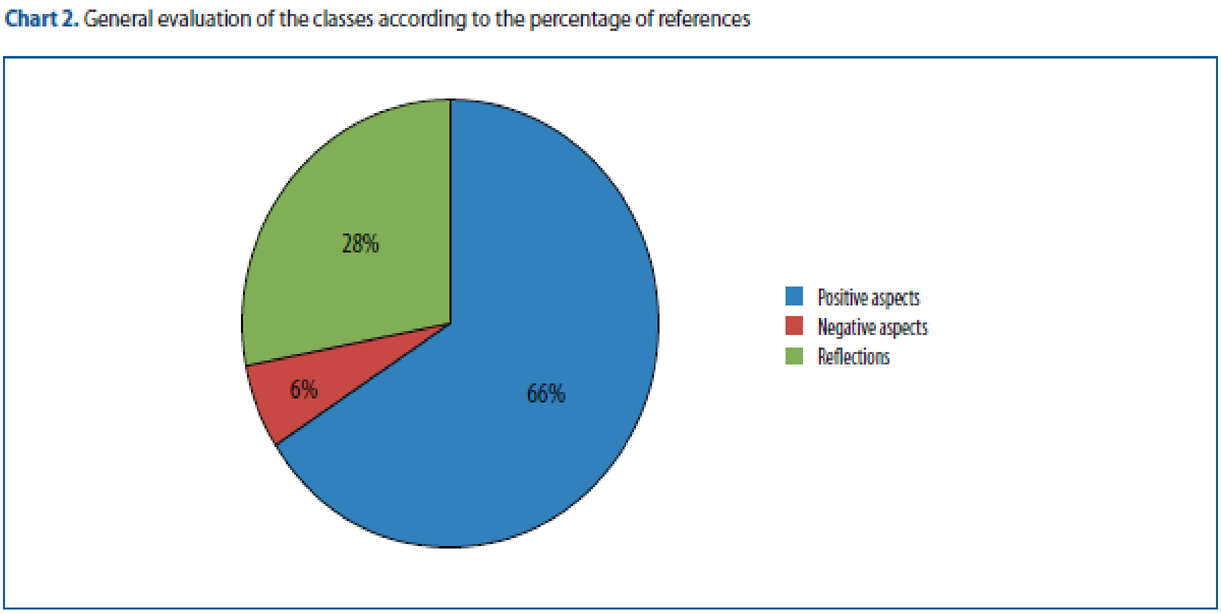


Figure 17: General evaluation of the classes according to the percentage of refrences (21)

2.1.2.3 A systematic identification and analysis of scientists on Twitter

This research proposes a systemic approach for identifying and analyzing scientists on Twitter based on the altimetric activities generated by scientists. Altmetrics are about quantitative investigations of learning activities on social media. They found the common disciplines in Twitter between scientists. The research combines both literatures methodologically and conceptually. They have created new methods that remove uncertainty and determining the actors on social media that describe the behavior of scientists to provide important information's for the use of indictors related to the social media metrics. Scientists use Twitter platform to share and discusses scientific topics in different disciplines. For example, biomedical students use the journal clubs on Twitter for writing daily activities. The source for analyzing altmetrics can be APIs on Twitter. They need to analyze the altmetrics to help identifying scientists on Twitter and to do that they must differentiate between who write tweets from the public and the tweet from the scientists. The method is a systematic study of scientist with different disciplines on Twitter that does not rely on external bibliographic databases; also, the method identifies scientists based on previous studies that used Twitter lists to identify user expertise. They classified the literature into two groups, one is names *product*, the second named *producer*-centric perspectives. The total analysis was 45, 867 identified scientists and they get 88, 412, 467 following pairs and 64, 449, 234 statuses (22).

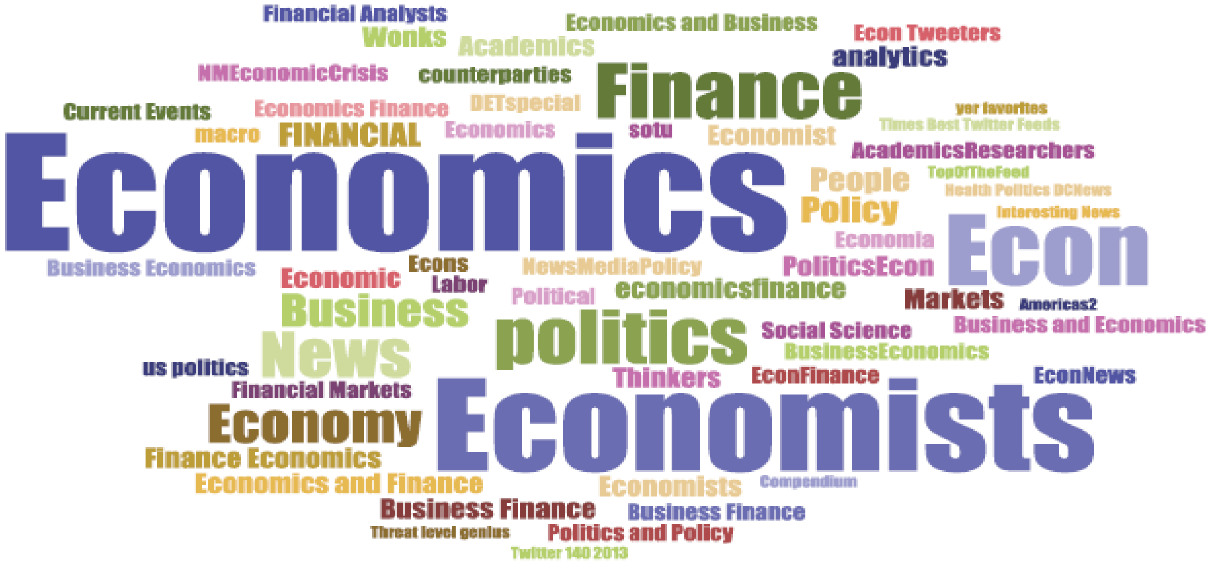


Fig 1. User identity recored from Twitter list names. We show the word cloud of Twitter lists containing @BetseyStevenson.

Figure 18: User identity record from twitter list names (22)

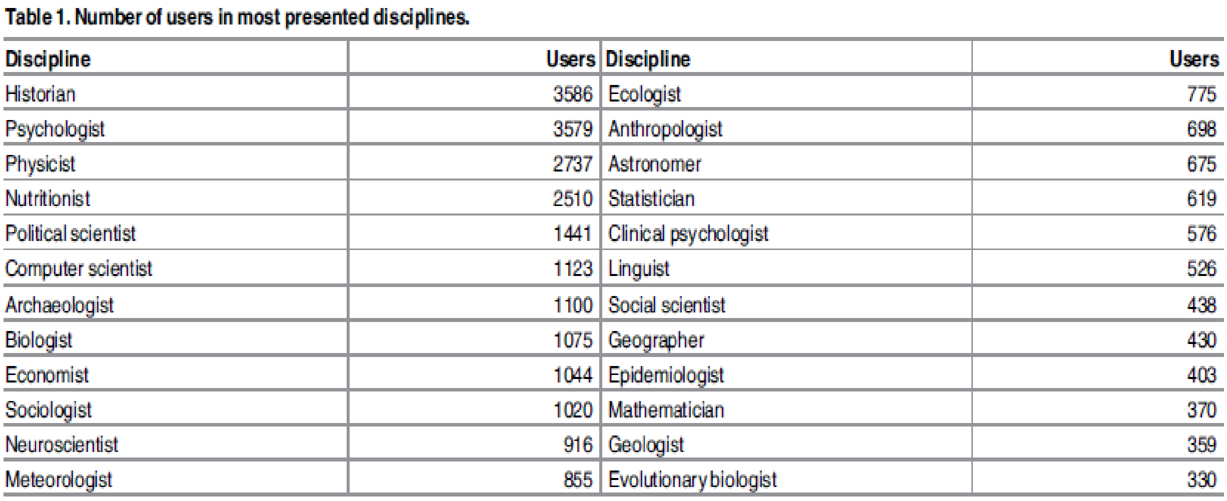


Table 1. Number of users in most presented disciplines.

Discipline	Users	Discipline	Users
Historian	3586	Ecologist	775
Psychologist	3579	Anthropologist	698
Physicist	2737	Astronomer	675
Nutritionist	2510	Statistician	619
Political scientist	1441	Clinical psychologist	576
Computer scientist	1123	Linguist	526
Archaeologist	1100	Social scientist	438
Biologist	1075	Geographer	430
Economist	1044	Epidemiologist	403
Sociologist	1020	Mathematician	370
Neuroscientist	916	Geologist	359
Meteorologist	855	Evolutionary biologist	330

Figure 19: Number of users in most presented disciplines (22)

2.1.3 Visualization

2.1.3.1 Visual Sentiment Analysis on Twitter Data Streams

In this poster, this study presents an approach to explore twitter, as shown in Figure 1. The approach attempts to automatically retrieve and analyze large number of tweets and classify those to tweets into two different categories (positive and negative). "Novel topic-based text stream analysis technique that automatically detects which attributes were frequently commented on in tweets, based on their density distribution, negativity, and influence characteristics" (23).

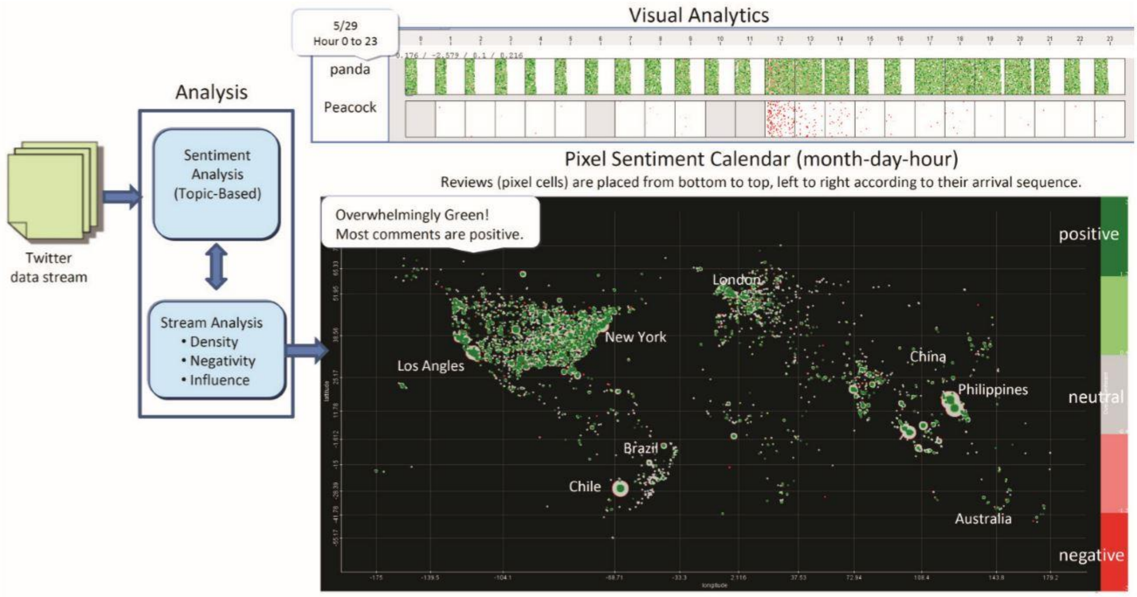


Figure 20 : Pixel sentiment geo map (23)

2.1.3.2 TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief

Microblogs provide huge and valuable information, this study presents a new application called HADR Humanitarian and Disaster Relief, this application helps organizations to analyze, track, and monitor tweets. This tool provides quick awareness prior to any crisis or a disaster. It can find the location from tweets so it results in finding the disaster location. HADR uses twitter API to retrieve tweets and store them for future analysis and a back-end database to stores incoming data (24).

2.1.3.3 Distilling Massive Amounts of Data into Simple Visualizations

Twitter Case Studies

This research data warehouse is built around a large Hadoop cluster. Data retrieved in the Hadoop Distributed File System (HDFS), Hadoop is implemented in Java, analytics are performed using Pig, "a high-level dataflow language that compiles into physical plans that are executed on Hadoop (Olstonetal.2008)".Pig uses common operations such as projection, selection, group, join, etc. This analysis comes at low cost (25).

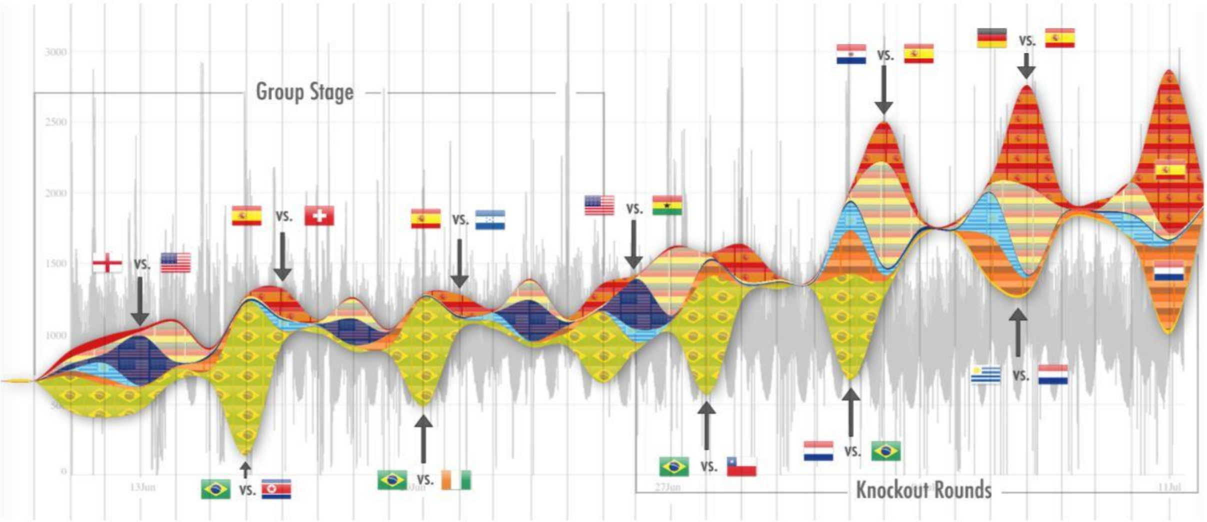


Figure 21: Background shows number of tweets per second and Foreground shows tweet volume of the hashtag (25)

2.1.3.4 ThemeStreams: Visualizing the Stream Of Themes Discussed in Politics

This study focusing on the political topics on social media that influencing public opinions and how it's increasing every day and there so much hot topics in the headlines that need to be analyzed to show which theme are being discussed. So, they propose ThemeStreams which is a tool that explains the political discussions to themes and the influencers and to show the mapping of interactive visualization. The participants of political discussion on social media are four types: those who have important position within the government, second, those who lobby for important issues, third, the journalists, forth, the public. Tweets are being collected for the four influencer groups since 2011 and at that time ThemeSpread index driving was 3.9 million tweets. ThemeStreams contains also analysis in other domains, such as newspaper archives (26).

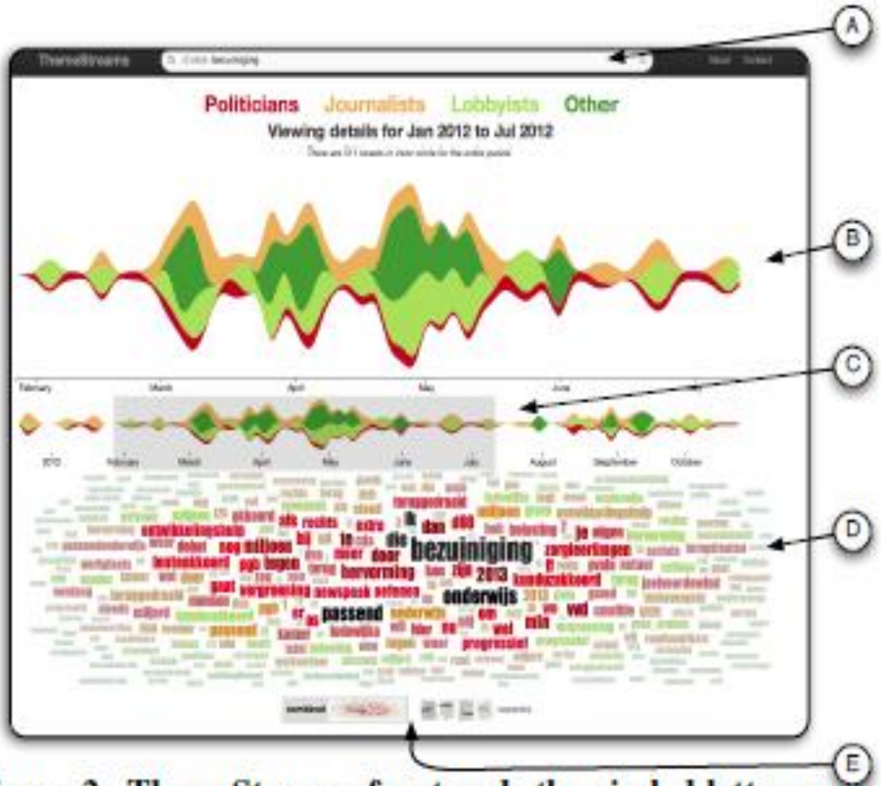


Figure 2: ThemeStreams front-end; the circled letters are explained in the text.

Figure 22: ThemeStreams front-end; the circled letters are explained in the text (26)

2.1.3.5 Document visualization: an overview of current research

This research shows multiple and important designs and ideas for visualization of documents and articles. **Single Document Visualization**, in this type of visualization the main focus is in the central features e.g. word, phrases etc. **Vocabulary-Based Visualization**, it helps the people understand the English vocabulary in visualized representation. **Visualization Based on Semantic Structure**, this type of visualization is interesting because it lets readers have a clear and brief view without going through the entire document, so it will save time for readers and let them read multiple topics (27)

.

2.2 Analysis of Related Work

Some previously shown researches don’t support real-time analysis furthermore others don’t support universities and networks, and some researches are lacking to provide multiple visualization techniques. One additional negative point is that they showed an old way to represent the data.

	Data Source	Visualization	Problem	Open Source	Evaluation
Paper 1	Streaming API	Sigmoid Design, Scatter Plot	Extracting real-time tweets and define related topics.	Twitter live detection framework (TLDF)	It is a good framework for real-time analysis of twitter data to detect related topics.
Paper 2	Filter API	Bar Charts, Line Graphs	Provides end to end solution for microblogs users	Microblogs System	Good use for large amount of data.
Paper 3	Twitter API, SmartFeeds	Volume Graphs	Analyze data to provide insight into trends for different range of applications	Scalable Reasoning System (SRS)	Lack of multiple visualization techniques

Paper 4	Twitter API	Scatterplots, Volume Graphs, word frequency charts, Bar Charts	Study about relevance of incorporating tweets	ScatterBlogs Visual analytics system	useful visual social media analytics
Paper 5	Streaming API	Scatterplot, Line graph	Finding different people interactions from Weibo and Twitter	Interactive analyzing software	Charts represent differences between weibo and twitter in an interactive way
Paper 6	Twitter API	Pie Chart, Bar Chart, Histogram	Organize big data to give a clear view about it and how to manage it	Distributed system, business intelligence software analytic framework	Twitter has unstructured data, this system makes it easier to retrieve it
Paper 7	ORA-NetScenes	Line Graphs, Volume Graphs	Develop real-time plans to deal with phenomenon of crisis	TweetTracker, REA,ORA-NetScenes	Deals perfectly with the time of crisis
Paper 8	Streaming API, REST API	Comparison between relational model and NoSQL model, Volume graph	Understand and analyze tweets	Visual Analytics, Open Authentication (OAuth), Twitter tweet object code	Open authentication provides users to access protected information

Paper 9	Streaming API		Increasing number of big data that needs to be processed and retrieve value from it	Quality analysis framework	Analyzing data about social media in general
Paper 10	Twitter API	Histogram, Line graphs	Analyzing geotagged tweets in Italian cities to estimate happiness	The Felicitt'a Framework	Finding positive reactions from people
Paper 11	Live Data Streams		Finding locations of social media activities	PySAL Python library, Geospatial analysis, Provenance for spatial weights generation code	Analyzing different types of social media activities
Paper 12	Twitter API	Volume graphs , Network graph	Protest led People to participate in the wrong hashtag	Data collection, Entropy and diversity	Interesting topic related to real life problem
Paper 13	Twitter API		Retrieve data based on geographical way	Multiple view interfaces	Provide detailed tweets locations

Paper 14	Streaming API	Flowcharts, Line Graphs	Understand and analyze tweets	REST, Twython	Old way to represent findings
Paper 15	Social media analytics (SMA)		To collect, summarize, analyze , and visualize data o extract useful information	Empirical analysis	Good way to find new useful applications based on social needs
Paper 16	Twitter Streaming API	Line graphs	Understanding People behavior and interaction on social media	COSMOS	Guide researchers to use the right tool for the right time
Paper 17	Twitter API		Retrieve large data to use it in business purposes	Tweetviz	Good tool help business achieve their goals
Paper 18	Facebook, Twitter API	Bar Chart, Pie Chart	Sharing academic knowledge between students and their universities in social media	Open Courseware (OCW)	Provides Ease of communication between students and universities
Paper 19	Twitter API	Pie Chart, Bar Chart	Evaluating teaching process through hashtags and	NVivo 10 Software	Improving teaching process through the students tweets will help elevate

			surveys from students		the teaching process
Paper 20	Twitter API	Word frequency chart	Differentiate between tweets from professional scientists and regular people, determining the different disciplines in scientists, number of users as well	Systemic study of scientist based on altemetrics	Save time in searching for right information from scientists
Paper 21	Twitter API		Classify tweets to two categories (Positive, Negative)	Visual Analysis Tool e.g. SAS, Polyanalyst	Great way of representing positive tweets
Paper 22	Twitter API	System Architecture of tweet tracker, Maps	Analyzing for disaster and crisis based on social media	TweetTracker	Perfect monitoring of tweets
Paper 23	Pig	Volume Graph	Locating high volume of data in social media	Hadoop Distributed File System (HDFS)	Great way to visualize different volumes of data
Paper 24	Twitter API	Volume Graph	Determining the political topics that	ThemeStreams	Good way to Find the most

			have been discussed in twitter		interesting political topics
Paper 25	Collecting Documents	Word frequency chart	Collecting multiple online documents, and providing visualization techniques to represent it	CMV	Provide readers with multiple techniques to save their time

2.3 Overview of Implementation Tools

The original source code of the previous studies is not freely available for everyone, so we will build an open-source system, also the system will deal with real data from universities supporting real-time analysis with multiple visualization techniques.

We will use the following software tools in our project:

- Spyder editor for python programming.
- Redis Server and Celery package for backend services.
- MongoDB to store tweets as collections.
- Robo 3T for accessing MongoDB.
- Bokeh library for visualization.
- Seaborn and Matplotlib for visualization.
- RapidMiner for more visualization.
- Pandas library to format and generate tables and statistics.

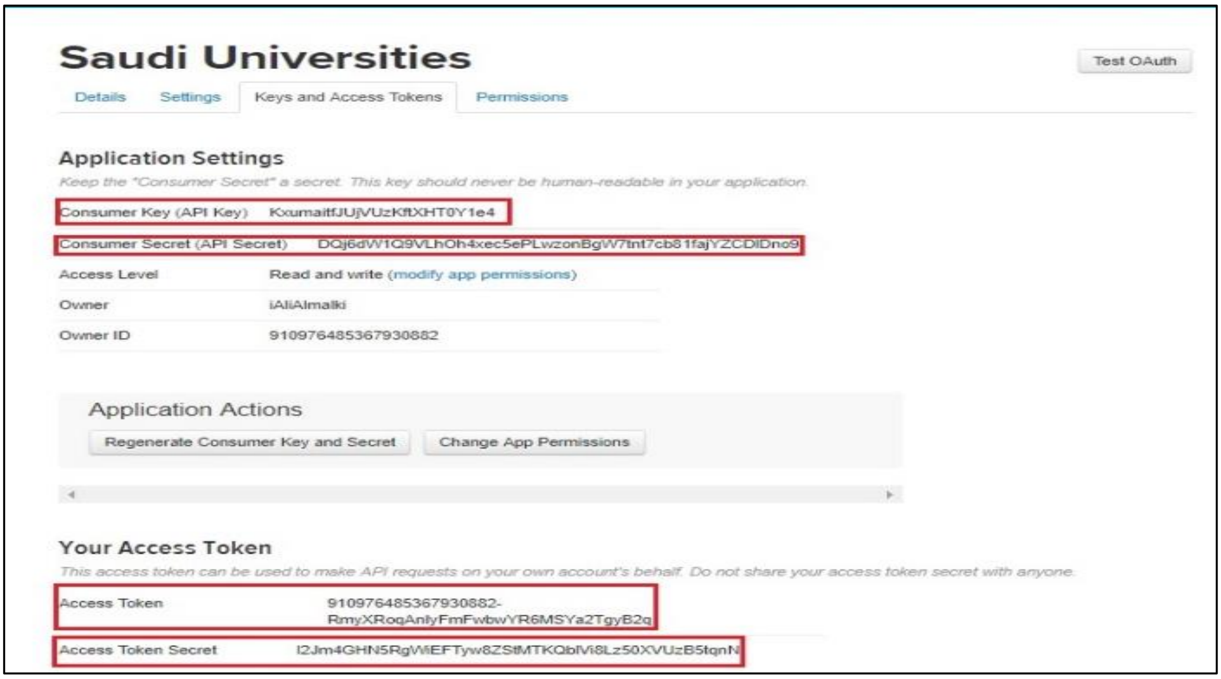
3. Chapter 3 | Analysis

3.1 Requirement Capture or Data Collection

The goal of our project is to analyze tweets. So, we have collected nearly 30000 tweets from some official public and private universities accounts on twitter (10 Accounts). Then we have stored and saved those tweets in Mongo database, to gain access to the saved tweets we used Robo 3T, Robo 3T allows us to view and analyze tweets by different queries, these queries vary depending on what type of analysis is needed

3.1.1 Tweet Gathering Process

First you have to create an app on twitter in https://apps.twitter.com, twitter will initialize four major keys that is specific and unique to the app, consumer key, consumer secret, access token, access token secret, when writing the code in Spyder we need to establish a successful connection between Robo3T and MongoDB.



Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	KxumaitfJUjVUzKftXHT0Y1e4
Consumer Secret (API Secret)	DQj6dW1Q9VLhOh4xec5ePLwzonBgW7tnt7cb81fajYZCDIDno9
Access Level	Read and write (modify app permissions)
Owner	iAliAlmalki
Owner ID	910976485367930882

Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	910976485367930882- RmyXRoqAnlyFmFwbwYR6MSYa2TgyB2q
Access Token Secret	I2Jm4GHN5RgWiEFTyw8ZStMTKQbIVi8Lz50XVUzB5tqnN

Figure 23: Four major Keys

When writing the code there are several libraries must be added:

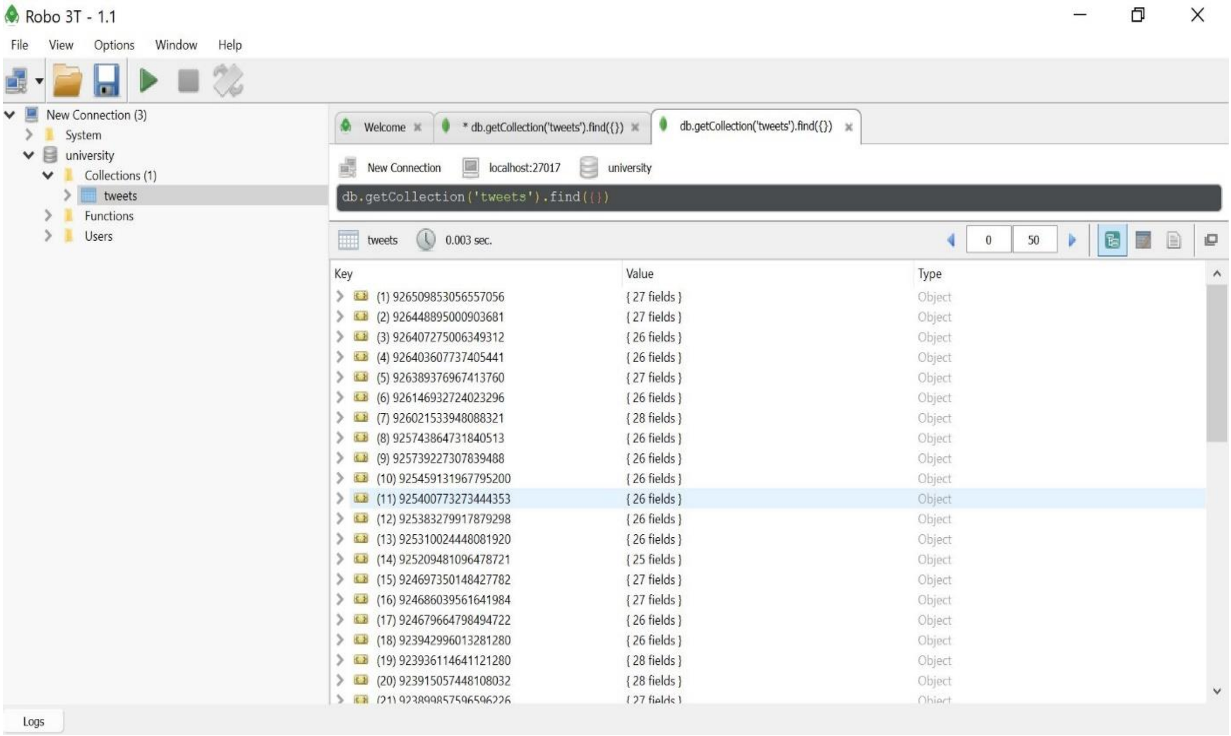
- Tweepy.
- Datetime.
- Pymongo.
- Mongoclient.
- Bokeh.

Now a database must connected through localhost server at port number 27017, Also an authentication must be added to the code.

There are three main functions in collecting tweets and saving it in a file or MongoDB:

- Api.user_timline: Collect tweets from a specific user.
- Api.search: Collect tweets based on a specific keyword.
- Filter: Uses real-time collection of tweets.

For our analysis phase we used api.user_timeline to gather tweets from 6 different universities.



```
db.getCollection('tweets').find({})
```

Key	Value	Type
(1) 926509853056557056	{ 27 fields }	Object
(2) 926448895000903681	{ 27 fields }	Object
(3) 926407275006349312	{ 26 fields }	Object
(4) 926403607737405441	{ 26 fields }	Object
(5) 926389376967413760	{ 27 fields }	Object
(6) 926146932724023296	{ 26 fields }	Object
(7) 926021533948088321	{ 28 fields }	Object
(8) 925743864731840513	{ 26 fields }	Object
(9) 925739227307839488	{ 26 fields }	Object
(10) 925459131967795200	{ 26 fields }	Object
(11) 925400773273444353	{ 26 fields }	Object
(12) 925383279917879298	{ 26 fields }	Object
(13) 925310024448081920	{ 26 fields }	Object
(14) 925209481096478721	{ 25 fields }	Object
(15) 924697350148427782	{ 27 fields }	Object
(16) 924686039561641984	{ 27 fields }	Object
(17) 924679664798494722	{ 26 fields }	Object
(18) 923942996013281280	{ 26 fields }	Object
(19) 923936114641121280	{ 28 fields }	Object
(20) 923915057448108032	{ 28 fields }	Object
(21) 923899857596596226	{ 27 fields }	Object

Figure 24: Robo 3T Displaying tweets

3.1.2 Analysis of Collected Data

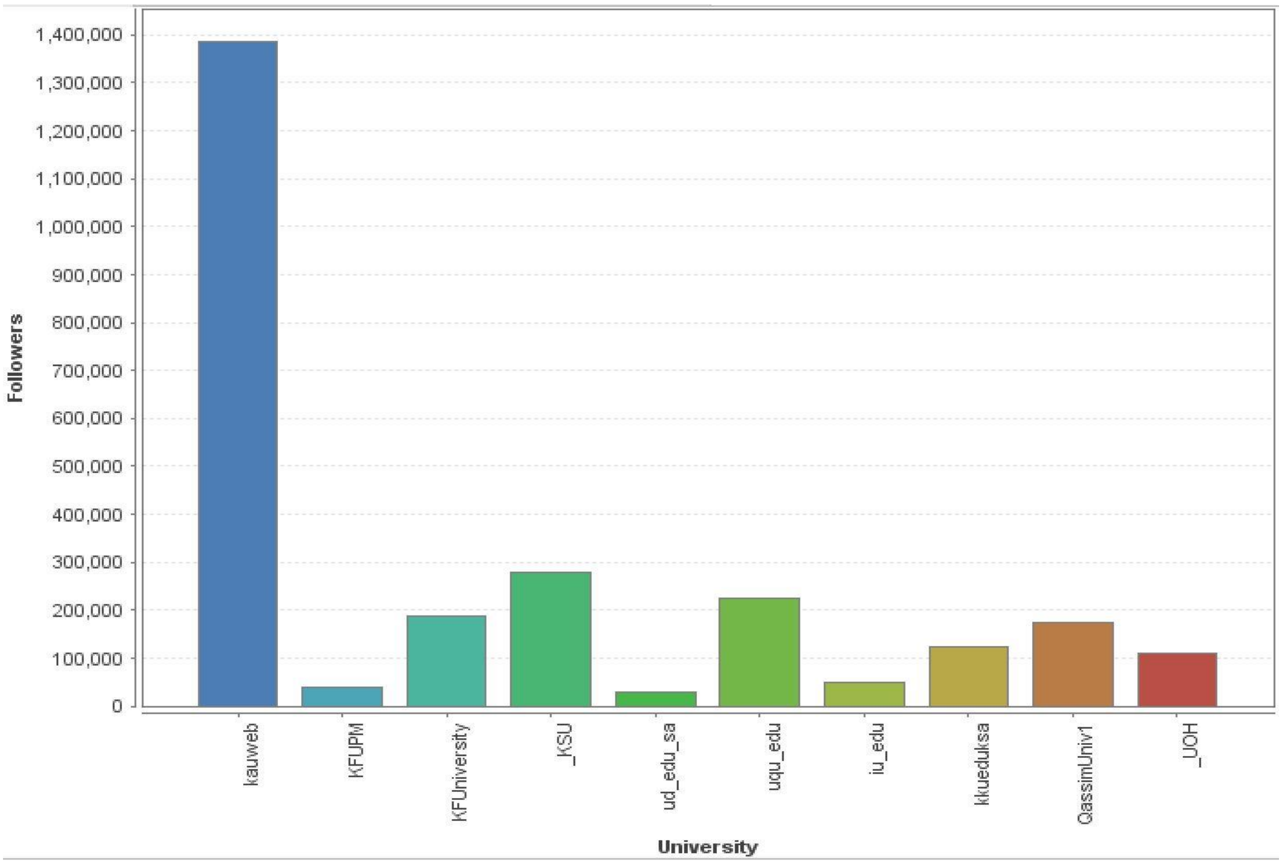


Figure 25: Number of Followers in Saudi Universities

Figure 25 shows a comparison between number of followers from 10 Saudi universities, coming in first place is king Abdulaziz university with nearly 1,400,000 followers and there is a huge difference between them and other Saudi universities, and in second place is King Saud university with 275,000 followers.

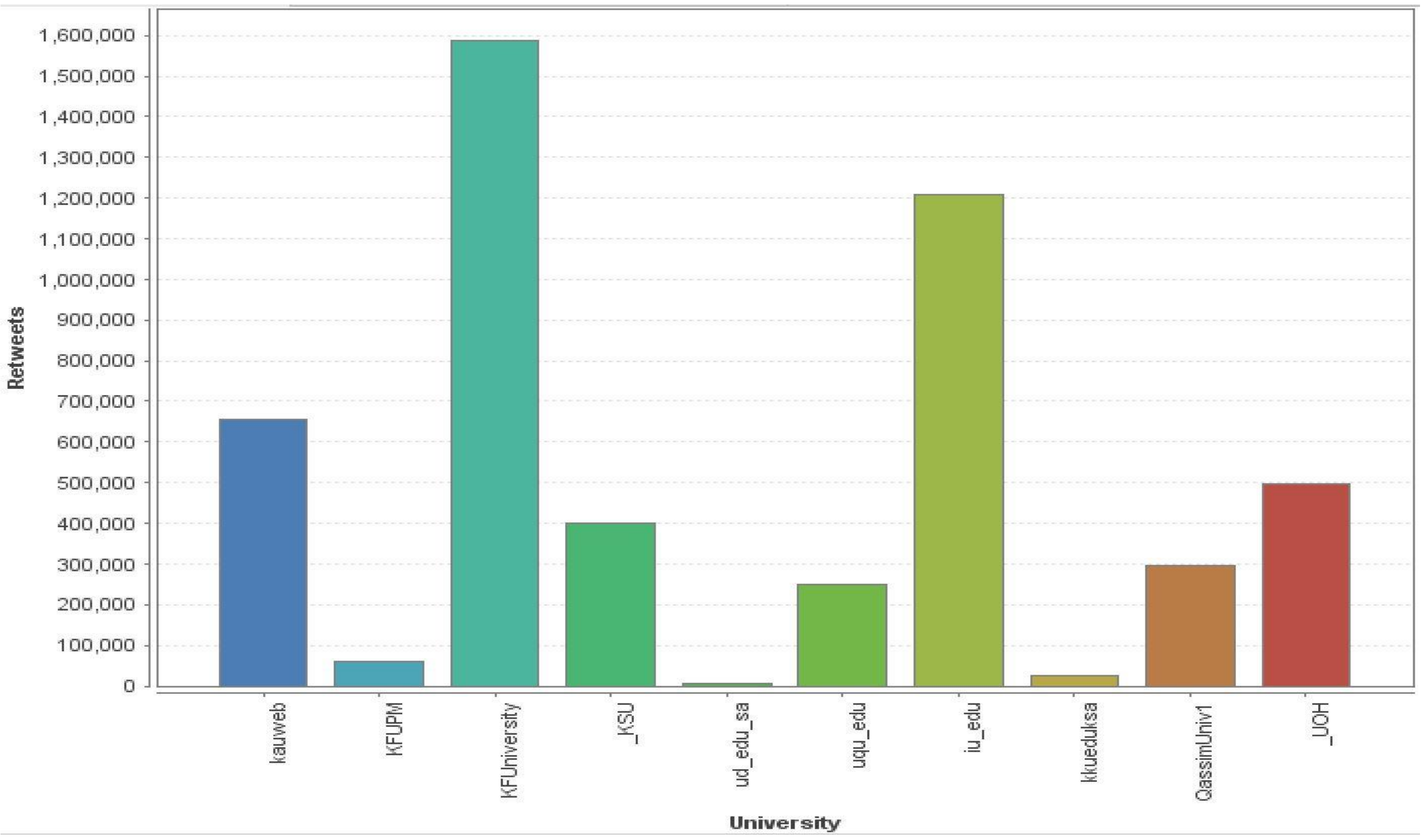


Figure 26: Number of Retweets

Based on figure 25 we have seen the number of followers, but figure 26 shows number of retweets from each of the 10 universities. This clarifies that it is not necessary that having a huge number followers can bring so many social response to tweets. KFUniversity with 1,600,000 retweets from more than 3000 tweets, the most number of retweets from our sample.

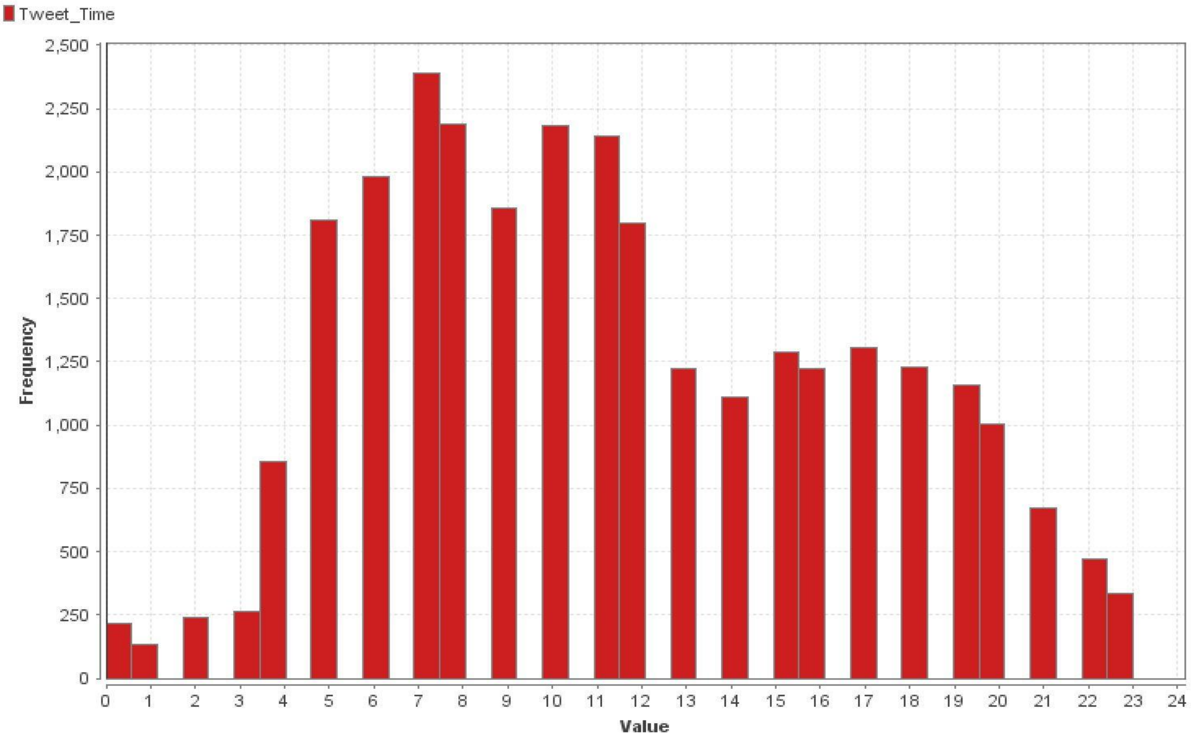


Figure 27: Most frequent tweet times

Figure 27 displays the frequent tweeting times of the 10 universities and it clearly shows that in the early hours of the day universities tweet a lot, while late hours shows decreased amount of tweeting and social activity.

3.1.3 Conclusion

We concluded from our analysis that the users of Twitter and especially users who are interested in educational purposes with their universities on Twitter need an application to save their time and to make their reading more objective. Likewise, to learn some scientific statistics and the magnitude of interaction, extract useful information from tweets and lots of stuff.

3.2 Requirement Specifications

Functional requirements are the main features of the application to be.

3.2.1 Functional Requirements

- Graphical user interface for the user to navigate the menu.
- Support for getting data from twitter using Filter, Search, and UserTimeline APIs.
- Application connects to twitter API and open an authorization gate automatically.
- Develop an interface to analyze the data using charts and visualizations
- Save graphs and charts in different formats.
- Provide time based rankings of universities based on their activities on twitter.
- Provide comparison of universities based on different attributes.
- Visualize in different graphs what people are talking about universities in real time.
- Display tweet patterns based on specific keywords for the tweets of the public.

3.2.2 Non-functional Requirements

- User can use many types of visualizations.
- Simple interface for user queries.
- Real-time data gathering.
- The response time of the application should not be more than 2 minutes.
- Able to refresh and recover all changes made up to one minute prior to the failure.
- Support for exceptions handling to avoid application crash.
- Application will not read or write your tweets and your private information will be safe.
- Running multiple tasks in backend which it does not appear to the user.

3.2.3 External Interface Requirements

Python programming language will be used to program the application.

- Python for programming
- Bokeh for visualization
- Flask.
- Mongo DB.
- Redis Server
- Celery for backend services.
- Robo 3T.

3.3 User Profile

Here we will specify which types of users will use the application, and what the characteristics of users are, and finally in what environment will the application be more useful for the users.

3.3.1 User Categories

- Teachers.
- Students.
- Administrative staff at universities.
- Anyone who has interest in universities.

3.3.2 Sample Specification

The users can use our application to collect, browse, store, or present tweets.

3.3.3 User Characteristics

- Menu navigation.
- Collecting tweets.
- User can conduct data analysis related to specific universities.
- Presenting tweet relationships in graph or charts.
- Saving graphs.

3.3.4 Environment

It will be useful for educational purposes, also knowing how much universities interact and use twitter.

Universities can know what people are saying about them, and people or students can have a deep background about universities.

3.4 Structuring System Requirements

- Data Gathering Module.
- Storage Module.
- Analysis Module.

3.4.1 System Use Case Steps

Use Case	Graphical user interface for the user to navigate the menu.
Actor	User, Database
Events	1. Open main page. 2. Choose between the three main choices. 3. Implement selected choice

Use Case	Support for getting data from twitter using Filter, Search, and User TimeLine APIs.
Actor	User, Twitter Timeline
Events	1. Choose from main interface "Crawl for Tweets". 2. Determine the collecting method. 3. Implement selected choice.

Use Case	Application connects to twitter API and open an authorization gate automatically.
Actor	User, Twitter API
Events	1. Open Login Page. 2. Open Authorization Gate. 3. Connect to Twitter API.

Use Case	Develop an interface to analyze the data using charts and visualizations and saving them in different formats.
Actor	User
Events	1. Choose from main interface "Visualization". 2. Select "Show Collected Data". 3. Determine graph/chart type. 4. Click "Save" 5. Choose the appropriate format.

Use Case	Provide time based rankings of universities based on their activities on twitter.
Actor	User
Events	1. Choose from main interface "View Stored Tweets". 2. Select "View Universities Information". 3. Determine the ranking type based on latest activity.

Use Case	Visualize what people are talking about universities in real time.
Actor	User , Twitter Timeline
Events	1. Choose from main interface "Visualization". 2. Select "Visualize tweets in real-time". 3. Choose desired attribute. 4. Determine graph/chart type. 5. Click "OK"

Use Case	Display tweet patterns based on specific keywords for the tweets of the universities and the public.
Actor	User, Twitter Timeline
Events	1. Search by keyword. 2. Write keyword. 3. Display statistics whenever universities or people talk about the given keyword.

4. Chapter 4 | System Design

4.1 Data Dictionary

This is the general view of our system architecture. It contains four major modules and each module have specific functions to conduct. Starting with the user interacting with the system and then choosing one of the main options of the system. Thus, conduct the methods specified within the module.

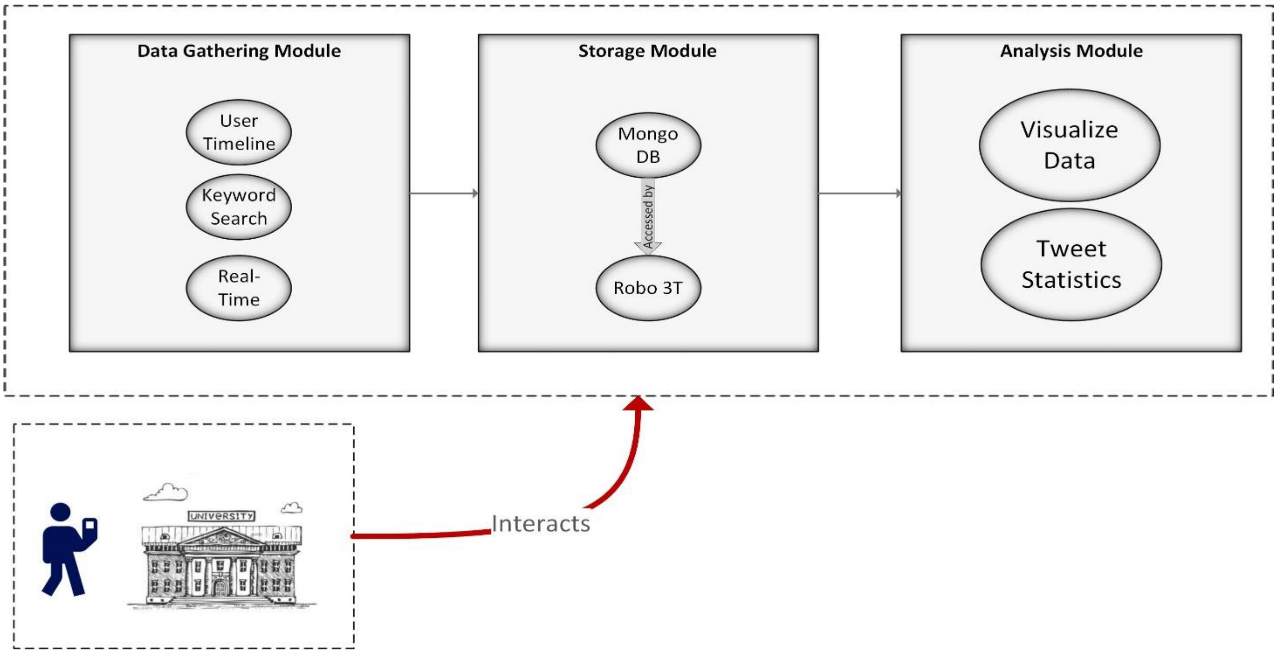


Figure 28: System Architecture

4.2 Interface Design

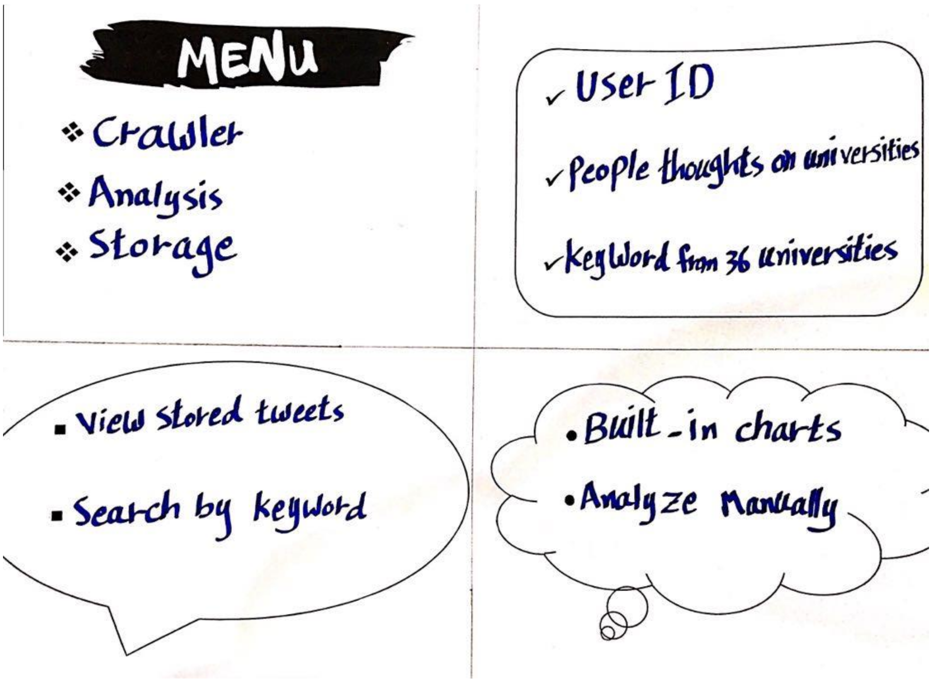


Figure 29: Interface 1

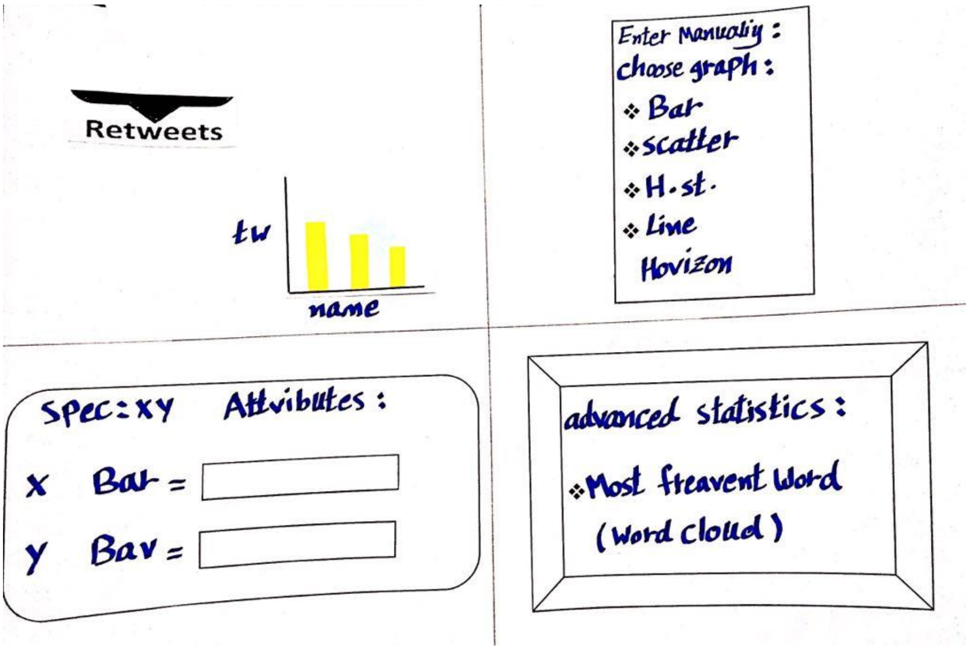


Figure 30: Interface 2

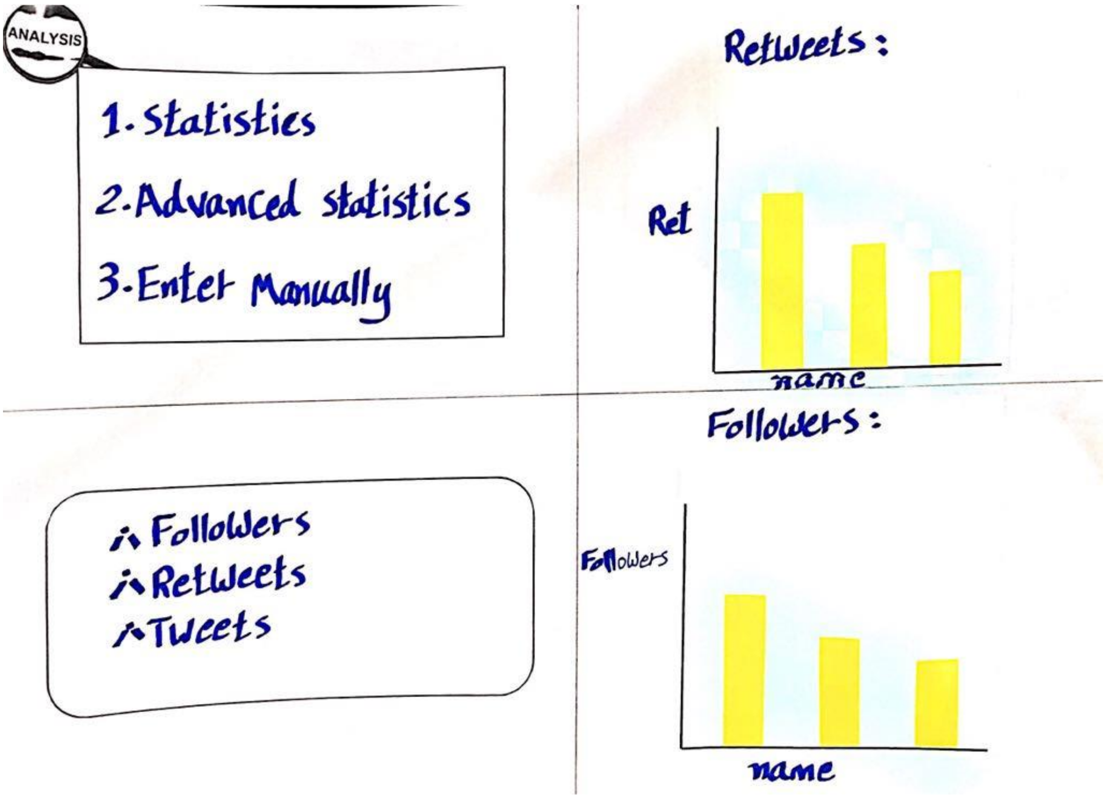


Figure 31: Interface 3

4.3 Use Case Diagrams

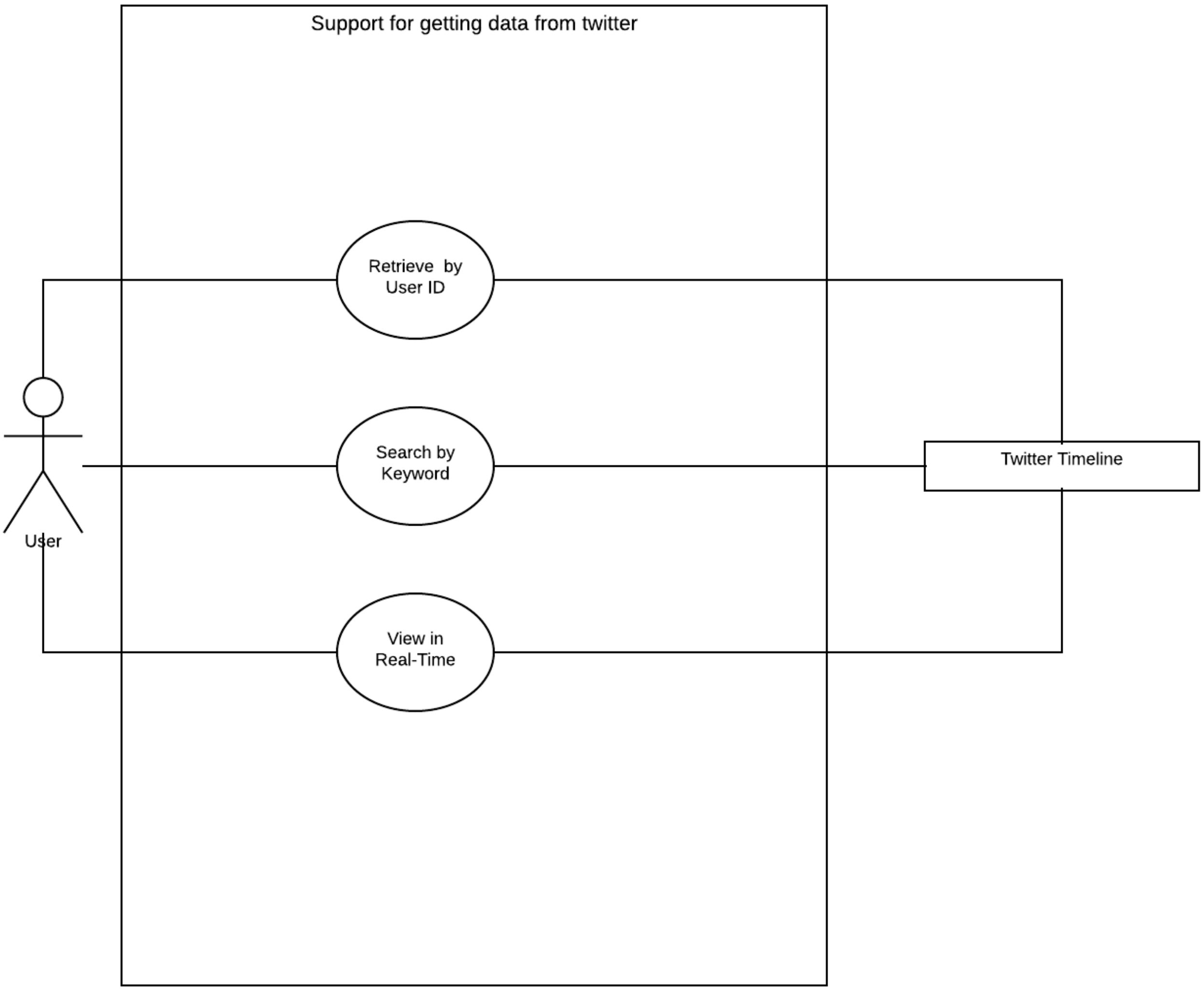


Figure 32: Use Case 1

This use case shows the steps that will be taken by the user in case he wants to collect tweets from twitter. First, the user will choose from the main interface collect tweets. Then, he will determine the collection method.

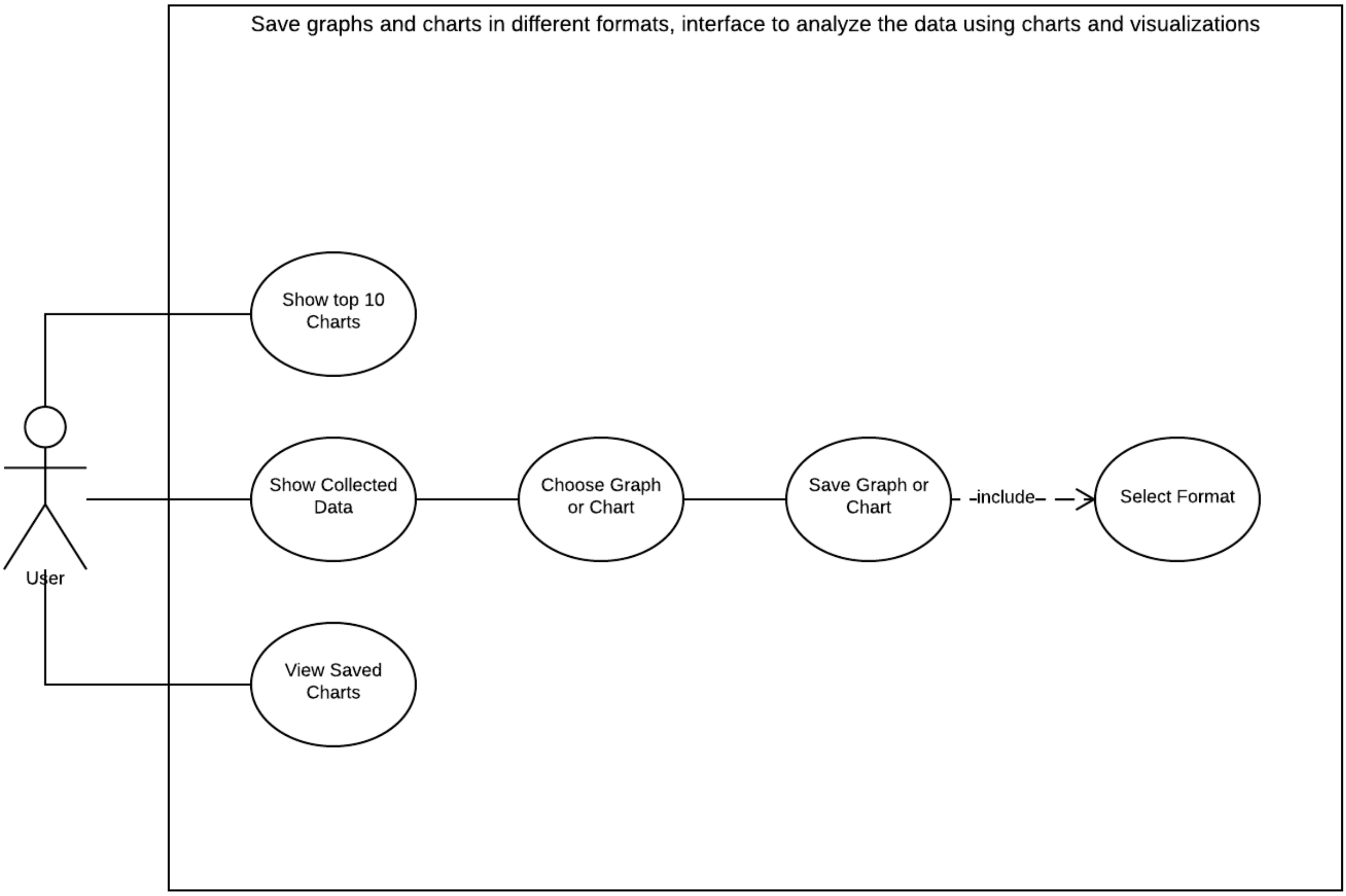


Figure 33: Use Case 2

This use case highlights on the process of representing tweets into charts or graphs and saving them to a specific format

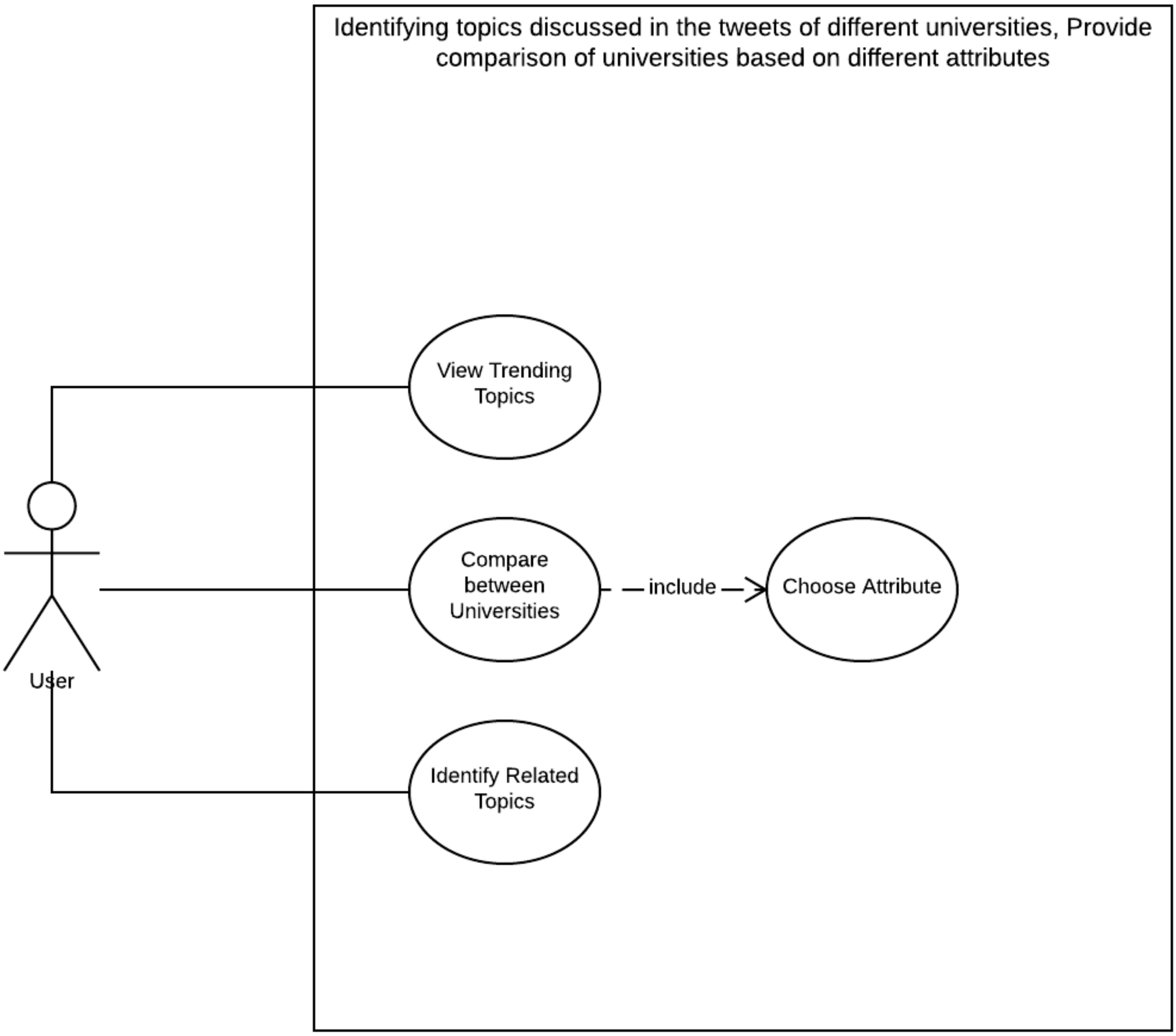

Identifying topics discussed in the tweets of different universities, Provide comparison of universities based on different attributes
View Trending Topics
Compare between Universities
include
Choose Attribute
Identify Related Topics
User

Figure 34: Use Case 3

This use case shows how to view universities topics whether it's a trending topic or a comparison between universities activities on Twitter or finding out the related topics between universities tweets if they all talk about a certain topic.

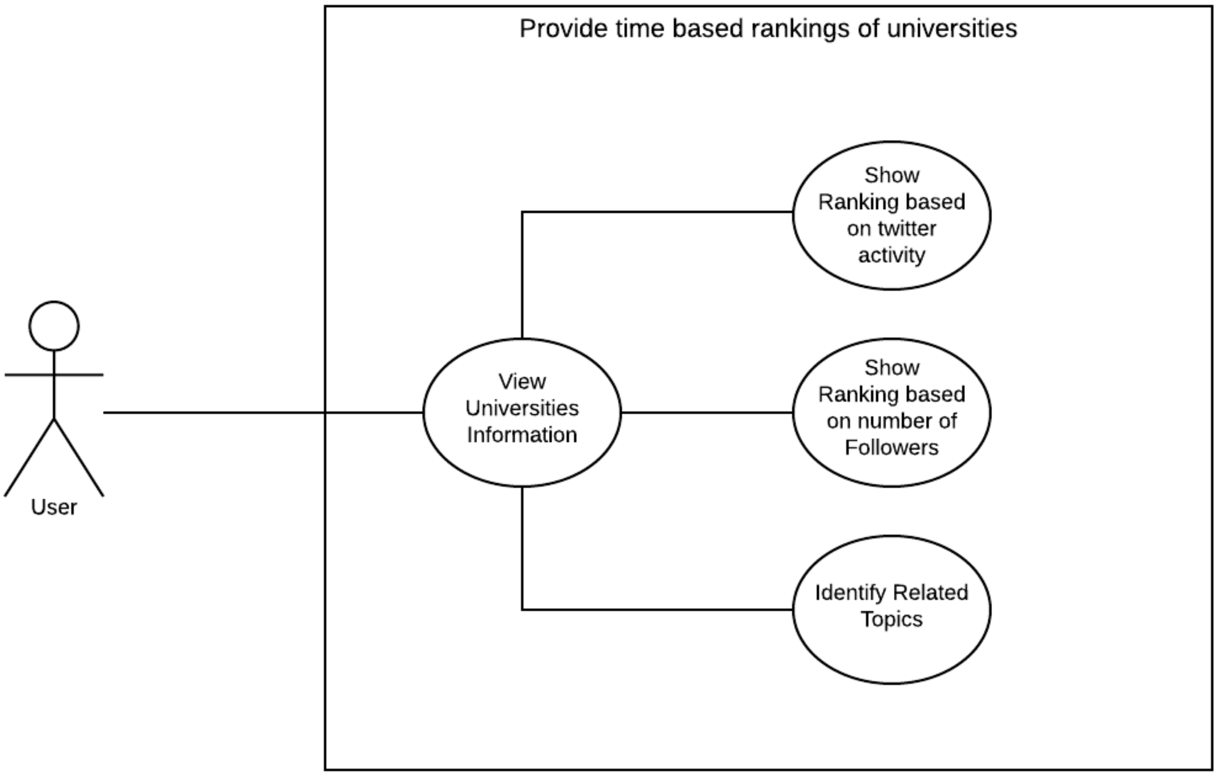

Provide time based rankings of universities
User
View Universities Information
Show Ranking based on twitter activity
Show Ranking based on number of Followers
Identify Related Topics

Figure 35: Use Case 4

This use case illustrates the methodology to the universities ranking based on their activities on Twitter such as the number of followers, number of retweets, and respond to students questions...etc.

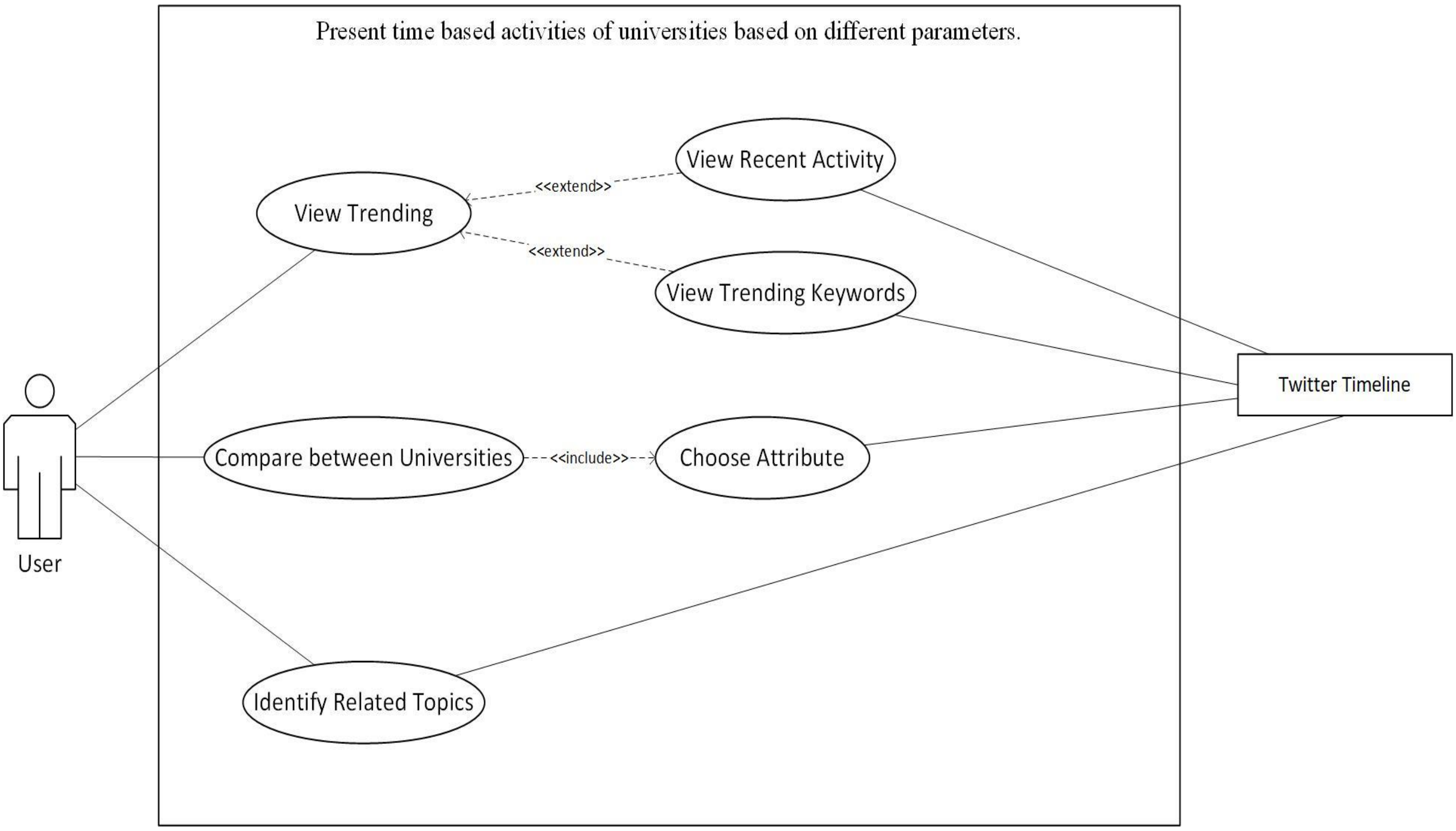


Figure 36: Use Case 5

This use case shows universities present time tweets and recent activities.

المملكة العربية السعودية
وزارة التعليم
جامعة الملك عبد العزيز
كلية الحاسبات وتقنية
المعلومات
قسم نظم المعلومات

Full Use Case Diagram

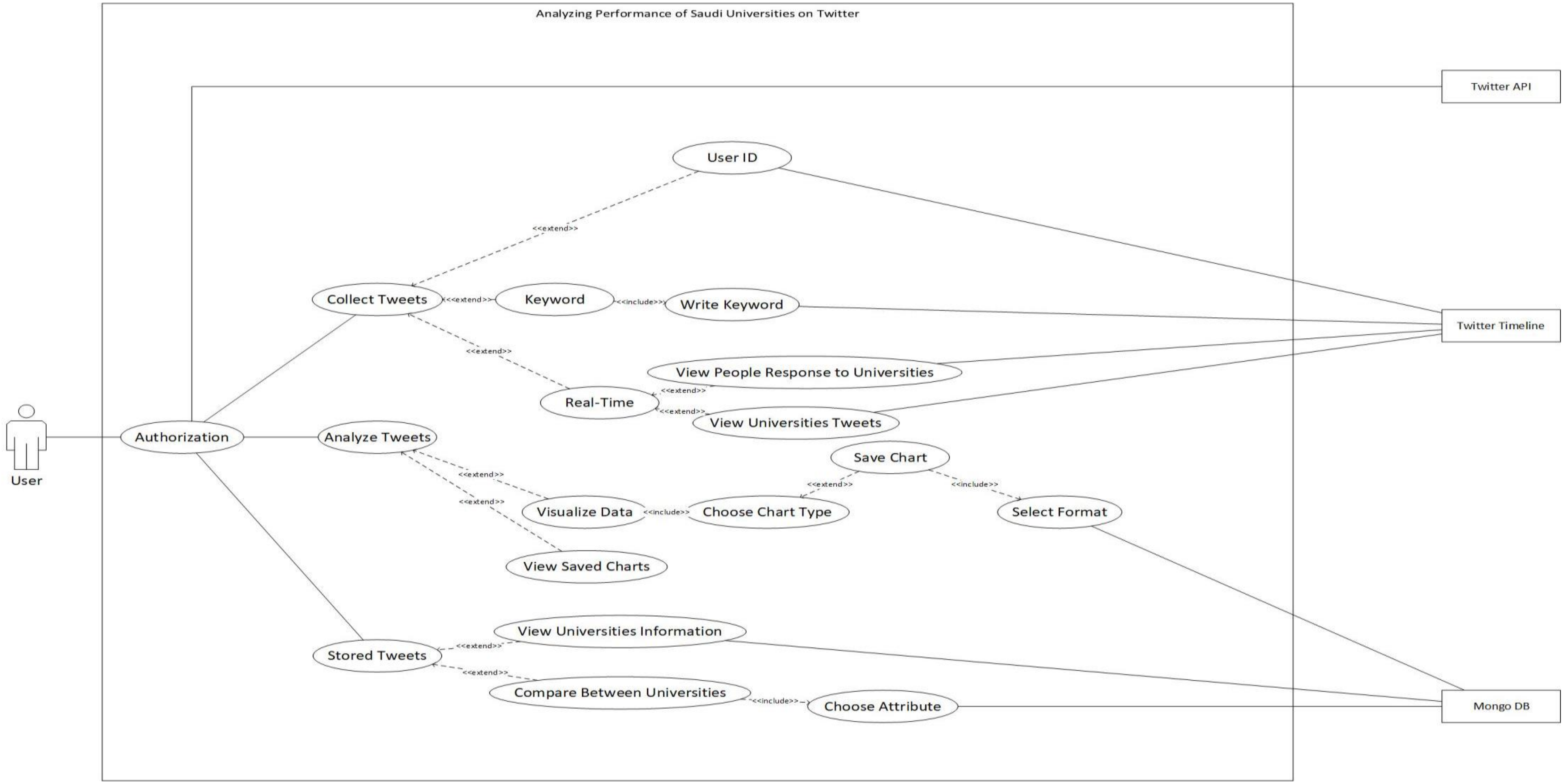


Figure 37:Full Use Case Diagram

4.4 Sequence Diagrams

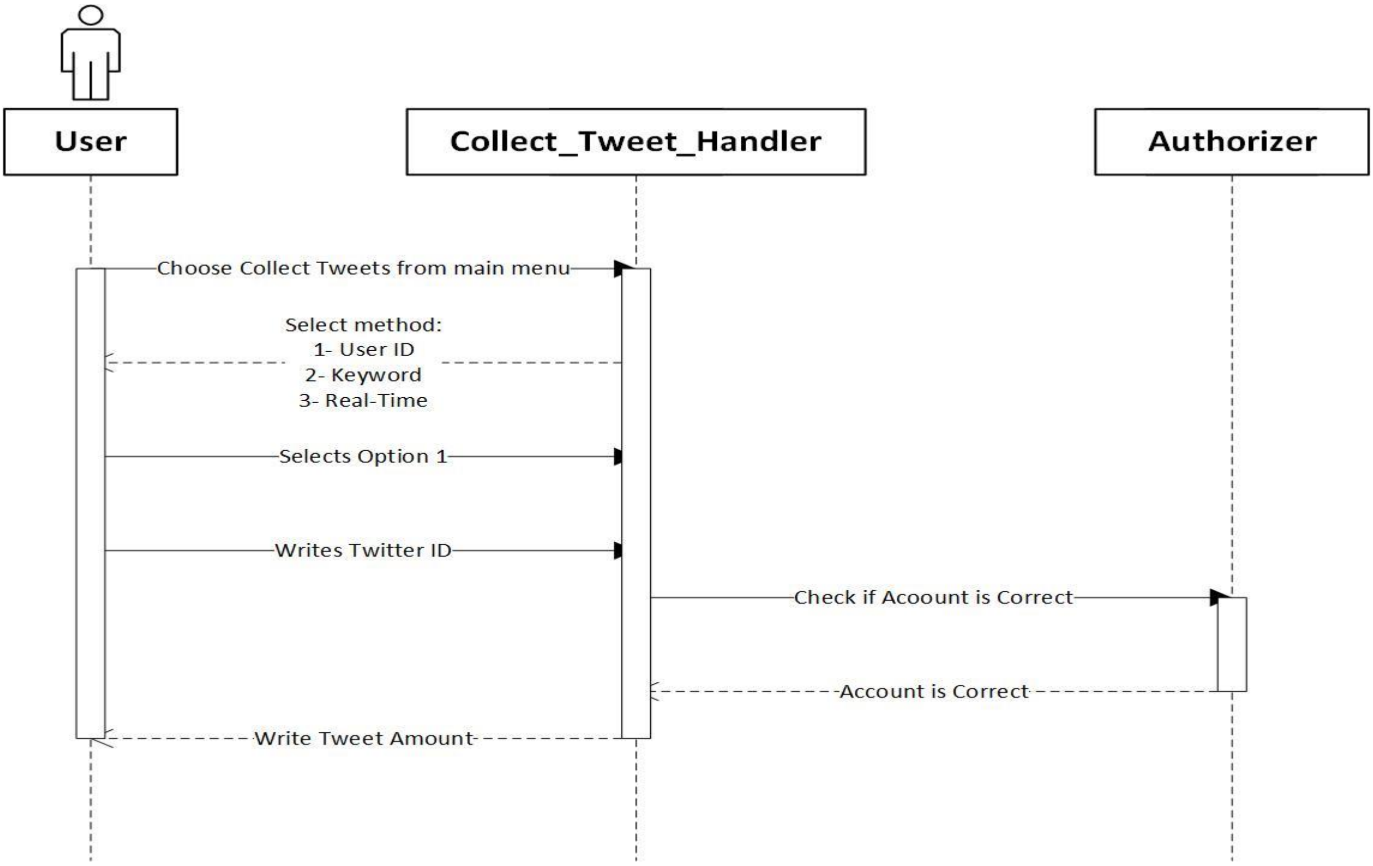


Figure 38: Sequence Diagram 1

This sequence diagram shows collecting tweets process. User will choose collect tweets option from the main menu and the collect tweets page will be provided by Tweet_Collector_Interface object, in addition, this object will return to the user a window about tweets collection types for the user to choose from them. After selecting option 1, user will write the university Twitter ID and the Query_Verifier object will make sure that the university account is correct and will return a window to the user to enter total tweets that will be collected.

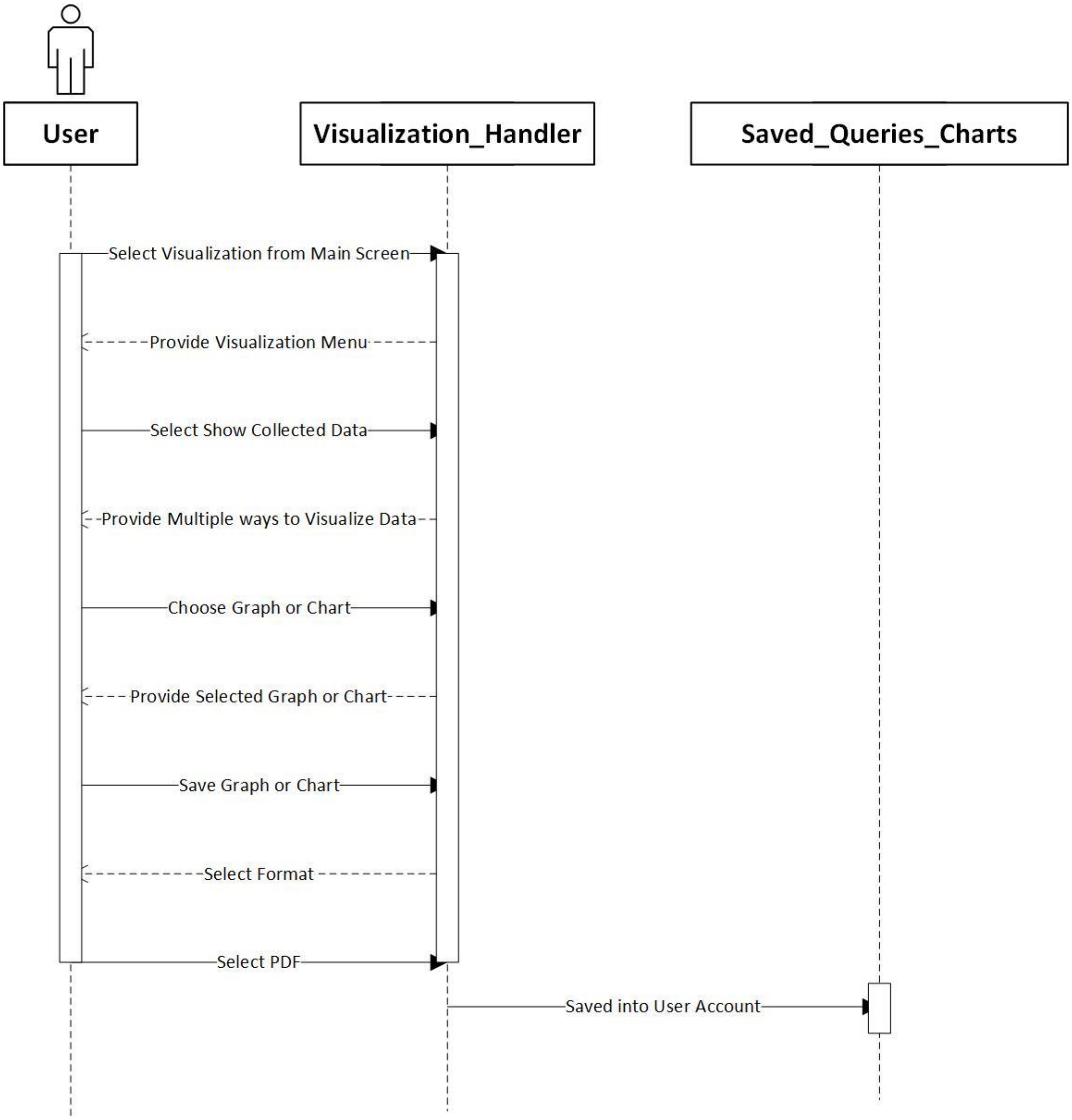

User
Visualization_Handler
Saved_Queries_Charts
Select Visualization from Main Screen
Provide Visualization Menu
Select Show Collected Data
Provide Multiple ways to Visualize Data
Choose Graph or Chart
Provide Selected Graph or Chart
Save Graph or Chart
Select Format
Select PDF
Saved into User Account

Figure 39: Sequence Diagram 2

This sequence diagram is related to tweets visualization. A user will select "Visualization" from the main menu. Then, Visualization_Handler object will return visualization options to the user. The user will determine a specific collected tweets to be visualized. Visualization_Handler object will return graphs/charts types (Histogram, Bar chart, Scatter Plot...etc.) to be chosen by the user. After a user has selected the graph type he needs to determine the attributes of the relationship to which the analysis will be conducted (E.g. Number of followers & Number of Retweets). The user can save and select specific format. After that, User_Saved_Queries_Charts object will save everything in the database.

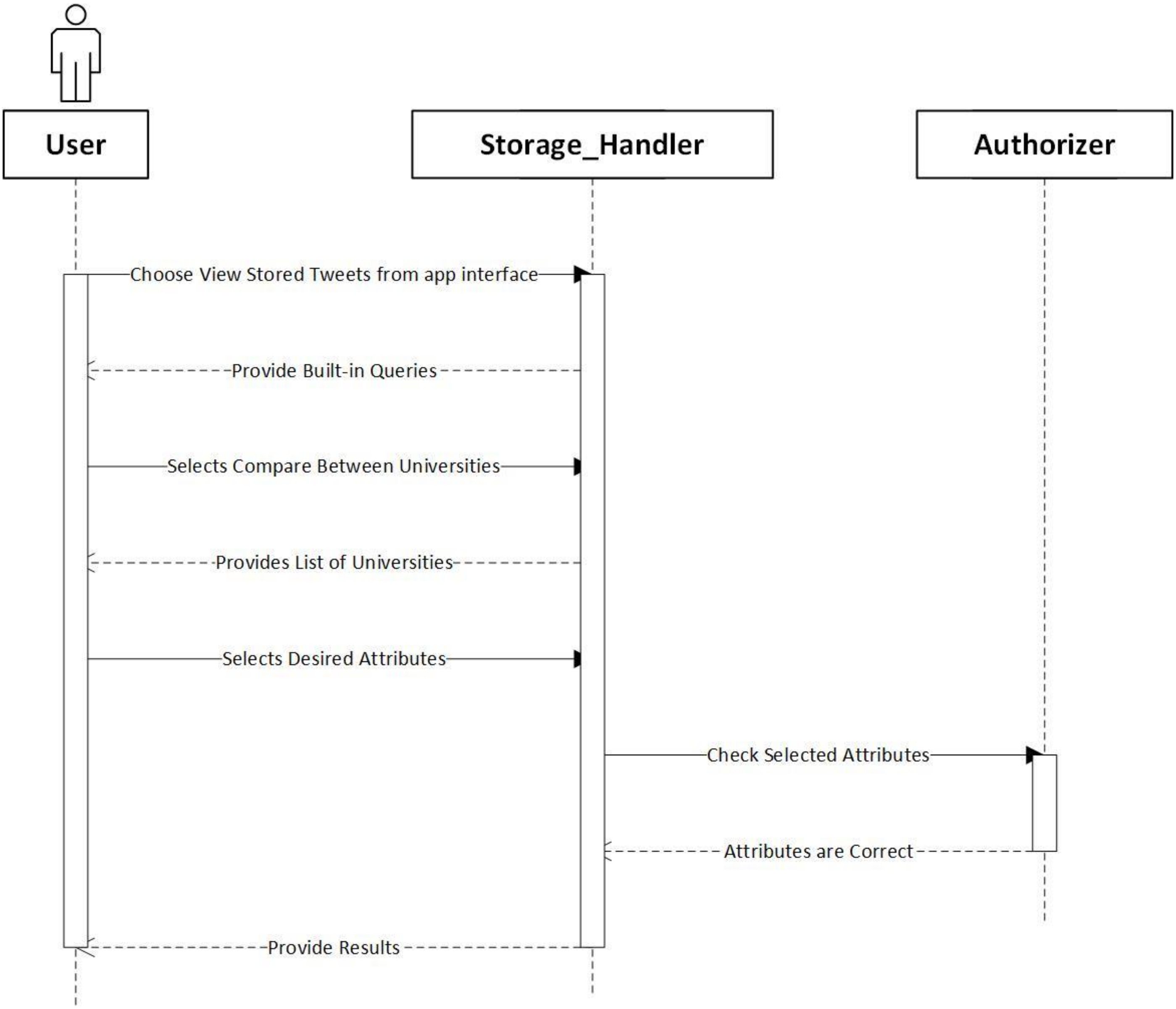

User
Storage_Handler
Authorizer
Choose View Stored Tweets from app interface
Provide Built-in Queries
Selects Compare Between Universities
Provides List of Universities
Selects Desired Attributes
Check Selected Attributes
Attributes are Correct
Provide Results

Figure 40: Sequence Diagram 3

This sequence diagram talks about comparisons between universities activities on Twitter based on specific parameters. A user will choose from the main interface options "View Stored Tweets". Then, Stored_Tweets_Handler object will return Built-in queries to the user. At this stage, a user will choose "compare between universities option". Stored_Tweets_Handler object will return a list of all 37 Saudi universities (both public and private) to the user to choose from them. After that, a user will determine the desired attributes for the comparisons (E.g. The most discussed topic, Number of followers...etc.).Query_Verifier object will check attributes correctness and provide the results to the user.

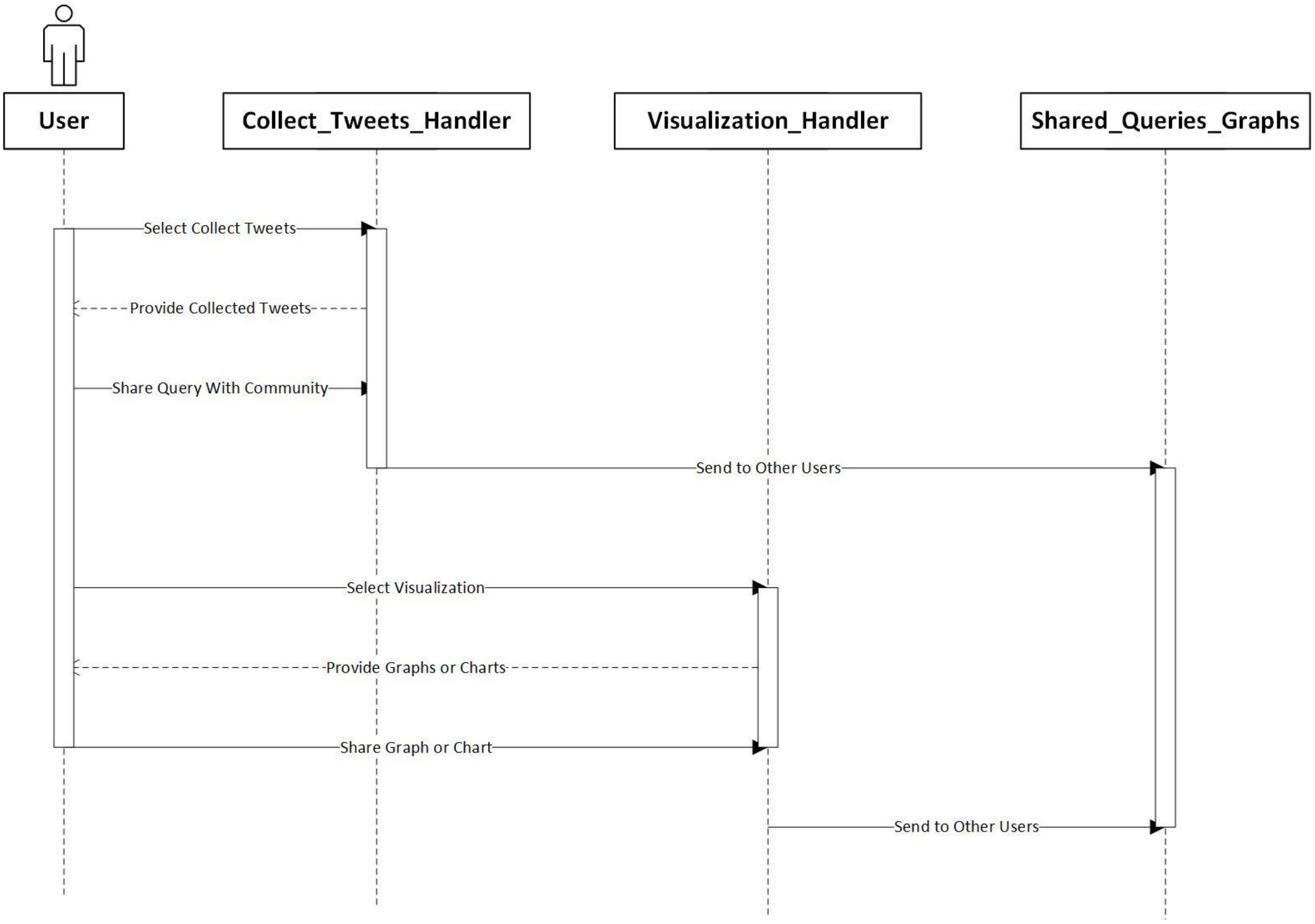


Figure 41: Sequence Diagram 4

This sequence diagram shows the approach of sharing queries and graphs with other users. For query sharing, the user will choose "Collect Tweets" from main app interface. Then, the page it gets provided by Collect_Tweets_Handler object to the user. A user will choose "Share Query with Community" option and the tweets will be sent to other users by Shared_Queries_Graphs object. Likewise, with visualization sharing but in the beginning the user will choose "Visualization" from main app interface.

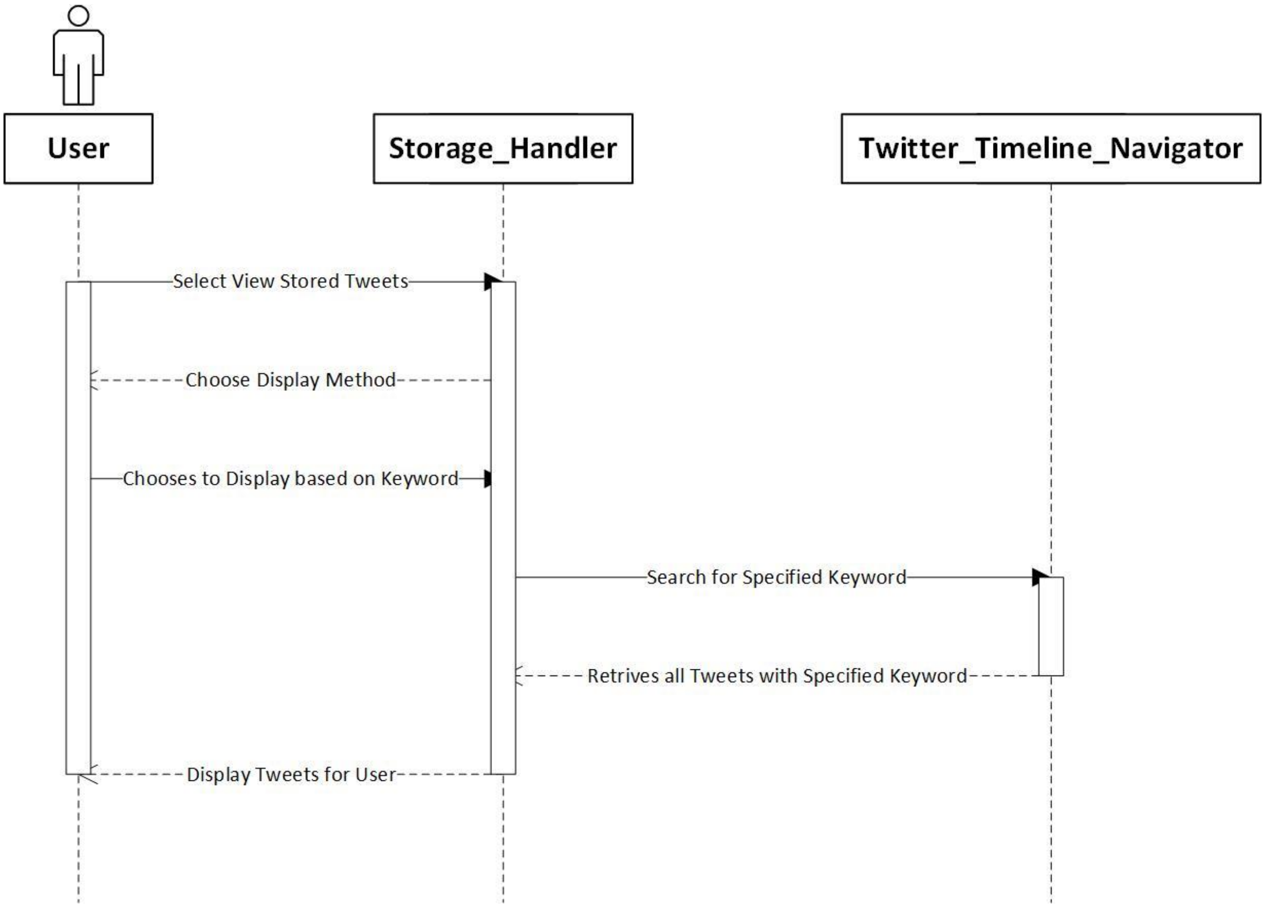

User
Storage_Handler
Twitter_Timeline_Navigator
Select View Stored Tweets
Choose Display Method
Chooses to Display based on Keyword
Search for Specified Keyword
Retrives all Tweets with Specified Keyword
Display Tweets for User

Figure 42: Sequence Diagram 5

This sequence diagram shows visualization about what people are talking about universities in real time. A user will choose "View Stored Tweets" from main app interface. Storage_Handler object will return the display method. A user will choose to display based on specific keyword. Twitter_Timeline_Navigator object will be searching for a specified keyword. Moreover, it will retrieve all tweets with specified keyword and display it to the user.

4.5 Class Diagram

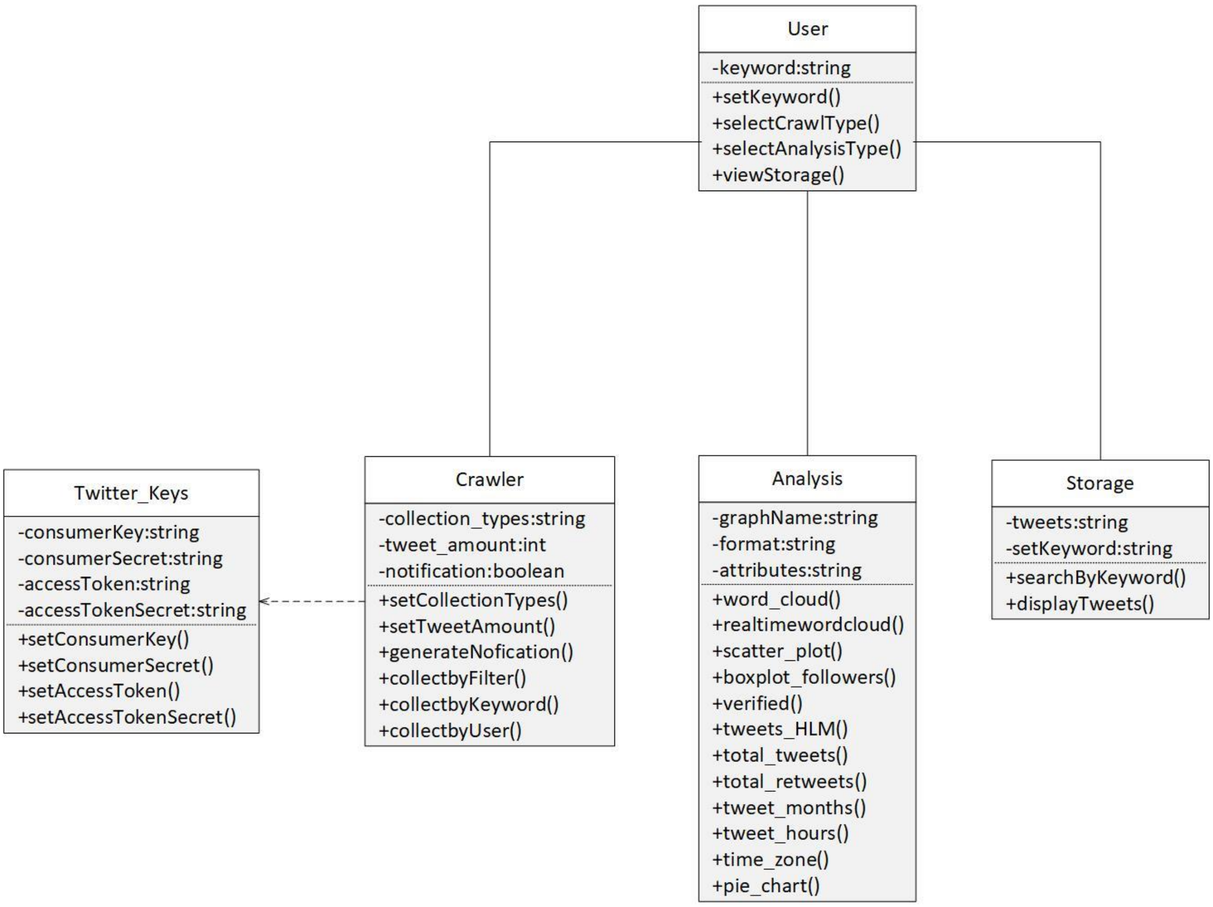


Figure 43: Class Diagram

This figure highlights the main classes in our application and how they interact with each other. Classes have parameters which is the activities conducted in the class and therefore turning them into Comprehensive functions.

5. Chapter 5 | Implementation

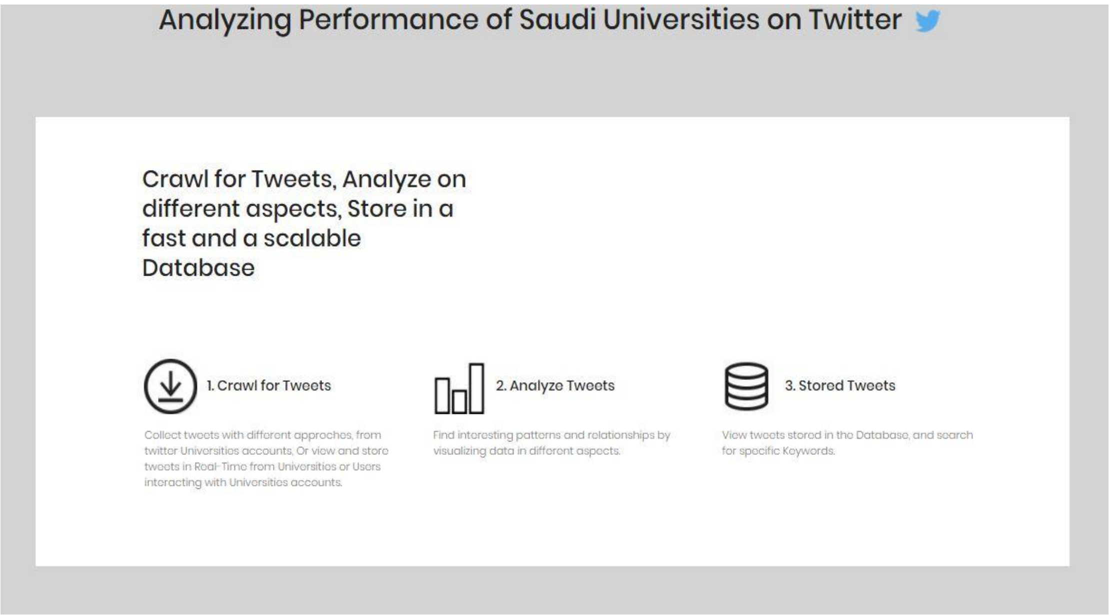
5.1 Tools

- Spyder Editor for python language programming.
- Mongo DB to store tweets.
- Robo 3T to access the database.
- Pymongo to connect to Mongo DB.
- Tweepy to open authorization gates to collect tweets.
- Bokeh for Charts.
- Wordcloud library to represent the most frequent words.
- Arabic_reshaper to read Arabic words in wordcloud.
- Pandas library to generate DataFrames and use it with Bokeh to visualize the data.
- Matplot and Seaborn for extra types of charts.

5.2 Interface Description

Three main options in the interface:

- Crawl for Tweets.
- Analyze Tweets.
- Storage.



Analyzing Performance of Saudi Universities on Twitter

Crawl for Tweets, Analyze on different aspects, Store in a fast and a scalable Database

1. Crawl for Tweets

Collect tweets with different approches, from twitter Universities accounts, Or view and store tweets in Real-Time from Universities or Users interacting with Universities accounts.

2. Analyze Tweets

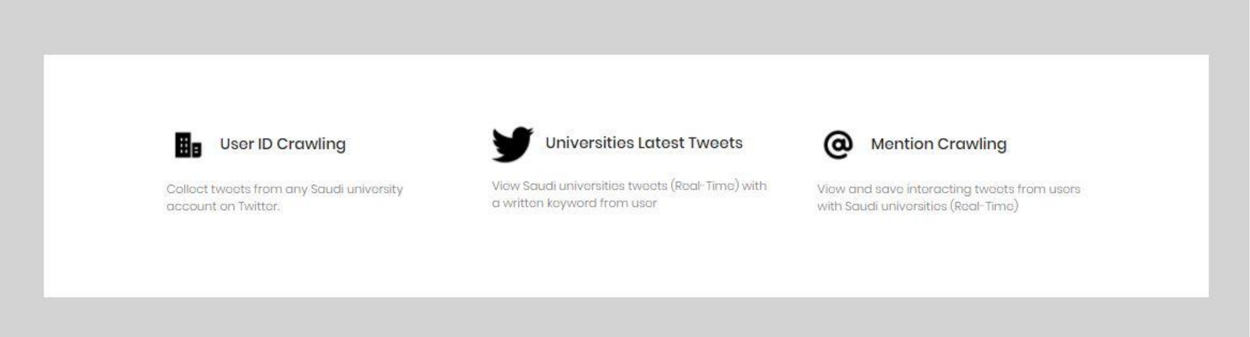
Find interesting patterns and relationships by visualizing data in different aspects.

3. Stored Tweets

View tweets stored in the Database, and search for specific Keywords.

Figure 44: Main_Menu

5.3 Walkthrough the System



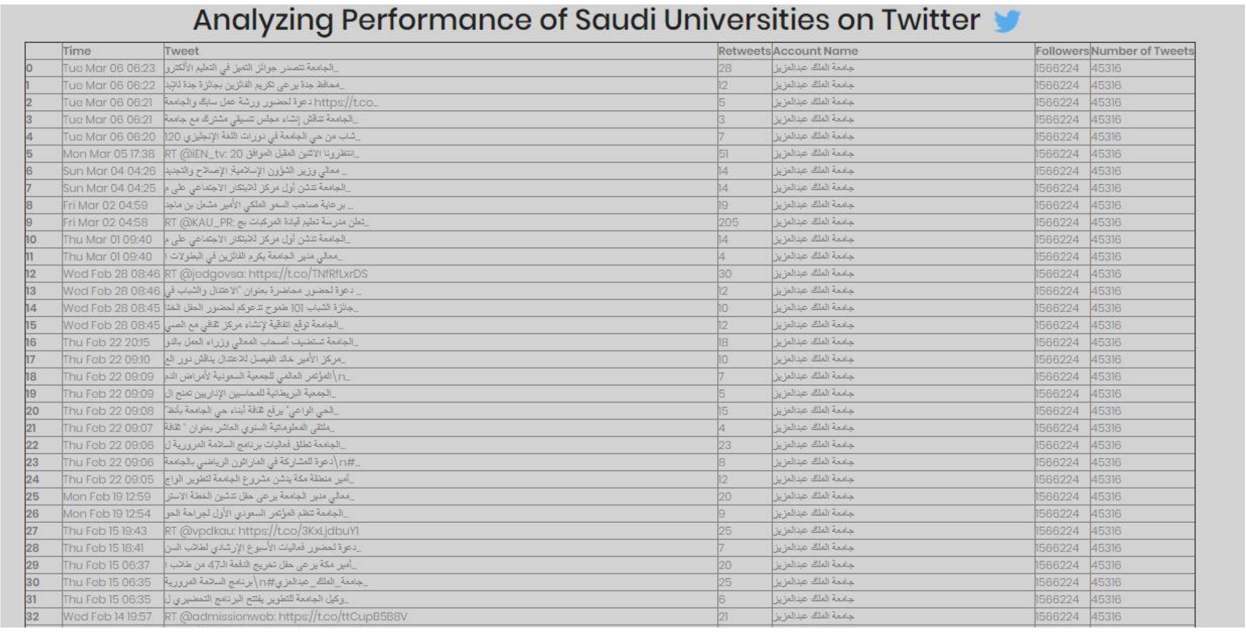
User ID Crawling
Collect tweets from any Saudi university account on Twitter.

Universities Latest Tweets
View Saudi universities tweets (Real-Time) with a written keyword from user

Mention Crawling
View and save interacting tweets from users with Saudi universities (Real-Time)

Figure 45: Crawling Options

First, user should crawl some tweets from the "Crawl for Tweets" option on the main menu, either by User ID or starting a real-time crawl in the backend.



Analyzing Performance of Saudi Universities on Twitter

	Time	Tweet	Retweets	Account Name	Followers	Number of Tweets
0	Tue Mar 06 06:23	_الجامعة تتصدر جوائز التميز في التعليم الألكترو	28	جامعة الملك عبدالعزيز	1566224	45316
1	Tue Mar 06 06:22	_محافظ جدة يرعى تكريم الفائزين بجائزة جدة للإبد	12	جامعة الملك عبدالعزيز	1566224	45316
2	Tue Mar 06 06:21	_https://t.co دعوة لحضور ورشة عمل سابك والجامعة	5	جامعة الملك عبدالعزيز	1566224	45316
3	Tue Mar 06 06:21	_الجامعة تناقش إنشاء مجلس تنسيقي مشترك مع جامعة	3	جامعة الملك عبدالعزيز	1566224	45316
4	Tue Mar 06 06:20	_شاب من حي الجامعة في دورات اللغة الإنجليزي 120	7	جامعة الملك عبدالعزيز	1566224	45316
5	Mon Mar 05 17:38	RT @iEN_tv: 20 _انتظرونا الاثنين المقبل الموافق	51	جامعة الملك عبدالعزيز	1566224	45316
6	Sun Mar 04 04:26	_ معالي وزير الشؤون الإسلامية الإصلاح والتجديد	14	جامعة الملك عبدالعزيز	1566224	45316
7	Sun Mar 04 04:25	_الجامعة تدشن أول مركز للابتكار الاجتماعي على م	14	جامعة الملك عبدالعزيز	1566224	45316
8	Fri Mar 02 04:59	_ برعاية صاحب السمو الملكي الأمير مشعل بن ماجد	19	جامعة الملك عبدالعزيز	1566224	45316
9	Fri Mar 02 04:58	RT @KAU_PR: _تعلن مدرسة تعليم قيادة المركبات بج	205	جامعة الملك عبدالعزيز	1566224	45316
10	Thu Mar 01 09:40	_الجامعة تدشن أول مركز للابتكار الاجتماعي على م	14	جامعة الملك عبدالعزيز	1566224	45316
11	Thu Mar 01 09:40	_معالي مدير الجامعة يكرم الفائزين في البطولات ا	4	جامعة الملك عبدالعزيز	1566224	45316
12	Wed Feb 28 08:46	RT @jedgovsa: https://t.co/TNfRfLxrDS	30	جامعة الملك عبدالعزيز	1566224	45316
13	Wed Feb 28 08:46	_ دعوة لحضور محاضرة بعنوان "الاعتدال والشباب في	12	جامعة الملك عبدالعزيز	1566224	45316
14	Wed Feb 28 08:45	_جائزة الشباب 101 طموح تدعوكم لحضور الحفل الختا	10	جامعة الملك عبدالعزيز	1566224	45316
15	Wed Feb 28 08:45	_الجامعة توقع اتفاقية لإنشاء مركز ثقافي مع الصي	12	جامعة الملك عبدالعزيز	1566224	45316
16	Thu Feb 22 20:15	_الجامعة تستضيف أصحاب المعالي وزراء العمل بالدو	18	جامعة الملك عبدالعزيز	1566224	45316
17	Thu Feb 22 09:10	_مركز الأمير خالد الفيصل للاعتدال يناقش دور الع	10	جامعة الملك عبدالعزيز	1566224	45316
18	Thu Feb 22 09:09	_n\المؤتمر العالمي للجمعية السعودية لأمراض الدم	7	جامعة الملك عبدالعزيز	1566224	45316
19	Thu Feb 22 09:09	_الجمعية البريطانية للمحاسبين الإداريين تمنح ال	5	جامعة الملك عبدالعزيز	1566224	45316
20	Thu Feb 22 09:08	_الحي الواعي" يرفع ثقافة أبناء حي الجامعة بأنظ"	15	جامعة الملك عبدالعزيز	1566224	45316
21	Thu Feb 22 09:07	_ملتقى المعلوماتية السنوي العاشر بعنوان " ثقافة	4	جامعة الملك عبدالعزيز	1566224	45316
22	Thu Feb 22 09:06	_الجامعة تطلق فعاليات برنامج السلامة المرورية ل	23	جامعة الملك عبدالعزيز	1566224	45316
23	Thu Feb 22 09:06	_#n\دعوة للمشاركة في الماراثون الرياضي بالجامعة	8	جامعة الملك عبدالعزيز	1566224	45316
24	Thu Feb 22 09:05	_أمير منطقة مكة يدشن مشروع الجامعة لتطوير الواج	12	جامعة الملك عبدالعزيز	1566224	45316
25	Mon Feb 19 12:59	_معالي مدير الجامعة يرعى حفل تدشين الخطة الاستر	20	جامعة الملك عبدالعزيز	1566224	45316
26	Mon Feb 19 12:54	_الجامعة تنظم المؤتمر السعودي الأول لجراحة الحو	9	جامعة الملك عبدالعزيز	1566224	45316
27	Thu Feb 15 19:43	RT @vpdkau: https://t.co/3KxLjdbuYl	25	جامعة الملك عبدالعزيز	1566224	45316
28	Thu Feb 15 18:41	_دعوة لحضور فعاليات الأسبوع الإرشادي لطلاب السن	7	جامعة الملك عبدالعزيز	1566224	45316
29	Thu Feb 15 06:37	_أمير مكة يرعى حفل تخريج الدفعة الـ47 من طلاب ا	20	جامعة الملك عبدالعزيز	1566224	45316
30	Thu Feb 15 06:35	_جامعة_الملك_عبدالعزي#n\برنامج السلامة المرورية	25	جامعة الملك عبدالعزيز	1566224	45316
31	Thu Feb 15 06:35	_وكيل الجامعة للتطوير يفتتح البرنامج التحضيري ل	6	جامعة الملك عبدالعزيز	1566224	45316
32	Wed Feb 14 19:57	RT @admissionweb: https://t.co/ttCupB5B8V	21	جامعة الملك عبدالعزيز	1566224	45316

Figure 46: Stored Tweets

Second, user can view the stored tweets while it being added from the "Storage" option in the main menu.

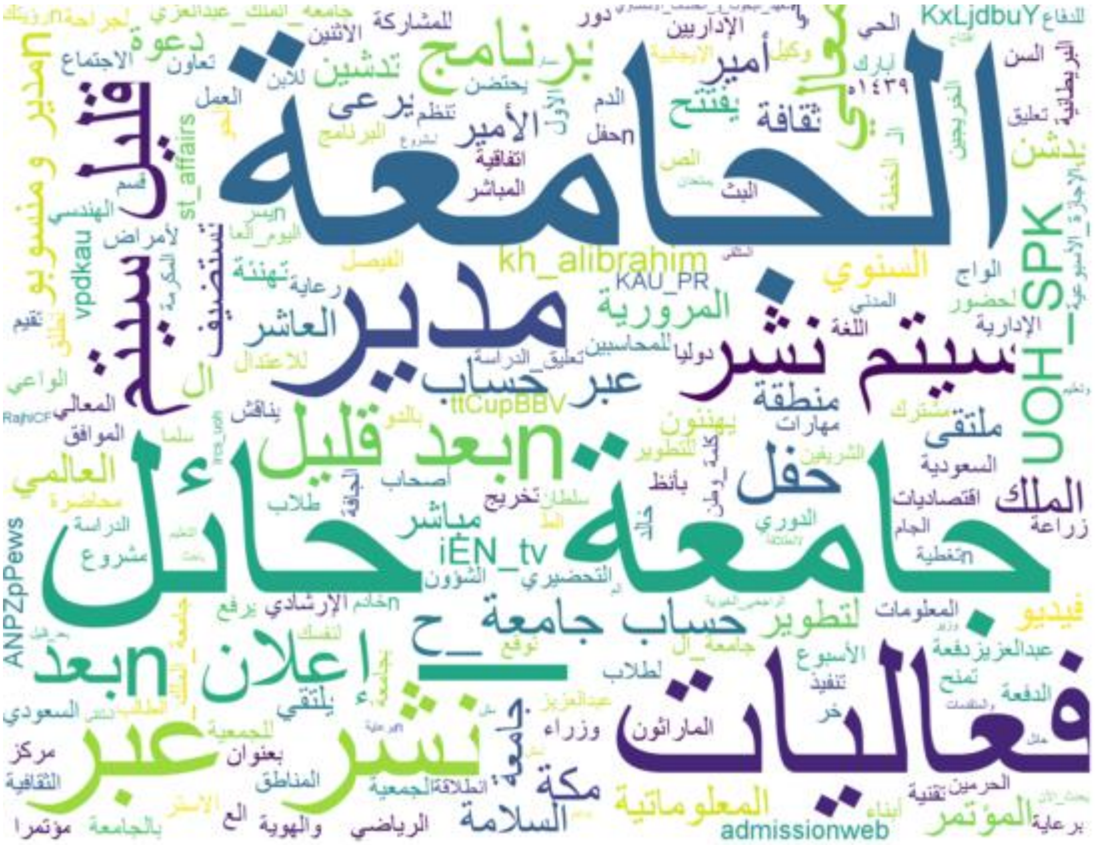


Figure 47: WordCloud

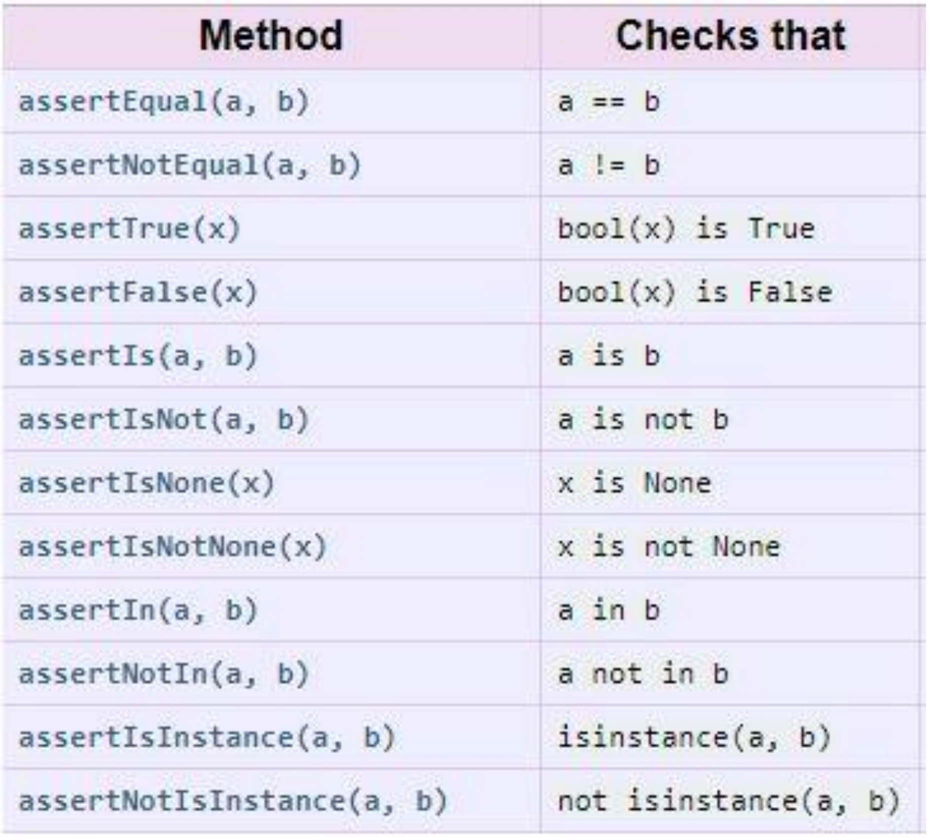
Third, user can now analyze the data collected in various charts and extract interesting relationships, this is an example of many charts that the user can generate from tweets that have been collected.

6. Chapter 6 | Testing

6.1 Methods Used in Usability Testing

Testing methods were conducted through unittest built-in package, which is a framework designed to test any type of functions in the program.

Unit test methodology starts with creating a given scenario to test a particular function and ensure if it is working fine, unittest package includes many functions that can be used based upon the program.



Method	Checks that
assertEqual(a, b)	a == b
assertNotEqual(a, b)	a != b
assertTrue(x)	bool(x) is True
assertFalse(x)	bool(x) is False
assertIs(a, b)	a is b
assertIsNot(a, b)	a is not b
assertIsNone(x)	x is None
assertIsNotNone(x)	x is not None
assertIn(a, b)	a in b
assertNotIn(a, b)	a not in b
assertIsInstance(a, b)	isinstance(a, b)
assertNotIsInstance(a, b)	not isinstance(a, b)

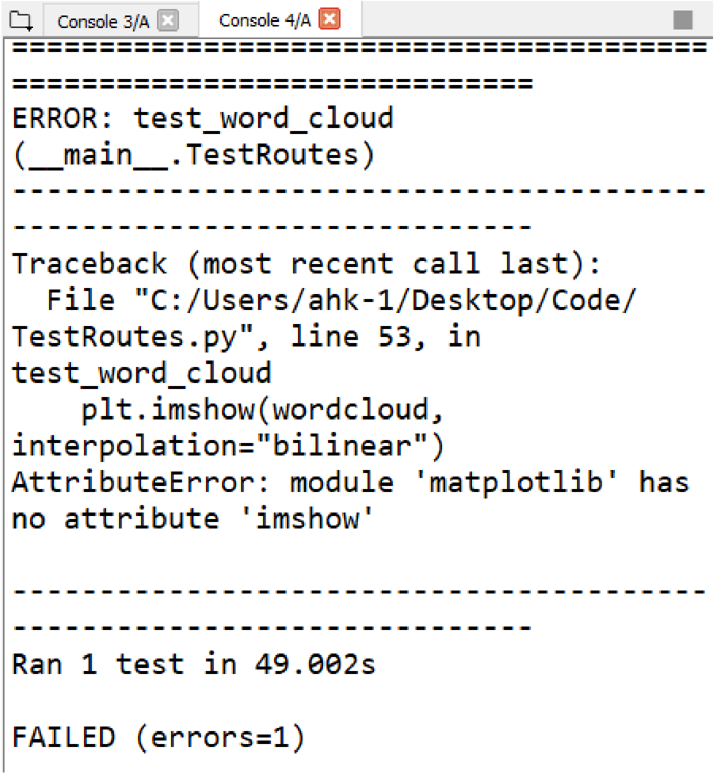
Figure 48: unittest Package Functions (29)

Since most of our program is about analysis that gives the user an insight of Saudi Universities data on twitter through charts/graphs, the tests have been conducted on the DataFrames which is the basic component on which these charts are based on, one function is used to test DataFrames, assert_frame_equal, and it checks whether the DataFrame value given is equal to the DataFrame which the function provides.

6.2 Test Case

Here is a Test of the WordCloud function that checks whether the tweets contains words from the stop words list or not and also checks the ability of the function to draw the desired chart.

An error occurred after unit test have been implemented.



```
========================================
==============================
ERROR: test_word_cloud
(__main__.TestRoutes)
----------------------------------------
------------------------------
Traceback (most recent call last):
  File "C:/Users/ahk-1/Desktop/Code/
TestRoutes.py", line 53, in
test_word_cloud
    plt.imshow(wordcloud,
interpolation="bilinear")
AttributeError: module 'matplotlib' has
no attribute 'imshow'

----------------------------------------
------------------------------
Ran 1 test in 49.002s

FAILED (errors=1)
```

Figure 49: Test Status 1

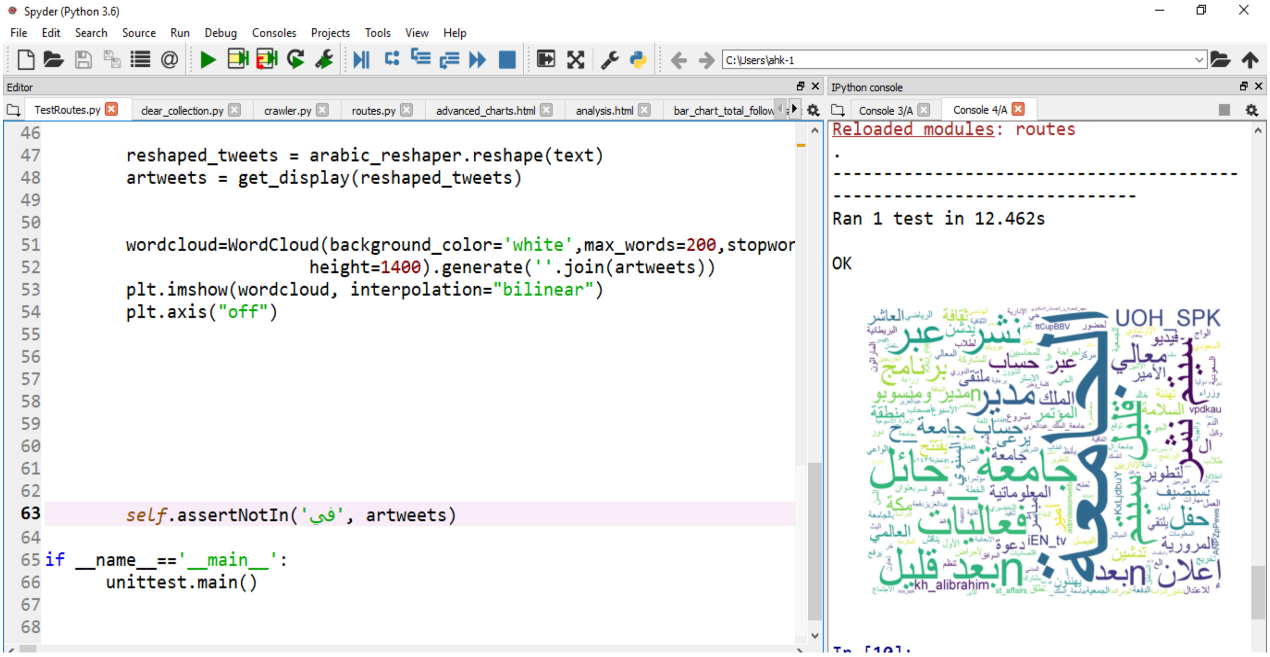
6.2.1 Tweepy Errors

One of the major errors that were encountered:

- 406 Error: Returned whenever there was an invalid format or an unclear parameter is given to the API request.
- 429 Error: Returned when the limit of requests is exceeded.
- 403 Error: Returns User is Not Found error.
- 404 Error: Returns Tweet Not Found error.

6.3 Results

- From the given test case above, errors have been handled properly using these solutions:



```
        reshaped_tweets = arabic_reshaper.reshape(text)
        artweets = get_display(reshaped_tweets)


        wordcloud=WordCloud(background_color='white',max_words=200,stopwor
                            height=1400).generate(''.join(artweets))
        plt.imshow(wordcloud, interpolation="bilinear")
        plt.axis("off")









        self.assertNotIn('في', artweets)

if __name__=='__main__':
    unittest.main()
```

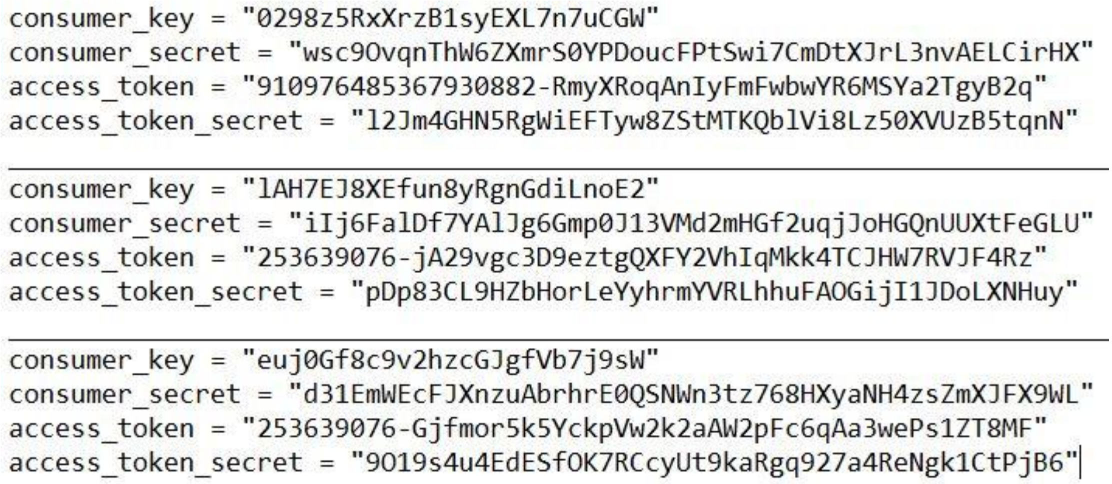
```
Reloaded modules: routes
.
----------------------------------------------------------------------
Ran 1 test in 12.462s

OK
```

Figure 50: Wordcloud Status 2

Error in Status1 Figure was in importing some of the libraries needed to generate the WordCloud and has been fixed.

- Handling Tweepy errors was solved by creating three different keys for each crawler, so error 429 is solved.



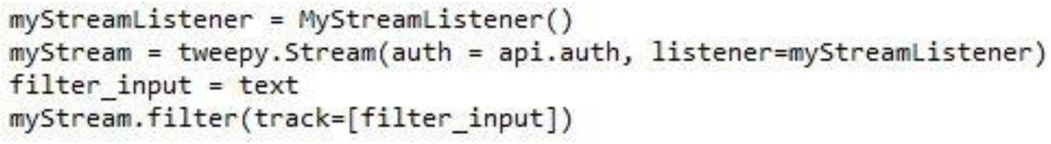
```
consumer_key = "0298z5RxXrzB1syEXL7n7uCGW"
consumer_secret = "wsc9OvqnThW6ZXmrS0YPDoucFPtSwi7CmDtXJrL3nvAELCirHX"
access_token = "910976485367930882-RmyXRoqAnIyFmFwbwYR6MSYa2TgyB2q"
access_token_secret = "l2Jm4GHN5RgWiEFTyw8ZStMTKQblVi8Lz50XVUzB5tqnN"
```

```
consumer_key = "lAH7EJ8XEfun8yRgnGdiLnoE2"
consumer_secret = "iIj6FalDf7YAlJg6Gmp0J13VMd2mHGf2uqjJoHGQnUUXtFeGLU"
access_token = "253639076-jA29vgc3D9eztgQXFY2VhIqMkk4TCJHW7RVJF4Rz"
access_token_secret = "pDp83CL9HZbHorLeYyhrmYVRLhhuFAOGijI1JDoLXNHuy"
```

```
consumer_key = "euj0Gf8c9v2hzcGJgfVb7j9sW"
consumer_secret = "d31EmWEcFJXnzuAbrhrE0QSNWn3tz768HXyaNH4zsZmXJFX9WL"
access_token = "253639076-Gjfmor5k5YckpVw2k2aAW2pFc6qAa3wePs1ZT8MF"
access_token_secret = "9O19s4u4EdESfOK7RCcyUt9kaRgq927a4ReNgk1CtPjB6"
```

Figure 52: Twitter Keys

- Error 406 was given because one of the parameters of Filter API was null, so after the user inputs a keyword for real time crawls, this error was handled.



```
myStreamListener = MyStreamListener()
myStream = tweepy.Stream(auth = api.auth, listener=myStreamListener)
filter_input = text
myStream.filter(track=[filter_input])
```

Figure 51: Filter API User Input

7. Chapter 7 | Conclusion

It was concluded from our project that the users of Twitter and especially users who are interested in educational purposes with their universities on Twitter need an algorithm to save their time and to make their reading more objective.

7.1 Problems and Difficulties

One of the difficulties during our project was in literature review regarding reading and fully understanding about a collection of studies throughout books, research, and articles related to the subject of our project and then summarize its ideas.

Moreover, another problem we faced in analysis phase, concerning visualize the data with different parameters and relations in diagrams, but it must appear from the file correctly so we used RapidMiner software for the visualization part.

Also, it was hard to deal with performing backend processing because it is time consuming for users to wait for the page to load infinitely.

Another difficulty was linking python code with the interface, it required an additional lines of code to integrate a python code and make it functional in a website**.**

Another issue was removing the Arabic stop words from the wordcloud.

In general social media analytics is brand new field these days, it has limited resources in the internet, and all of the work that have been done by personal effort.

7.2 Findings

The analysis conducted on 10 Saudi universities and the results were the following:

- King Abdulaziz University is the most follow-up on Twitter with nearly 1,400,000 followers.
- KFUniversity have the most number of retweets from our sample with 1,600,000 retweets from more than 3000 tweets.
- Frequent tweeting times of universities are early hours of the day, while late hours tweeting and social activity decreases.
- Tweeting times based on months to know which months were Saudi Universities were the most active.
- Knowing the most active University from the collected data.
- Generating a Wordcloud to represent the most frequent words in tweets of Saudi Universities**.**

7.3 Future Work

The future work of our project will focus on developing a desktop application that analyzes unstructured data on Twitter regarding Saudi universities, and designing methodologies.

- Adding unique records or set of data for different users.
- Deploying the algorithm to a different categories.
- Make it as an IOS application.
- Expand the tweet capacity to make it considered as Big Data Analysis.

Taken into consideration the issues that we have discussed in this report.

References

1. **[Online] https://en.wikipedia.org/wiki/Pandas_(software).**

2. [Online] https://en.wikipedia.org/wiki/NumPy.

3. ***A framework for real-time Twitter data analysis.*** **Gaglio, Salvatore, Giuseppe Lo Re, and Marco Morana. 2016, Vol. 73.**

4. ***Taghreed: a system for querying, analyzing, and visualizing geotagged microblogs.*** **Magdy, Amr, et al. In Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2014, November. 163-172.**

5. ***Web-based visual analytics for social media.*** **Best, Daniel M., et al. 2012 May 20. (pp. 2-5).**

6. ***Visual analytics: Definition, process, and challenges.*** **Keim, Daniel, et al. Lecture notes in computer science, 2008, Vol. 4950. 154-176..**

7. ***A comparative analysis on weibo and twitter.*** **Han, Wentao, et al. Tsinghua Science and Technology, 2016 , Vol. 21.1. 1-16.**

8. ***A Data Analytic Framework for Unstructured Text.*** **Al-Barhamtoshy, Hassanin M., and Fathy E. Eassa. Life Science Journal , 2014, Vol. 11.10. 339-350.**

9. ***Near real time assessment of social media using geo-temporal network analytics.*** **Carley, Kathleen M., et al. Advances in Social Networks Analysis and Mining (ASONAM), 2013. pp. 517-524.**

10. ***Twitter data analytics.*** **Kumar, Shamanth, Fred Morstatter, and Huan Liu. 2014. pp. 1041-4347.**

11. ***Framework for social media big data quality analysis.*** **Jaafar, Nouf, Manal Al-Jadaan, and Reem Alnutaifi. New Trends in Database and Information Systems, 2015. 301-314.**

12. ***Felicittà: Visualizing and Estimating Happiness in Italian Cities from Geotagged Tweets.*** **Allisio, Leonardo, et al. 2013, Vol. 1096. 95-106.**

13. ***Open geospatial analytics with PySAL.*** **Rey, Sergio J., et al. ISPRS International Journal of Geo-Information, 2015, Vol. 4.2. 815-836.**

14. ***Divergent discourse between protests and counter-protests.*** **Gallagher, Ryan J., et al. # BlackLivesMatter and# AllLivesMatter." arXiv preprint arXiv, 2016. 1606.06820.**

15. ***Geo-Twitter Analytics: Applications in Crisis Management.*** **MacEachren, Alan M., et al. 25th International Cartographic Conference, 2011. pp. 3-8.**

16. ***Twitter Analytics: Architecture, Tools and Analysis.*** **Perera, Rohan DW, et al. In MILITARY COMMUNICATIONS CONFERENCE, 2010-MILCOM, 2010. pp. 2186-2191.**

17. ***Novel applications of social media analytics.*** **Fan, Weiguo, and Xiangbin Yan. Information and Management, 2015, Vol. 52.7 . 761-763..**

18. ***COSMOS: Towards an integrated and scalable service for analyzing social media on demand.*** **Burnap, Peter, et al. International Journal of Parallel, Emergent and Distributed Systems , 2015, Vol. 30.2. 80-100.**

19. ***Tweetviz: Visualizing Tweets for Business Intelligence.*** **Sijtsma, Bas, Pernilla Qvarfordt, and Francine Chen. InProceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, 2016. pp. 1153-1156.**

20. ***A case study of Israeli higher-education institutes sharing scholarly information with the community via social networks.*** **Forkosh-Baruch, Alona, and Arnon Hershkovitz. The Internet and Higher Education , 2012, Vol. 15.1. 58-68.**

21. ***Twitter as a resource to evaluate the university teaching process.*** **Suárez, Jonatan García, Carmen Trigueros Cervantes, and Enrique Rivera García. International Journal of Educational Technology in Higher Education , 2015, Vol. 12.3 . 32-45.**

22. ***A systematic identification and analysis of scientists on Twitter.*** **Ke, Qing, Yong-Yeol Ahn, and Cassidy R. Sugimoto. PloS one , 2017, Vol. 12.4. e0175368.**

23. ***Visual Sentiment Analysis on Twitter Data Streams.*** **Hao, Ming, et al. Visual Analytics Science and Technology (VAST), 2011. pp. 277-278.**

24. ***TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief.*** **Kumar, Shamanth, et al. ICWSM, 2011.**

25. ***Distilling Massive Amounts of Data into Simple Visualizations.*** **Rios, Miguel, and Jimmy Lin. Workshop on Social Media Visualization at ICWSM, 2012.**

26. ***Themestreams: Visualizing the stream of themes discussed in politics.*** **De Rooij, Ork, Daan Odijk, and Maarten De Rijke. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013. pp. 1077-1078.**

27. ***Document visualization: an overview of current research.*** **Gan, Qihong, et al. Wiley Interdisciplinary Reviews: Computational Statistics , 2014, Vol. 6.1. 19-36.**

28. [Online] www.researchgate.net/publication/271550479_Sentiment_Analysis_in_Arabic_tweets.

29. Docs of Python. [Online] https://docs.python.org/2/library/unittest.html.

Kingdom of Saudi Arabia
Ministry of Education
King Abdulaziz University
Faculty of Computing and Information Technology
Department of Information Systems

المملكة العربية السعودية
وزارة التعليم
جامعة الملك عبد العزيز
كلية الحاسبات وتقنية
المعلومات
قسم نظم المعلومات

Copyright Form

Project Title: Analyzing Performance of Saudi Universities on Twitter

Term of the Project: Semester 1st Academic Year 2017

We the team (Team Members Names):

	Student’s Name	Student’s ID	Student Signature
1	Amjad Dahlawi	1406448	
2	Ali Almalki	1408243	

Hereby assign our copyright of this report and of the corresponding executive summary to the Information Systems (IS) Department of King Abdulaziz University. We also hereby agree that the report or demo from our oral presentations is becoming full property of the IS Department.

Publication of this report does not constitute approval by King Abdulaziz University, the IS Department or its faculty members of the findings or conclusions contained herein. It is published for the exchange and stimulation of ideas.

Kingdom of Saudi Arabia
Ministry of Education
King Abdulaziz University
Faculty of Computing and Information Technology
Department of Information Systems

المملكة العربية السعودية
وزارة التعليم
جامعة الملك عبد العزيز
كلية الحاسبات وتقنية
المعلومات
قسم نظم المعلومات

Plagiarism Sheet

To plagiarize is to present a writing or invention belonging to another without proper acknowledgement. It usually takes one of the following three forms: the inclusion of another person's writing in one's own project or essay, paraphrasing of another person's work, or presentation of another person's original theories and/or views. If an allegation of plagiarism exists, disciplinary proceedings may be initiated and carried out within the academic program of the department in which the alleged offense occurred. In the case that a student or project team is alleged to have committed plagiarism, the Senior Project committee and faculty advisor may refuse to grade the project and record it as zero grade. Penalties may include failure in the course as well as recommendation for disciplinary action based on University regulations regarding such cases which may include a “disciplinary failure” grade or the disciplinary failure of all courses of the semester in which the offense was carried-out.

The Information systems department at King Abdulaziz University will STRICTLY enforce this policy on plagiarism.

NOTICE TO STUDENT:

Student acknowledges that he has read the above policy pertaining to plagiarism and understands that the penalty for such an act could result in disciplinary action.

Project Title: Analyzing Performance of Saudi Universities on Twitter

Term of the Project: Semester 1st Academic Year 2017

	Student’s Name	Student’s ID	Student Signature	Date
1	Amjad Dahlawi	1406448		10/12/17
2	Ali Almalki	1408243		10/12/17

(Please sign and return to the IS Department – Senior Project committee)

Kingdom of Saudi Arabia
Ministry of Education
King Abdulaziz University
Faculty of Computing and Information Technology
Department of Information Systems

المملكة العربية السعودية
وزارة التعليم
جامعة الملك عبد العزيز
كلية الحاسبات وتقنية
المعلومات
قسم نظم المعلومات

Student Commitment Form

System's Functional Requirements:

1. Graphical user interface for the user to navigate the menu.
2. Support for getting data from twitter using Filter, Search, and UserTimeline APIs.
3. Application connects to twitter API and open an authorization gate automatically.
4. Develop an interface to analyze the data using charts and visualizations
5. Save graphs and charts in different formats.
6. Provide time based rankings of universities based on their activities on twitter.
7. Provide comparison of universities based on different attributes.
8. Visualize in different graphs what people are talking about universities in real time.
9. Display tweet patterns based on specific keywords for the tweets of the public.

We the group members who have signed below commit ourselves to uphold and complete the Functional Requirements listed above of the senior project system. We understand the risks taken in not completing and augmenting those functional requirements into the second phase of the senior project.

We hereby sign below as agreement and commitment for the above:

Student	Student Name	Signature
1406448	Amjad Dahlawi	
1408243	Ali Almalki	

Supervisor

Signature