Analysis and Visualization of Social Data

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Computer Science and Engineering

Submitted By

Puneet Garg
(801332020)

Under the supervision of:

Dr. (Mrs.) Rinkle Rani
Assistant Professor

Mr. Sumit Miglani
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2015

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Analysis and Visualization of Social Web Data*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of **Dr. Rinkle Rani, Mr. Sumit Miglani** and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

(Puneet Garg)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Rinkle 12/07/15

(Mr. Sumit Miglani)
Assistant Professor
Computer Science and Engineering
Department

(Dr. Rinkle Rani)
Assistant Professor
Computer Science and Engineering
Department

Countersigned by

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. S. Bhatia)
Dean (Academic Affairs)
Thapar University
Patiala

Acknowledgement

The successful completion of any task would be incomplete without acknowledging the people who made it possible and whose constant guidance and encouragement secured the success.

First of all I wish to acknowledge the benevolence of omnipotent God who gave me strength and courage to overcome all obstacles and showed me the silver lining in the dark clouds. With the profound sense of gratitude and heartiest regard, I express my sincere feelings of indebtedness to my guide **Dr. Rinkle Rani** and co-guide **Mr. Sumit . Miglani** for their positive attitude, excellent guidance, constant encouragement, keen interest, invaluable cooperation, generous attitude and above all their blessings. They have been a source of inspiration for me.

I am grateful to **Dr. Deepak Garg**, Head of Department and **Dr. Ashutosh Mishra**, P.G. Coordinator, Computer Science and Engineering Department, Thapar University for the motivation and inspiration for the completion of this thesis. I will be failing in my duty if I do not express my gratitude to **Dr. S. S. Bhatia**, Senior Professor and Dean of Academics Affairs in the University, for making provisions of infrastructure such as library facilities, computer labs equipped with internet facility, immensely useful for the learners to equip themselves with latest in the field.

Last but not the least I would like to express my heartfelt thanks to my parents and my friends who with their thought provoking views, veracity and whole hearted cooperation helped me in doing this thesis.

Puneet Garg

(801332020)

Abstract

Social media has become very popular communication tool among internet users in the recent years. Humans have an ingrained tendency to share their ideas, experiences and knowledge, which associate them with the rest of the world, so that they can be recognized and can also identify their importance and worth. They are eager to know about happenings around them, that is why they communicate in order to share their ideas, observations and queries. Social media is one such communication medium that made people to be heard and satisfy their curiosity to know about rest of the world. A large unstructured data is available for analysis on the social web. The data available on these sites have redundancies as users are free to enter the data according to their knowledge and interest. This data needs to be normalized before doing any analysis due to the presence of various redundancies in it. Analyzing these huge social datasets and predicting the opinions of individuals plays an important role in business and academics. In this research, LinkedIn data is extracted by using LinkedIn API and normalized by removing redundancies. Further, data is also normalized according to locations of LinkedIn connections using geo coordinates provided by Microsoft Bing. Then, clustering of this normalized data set is done according to job title, company names and geographic locations using Greedy, Hierarchical and K-Means clustering algorithms and clusters are visualized to have a better insight into them. Secondly, we extract tweets about "AAM AADMI PARTY" (a recently grown political party) and build a development environment using python to induce analytical insights from these tweets using frequency analysis and sentiment analysis.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

Now-a-days, social networking sites have become not only very popular but also a need for the new generation. Through these sites, the people can interact with others for their common interests, actions, attitude, thinking or awareness. These sites have eliminated the barriers of time, country, language, gender and money. The users are surfing these sites for various purposes like making new friends, hunting for new job, advertising their business and finding siblings etc. These sites provide a platform of the real world at home for the users. Every time, everywhere connectivity makes these sites at boom rapidly in the technocrats. Recently, Facebook, Twitter, LinkedIn, Google^{+} become very popular social websites. Facebook has more than 800 million active users. These users are increasing 83% annually since 2008 [1]. Hundreds of new websites are impending every day and attract users with various offers. Social networks are the central key to comprehend common phenomenon, envisage human deeds and examine social constitution. By studying the behavior and concept of social networks, one can understand the social phenomena. The study of social networking also helps in optimization, execution and organization of problems in reality [2].

The current advances in information technology and increasing use of social media make it very easy to generate a bunch of data. Therefore the big challenges in front of us are to collect and integrate these bunches of huge data from the data sources which are universally distributed. If we have a look on the past 20 years, we can see the data has increased in a large scale. In 2011, IDC(International Data Corporation) submitted a report in which, he said that the overall data volume which was created and copied in this world was 1.8 ZB and within 5 years, it is supposed to be increased by nine times [3]. If we think about YouTube, every minute videos of about 72 hour are uploaded [4]. This Big Data can be understood as huge amount of unstructured data which needs a lot of real time analysis. Various Companies generates and process lot of such data such as Google which processes hundreds of PB of data, Facebook which generates nearly 10 PB per month of log data, a Chinese company Baidu which processes tens of PB of data and Taobao which processes data of tens of PB per day for online trading [5].

Fig. 1.1. Online Social Network as a big source of Big Data

Today, the researchers are focusing to investigate the structures and associations of various social networks. The main research area is study and analysis of compactness, centrality and grouping with the promise of security and privacy of the users data in these networks. The social networks are created and knockdown by studying daily basis and continuous communication among users. These studies incorporate with individual's relationships with others in terms of different parameters of human behavior [6]. In this way, these networks are proven an important and critical tool to study the communal phenomenon, composition, relationships and behaviors of individuals and/or group. The researchers are aiming to investigate these networks for various issues like number of users, increment in number of users, various applications provided by the sites, user's applications, and interest due to the reputation of these online social networks, ease in graphical investigation, the accessibility of huge amount of chronicle data, business and marketing welfares etc. Moreover, these Social networks show attractive and motivating research confronts like finding out and toning of identical sub-graphs, neighborhood relationship examination and representation, subscribers grouping, characterizing, cataloging and information proliferation etc. Therefore, investigation of data analysis of these online websites of social networking has an immense prospective for researchers in an assortment of branches [7].

1.1 Introduction to Big Data

Big data can be considered as the datasets which can not be perceived, acquired and processed by present IT software and hardware tools within a tolerable time. Big data can be considered as the next frontier for great innovation, lot of competition and more and more productivity. To acquire, store and manage big data is beyond the scope of classic database software.

The amazing growth of cloud computing, social networking and IOT (Internet of Things) is continuously promoting the huge growth of data. Cloud computing play its role by safeguarding data assessment from accessing sites and channels. On the other hand, the internet of things will create sensors in cities all over the world which collects and transmits the data which is to be stored and processed in that cloud. With consideration of various aspects of big data such as heterogeneity, real time analysis, complexity and privacy there is a need to mine and analyze this data, so that we can reveal the actual intrinsic property of this big data and the process of decision making can be improved.

An analyst of META (Gartner), Doug Laney [8] elaborated big data with **3Vs Model,** increase in volume, velocity and variety.

1.1.1 The 3Vs Model

Researchers summarized the 3 aspects of big data which are Velocity, Volume and Variety, also known as 3Vs [9].

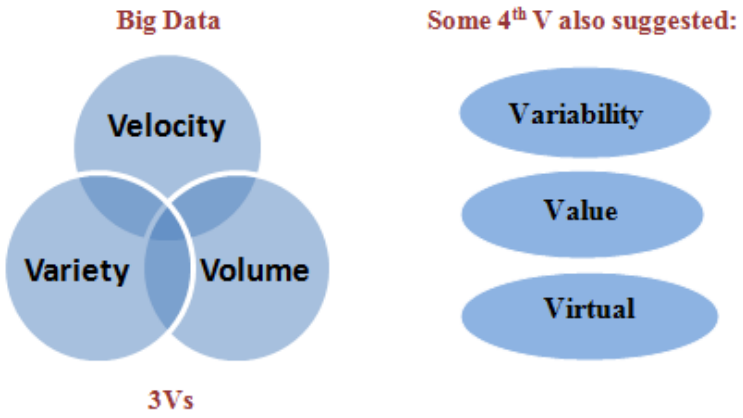


Fig. 1.2. Various V's of Big Data

The 3Vs can be illustrated as:

- **Volume:** The increasing volume of data is one of the most noticeable challenges. The scientist in various fields of science also observe the increasing data volume and they encounter computing limitation due to this increasing volume. These types of issues due to this big volume can also be found in web applications, business informatics, finance etc. Despite of various such types of big volume issues, there exists no quantification criteria about big data. This quantification of big data depends on the data structure complexity and the requirements of the target application. Data volume is increasing exponentially and it is estimated that there will be 44 times increase in data from 2009 to 2020 from 0.8 ZB to 35 ZB.
- **Velocity:** velocity can be viewed here as the speed with which new data is created and the existing data is updated. Now a days most of the data is machine generated data and we can think of a system which can sense the data immediately when it is created. Both the layers i.e., query processing layer and the storage layer of the database management system need to be fast and scalable that can match with the speed of data creation.

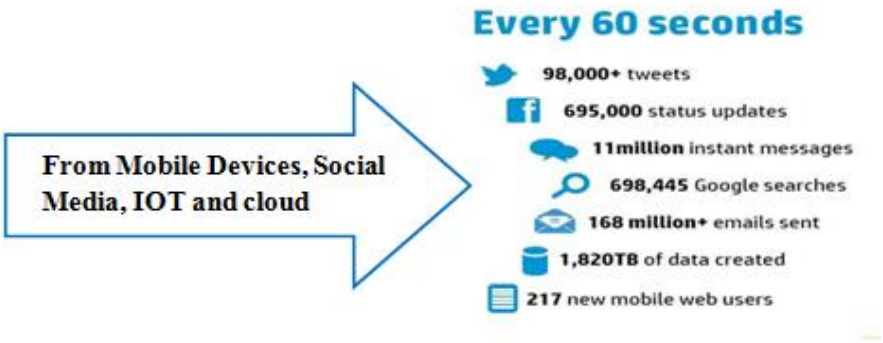


Fig. 1.3. Velocity

- **Variety:** Due to the environment of real world applications, the data which we are receiving is not coming from a single source. Therefore the data can be of various models and formats. For big data mining and analysis of such data there is a need of integration of such a huge variety of data. From this perspective the challenge of variety comes along with the implementation of big data. This huge variety of data can be a combination of structured, semi structured and unstructured data.

This huge data can be user generated contents, transactional data, scientific data, Web data, Graph data etc. [9].

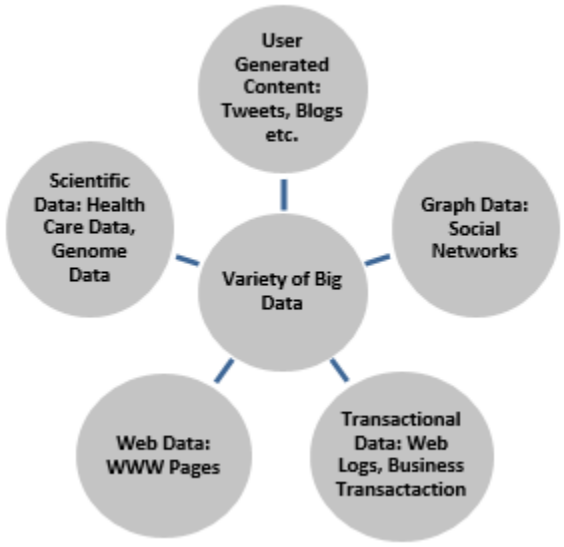


Fig. 1.4. Various Sources of Big Data

Recently some more 4th V have been suggested which are Variability, Value and Virtual which refer to some other different aspects of big data. The most important among these is Value.

- **Value:** The value that big data could create is the effect which we can obtain by analyzing the big data i.e., we can improve the productivity of the enterprise and competitiveness and can create lots of benefits for the customers also. Mc Kinsey [10] provided some facts that the expenditure of US healthcare can be reduced by 8 % by effectively utilizing big data. The efficiency of various government operations can also be increased by utilizing big data efficiently. So, we can conclude that there is tremendous need to analyse big data so that the process of decision making can be improved and its benefit can be taken by companies working in various fields.

1.1.2 Big Data Challenges

As we have seen there is sharp increase in big data volume day by day, resulting increased challenges that is how to acquire big data, how to store, manage and analyze the big data. The traditional RDBMS has a big limitation that it can be used with only

structured data and requires expensive hardware, therefore it cannot handle the big volume and big variety of big data generated at a high speed. Some researchers [11-13] discuss various difficulties while developing big data. Some of the key challenges which are analyzed are discussed below:

- **Data Representation:** As there are various types of heterogeneity found in big data such in data types, in structure and some other factors such as accessibility, organization and structure etc., the data representation become a crucial task due to all these reasons. Effective data representation will lead to better integrated technologies and better analysis of big data.
- **Data Confidentiality:** Till now we are dependent on some professional tools and frameworks for analyzing big data, due to this dependency risk of security of data is increased. The datasets can contains some confidential information such as transaction log of banking system can contain credit card information of various customers, providing this type of information to the third party is a big threat to confidentiality.
- **Reduction in Redundancy and Compressing the Data:** As we know the data is coming from a variety of resources so there can be big redundancy among all the datasets. By eliminating such redundancy the cost of analyzing big data can be reduced to a greater extent because this redundant information creates a lot of problems in analyzing. Further data compression is also a main task to compress the huge amount of big data.
- **Scalable and Expendable:** The system which we create for analysis of big data must be able to handle both the present and future data. As we already know that data is drastically increasing day by day so our big data analysis system must be scalable to handle the growing data.
- **Managing Data Life-Cycle:** Growth of advancement in storage systems is less as compared to the growth of big data. To make existing storage systems to be capable with big data we have to think about which data to be stored and which data should not be stored.

1.1.3 Big Data Storage

The databases which are used to store the big data are the NoSQL databases. NoSQL databases do not use SQL statements to manipulate the data because these are different from RDBMS databases in the sense of storing and manipulating the data. These type of databases not primarily built on the concept of tables that is they can also store unstructured data.

In the starting of 21st century the movement of NoSQL databases starts. RDBMS is totally based on the ACID properties and have a great attachment with ACID properties instead of this NoSQL databases are based on BASE properties.

- **Basic Availability**: Each and every request is guaranteed a response either successful or unsuccessful.
- **Soft State**: The state of the system can be changed over time, at times without any input**.**
- **Eventual Consistency**: The database can be momentarily inconsistent.

The NoSQL databases are based on the CAP theorem. CAP theorem means three things that are consistency, availability and tolerance to network partition. CAP theorem states that it is impossible for distributed system to achieve these three things simultaneously. As we are talking about distributed systems so there is need of tolerance to network partition but among consistency and availability only one can be achieved at a time. Big table and HBASE are two popular CP systems which provide consistency and tolerance to network partition. Cassendra and dynamo are the two most popular AP systems which provide the availability along with tolerance to network partition.

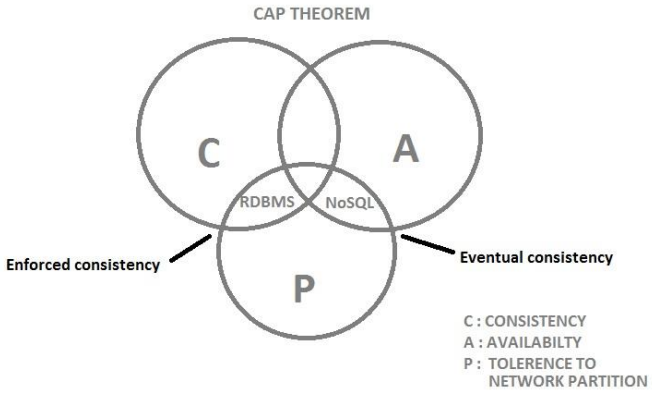


Fig. 1.5. CAP Theorem

There are various NoSQL databases such as big table and HBASE etc. depending on various types of databases such as document type databases, key-value type databases, xml type databases, column type databases and graph type databases which are illustrated in a table given below:

TABLE 1.1: Various NoSQL Databases

Document	Key-Value	XML	Column	Graph
MongoDB	Redis	BaseX	BigTable	Neo4J
CouchDB	Membase	eXist	Hadoop/HBase	FlockDB
RavenDB	Voldemort		Cassandra	InfiniteGraph
Terrastore	MemcacheDB		SimpleDB	
			Cloudera	

1.1.4 Big Data Processing

There is a big variety of data that is available to us for analysis. First of all data integration on data collected from various sources is performed and then data cleaning is done that removes some unwanted data. After that data reduction is performed to remove noise and to keep only meaningful data. Then we manage data query and indexing and finally data analysis and mining is performed.

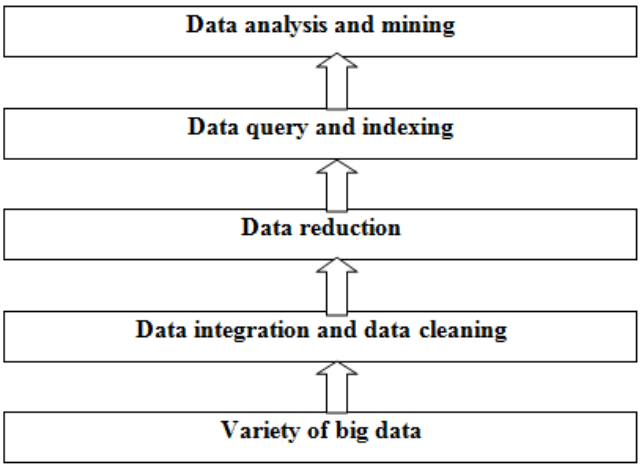


Fig. 1.6. Big Data Processing Steps

The collected data cannot be analysed as such so we have to perform some data pre-processing steps including data integration, cleaning of data, and elimination of redundancy. These are described below:

- **Integration of Data:** The data integration is the main part of big data processing as we know big data is heterogeneous in nature and is obtained from a variety of sources. So we have to integrate data collected from different sources. The main techniques used for data integration are data warehousing and data federation. These techniques are based on ETL that is to extract, transform and load the data. To query and aggregate the big data from different data sources a virtual database can be constructed.
- **Data Cleaning:** As data is coming from various sources, there is a lot of data which is inaccurate and inconsistent, which is incomplete and of no use. So, we have to delete or modify such unimportant data from the bunch of big data. Data cleaning consists of these operations[14]: all the error types are determined and defined, then search for the errors and identify them, then correct these identified errors, then creating a document indicating the various error types identified and further we have to prevent these types of errors in future.
- **Redundancy Elimination from Data:** due to heterogeneous nature of big data, a lot of redundant data is also present which can make our data transmission very expensive, storage space is wasted and it leads to inconsistency problem also. So, to remove this redundancy first of all we have to detect the redundancy, then we have to filter the data and then to compress the data.

Once the data become suitable for analysis, we can proceed for querying and indexing big data and then analysis of such huge datasets available.

- **Querying and Indexing Big Data:** A lot of challenges arise when we think about query and indexing of big data. The main challenge is its size which is so huge that is not easy for people and most software to process and manage that data. The second challenge is to store big data because it can not be stored in a single machine so a distributed environment is needed. Therefore different from the traditional index structure, the indexing of big data will be made on the basis of the distributed system. Tree like index also does not work well in collaboration with big data because of bottleneck problem. Fault tolerance is also one of the factors that can not be neglected in case of indexing big data.

The distributed B tree method is one method that can be used to index big data. With the use of optimistic strategy for concurrency control it can provide transactional access by performing consistent concurrent updates. Another strategy which can be used to index big data is BATON overlay [48] which is used to support the range queries. An N node BATON overlay network can answer the both types of queries that are exact query and rang query in O (log N) time.

- **Analysis and Mining of Big Data:** To obtain the big value from big data it is necessary that we have to analyse and mine that big data properly with appropriate techniques. The analysis and mining of big data is a big challenge due to its complex and heterogeneous nature. In order to utilize the concept of big data in decision making process, a deep analysis of that data is needed which is difficult to be done by SQL statements. People in various fields must be able not only to know the current happenings but also can predict what can be happened in the near future by analyzing the big data.

  By analyzing the big data we can take various effective actions to retain the customers which can overcome the risk of losing the customers [49]. The simple OLAP data processing techniques are not enough for such type of analysis. We can try some other types of analysis which are complex such as path analysis, what-if analysis, graph analysis etc. [49]. In 2004, Google provided a solution that is **Map- Reduce** framework which is parallel computing model that can be used to do big data analysis and mining. Now a days Map-Reduce framework is also integrated with some typical analysis software such as R and Hadoop [50]. R is an open source software used for statistical analysis. Weka is also an open source software that can also be used for data mining and machine learning.

1.2 Mining and Analyzing the Social Web

In recent years the escalating tendency of humans to use social networks such as Twitter, Facebook and LinkedIn, have resulted in making various kinds of relationships and interactions which, have led to the availability and generation of massive amount of precious data that has never been accessible before. Such massive valuable data can be

utilized in some varied, new, useful and eye-catching research areas to lot of researchers [15]. Although the area social web mining has received an enormous deal of consideration in the last decade, yet various problems related to mining social web is still in its immaturity and needs some more techniques to be explored in the future for further upgrading. However, since the huge data generated from social web is vast, noisy, dynamic and distributed, this requires suitable data mining procedures to analyze such complex, large and repeatedly changing social data. These days, the main focus of analysts is to observe the association and structure of various social networking sites. To study and analyze compactness, centrality and grouping with the promise of security and privacy of the users data in these networks is the main research area in social web analytics. Mining and analyzing the social web helps in identifying various trends of the society and perspective of society towards different products, events, political party etc. Various research issues associated with social web analytics such as predicting future trends, community detection, link prediction, recommender system, frequency analysis, mood analysis, opinion mining, influence propagation etc. are shown in Figure 1.7. These are the latest research areas which emerged from social web analysis. The digital data generated from these online social networks is noisy, vast, dynamic and distributed [15]. Efficient data mining techniques are required to analyze such complex, large and rapidly changing social data.

Fig. 1.7. Research Issues associated with Social Web Analytics

Some of the main research issues are discussed below:

1.2.1 Influence Propagation

Nowadays, as online social networks are attracting majority of people, latter rely on building decisions on the basis of influence of such social networking sites [16]. For example, influence propagation may help in making a decision like which product to buy, which video/audio to watch, which social community to connect, and so on. Thus, influence propagation has played very important role for valuable viral marketing, where different companies can convince their customers to purchase their products with the help of those active persons in these online social networks who can play an important role in influencing other persons. Hence this requires necessitate for researchers to study the influence propagation in online social networks which is currently a sizzling topic related to social network mining functionalities.

1.2.2 Community or Group Detection

Community detection or group detection in online social networks is based on studying the structure of online social network to cluster individuals into different groups by finding which persons correlate more with each other in comparison to other users. Such detection of groups/communities can help to further make a judgment about what services, products and activities a person might be interested in. Hence, investigating structures of these online social communities is of prime significance that attracts the researchers in data mining field to make a deep study of group detection in online social networks. An illustration of 3 communities in a online social network is depicted in Figure 1.8. Two nodes may be linked by an edge if and only if any relationship exists between these nodes according to their role or participation in the online social networks. Connections of various nodes in this situation are binary. Such representation of groups/communities can be best visualized with the help of graphs. In fact, one of the most appropriate features of graphs is community structure for representing real systems.

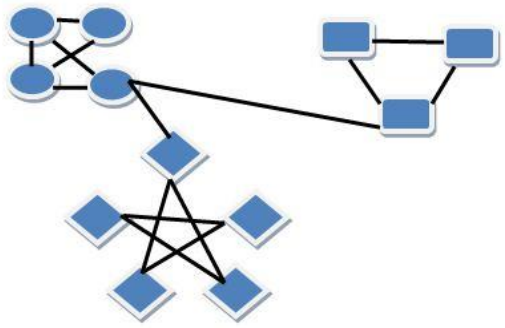


Fig. 1.8. Online Social Network showing three Communities

1.2.3 Link Prediction

Link prediction in online social networks is all about predicting the chances of a future relationship between two nodes, knowing that there is presently no relationship between these nodes. This is achievable by mining the huge amount of data available in these online networking sites to find out "which products have a relationship with other products" or "who is a buddy of whom". By gathering and analyzing useful information about a product or individual, Social web can infer new connections among nodes/members of social networks that are probable to occur in the near future. Thus, link prediction is an important research sub area for social web mining.

1.2.4 Sentiment Analysis or Opinion Mining

Web sentiment sources such as web forums, merchant sites, groups, discussion and blogs are rapidly growing containing valuable information useful for both manufacturers and customers. Due to ease of access of these web resources, sentiment sources are utilized as a platform by individual persons to share their opinions or experiences. User's sentiment plays a key role both for manufacturers as well as customers for taking proper decisions. The feedbacks of various existing users are helpful for new customers in selecting a right product, whereas on the other hand, it helps manufacturers to know the weaknesses and strength of their products from the viewpoint of end-users. Such sentiment analysis is very informative in developing product development plans and marketing. Thus sentiment analysis or opinion mining can be defined as identifying and extracting the subjective information in the source materials with the use of natural language processing, computational linguistics and text analysis [17].

1.2.5 Recommender Systems

Recommender systems are a sub class of information filtering systems which seek to predict the 'preference' or 'rating' that an individual would give to a product. They have become really common in recent years, and are applied in a lot of applications. The most popular applications are probably music, movies, books, news, search queries, research articles, products, and social tags in general. However, there are lot of recommender systems can be made for jokes, experts, financial services, restaurants, persons (online dating), and life insurance and Twitter followers.

1.2.6 Frequency Analysis

Frequency Analysis is to calculate some simple statistics from this huge amount of data generated every day. For example calculating number of users to a particular website, how many followers are there on a Twitter account, number of students studied from Thapar University Patiala in my LinkedIn contacts, number of likes for a particular post, number of users from a particular area etc.

1.3 LinkedIn: The Professional Social Web

LinkedIn is an online social networking site concentrated on professional and business associations. At the first look, LinkedIn is also looks like any other social website, but in reality, it is pretty dissimilar in its API data with others. For example, we can compare Twitter with an hectic open public forum broadcasted in the city and Facebook with a hall overflowed with friends and family members talking regarding belongings which are (generally) suitable for dinner discussion, then LinkedIn will be a private cabin having access with some specific professional with a semiformal dress code and a meaningful motive and excellent manners who are demanding to express the precise assessment, significance, knowledge and proficiency to promote their business ethics and values in the marketplace.

Due to the responsive and professional characteristics, LinkedIn's data and its API is quite different from data available at other social network sites and has its own senses. The main motive of the people who sign up this social network is predominantly

concerned in the commerce prospect of their business details like business relationships, previous employment etc.

LinkedIn social network provides a large data of the millions of subscribers. The conventional database management techniques are not able to manage or handle this huge datasets. The datasets, available before 2000, are very simple and in a few amount. Now, the number of subscribers has increase in multiple of thousands. This makes the datasets very complex. To store, compute and handle these datasets, various powerful tools and techniques have proposed by the researchers. As per the advancement in the technology, various data handling tools have been developed. The research is going on for searching and developing such type of tools to handle and provide the sensitive information for different purposes continuously in a drastically manner. These social sites are the biggest source of the data where the subscribers provide their information themselves to get the benefits of the various services.

It is a difficult task to manipulate these complex datasets at a big scale. This can be possible by using the specific system tools and methods to divide these datasets into small subsets (for example subgroups of personal contacts) for potentially constructive outcomes. Due to the soberness and professionalism in the data of subscribers, LinkedIn is a predominantly attractive goal. By investigating the data of LinkedIn social network, we can collect lot of important information about individuals and groups. We can achieve this target at a micro diagnostic approach by analyzing the profiles of the individuals for oddments of information and at a macro diagnostic approach by analyzing the prototypes of the data. The LinkedIn profile provides the information (profession, employer, school and college information, past history of the job etc.) of the individuals to the users even they are not connected to the network [27].

The data provided by the LinkedIn is also important because the subscribers provide their previous job details in the time series so that we can analyze the information regarding colleagues at that period. We can also build some basic and essential data of the users to find out the answers the following questions:

• The people who are interested or working in the same area of yours for example occupation or employer?

• The people who are worked in organizations you want to work for?

• The people who are residents of the same geographical area where you reside?

LinkedIn is a predominantly appealing and exciting objective for analyzing the social web, specified the proficient environment of its subscribers. By investigating LinkedIn network information and data, we can understand the behavior and other properties of individuals and group [27]. With analysis of data of LinkedIn, we can explore the relations of individuals and group in professional life. By collecting all the LinkedIn relations, the previous background of occupation, employer, designation, locality etc., we can exactly picture the personality of the users. We can also compare the individuals by setting some parameters at LinkedIn.

1.4 Twitter: A Microblogging Service

Twitter, a micro blogging amenity empowers people to communicate with precise message with a limit of 140 characters, which represents their thoughts and ideas. Twitter is a predominantly appealing and exciting objective for an analyst that specifies the proficient environment of its subscribers. By analyzing tweets, one can understand the behavior and opinion of individuals and groups [22]. Twitter can be defined as a free, global and high speed microblogging service that helps people to share their ideas with others by short and precise, 140 characters long messages, providing easy and rapid communication. More than 500 million curious users have already registered on twitter; most of them actively engage their curiosity on regular basis. It is a great source of social data as it is open for public consumption and it provides well documented and clean API. Tweets are particularly interesting as they happen at the "velocity of thoughts" and are accessible in the nearest real time. Twitter provides a simple and asymmetric "Following Model" that satisfies human's curiosity. The following mechanism of twitter is asymmetric because it allows one user to follow other without mutual acceptance. Huge amount of social digital data is being produced on regular basis in the form of tweets. Analyzing user's opinions from this large social dataset leads to large set of possibilities in various fields such as business, marketing, advertising, politics, education etc. Different persons have different opinions about a political party; it can be either a criticism or an appraisal [28]. Analyzing combined view of all the persons about a

political party helps in building an overall image of the party out of these individual opinions.

Since the arrival of twitter, twitter has gained remarkable growth in the number of users and has become one of the most famous microblogging website [29]. twitter has a very simple interface that allow users to read or send messages called 'tweets' not more than 140 character. Everyday 400 million or more tweets are posted worldwide [30].

Enough tools are available for researchers and developers to design interesting applications. Twitter provides different APIs to extract tweets, retweets, followers etc. REST API helps programmers to access, write and read twitter data. Streaming API allow continuous access to twitter data for a search query and it runs continuously until it is killed or till there is an internet connection. Twitter provides various streaming endpoints to access twitter data [31].

- **Public Streams**: These streams are used for accessing publically available twitter data. They are used for mining the data for particular research interest by an individual.
- **User Streams:** These APIs are used for accessing the data that is related to a particular user. Tweets, retweets and followers of a particular user can be accessed using these streams.
- **Site Streams:** These streams are used predominantly for those servers that are associated to twitter server as a proxy server for lots of users.

Tweet is a short 140 character status update posted by users. Tweet not only contains the actual text message but also bundled with lots of metadata like entities and places. Tweet entities are in the form of user mentions, URLs, media and hashtags. Tweets may also contain references to various places which map various locations in the real world.

Tweets and Twitter's relation model or we can say its "following" mechanism connect people in multiple ways, ranging from precise (but often meaningful) conversational dialogues to the interest graphs that link people and various things that they care about. Twitter's relationship model allows us to keep up with the newest happenings of *any* other user of twitter, even though that user may not decide to follow you back or even know that you survive, whereas some social networking websites like

LinkedIn and Facebook require the mutual approval of a connection between various users (which usually implies a real-world association of some type). Twitter's *following* model is very simple but exploits an elementary characteristic of what makes us human beings: our curiosity. Whether it is an obsession with celebrity gossip, an intense interest in a particular political party and political topic, an urge to keep up with a most wanted sports team, or a desire to hook up with someone new, Twitter provides you with endless opportunities to satisfy our curiosity.

Chapter 2 Literature Survey

2.1 Big Data

In the last 20 years, data has increased in a very large scale in various fields. International Data Corporation (IDC) submitted a report in 2011. According to that report, the overall copied and created data volume in the whole world was about 1.8ZB ($\approx 1021B$), which increased by nearly 9 times within 5 years. This amount of data generated will double at least after every other two year in the coming future. Under this explosive growth of world's global data, the term "big data" is used to define enormous datasets. Big data typically contains masses of unstructured data that is why big data need more real-time analysis in comparison to traditional datasets. In addition to this, big data also brings a lot of new opportunities for discovering bundle of new values, which helps us to gain a deep understanding of various hidden values, and also incurs lot of new challenges, e.g., how such datasets can be effectively managed and organized. Recently, lot of industries are showing their interest in the high potential of growing big data, and various government agencies also announced some major plans that can accelerate big data research and applications. In addition to all this, issues on big data are also normally covered in public media, such as New York Times, The Economist and National Public Radio. Two premier scientific journals named *Nature* and *Science* also opened some special columns that discuss the various impacts and challenges of big data. The era of big data has come, no doubt about that. Nowadays, big data that relates to the services of various internet companies grow rapidly. For example, Google, the most famous search engine, processes data of hundreds of Petabyte (PB) and Facebook, one of the most growing social media generates log data of over 10 PB per month.

M. Chen et al. [5] reviewed the state-of-the-art and background of big data. They first introduced the background of big data and reviewed its related technologies, such as Internet of Things, could computing, Hadoop and data centers. They focused on the various phases of the value chain of big data, i.e., big data generation, big data acquisition, big data storage, and big data analysis. For each phase, they introduced the general background, discussed various technical challenges, and reviewed the latest

advances. They finally examined lot of representative applications of big data, including Internet of Things, enterprise management, online social networks, collective intelligence, medical applications and smart grid. These discussions aim to provide the readers a complete overview and big-picture of this exciting area. This survey is concluded with a discussion of lot of open problems and lot of future directions.

J. Chen et al. [9] reviewed various big data challenges from a data management respective. They discussed data diversity, data reduction, data integration and cleaning, data indexing and query, and finally data analysis and mining with respect to big data.

While the amount of such large datasets is drastically growing, it also brings about a lot of challenging problems that requires prompt solutions:

- The latest advances of IT make it quite easy to generate huge amount of data. For example, on an average, about 72 hours of videos are uploaded every minute to YouTube. Therefore, researchers are confronted with the great challenge of collecting and integrating this huge massive unstructured data from lot of widely distributed data sources.
- The increasing growth of the Internet of Things (IoT), cloud computing and social media usage further promotes the very sharp growth of big data. Cloud computing provides various safeguarding, channels and access sites for data asset. In paradigm of IoT, lot of sensors present all over the world are transmitting and collecting data to be stored and processed this huge dataset in the cloud. Such dataset in both mutual relations and quantity will far surpass the ability of the IT infrastructure and architectures of various existing enterprises and its real time constraint will also greatly stress the currently available computing capacity. The increasingly rising data cause a problem that how to manage and store such large heterogeneous data with moderate requirements on software and hardware infrastructure.
- In consideration of the scalability, heterogeneity, complexity, real time and privacy of big data, we have to effectively mine the these huge datasets at different levels during the modelling, analysis, forecasting, and visualization, so as to expose its intrinsic property and improve the process of decision making.

2.2 Social Web Analysis

A social network can be defined as a social community of individuals, who are related to each other based on a common relation of concern, e.g. trust, friendship, etc. Thus social network analysis can be defined as the study of online social networks to understand their behavior and structure. Social web analysis has gained importance due to its usage in different applications - from organizational dynamics (e.g. management) and search engines to product marketing (e.g. viral marketing). Recently there has been a brisk increase in interest about online social network analysis in the present data mining community. The basic inspiration is the requirement to exploit knowledge from abundant amounts of data extracted, pertaining to the social behavior of individuals in such online environments. Various data mining based methodologies are proving to be valuable for analysis of the huge social network data, especially for large data that cannot be handled by various traditional methods.

Mining and analyzing the social web has been an emerging area of research in the recent era mainly due to the enormous increase in the popularity and usage of such social networks. However, the vast amount of resources presented in these online Social Networks poses a big confront for researchers to analyze these social networks. Also, the data produced from these social networks is dynamic and requires intelligent mining to analyze this data.

Recently the increasing affinity of citizens to use these Social Networks such as Facebook, LinkedIn, and Twitter have resulted in building different kinds of relationships and interactions which have led to the availability of an enormous amount of valuable data that has never been accessible before. Such enormous valuable data can be utilized in some new, eye-catching, varied and valuable research areas to lot of researchers. Although the area Social web analysis has received an enormous deal of consideration in last few years, yet a lot of problems related to social web analysis is still in its immaturity and needs some more methods to be developed in the future for additional improvement.

However, since data generated from these social networks is noisy, vast, dynamic and distributed, this requires suitable data mining methods to analyze such complex, large, and rapidly changing social data. Research is being carried out associated with

various research issues in analyzing the social web such as, expert finding, influence propagation, link prediction, recommender systems, opinion mining, community detection, mood analysis, etc.

Social network analysis can be defined as the mapping and measuring of interactions, relationships and flows between people, organizations, groups, URLs, computers, and other linked information entities. The nodes in the social network are the individuals and groups while the links represent the relationships/flows between the connected nodes. Social network analysis provides both a mathematical and visual analysis of the human relationships. Management consultants use the methodology of social web analysis with their clients and name it as organizational network analysis.

P. Mika [44] presented the Flink system for the aggregation, extraction and visualization of social networks. The system employs semantic methodology for reasoning with lot of personal information which are extracted from huge amount of electronic information sources including emails, web pages, FOAF profiles and publication archives. They use this acquired knowledge for the purpose of social network analysis and for producing a web-based presentation of the social community. They demonstrated novel approach to social science that is based on vast electronic data using semantic web research community.

F. Fu et al. [45] presented empirical analysis of various statistical properties of two important Chinese social networks available online, one is a blogging network and other is SNS network open to college students. These are the social networks which are emerging in the age of Web 2.0. They demonstrated that both social networks possess scale free and small world features that are already observed in artificial networks and real-world. In addition, they investigated the distribution of topological distance. They also study the correlations between degree (in/out), clustering coefficient and its degree, in-degree (for the blogging network) and popularity respectively. They find that various blogging networks represent disassortative mixing pattern, whereas the student's SNS social network is an assortative one. Their research may help us to clarify the self-organizing structural features of such online social communities embedded in technical forms.

A. Mislove et al. [46] presented a large-scale measurement, analysis and study of the structure of various online social networks. They examined huge amount of data generated from four these online social networks: Flickr, Live Journal, YouTube, and Orkut. They crawled the various publicly accessible user links on these sites, thus obtaining a very large portion of these social network's graphs. Their dataset contains about 328 million links and 11.3 million users. This is the first research to examine and analyze multiple online social networks at one scale. They observed that the in-degree of various user nodes tends to match the out-degree of those nodes; that the social networks contain a tightly connected core of nodes that have high-degree; and that this core links some small groups of strongly-clustered communities, low-degree nodes are present at the fringes of the network.

2.3 Data Mining Techniques: Clustering and Classification

With the huge amount of generated from social media and other repositories, it is progressively more important to develop great means for interpretation and analysis of data and for the mining of remarkable knowledge that could assist in various decision-making process. Data Mining, also known as knowledge discovery in databases can be defined as "the nontrivial process that identifies novel, valid, potentially useful and understandable pattern in data".

Researchers identify two elementary objectives of data mining: description and prediction. Prediction utilizes various existing variables in the database in order to predict the future values of interest and description mainly focuses on finding various patterns that describes the data and the subsequent presentation for individual interpretation. The relative prominence of both description and prediction differ with respect to fundamental technique and the application. There are several data mining techniques fulfilling these objectives: classification mining, association rule mining and clustering using the techniques such as genetic algorithms, decision tree, neural networks and machine learning [47].

B. H. Park and H. Kargupta [52] presented a brief overview of the Distributed Data Mining algorithms, applications, systems and the emerging research areas. They first presented the related research of distributed data mining and illustrated various data

distribution scenarios. Then they reviewed the various distributed data mining algorithms. Subsequently, they discussed various architectural issues in distributed data mining systems.

J. Han et al. [55] designed a spatial data mining system prototype that is called GeoMiner. The spatial data mining power of GeoMiner includes mining mainly three kinds of rules: comparison rules, characteristic rules and association rules, in geo-spatial datasets, with a designed extension to include mining clustering rules and classification rules. GeoMiner includes the spatial on-line analytical processing (OLAP) section, spatial data cube construction section, and spatial data mining section. A spatial data mining language, Geo-Mining Query Language, is designed by them and implemented as an extension to Spatial SQL, for spatial data mining. Moreover, they constructed a user-friendly, interactive data mining interface and implemented the tools for visualization of discovered spatial knowledge.

The digital data generated from various online social networks is noisy, vast, dynamic and distributed. Efficient data mining techniques are required to analyze such complex, large and rapidly changing social data. Some of the data mining concepts are described below:

2.3.1 Clustering

Most of the time, an industry maintains its database to collect various types of information, but these datasets are not valid for solution of each and every problems. So, this data can not be assumed as universal datasets for all the problems e.g. there can be lacking in the design of the application's user interface, blank or over filled columns, abbreviations, misspelled etc. Social networking sites provide the text free service to the users to enter their information. To analyze the data sets in an effective manner, the data have to be normalized first, after that efficient clustering technique should be used [18]. Clustering is an unsupervised machine-learning method which is used to fastener in every data mining toolkits [19]. In this, all the information is first collected and then partitioned into a number of small clusters based on similarities in their properties. Various methods for clustering of the data sets can be legitimate as data mining tool kit [20]. The outcomes of the clustering depend upon the choice of clustering technique and type which is used.

There are a number of clustering techniques available, but that technique should be adapted which gives the best solution for our problem [21]. There are some of the general similarity metrics which are very useful to compare the company or designation names such as Levenshtein Distance, N-gram similarity, jaccard distance etc.

Clustering is a technique that can be used for the objective of division of data into various groups of similar objects. Each group which is called cluster consists of objects that are analogous among themselves and dissimilar to various objects present in other groups. Data modeling puts the technique of clustering in a historical viewpoint rooted in statistics, mathematics, and numerical analysis. Clusters from a machine learning point of view correspond to some hidden patterns and the search for such clusters is an unsupervised learning approach, and the resulting clustered system characterizes a data concept. From a practical point of view clustering plays an excellent role in data mining applications such as information retrieval and text mining, scientific data exploration, spatial database applications, CRM, marketing, medical diagnostics, Web analysis, computational biology, and many others.

- **Data Normalization:** Various subscribers fill their data on these online social networking sites such as LinkedIn with their potential. That is why the data contents are not in the same format. We are collecting data by requesting the API. Some of the details are missing or overfilled by the subscribers. So, to make this data useful for our project, we have to manage this data into desire format using some tools. For an example, In an API, users are filling job title. But, we cannot assume it as final data contents for job titles. As the user fill the job title according to his knowledge. Suppose, we require "Quality Inspection Officer", then some will write it as it is. But, some will fill it as "Q. I. O." or "Quality Officer" or "Quality Manager" or "Quality Inspector" etc. So, we have to go through normalization of the data and arrange them in the desired format to assure the best results from these social datasets.
- **Similarity Computation:** The next target in the data analysis is to compare the normalized data with their similarity base. Even after, normalization of the data, it is a difficult task to find out the best results based on performance basis from the social datasets. Some comparison or analyses are as easy as simple arithmetic

addition. But, some comparisons or analyses are very complex. For Example, the comparison can be done for finding out the similarity between various basic information like name, present and previous job, employers, and locations etc. for people who are subscribers of LinkedIn. So, we have to apply some heuristic approach to find out the best suitable option or comparison between two data contents. The Natural Language Tool Kit (NLTK) is a python tool kit that is very useful in managing or computing the datasets of various social network websites. In Python packages, firstly install pip install nltk. In this, we have some of the general similarity metrics which are very useful to compare the company or designation names.

- **Edit Distance:** It is also known as Levenshtein distance. In this, we measure the number of additions, cutting and substitutes inserted to switch one array or data content into another. Suppose, we have to convert "ball" into "wall", then we replace only the first letter of "ball" i.e. "b" into "w". The NLTK value will be "1" for this operation due to replacing of the single letter. The function to perform this task in NLTK is nltk.metrics.distance.edit_distance. The number of steps to perform the edit distance task is different from the number of edited letters.
- **N-gram Similarity:** The *n*-gram similarity can be defined as number of ways to express the all possible successive sequences of *n* tokens from a text reference. In this type of approach, we count the number of similarities between two entities. In NLTK, there are two types of scoring mechanisms viz. bigram and trigram. These can be analysed in NLTK with two classes using the Bigram Association Measures and Trigram Association Measures using nltk.metrics.association module.
- **Jaccard Distance:** To compute the data sets of LinkedIn for the similarities, we compare two sets where set can be expressed as an asymmetric group of data. In Jaccard distance measure, we define these data sets as similarity metrics. This distance can be expressed as ratio of intersection of the two sets to the union of the sets i.e.

$$\frac{|\boldsymbol{Set\ 1} \cap \boldsymbol{Set\ 2}|}{|\boldsymbol{Set\ 1} \cup \boldsymbol{Set\ 2}|}$$

This can also be represented as the ratio of cardinalities of intersection to the union of the sets. The outcome of this ratio is the common items between two sets. Generally, we express the similarity between two sets in terms of Jaccard distance which is obtained by subtracting the final values of the division from 1.0.

- **Dimensionality Reduction:** Dimensionality reduction is a method to reduce the complexity of the data contents to materialize in an effective manner. The datasets are divided into smaller subsets so that the comparison can be done easily and timely. Otherwise, the huge data will become nontrivial. For this purpose, we analyse or compare each member's data content to all other members. For example, if there are n numbers of data sets, then we have to compare each data set with n-1 data sets. This problem is well known as an *n-squared problem* in the region and denoted as $O\ (n^2)$ problem formulation. This problem have no meaning for very less number of data sets i.e. for small n, but for very large value of *n*, this problem become gigantic. Generally, we have to wait or go through a long process to analyse such type of data contents. For very complex system, this process consumes time in months, years or much more for getting the results.

  So, we use dimensionality reduction method to find out the solution of such problems. In this technique, we analyse some functions to systematize the data contents as "similar enough" into a pre-set number of sets in a way that these contents can be compared to all others easily. Various companies have made their own data sets using this dimensionality reduction technique. The researchers are carrying out investigation after investigations to resolve such type of problem using this technique.

Various methods for clustering of the data sets can be legitimate as data mining tool kit as it is required in each division or area of any company mostly. Most of the time, an industry maintains its database to collect various types of information, but these datasets are not valid for solution of each and every problems. So, this data can't be

assumed as universal datasets for all the problems. There are many reasons or parameters for existence of such type of problems.

Even, the most of the users are professional sensitive in nature and entered the data very accurately. But, due to different understandings and/ or mind sets, the people filled the same information into many ways. LinkedIn provides the text free service to the users to enter their information. So, the users filled the same information in thousands of ways. For example, we want to collect all the members of "Thapar University" as their company or educational Institute. In the data sets, some users will fill it as "Thapar University". But some will filled it as "T. U." or "Thapar University, Patiala" or "Thapar Institute of Engineering & Technology" or "Thapar Institute of Engg. & Tech." or "Thapar Institute of Engineering & Technology, Patiala" or "Thapar Institute of Engg. & Tech., Patiala" or "T. I. E. T., Patiala" or "T. I. E. T." and many more. So, to analyze the data sets we have to manage the data sets with considering various mind sets. First of all this dataset is normalized so that it can be prepared for analysis then efficient clustering technique can be applied on that. Some of the main clustering techniques are described below:

2.3.1.1 Greedy Clustering

In Greedy clustering, a nested loop iterates over whole dataset and groups them together according to a threshold value defined by the similarity metric such as Jaccard Distance. If the Jaccard distance between two job titles is "close enough", they are greedily grouped together [22].

2.3.1.2 Hierarchical Clustering

It is a deterministic clustering technique in which it computes distances between all the items, stores them in a matrix, then traverses the whole matrix and clusters items that has a minimum threshold distance defined by the parameter such as Jaccard Distance. It is considered as hierarchical because it traverses over the whole matrix and clusters items together which creates a tree based on the relative distances between different items [23]. This is also called agglomerative clustering because it creates a tree by arranging each

data element into several clusters, which hierarchically conflate into other top level clusters till the whole set of data clusters at the root of the tree. The leaves of the tree represent the data elements which are being clustered, whereas intermediate nodes hierarchically agglomerate these data items into clusters.

Y. Zhao and G. Karypis [51] evaluated different agglomerative and partition approaches for hierarchical clustering. Their experimental assessment showed that partition algorithms always provide better clustering results than agglomerative clustering algorithms, which suggests that partition clustering approaches are well-suited for clustering very large social document datasets due to not only their comparatively small computational requirements, but also analogous or even better performance of clustering. They presented a new category of clustering techniques called 'constrained agglomerative clustering algorithms' that combine the features of both agglomerative and partition algorithms. Their experimental results showed that every time their results lead to better hierarchical solutions for clustering than partition or agglomerative algorithms alone.

2.3.1.3 K-Means Clustering

Hierarchical clustering is an expensive technique having time complexity of the order O (n3), whereas K-means clustering normally executes with a time complexity of the order O (k*n). Even if the value of k is very small, the savings are very substantial. The savings in performance comes at the expense of results that are approximate, but they still have the potential to be quite efficient. The idea is that a multidimensional space containing n points is clustered into k clusters. Although K-means clustering can be run in two dimensional or two thousand dimensional space, the frequently used range is somewhere between tens of dimensions [22] [24]. When the dimension of the space is relatively small, this clustering technique can be effective because it runs fairly quickly and produces reasonable results. However we have to choose an appropriate value of k, but this is not always possible [25]. In K-means clustering, k points are picked randomly in the whole data space as initial seeds which are used to compute k clusters: K_1, K_2, K_3, ... K_k. Then each of n points present in data set is assigned to a cluster by searching for the nearest K_n. Then centroid of each cluster is calculated and K_i value of that cluster is

reassigned to its centroid value. This procedure is repeated until the saturation comes and members of clusters become stable between iterations. The name "K-means" is derived from the fact that K means are calculated in each iteration.

J. A. Hartigan and M. A. Wong [53] presented the efficient K-means clustering algorithm. The main goal of K-means clustering algorithm is to divide M points in n dimensional space into k number of clusters the within cluster sum of square error is minimized.

C. Ordonez [54] presented three different variants of the K-means clustering algorithm to cluster the binary data streams. The variants proposed by them include on-line K-means, incremental K-means and scalable K-means; a proposed variant is also introduced by them that find very high quality solutions in less time. Higher quality of clustering solutions was obtained with a mean-based initialization method and incremental learning. This speedup is achieved by them through a basic set of sufficient operations and statistics with sparse matrices. They compared various K-means variants with respect to speed and quality of results. The algorithms proposed by them can be used to monitor various transactions.

2.3.2 Classification

Classification can be defined as predicting a certain output based on a given input. To predict the output, the classification algorithm processes the training dataset containing a set of input attributes and the respective output, usually called class/goal attribute. The classification algorithm tries to discover relations between the input attributes that would make it possible to predict the output. After that the classification algorithm is provided an unknown data set, called prediction dataset, containing the same set of input attributes, except for the class/prediction attribute – not yet known. The classification algorithm analyses the input attributes and predicts the outcome. The prediction accuracy of the algorithm can be defined as how good the algorithm is. For example, consider a medical database training set that would have relevant information about patients recorded previously, where the output/prediction attribute is whether the patient had a heart disease or not. Table 2.1 and 2.2 given below illustrates the training dataset and prediction datasets of such a medical database.

Table 2.1 Sample Training set for Medical Dataset

Age	Heart Rate	Blood Pressure	Heart Problem
66	76	151/71	Yes
37	82	113/77	No
72	68	109/66	No

Table 2.2. Sample Prediction set for Medical Dataset

Age	Heart Rate	Blood Pressure	Heart Problem
42	99	148/70	?
45	59	111/75	?
84	79	110/69	?

There are various types of knowledge representation exist in our literature; among them classification normally utilizes decision rules to express the knowledge. These Prediction rules are represented in the form of "IF-THEN" rules, where the IF part (antecedent) consists of a AND (conjunction) of conditions and the rule's THEN part (consequent) predicts a certain output/class attribute value for an item which satisfies its antecedent condition. If we consider the example given above, a rule which predicts the first row in the given training dataset may be expressed as following:

"IF (Age=66 AND Heart rate>75) OR (Age>62 AND Blood pressure>141/70) THEN Heart problem=yes".

In most cases the decision rule is immensely bigger than the example given above. Conjunction has a good property for problem of classification; each condition divided by OR's defines quite smaller rules that captures relationship between various attributes. Satisfying any rule from these smaller rules set means that the consequent part is the prediction of outcome. Each smaller rule is built with AND's which makes possible narrowing down relationship between various attributes. How well output predictions are done is calculated as the percentage of various predictions hit against the total predictions

performed. A decent rule must to have a hit rate more than the happening of the output attribute. Thus we can say, if the classification algorithm is looking to predict rain in a season and it actually rains 80% of that season, the classification algorithm could have a prediction hit rate of 80% by just predicting the rain in all the season. The optimal solution of classification algorithm is a rule that gives 100% prediction hit rate that is very difficult to achieve, not impossible. Therefore, except for few specific problems, classification problem can only be answered by approximation techniques. There are different classification techniques, such as: -

- Statistical classification is a technique in which the individual items are grouped together on the basis of some quantitative information of the characteristics which are inherent in the items present in the dataset (referred to as characters, variables, etc.) and based on a provided training set of items which is previously labelled. Some of statistical algorithms are least mean square quadratic, linear discriminant analysis, kernel and the k nearest neighbours.
- A decision tree algorithm produces a set of decisions rule/conditions organized in a hierarchical manner. This is a predictive model that classifies an instance by following the pathway of satisfied conditions starting from the root of decision tree until the leaf is reached, which will correspond to an output class label. A decision tree produced by classification algorithm can be converted into a set of classification rules easily. Some of decision tree algorithms are CART and C4.5.
- Rule Induction is a part of machine learning techniques in which IF-THEN decision making rules are constructed from a set of interpretation. This algorithm can be thought as heuristic state space search. In this algorithm, a state which represents a candidate rule and operators represents specialization and generalization operations that convert one candidate rule into another. Some rule induction algorithms are AprioriC, CN2, Supervised Inductive Algorithm (SIA), XCS, a Grammar-based genetic programming algorithm (GGP) and a genetic algorithm using real-valued genes (Corcoran).
- Fuzzy rule induction algorithm uses fuzzy logic in order to understand the underlying dataset linguistically. To define a fuzzy system fully, fuzzy partitions and a rule base (structure) have to be designed (parameters) for all the variables.

Some common fuzzy rule learning algorithms are MaxLogitBoost, LogitBoost, Grammar based genetic Programming (GP), AdaBoost, a hybrid Simulated Annealing/genetic Programming algorithm (SAP), a hybrid Grammar based genetic Programming/genetic Algorithm method (GAP) and an adaptation of the Wang Mendel algorithm (Chi).

- Neural Networks could also be used for rule induction classification algorithm. A neural network (parallel distributed processing network) is a computing procedure that is loosely structured after cortical structures present in the brain. It consists of various interconnected processing nodes or neurons which work together and fabricate an output function. Some common examples of neural network algorithms are a radial basis function neural network (RBFN), multilayer perceptron, a hybrid Genetic Algorithm Neural Network (GANN), incremental RBFN, decremental RBFN and Neural Network Evolutionary Programming (NNEP).

Another classification procedure can be considered as if we are given a particular set of objects, in which each object belongs to a previously known class, and each of which possesses a previously known vector of variables and our aim is to design a prediction rule which will helps in assigning future objects to a particular class, having only vectors of present variables that describe the future objects. This kind of problems is called supervised classification problems that are ubiquitous and a lot of procedures for constructing such types of rules have been developed so far. One very important procedure among this is the naive Bayes classification method, also called independence Bayes algorithm [26]. This method is important because it is easy to construct, not requires any complex iterative parameter evaluation schemes. This means it can be easily applied to the large data sets.

2.4 Sentiment Analysis or Opinion Mining

Sentiment analysis or opinion mining is an emerging area of natural language processing with lot of research areas ranging from learning the polarity of phrases and words to document level classification. B. Pang et al. [32] done sentiment analysis using document level classification and A. Esuli et al. [34] and V. Hatzivassiloglou et al. [33] used the

approach of learning polarity of words for twitter sentiment analysis. H. Yu and V. Hatzivassiloglou [35] presented a work in which they concluded that due to limitation of the size of tweet, sentiment classification of the tweets is most similar to the sentence level sentiment classification. However the specialized and informal language used in the tweets as well as the varying nature of these microblogging domains, made the opinion mining from twitter a very different task. It is an open problem that well the techniques and features which are used previously on the well-formed data will transfer to this microblogging domain. Just in the last some years, there have been a lot of research works done in the field of twitter sentiment analysis. Some researchers have begun to utilize the part-of-speech features but their results are still remain fixed.

Researches have also begun to find out various techniques that collects training data automatically. Several researches depends on emoticons for preparing training datasets, such as A. Bifet and E. Frank [36] and A. Pak and P. Paroubek [37] utilized various existing twitter sentiment websites for collecting their training dataset. L. Barbosa and J. Feng [38] have also exploited the lot of existing twitter sentiment websites for collecting their training dataset. D. Davidov et al. [39] made use of hashtags for building their training dataset but they performed only 2-way polarity classification not the 3-way polarity classification.

R. Probowo et al. [17] combined the rule based sentiment classification and supervised machine learning approach into a new hybrid method and tested this combined approach on product reviews, movies reviews and various Myspace comments. They deduce that this hybrid sentiment classification model can improve the effectiveness of classification. They also proposed a complimentary approach that is a semi-automatic classification approach in which every classifier contribute to other to achieve a better level of effectiveness.

Kouloumpis E. et al. [29] investigated the utility of various linguistic features predicting the opinion of twitter messages. They have evaluated the usefulness of various existing lexical resources as well as those features that capture the information about the creative and informal language used in these microblogging websites. They took a supervised approach to solve the problem.

Y. Choi. Et al. [40] focused on extracting the various sources of sentiment instead of carrying out an opinion extraction or sentiment classification, for example the organizations or individual that plays a very important role in influencing the opinion of other individual. A lot of data sources other than twitter have been used for opinion mining such as product reviews, the corpus of DUC (Document understanding Conference), customer feedback, Wall Street Journal Corpus etc.

To automate the process of opinion mining, various approaches have been applied so that the sentiment or opinion of expression, words, sentence or documents can be predicted. K. Hiroshi et al. [41] done this task with natural language processing and some pattern based machine learning procedures such as Naïve Bayes (NB), Support Vector Machines (SVM), Maximum Entropy and unsupervised learning etc.

Go et al. [43] used distant learning approach to obtain sentiment data. They use a lot of tweets that ends in positive emoticons such as ":)" ":-)" as positive and various negative emoticons such as ":(" ":-(" as negative. They build sentiment analysis models using Naive Bayes classifier, MaxEnt classifier and Support Vector Machines (SVM), and they concluded that SVM outperforms all other classifiers. In terms of feature space, they tried a Unigram model, Bigram model in conjunction with POS features. They concluded that the unigram model outperforms all other classifier models.

A. Younus et al. [28] took up a study of lot of social web engagement patterns of various users from this developing world through a deep study of Twitter's role during the current Tunisian uprising. They are motivated by the results which comes from a user survey which is conducted mainly for various users from this developing world who tweeted in bulk during uprisings in the Arab world, they proposed a novel based approach for subjectivity analysis of various tweets which corresponds to political events in this developing world. The method proposed by them differs from the existing subjectivity analysis approaches because this method takes into account various social features of social web platforms for this subjectivity classification task. Through lot of experimental evaluations, they observed the accuracy of the proposed approach to be 83.3% which demonstrates a quite promising outcome for very large-scale application of their proposed subjectivity analysis method.

Chapter 3 Research Problem

3.1 Problem Statement

Social media has become very popular communication tool among internet users in the recent years. Humans have an ingrained tendency to share their ideas, experiences and knowledge, which associate them with the rest of the world, so that they can be recognized and can also identify their importance and worth. They are eager to know about happenings around them, that is why they communicate in order to share their ideas, observations and queries. Social media is one such communication medium that made people to be heard and satisfy their curiosity to know about rest of the world. The rapidly escalating coverage of the social media user bases and user engagement provide the social media an immense potential to study the logic under the users social behavior and organization, and finally to transfer this knowledge to real time business profit.

Social media is one of the important sources of real time big data. A large unstructured data is available for analysis on the social web. The data available on these sites have redundancies as users are free to enter the data according to their knowledge and interest. This data needs to be normalized before doing any analysis due to the presence of various redundancies in it. Analyzing these huge social datasets and predicting the opinions of individuals plays an important role in business and academics. By mining this social data, a lot of useful information about our social contacts can be analysed and visualized, which cannot be discovered manually and this extracted knowledge can be used in various decision making purposes.

In social networking sites, in addition to friendship relationships, individuals can have lot of family ties with other individuals, join several social groups, be associated with a variety of organizations, or belong to diverse geographical sub-networks. By considering the complex composition of these social networks rather than restraining the analysis to a single relationship type or entity, we can build lot of richer models that endow with better indulgent of the dynamics in these websites and thus help in providing extremely accurate predictions about upcoming events. These days, the main focus of analysts is to observe the association and structure of various social networking sites. To study and analyze compactness, centrality and grouping with the promise of security and

privacy of the users data in these networks is the main research area in social web analytics. Mining and analyzing the social web helps in identifying various trends of the society and perspective of society towards different products, events, political party etc. Various research issues associated with social web analytics are predicting future trends, community detection, link prediction, recommender system, frequency analysis, mood analysis, opinion mining, influence propagation etc. Use of efficient data mining techniques, machine learning approaches, various statistical tools and network analysis concepts made the social data analysis, a challenging and more effective goal.

3.2 Research Gaps

A lot of research is being carried out in the area of social web analytics. The social networks are created and knockdown by studying daily basis and continuous communication among users. These studies incorporate with individual's relationships with others in terms of different parameters of human behavior. In this way, these networks are proven an important and critical tool to study the communal phenomenon, composition, relationships and behaviors of individuals and/ or group. The preprocessing steps such as normalization of data and similarity measurement are carried out to make the social data set suitable for analysis. A lot of research on data clustering has been proposed but there is lack of effective study to analyze and utilize the professional behavior of human beings by using various clustering techniques on the data present on different professional communities available online such as LinkedIn. Every social dataset contains a lot of spatial information about the location of users and location from where they send messages or post various statuses updates. A very few research work in the area of social data analytics is carried out which uses this spatial data present in the social datasets. This spatial data can lead to one important analysis and visualization concept that is geo-clustering. Frequency analysis and effective visualization of the analyzed data is another main concern in the area of social web analytics such as how many likes, posts, retweets are there for particular statuses updates, posts and tweets respectively. Our research work will focus on the use of clustering techniques for detection of various online professional communities according to particular job/profession, location or company. Moreover the research will focus on the use of

classification techniques for doing frequency analysis and opinion mining/sentiment analysis using the extracted and preprocessed datasets of tweets.

3.3 Research Objectives

In the light of above discussed research gaps following objectives have been formulated.

- To study various research issues associated with social web analysis and efficient data mining techniques for analyzing and visualizing the social data.
- To remove different redundancies present in the extracted social dataset using normalization to make this dataset suitable for further processing.
- To develop a social web analytic application for analysis and visualization of professional's LinkedIn data using different clustering techniques.
- To develop a geo-clustering application using spatial data present in the LinkedIn dataset.
- To perform basic frequency analysis, retweet analysis and sentiment analysis on Twitter data.

3.4 Research Methodology

In our research work we will python language, which is an excellent scripting language for manipulating text. The dataset required for analysis will be extracted using LinkedIn REST APIs and Twitter REST APIs. Twitter uses OAuth 1.0a authorization mechanism and LinkedIn uses OAuth 2.0 authorization mechanism to provide its data for development purposes. Python "LinkedIn" package and authorization credentials (OAuth 2.0) obtained by registering for an application on LinkedIn development environment, will be used to explore the LinkedIn REST API. Similarly Python "twitter" package and authorization credentials (OAuth 1.0a) obtained by registering for an application on Twitter development environment, will be used to explore the Twitter REST API. We will use python natural language toolkit (NLTK) module for analyzing our social dataset, which is particularly good for text analytics. For normalization of spatial data present in our Social dataset, an API key from Microsoft Bing web service will be requested to find out the geo-coordinates of every location. Various clustering techniques such as greedy, hierarchical and K-means clustering will be used to detect various online professional

communities according to particular job/profession, company or location. Sentiment analysis on the tweets will be performed using Alchemy API. Alchemy API provides convenient mechanisms to identify positive, negative and neutral sentiment within a web page or document. Python “matplotlib” package and R language will be used for visualization of analyzed data.

Chapter 4 Analysis and Visualization of Social Data

For analysis and visualization of social web data, two most common social networking sites are used in this research. One is LinkedIn (The professional social web) and other is a microblogging web service, Twitter.

LinkedIn data is extracted by using LinkedIn API and normalized by removing redundancies to make this dataset suitable for analysis. Further, data is also normalized according to locations of LinkedIn connections using geo coordinates provided by Microsoft Bing. Then, clustering of this normalized data set is done according to job title, company names and geographic locations using Greedy, Hierarchical and K-Means clustering algorithms and clusters are visualized to have a better insight into them. So by analyzing LinkedIn dataset, various types of professional communities can be detected such as communities on the basis of job title, companies and geographic locations.

As second objective, tweets about "AAM AADMI PARTY" (a recently grown political party) are extracted and a development environment is developed using python to induce a lot of analytical insights from these tweets using frequency analysis and sentiment analysis. Different persons have different opinions about a political party; it can be either a criticism or an appraisal. Analyzing combined view of all the persons about a political party helps in building an overall image of the party out of these individual opinions.

4.1 Analysis and Visualization of Professional's LinkedIn Data

LinkedIn is an online social networking site concentrated on professional and business associations. At the first look, LinkedIn is also looks like any other social website, but in reality, it is pretty dissimilar in its API and data with others. The main motive of the people who sign up this social network is predominantly concerned in the commerce prospect of their business details like business relationships, previous employment etc.

LinkedIn social network provides a large dataset of the millions of its subscribers. The conventional database management techniques are not able to manage or handle this huge datasets. The datasets, available before 2000, are very simple and in a few amount. Now, the number of subscribers has increase in multiple of thousands. This makes the

datasets very complex. To store, compute and handle these datasets, various powerful tools and techniques have proposed by the researchers. As per the advancement in the technology, various data handling tools have been developed. The research is going on for searching and developing such type of tools to handle and provide the sensitive information for different purposes continuously in a drastically manner. LinkedIn is the biggest source of the data where the subscribers provide their information themselves to get the benefits of the various services.

LinkedIn is a predominantly appealing and exciting objective, specifying the proficient environment of its subscribers. By investigating LinkedIn network information and data, one can understand the behaviour and other properties of individuals and group. By collecting all the LinkedIn data i.e. occupation, employer, designation, locality etc., personality of the users can be exactly visualized. At first glance, LinkedIn appears like other social website, but, due to the responsive and professional characteristics, LinkedIn's data are quite different from data available at other social networking sites and have its own senses. By analyzing LinkedIn dataset, various types of professional communities can be detected such as communities on the basis of job title, companies and geographic locations.

4.1.1 Implementation Methodology

The implementation methodology is described in Figure 4.1. In the first step, the data set in JSON format is obtained from LinkedIn REST API. To make data useful for further analysis, Normalization of data set is done in the second step. In third Step, similarity measurement and frequency analysis of job titles and company names are carried out. Clustering algorithms such as Greedy Clustering, Hierarchical Clustering and K-Means Clustering for clustering the data set according to different similarity criteria such as job title, company names and geographic coordinates are programmed in step 4. Finally, clusters are obtained and visualization of clusters for analysis is done. Geographic clusters obtained are visualized on Google Map [56].

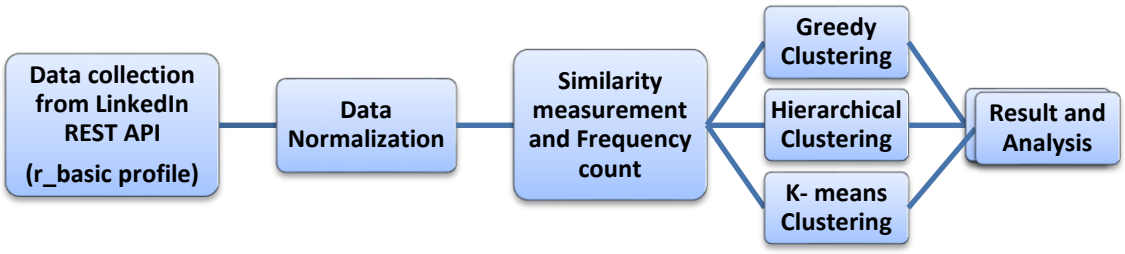


Fig 4.1. Implementation Methodology for Analyzing the LinkedIn Data

Various packages in Python are available for analysis and visualization of LinkedIn data. These Python packages can be easily downloaded and installed using *pip*. Table 4.1 shows various python packages used here in this development environment and their purpose.

Table 4.1 Python Packages used for Analysis and Visualization of LinkedIn Data

Python Packages	Purpose
Linkedin	Python wrapper that wraps the LinkedIn API.
Json	Encode and decode tweets into JSON format.
Prettytable	Nicely displaying tabular dataset in ASCII table format.
Csv	Implements various classes to write and read tabular data in CSV format.
Collection	Implements specialized container data types.
Geopy	Python client for several popular geocoding web services.
Re	Provides regular expression matching procedures to search for patterns.
Nltk	Python package used for NLP (natural language processing).
HTTPError	This is useful when handling exotic HTTP errors, such as requests for authentication.
Cluster	Allows creating several groups (clusters) of objects from a list and implements various clustering algorithms.
xml.dom.minidom	Provides implementation of the Document Object Model interface
IFrame	Demonstrate the use of iframes and nbviewer to embed IPython notebooks in WordPress.
Display	Various display related classes are present in this module.

4.1.2 Making LinkedIn API Requests and Data Extraction

First of all, we have to register on LinkedIn application developer platform through an application form using the link https://www.linkedin.com/secure/developer; and generate the details of our application's API key, Secret Key, OAuth User Token, and OAuth User

Secret identifications which will be used to access the API programmatically. Figure 4.2 display various authorization keys which we got after filling the application form at LinkedIn developer platform.

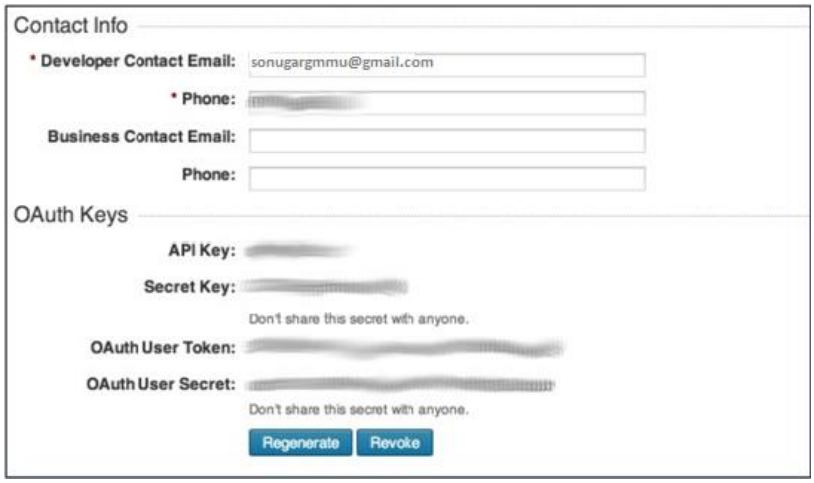


Fig. 4.2 Getting Access to LinkedIn API

Now, the python-LinkedIn package is installed by running the command **pip install python-linkedin** in a terminal. Then all the LinkedIn credentials obtained from LinkedIn developer platform are used to ultimately create an instance of the class "LinkedInApplication" so that our LinkedIn account data can be accessed. Then LinkedIn basic profile information are obtained, which includes the name of the account holder and headline by using LinkedIn REST API as shown in following code snippet:

```
Auth = linkedin.LinkedInDeveloperAuthentication (CONSUMER_KEY,
                                                 CONSUMER_SECRET,
                                                 USER_TOKEN,
                                                 USER_SECRET,
                                                 RETURN_URL,
                        Permissions =linkedin.PERMISSIONS.enums.values ( ))
app = linkedin.LinkedInApplication(auth)
app.get_profile ( )
```

Basically, there are two APIs viz. Connections API and the Search API for accessing the data at LinkedIn. The former API provides a directory of our connections

so that information of the profile can be obtained through jumping-off point. The second API endows with a questionnaire of basic information like name, present and previous job, employers, and locations etc. for people who are subscribers of LinkedIn. The data obtained from LinkedIn is very important and remarkable which contains a full present and past history of individuals.

After that, all the profile information of our LinkedIn connections is retrieved by utilizing the 'app' which is an instance of LinkedIn Application. After collecting all the information, this data is downloaded in to a file so that any other unwanted API requests can be avoided. The downloaded file of all our connections information is a saved in JSON format. This file 'connections_data.json' contains information of our LinkedIn connections in JSON format as shown in Figure 4.3.


```
{
 "_total": 126,
 "_count": 125,
 "_start": 0,
 "values": [
  {
   "firstName": "Lovedeep",
   "headline": "Corporate HR at Suffescom Solutions Pvt. Ltd.",
   "lastName": "Behl",
   "industry": "Human Resources",
   "siteStandardProfileRequest": {
    "url": "
    https://www.linkedin.com/profile/view?id=265836582&authType=name&au
    thToken=o5PI&trk=api*a3926274*s3995014*"
   },
   "pictureUrl": "
   https://media.licdn.com/mpr/mprx/0_imOA6V7cs7nRkGRMSHfC6ZfMsSisXTIM3
   7om6ZaQ5mcERLpJ7eaje4seNL_5bQovGSpfd0rV3rAs",
   "location": {
    "country": {
     "code": "in"
    },
    "name": "Chandigarh Area, India"
   },
   "apiStandardProfileRequest": {
    "url": "https://api.linkedin.com/v1/people/syDbv6hnuW",
    "headers": {
     "_total": 1,
     "values": [
      {
       "name": "x-li-auth-token",
       "value": "name:o5PI"
      }
```


Fig. 4.3. A Snippet from 'linkedin_connection.json' File

Now, desired information can be retrieved from the downloaded data and formulated into a table format using the python **prettytable** package. Figure 4.4 shows a similar data in a table format. In this Figure, the various subscribers and their location are shown. The complete profile information of the subscribers connected to the network can be retrieve in the table format by using field selectors as drowned in the Profile Fields online documentation, constraint to the availability. At this instant, the professional identity with designation at present and past history for our profile and the entire subscriber's profile connected to our network is retrieved with the help of this code snippet:

```
my_positions = app.get_profile (selectors = ['positions'])
print json.dumps (my_positions, indent = 1)
connection_id = connections ['values'] [0] ['id']
connection_positions=app.get_profile (member_id = connection_id, selectors = ['positions'])
Print json.dumps (connection_positions, indent = 1)
```

Name	Location
Lovedeep Behl	Chandigarh Area, India
Harsha Mehta	Chandigarh Area, India
Ankit Mahato	Bengaluru Area, India
Gautam Anand	Singapore
Shivam Goyal	Hyderabad Area, India
Animesh Swain	Mumbai Area, India
Sourabha Shakya	Noida Area, India
FatehHR Consultancy	Chandigarh Area, India
Rahul Juneja	Gurgaon, India
Shubham thakur	Karnal Area, India
kulbir singh	India
Nitansh bareja	Gurgaon, India
Priya Anand	Leicester, United Kingdom
Rohit Thakur	Gurgaon, India
Anjali Magoo	New Delhi Area, India
InteliGenes Technologies	New Delhi Area, India
Ravi Lathar	Pune Area, India
Siddhant Sahi	Noida Area, India
ANKIT TIWARI (TechSupport PPC Consultant)	India

Fig. 4 4. Prettytable showing the Names and Location of the Subscribers

Sample output of the above code snippet reveals a number of interesting details about job position of every connection, including the company name, industry, summary

of efforts, and employment dates as shown in Figure 4.5. The subscribers fill their data at LinkedIn according to their requirements, interest and potential. So the information downloaded from LinkedIn will not be in the same format. It is not necessary that all the fields are filled according to our requirements. Even some columns will be empty or filled with needless information. Also, there can be more information in some subscriber's profile which is not required. To make the available information useful i.e. according to our requirements, we can use **field selector syntax** and convert in the desired format without requesting more API calls. For example, we have selected the name, company and id details for industries to fill up the positions with qualified professionals as shown in the following code:

```
my_positions = app.get_profile (selectors = ['positions: (company: (name, industry, id))'])
```


```
{
 "positions": {
  "_total": 1,
  "values": [
   {
    "startDate": {
     "year": 2014,
     "month": 1
    },
    "title": "Jr.Developer",
    "company": {
     "id": 14096,
     "name": "Net Solutions"
    },
    "summary": "A focused, ambitious and innovative Software Professional with 1
+ years of experience of working with various Python and Microsoft technologies.
\nExperience with DJANGO, Web.py , python ,OOPS, C#, ASP.NET 2.0, 3.5, 4.0, jQue
ry, MVC, SQL Server 2005 and 2008.\nExperience with handing project Risks\nExper
ience on Project Delivery and implementation\nExperience with N-tier Architectur
e.\nLooking for a challenging position where I will best utilize my skills.",
    "isCurrent": true,
    "id": 599205123
   }
  ]
 }
}
```


Fig. 4.5. Job Position History for My Profile

4.1.3 Downloading LinkedIn Connections as a CSV File

With the help of Linkedin REST API, all the information of the subscribers can be accessed and retrieved after creating authenticated account at http**://**linkedin**.**com**.** All the information like position, designation, employer etc. can also be stored by exporting our LinkedIn connections into the address book in a CSV file format. For this purpose, first go to the Connections menu and select connections menu items to steer to our LinkedIn connections page. Now, select the "Export connections" link. However, the LinkedIn Connections dialog can also be exported directly as shown in Figure 4.6.

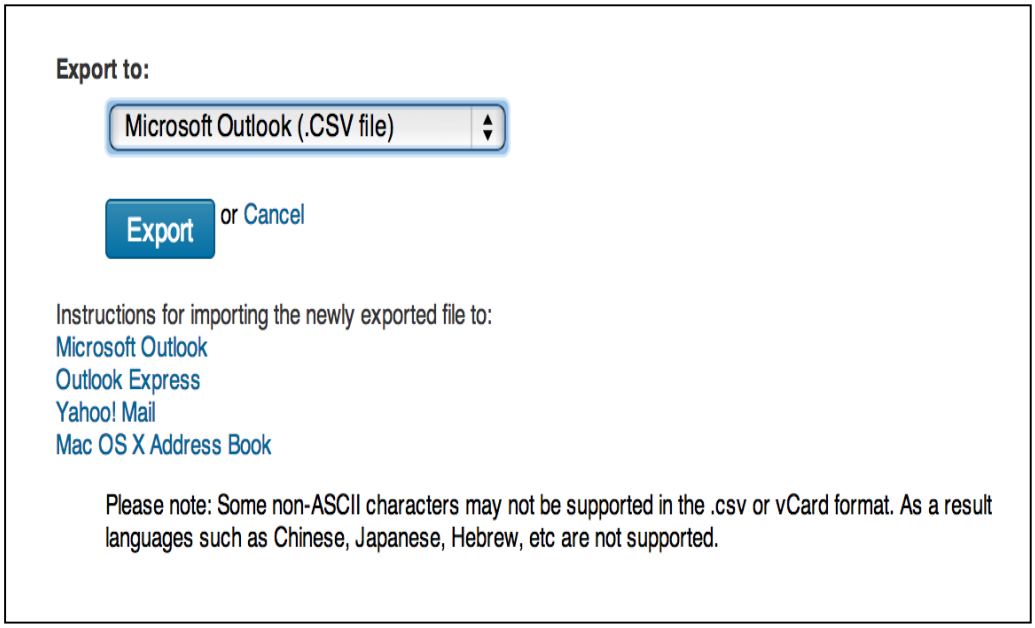


Fig. 4.6. Getting LinkedIn connections in a CSV File

4.1.4 Extracted Dataset and its Features

The dataset in JSON and CSV format is obtained from LinkedIn API by exploring the API to get application's API key, Secret Key, OAuth User Token, and OAuth User Secret identifications [11]. Table 4.2 shows various features present in the dataset obtained from LinkedIn API and corresponding description of each field.

Table 4.2. Description of Features

Features	Information
First name	First name of Individual
Headline	Specifying position in particular company
Last name	Last name of the Individual
Industry	Industry detail of the Individual
Location name	Location details of work area
Country	Country code
Picture URL	URL of the profile picture
ID	Unique id provided by LinkedIn

4.1.5 Clustering the LinkedIn Connections

Clustering is an unsupervised machine-learning method which is used to fastener in every data mining toolkits. In this, first all the information is collected and then partitioned into a number of small groups (named as clusters) based on their properties which are required to evaluate the information data of the anthology. For an instance, if we want to relate the information of the people according to their geographic relocation, we have to design or cluster the information into groups of our LinkedIn connections geographic region wise. This clustering analysis of LinkedIn or any other social network site can be done into three main steps viz. Data Normalization, Similarity computation and Dimensionality Reduction. Various methods for clustering of the data sets can be legitimate as data mining tool kit as it is required in each division or area of any company mostly. Most of the time, an industry maintains its database to collect various types of information, but these datasets are not valid for solution of each and every problems. So, this data can not be assumed as universal datasets for all the problems. There are many reasons or parameters for existence of such type of problems. For example, there can be lacking in the design of the application's user interface, there can be some columns which are unfilled or over filled by the users, users filled the data into abbreviations, misspelled the information, users put up different information as per their potential etc. In this way, the data sets become more complex to analyze. Even, the most of the users are professional sensitive in nature and enter the data very accurately. But, due to different understandings and/ or mind sets, the people filled the same information into many ways. LinkedIn provides the text free service to the users to enter their information. So, the

users filled the same information in thousands of ways. For example, we want to collect all the members of “Thapar University” as their company or educational Institute. In the data sets, some users will fill it as it is i. e. “Thapar University”. But some will fill it as “T. U.” or “Thapar University, Patiala” or “Thapar Institute of Engineering & Technology” or “Thapar Institute of Engg. & Tech.” or “Thapar Institute of Engineering & Technology, Patiala” or “Thapar Institute of Engg. & Tech., Patiala” or “T. I. E. T., Patiala” or “T. I. E. T.” and many more. So, to analyze the data sets we have to normalize the data sets with considering various mind sets.

To analyze the data sets in an effective manner, dataset have to be normalized efficiently. The main objective of such type of data sets is to compute them according to the similarities. The clustering is the best approach to solve such type of problems. The outcomes of the clustering depend upon the choice of clustering technique and its type. There are a number of clustering techniques available, but we have to adopt the technique which should be the best solution for our objective.

Now few real world data from LinkedIn is accumulated and analyzed using meaningful clusters to obtain a few more insights into the dynamics of our professional network. Whether we want to take an honest look at whether our networking skills have been helping us to meet the “right kinds of people,” we want to approach contacts who will most likely fit into a certain socioeconomic bracket with a particular kind of business inquiry or proposition, or we want to determine if there is a better place we could live or open a remote office to drum up business, there is bound to be something valuable in a professional network rich with high-quality data. Here, a few different clustering approaches are implemented by further considering the problem of grouping together job titles that are similar.

4.1.5.1 Normalizing Data before Analysis

A common pattern for normalizing company names and job titles will be implemented. Problem of disambiguating and geocoding geographic references from LinkedIn profile information will also be implemented. (In other words, we will attempt to convert labels from LinkedIn profiles such as “Chandigarh Area” to coordinates that can be plotted on a map.) The chief artifact of data normalization efforts is that we can count and

analyze important features of the data and enable advanced data mining techniques such as clustering. In the case of LinkedIn data, entities such as companies, job titles and geographic locations will be examined.

- **Normalizing and Counting Companies**: We can analyse the selected data contents stored in the CSV file of LinkedIn connection's contacts. Then, we can standardize and exhibit these data sets in a predominant fashion. Firstly, we select a set of information which can be converted in to number of sub sets of information. For example, for finding the company name "Google", we have to strip off general suffix like "Inc." or "LLC" or "LLP" and many more as given in the below code strip :

  transforms = [(', Inc.', ''), (', Inc', ''), (',LLC', ''), (', LLP', ''), (' LLC', ''), (' Inc.', '')]

- **Normalizing and Counting Job Titles:** In the next step, we want to analyse the data according to the job titles. Here also, various users have entered the data into a number of ways. To make an effective way of analysis, we have to manage this data into number of subsets. Same as per the company name, the job title can be also arranged in to a specific manner. The specified job titles are in common reference with little bit difference in the information. Some extent of this problem can be solved by defining all the well-known abbreviations like "C. E. O.", "T. P. O." etc. But, this is not the final solution of the problem because there are a number of titles and job specification which can not be defined effectively and surely for example manager, engineer and helper etc. are the common job titles which is not defined clearly the specific job titles. So, we have to define a solution which can further classify such job titles into useful information for the review purposes. To solve this type of problem, we have done firstly normalization of data and then we apply a fundamental frequency analysis for all such type of job titles to represent them as clusters.
- **Normalizing and Counting Locations:** In the next step, we want to analyze the data contents of the subscribers of LinkedIn according to their

geographical area. It is a difficult task to differentiate the users on the basis of their geographic references. This can be done in different ways. We can locate all the users' state wise and then district wise. But, differentiating them according to their exact location is quite difficult. Generally, LinkedIn user's data has the advantage that most of the users are belongs to metropolitan areas or well-known places. But, we can not assume that all the users are from the metropolitan's cities or well-known places. So, we have to find out another solution to compare or make a relationship between them according to their geographical location. A solution of this problem is to represent the users according to the geographical coordinates.

To find out graphical coordinates or to locate the users region wise, we use a Python package called "goopy" via pip install goopy; this helps in providing the information of all the users location wise with their geographical coordinates. This package will be used itself as a proxy available at the various current web service providers like Google, Yahoo, Bing and many more search engine websites using geocoding. For this purpose, we request an API and interface with any of the geocoding services to avoid the manual crafting. In this project, we have requested an API key with Bing web service.

4.1.5.2 Measuring Similarity

After normalizing the dataset, we proceed to the problem of computing similarity, which is the principal basis of clustering. The most substantive decision we need to make in clustering a set of strings (such as job titles) in a useful way in which underlying similarity metric to use. There are myriad string similarity metrics available, and choosing the one that is most appropriate for our situation largely depends on the nature of our objective. Here are a few of the common similarity metrics that might be helpful in comparing job titles that are implemented in NLTK such as Edit Distance, Jaccard Distance and n-gram similarity metrics [19]. Now a few different clustering approaches

are implemented by further considering the problem of grouping together job titles, company names and geographic location of the users that are similar.

- **Greedy Clustering:** Now we try to cluster job titles by comparing them to one another using Jaccard distance. We clusters similar titles and then display our contacts accordingly. We start by separating out combined titles using a list of common conjunctions and then normalize common titles. Then, a nested loop iterates over all of the titles and clusters them together according to a thresholded Jaccard similarity metric as defined by DISTANCE. If the distance between any two titles as determined by a similarity heuristic is "close enough," we greedily group them together. We often need to assign each cluster a meaningful label. The working implementation computes labels by taking the set wise intersection of terms in the job titles for each cluster. The types of results from this code are useful in that they group together individuals who are likely to share common responsibilities in their job duties. As previously noted, this information might be useful for a variety of reasons, Whether we are planning an event that includes a "CEO Panel" trying to figure out who can best help us to make the next career move, or trying to determine whether we are really well enough connected to other similar professionals given our own job responsibilities and future aspirations**.**

  The greedy clustering is an O (n^2) algorithm and what we really want to do is come up with an algorithm that is on the order of O ($k*n$), where k is much smaller than n. one variation that we might try is rewriting the nested loops so that a random sample is selected for the scoring function, which would effectively reduce it to O ($k*n$), if k were the sample size. As the value of the sample size approaches n, however, you would expect the runtime to begin approaching the O (n^2) runtime. Another approach you might consider is to randomly sample the data into n bins (where n is some number that is generally less than or equal to the square root of the number of items in your set), perform clustering within each of those individual bins, and then optionally merge the output.

- **Hierarchical Clustering:** Now, job titles are also clustered by using hierarchical clustering. An excellent implementation of both of approaches hierarchical and K-means clustering is available via the cluster module that can be installed via **pip install cluster**. Hierarchical clustering is a deterministic technique in that it computes the full matrix of distances between all items and then walks through the matrix clustering items that meet a minimum distance threshold. It is *hierarchical* in that walking over the matrix and clustering items together produces a tree structure that expresses the relative distances between items. This is also called *agglomerative* because it constructs a tree by arranging individual data items into clusters, which hierarchically merge into other clusters until the entire data set is clustered at the top of the tree. The leaf nodes on the tree represent the data items that are being clustered, while intermediate nodes in the tree hierarchically agglomerate these items into clusters. Individual peoples are leaves on the tree that are clustered, while nodes such as "Chief Technology Officer" agglomerate these leaves into a cluster. Here we draw two graphs of our connections after doing hierarchical clustering using "D3.js", one is dendogram and other is node link tree.
- **K-Means Clustering:** Whereas hierarchical clustering is a deterministic technique that exhausts the possibilities and is often an expensive computation on the order of O(*n*3), *k* means clustering generally executes on the order of O(*k***n*) times. For even small values of *k*, the savings are substantial. The savings in performance come at the expense of results that are approximate, but they still have the potential to be quite good. The idea is that you generally have a multidimensional space containing *n* points, which you cluster into *k* clusters. Now we use K-means clustering algorithm to cluster our professional LinkedIn network by using geographic locations as our similarity criteria and visualize the clusters by plotting them on Goggle Earth.

4.1.6 Visualizing Geographic Clusters with Google Earth

Now we use K-means to visualize and cluster our professional LinkedIn network by plotting it on Goggle Earth. We already the fetch location information through LinkedIn API that describes the major metropolitan area, such as "New Delhi Area," we will be able to geocode the locations into coordinates and emit them in an appropriate format (such as KML) that we can plot in a tool like Google Earth, which provides an interactive user experience. The primary things that we must do in order to convert our LinkedIn contacts to a format such as KML include parsing out the geographic location from each of our connection's profiles and constructing the KML for visualization such as Google Earth. The KMeansClustering class of the cluster package can calculate clusters for us, then we munge the data and clustering results into KML. We group our contacts by location, cluster them, and then use the results of the clustering algorithm to compute the centroid of each cluster.

4.2 Analysis and Visualization of Twitter Data

Twitter is a predominantly appealing and exciting objective for an analyst that specifies the proficient environment of its subscribers. By analyzing tweets, one can understand the behavior and opinion of individuals and groups [4]. Twitter can be defined as a free, global and high speed microblogging service that helps people to share their ideas with others by short and precise, 140 characters long messages, providing easy and rapid communication. More than 500 million curious users have already registered on twitter; most of them actively engage their curiosity on regular basis. It is a great source of social data as it is open for public consumption and it provides well documented and clean API. Tweets are particularly interesting as they happen at the "velocity of thoughts" and are accessible in the nearest real time.

Twitter provides a simple and asymmetric "Following Model" that satisfies human's curiosity. The following mechanism of twitter is asymmetric because it allows one user to follow other without mutual acceptance. Huge amount of social digital data is being produced on regular basis in the form of tweets. Analyzing user's opinions from this large social dataset leads to large set of possibilities in various fields such as business, marketing, advertising, politics, education etc. Different persons have different

opinions about a political party; it can be either a criticism or an appraisal [5]. Analyzing combined view of all the persons about a political party helps in building an overall image of the party out of these individual opinions. Figure 4.7 shows some of the tweets from different persons about "Aam Aadmi Party".

Fig. 4.7 Some Example of Tweets

Frequency analysis of various facts presents in a tweet structure, retweet analysis and sentiment analysis plays an important role in business analytics. Sentiment analysis or opinion mining can be defined as identifying and extracting the subjective information in the source materials with the use of natural language processing, computational linguistics and text analysis.

4.2.1 Implementation Methodology

Twitter uses OAuth 1.0a authorization mechanism to provide its data for development purposes. The implementation methodology is shown in Figure 4.8. In first step, Python twitter package and authorization credentials (OAuth 1.0a) obtained by registering for an application on Twitter development environment, are used to explore the Twitter REST API. After that the tweets about Aam Aadmi Party are extracted by using search query with parameter "Aam Aadmi Party". All the extracted tweets are stored in MongoDb (NoSQL database). Then in the next step, tweets are analyzed by performing frequency analysis, entities extraction and Retweet analysis etc. Further in next step, sentiment analysis on these tweets is performed using Alchemy API. Alchemy API provides convenient mechanisms to identify positive, negative and neutral sentiment within a web page or document. Finally, visualization of the analyzed tweets is done using R language.

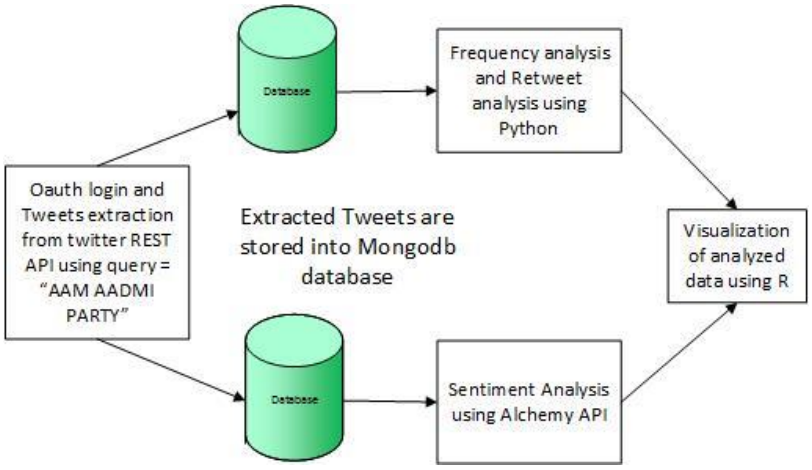


Fig 4.8. Implementation Methodology for Analyzing the Twitter Data

Various packages in Python are available for tweets analysis and visualization. These Python packages can be easily downloaded and installed using *pip*. Table 4.3 shows various python packages used here in this development environment and their purpose.

Table 4.3. Python Packages used for Analysis and Visualization of Tweets

Python Packages	**Purpose**
Twitter	Python wrapper that wraps the Twitter API.
Json	Encode and decode tweets into JSON format.
Counter	A dictionary subclass used for counting the objects which are hashable.
Prettytable	Nicely displaying tabular dataset in ASCII table format.
Pyplot	To produce 2D graphics suitable for publication purposes.
Pymongo	Helps in interacting with MongoDb from python.
Re	Provides regular expression matching procedures to search for patterns and strings.
URLlib2	Helps in opening URLs that may be extended by establishing custom protocol handlers.
Time	Deals with various time conversions between universal and arbitrary time zones.
Nltk	Python package used for NLP (natural language processing).
Numpy	Array processing for strings, numbers, objects and records.
Extractor	Simple python package used for keyword extraction.
AlchemyAPI	It provides access to AlchemyAPI from python for natural language processing and unstructured text analysis.

Table 4.4. Functions defined in Python with their Objective

Function Name	Objective
Oauth_Login()	Use python's twitter package and the OAuth 1.0a credentials to gain API access to twitter account without HTTP redirects.
Twitter_search()	Use the twitter search API to execute a custom search query.
Save_to_MongoDB(), Load_from_mongoDB()	These functions use MongoDB (document-oriented database) to store and access the dataset in a suitable JSON format.
Extract_Tweet_Entities()	Extracts the various tweet entities from the tweets such as #hashtags, *@username* mentions and URLs for analysis.
Find_Popular_Tweets()	Find out most popular tweets by analyzing the retweet_count field to determine how many times a particular tweet was retweeted.
Get_Retweet_Attribution()	Analyze tweet content by regular expression heuristics to check for the existence of various conventions such as "RT @AamAadmiParty" or "via @AamAadmiParty".
Handle_Twitter_Http_error()	Handles various Http error codes by providing abstract logic.
Analyze_Tweet_Content()	Uses simple statistics like average count of words per tweet and lexical diversity, to gain simple insight into tweets.
Analyze_Favorites()	Uses the GET favorites/list API to retrieve favorite tweets of a user and then apply various techniques to extract, detect and tally tweet entities to illustrate the content.
Get_Text_sentiment()	Uses Alchemy API to carry out sentiment analysis of the tweets.

4.2.2 Making LinkedIn API Requests and Data Extraction

Twitter has taken immense care to craft an elegantly straightforward RESTful API that is perceptive and easy to use. Even so, there are enormous libraries available to further diminish the work involved in making various API requests. A predominantly beautiful Python package that squashing the Twitter API and imitates the public API semantics nearly one-to-one is twitter. Like most additional Python packages, you can install it with pip by typing **pip install twitter** in a terminal.

Before making API requests to twitter, we will need to create an application at *https://dev.twitter.com/apps*. Creating an application is the ordinary way for developers to get API access and for Twitter to interact and monitor with third party platform developers as needed. The process for creating a developer application is pretty normal, and all that is needed is read only access to the Twitter API.

Tweets are extracted from Twitter API in JSON format after getting application's API key, Secret Key, OAuth User Token, and OAuth User Secret identifications [12]. A particularly amazing python package "twitter" is used here for extraction of tweets that encases the Twitter API and imitates the public Twitter API semantics. Figure 4.9 shows the creation of new twitter application and getting the OAuth credentials and API access.

Then we authorize our application to access twitter account data by using the following code snippet:

```
auth = twitter.oauth.OAuth (OAUTH_TOKEN,
                            OAUTH_TOKEN_SECRET,
                            CONSUMER_KEY,
                            CONSUMER_SECRET)

twitter_api = twitter.Twitter (auth = auth)

print twitter_api
```

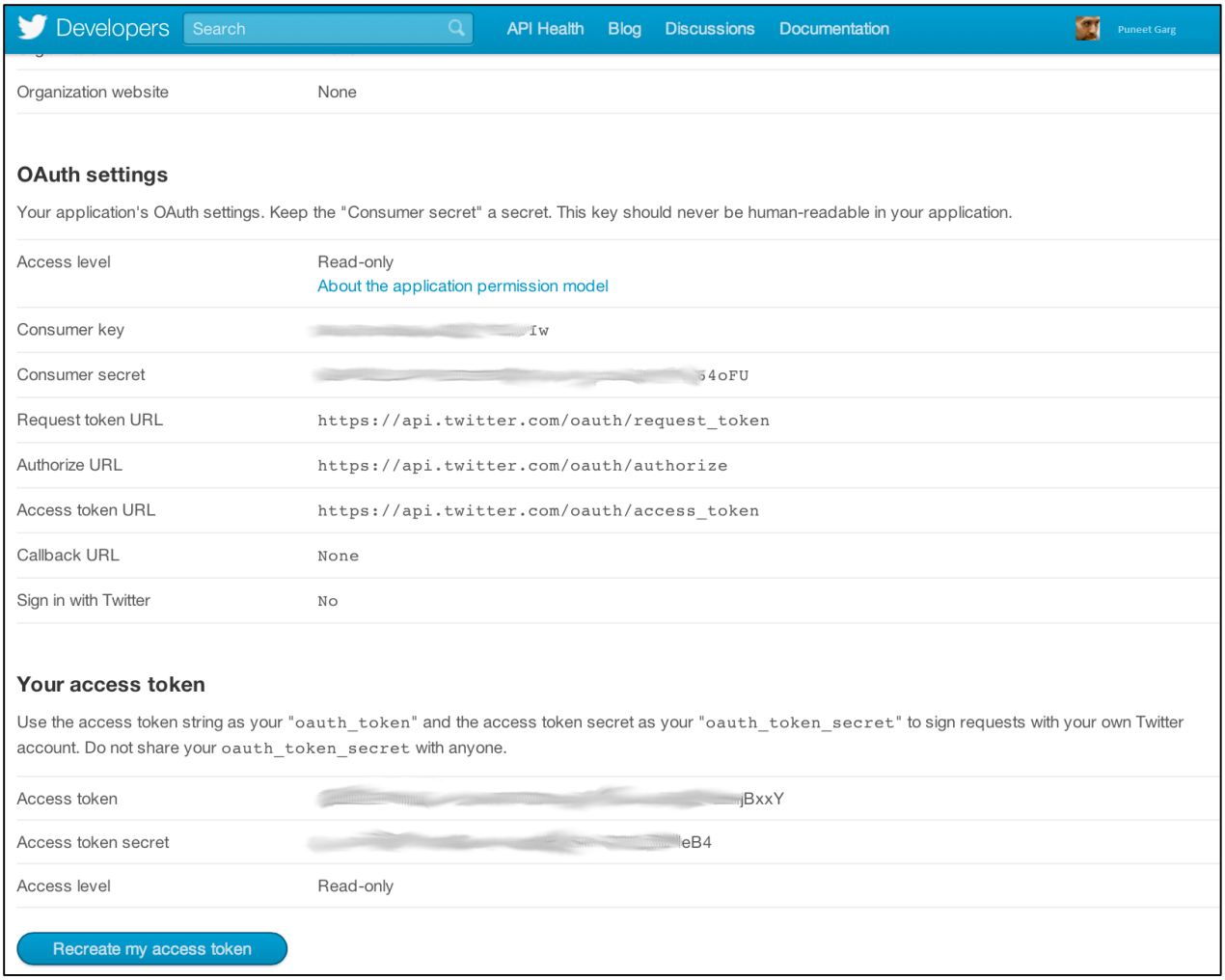


Fig. 4.9. Getting OAuth Credentials and API Access

Now we use the GET search/tweets source for a meticulous query of awareness, including the capability to use a particular field that is incorporated in the metadata for the search outcomes to easily make supplementary requests for further search results. A tweet encoded in JSON format is shown in Figure 4.10 which clearly shows the various fields in a tweet.

```
{
 "contributors": null,
 "truncated": false,
 "text": "RT @khushpreetisa: @Gurmeetramrahim #MSGDoing111WelfareWork
s G Aap G Ki Rhmat 100 % Sure G My Care Taker PAPAG Thank U sooooo mu
ch I Lov U \u2026",
 "is_quote_status": false,
 "in_reply_to_status_id": null,
 "id": 611575018334457856,
 "favorite_count": 0,
 "source": "<a href=\"http://twitter.com/download/android\" rel=\"nof
ollow\">Twitter for Android</a>",
 "retweeted": false,
 "coordinates": null,
 "entities": {
  "symbols": [],
  "user_mentions": [
   {
    "id": 2854534602,
    "indices": [
     3,
     17
    ],
    "id_str": "2854534602",
    "screen_name": "khushpreetisa",
    "name": "\u2764\u2668\u0136h\u0171\u0161hp\u0159\u00eb\u00eb\u016
5 \u00ee\u00f1\u0161\u00e0n\u2668\u2764"
   },
   {
    "id": 2852359916,
    "indices": [
     19,
     35
    ],
    "id_str": "2852359916",
    "screen_name": "Gurmeetramrahim",
    "name": "GURMEET RAM RAHIM"
   }
  ],
```


Fig. 4.10. A Tweet Encoded in JSON Format

4.2.3 Fundamental Twitter Terminologies

Tweet is a short 140 character status update posted by users. Tweet not only contains the actual text message but also bundled with lots of metadata like entities and places. Tweet entities are in the form of user mentions, URLs, media and hashtags. Tweets may also

contain references to various places which map various locations in the real world. The fundamental terminologies used in twitter are shown in table 4.5.

Table 4.5. Fundamental Twitter Terminologies

Twitter Terminology	Description
Tweets	Short status updates (maximum 140 characters) posted by users.
Timeline	Collection of tweets chronologically sorted.
Tweet Entities	Metadata in the form of hashtags, user mentions, media and URLs that can be associated with a tweet.
Places	Metadata in the form of locations in real world can be associated with a tweet.
User Mentions	Mentions the particularly targeted users in a tweet, starting with a @ symbol
Hashtags	Cross-reference topics are mentioned with hashtags, starting with a # symbol.
Media	Any media file associated with a tweet.
URLs	Links to various web pages may be part of tweet in form of tweet entities.
RT	Retweet information.

4.2.4 Analyzing Tweets and Tweet Entities with Frequency Analysis

Virtually all analysis constitutes the simple work out of counting things on several levels and much of what we will be doing in this is manipulating twitter dataset so that it can be counted and further manipulated in some meaningful ways. Among the more convincing reasons for mining and analyzing Twitter data is to attempt to answer the query of what people are discussing about right now. One of the simplest approaches we could apply to reply this question is basic frequency analysis. As of Python 2.7, a collection module is presented that gives a counter that builds computing a frequency distribution rather trivial. The result obtained from the frequency distribution is a map of various key/value pairs equivalent to terms and their corresponding frequencies. This result can be represented in a nice tabular format with the help of Python prettytable package.

4.2.5 Computing the Lexical Diversity of Tweets

A slightly more sophisticated measurement that involves calculating straightforward frequencies and can be applied to unstructured text is a statistical metric known as *lexical diversity*. Lexical diversity can be defined as number of distinct tokens in the tweets divided by the overall number of tokens present in the tweets. Lexical diversity is remarkable concept in the field of inter personal communications as it gives a quantitative assess for the diversity of a group's or individual's vocabulary. Now Lexical Diversity of the tweets and average number of word present per tweet is calculated.

4.2.6 Examining Patterns in Retweets

Even though the many Twitter clients and user interface have long since implemented the indigenous Retweet API used to calculate status values such as retweeted_status and retweet_count, some Twitter users may favor to write a tweet, which involves a workflow containing copying and pasting the text and suffixing "/ via *@username*" or prepending "RT *@username*" to provide attribution. We further analyze the dataset to find out if there was a particular tweet that was extremely retweeted or if there were just plenty of "one-off" retweets. The procedure we will take to discover the most admired retweets is to just iterate over each status update and store out the corresponding retweet count, text of the retweet and originator of the retweet, if the status update is a retweet.

4.2.7 Sentiment Analysis of Tweets

Sentiment analysis or opinion mining can be defined as identifying and extracting the subjective information in the source materials with the use of natural language processing, computational linguistics and text analysis [17]. For doing sentiment analysis we use Alchemy API for sentiment analysis which is given by IBM. It provides an easy to use procedure to classify positive/negative opinion within any web page or document. Alchemy API's Sentiment Analysis API is capable of computing user specified sentiment targeting, document level sentiment, entity level sentiment, keyword level sentiment and emoticons sentiment. This API provides multiple modes of opinion mining for a variety of use cases like social media monitoring. Alchemy API provides endpoints for performing opinion mining on URLs and HTML files. Alchemy API uses sentiment

analysis algorithm which looks for words that hold a positive/negative connotation and then find out which place, person or thing they are pointing to. It also recognizes negations (i.e. "this bike is good" vs. "this bike is not good") and various modifiers (i.e. "this bike is good" vs. "this bike is really good"). The sentiment analysis API works on both large and small documents, including blog posts, news articles, comments, product reviews and Tweets.

Analyzing user's opinions from this large social dataset leads to large set of possibilities in various fields such as business, marketing, advertising, politics, education etc. Different persons have different opinions about a political party; it can be either a criticism or an appraisal. Analyzing combined view of all the persons about a political party helps in building an overall image of the party out of these individual opinions.

Chapter 5 Results and Discussion

5.1 Analysis and Visualization of LinkedIn Data

LinkedIn data is extracted by using LinkedIn API and normalized by removing redundancies. Data is further normalized according to locations of LinkedIn connections using geo coordinates provided by Microsoft Bing. Then, clustering of this normalized data set is done according to job title, company names and geographic locations using Greedy, Hierarchical and K-Means clustering algorithms. Then clusters are visualized to have a better insight into them.

5.1.1 Data Normalization along with Frequency Analysis

First of all, the extracted data from LinkedIn API is preprocessed to make it useful for further analysis. For this purpose data is normalized by considering various mindsets so as to remove the redundancies present in job title, company names and locations. All the collective titles are split using a slash like convener/ organizer and replaced with all the well-known short forms. The data set is also normalized by removing redundancies present due to location of the users by getting geographic coordinates of various locations from Microsoft Bing. The final results are obtained after processing these data sets through frequency analysis. Figure 5.1 shows the result of LinkedIn connection profiles after analyzing the various company names:

Company	Freq
Infosys	4
Amdocs	2
GENPACT	2
TechAhead	2
Self-employed	2
smartData Enterprises (I) .	2

Fig. 5.1. Normalizing and Counting Companies

Figure 5.2 shows a sample space after analyzing the same set through data normalization and frequency analysis with job titles. First, we split all the collective titles using a slash like convener/ organizer and then replace all the well-known short forms. The final results are obtained after processing these data sets through frequency analysis as shown in Figure 5.2. In Figure 5.3 various token present in the job titles are shown with their frequency count.

Title	Freq
System Engineer	5
Chief Executive Officer	2
Software Engineer	2
Software Developer	2
Subject Matter Expert	2
Business Development Executive	2

Fig. 5.2. Normalizing Job Titles and Corresponding Frequency Count

Token	Freq
Engineer	21
Software	14
Developer	9
Senior	7
System	7
Executive	6
Manager	6
Associate	5
Analyst	4
Project	4
Assistant	4
Development	4
Business	3

Fig. 5.3. Frequency Count of Tokens present in Job Title

One thing that is remarkable about the sample outcome is that the most common job title on the basis of exact matches is “System Engineer” which is intimately followed by other senior job positions such as “Chief executive officer”. Hence, the ego of this social professional network has logically excellent access to business leaders and entrepreneurs. The most common tokens present within the various job titles are “Engineer” and “Software”. Although the token “Engineer” is not a constituent token

present in the most common job title, it does come out from a large number of job titles such as “Software Engineer”, “Senior Software Engineer” and “System Engineer”, which show up close to the peak of the job title’s tokens list. Therefore, the ego of this social professional network appears to have a lot of connections to the technical practitioners as well.

Now we want to analyze the data contents of the subscribers of LinkedIn according to their geographical area. The data set is also normalized by removing redundancies present due to location of the users by getting geographic coordinates of various locations from Microsoft Bing. For this purpose, we have requested an API key with Bing web service. The sample outcome by using this API has been shown in Figure 5.4 which gives the geographic coordinates obtained from Bing API for some given location.


```
{
 "Pilibhit Area, India": [
  [
   28.7347297668,
   79.8858337402,
   0.0
  ]
 ],
 "Georgia": [
  [
   32.6483230591,
   -83.4445343018,
   0.0
  ]
 ],
```


Fig. 5.4. Geocoding Locations of LinkedIn Connections with Microsoft Bing

Now, we will use these geographic coordinates returned from geocoding service as part of a K-means clustering algorithm that can be a superior way to analyze this professional network.

5.1.2 Clustering Job Titles and Visualization of obtained Clusters

Now, LinkedIn data is clustered according to job titles by finding similarity between them using similarity metric Jaccard distance. The similar titles are clustered and displayed accordingly using greedy clustering technique as shown in Figure 5.5**.** This

approach is useful to manage and filter data contents into sub parts as per various analyzing criterion. We often need to allocate each cluster a significant label. We compute significant labels by taking the set wise intersection of various terms present in the job titles for every cluster.

The types of results obtained here are useful when they cluster the users who are likely to share same responsibilities in their job duties. This information might be useful for a variety of reasons, whether we are trying to find out who can best help us to make the next career move or trying to decide whether you are actually well adequate connected to other related professionals given our own job duties and the future aspirations.

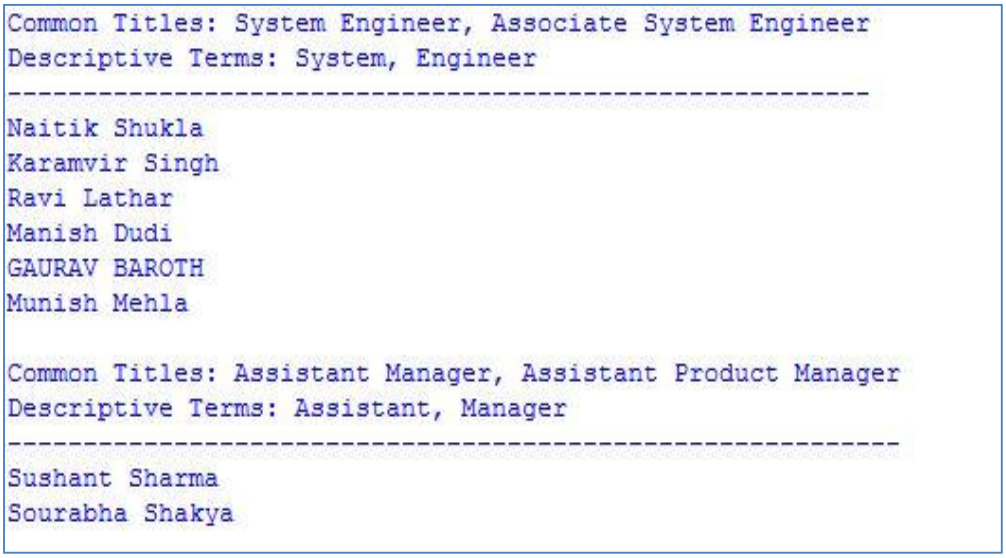
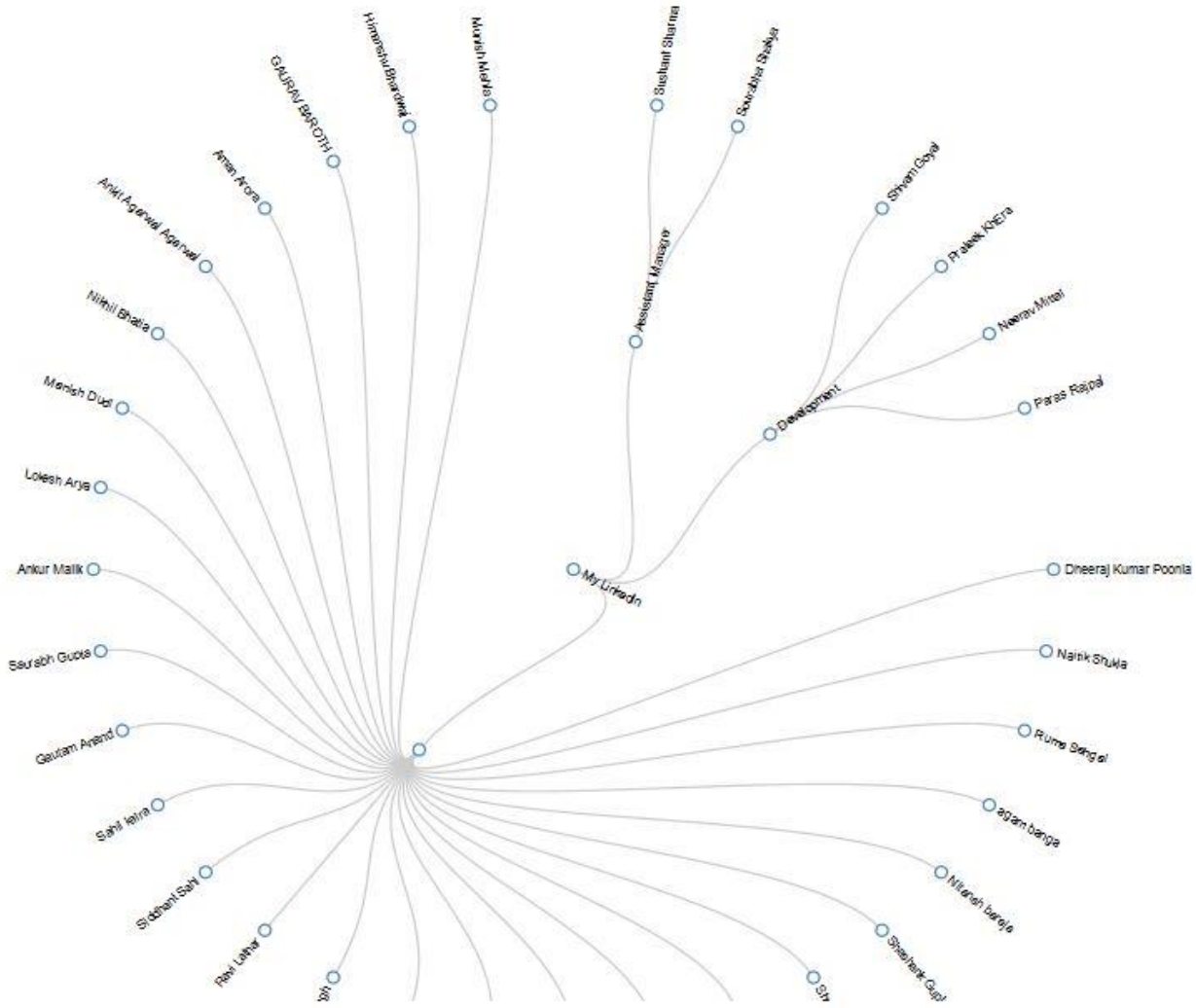


Fig. 5.5. Clustering Job Titles using Greedy Clustering

After that, Hierarchical clustering technique is used for clustering similar job titles. Two graphs of our connections after doing hierarchical clustering are drawn using D3.js (the state of art visualization toolkit), one is node link tree layout and other is a dendogram as shown in Figure 5.6 and Figure 5.7 respectively, the individual people are on the leaves of the tree, while the nodes such as "Assistant Manager" agglomerate all the leaf nodes into a cluster (small clusters having only one contacts are ignored in this graph). A remarkable amount of information becomes noticeable from these visualization diagrams when we are able to stare at a simple image of our professional network.

Although the tree in the dendogram is only two levels deep, it is not hard to imagine an additional level of agglomeration that conceptualizes something along the lines of a business executive with a label like “Chief, Officer” and agglomerates the “Chief, Technology, Officer” and “Chief, Executive, Officer” nodes. Oftentimes, agglomerative



clustering is not appropriate for large data sets because of its impractical runtimes.

Fig. 5.6. Node-Link Tree Layout of Contacts Clustered by Job Title

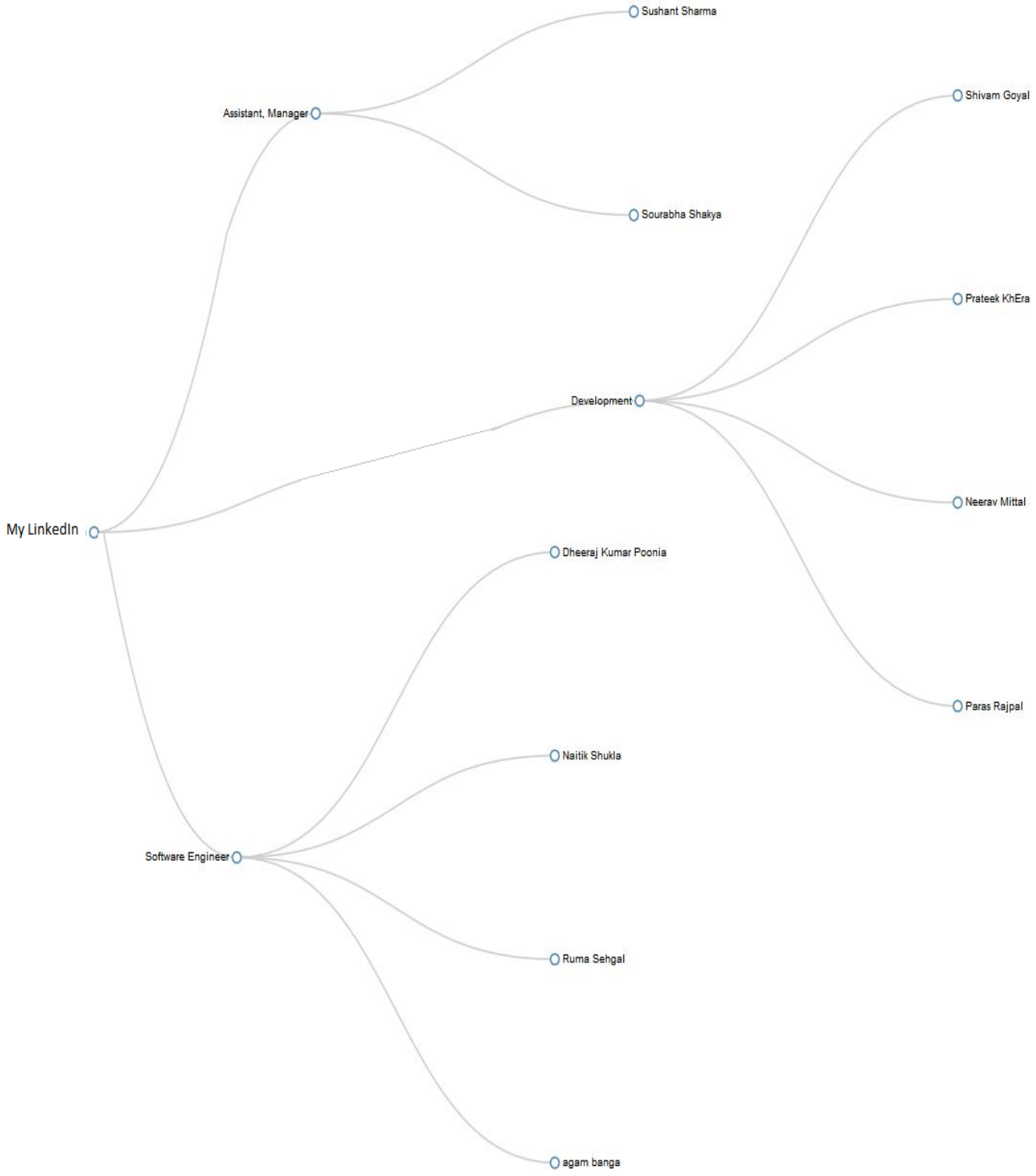


Fig. 5.7 Dendogram Layout of Contacts Clustered by Job Title

5.1.3 Geo-Clustering and Visualization of Geographic Clusters on Google Earth

Here data contents of LinkedIn connection are analyzed according to their geographical area. For this, location of users is represented with geographical coordinates and geo-clustering is performed on the LinkedIn connections. An API key from Bing web service is requested to find out the geo-coordinates of every location. K-means clustering algorithm is used to cluster and visualize our professional's LinkedIn data and results are plotted on Goggle earth as shown in Figure 5.8. After grouping our LinkedIn contacts according to geo-location, centroid of each cluster is also computed and plotted on the map as shown in Figure 5.9.

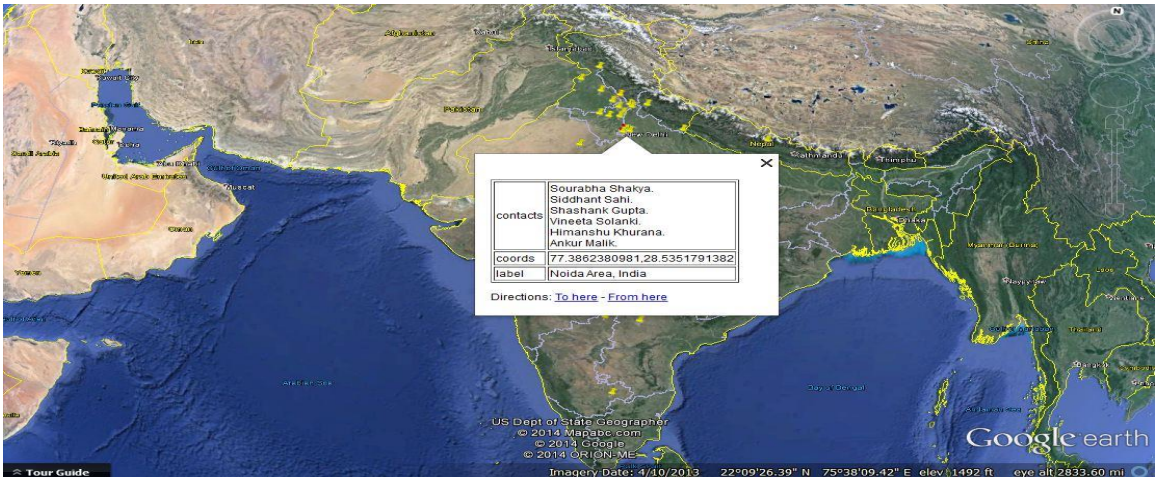


Fig. 5.8. Visualization of Geographic Clusters on Google Earth

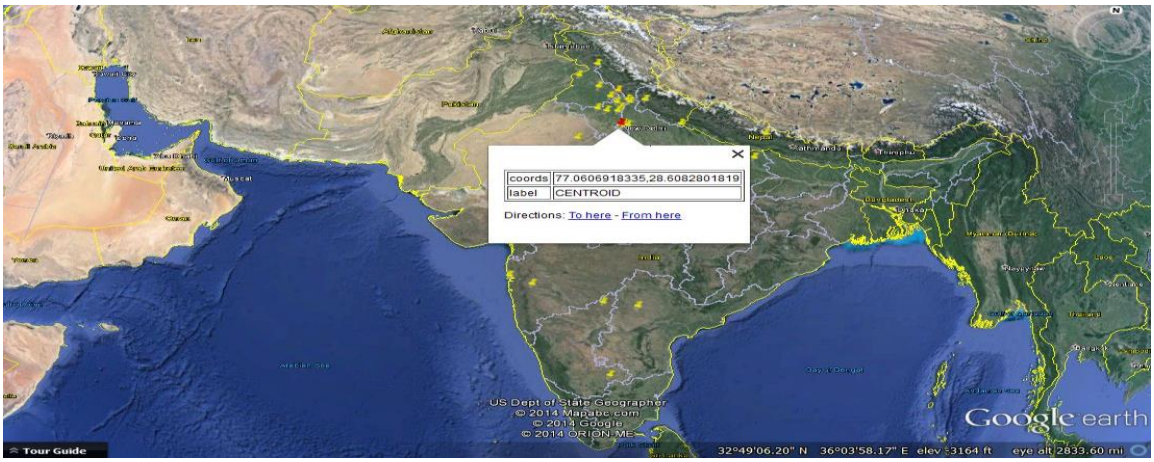


Fig. 5.9. Centroids of the Clusters computed by K-Means

5.1.4 Role of Clustering in Enhancing User Experiences

Straightforward and easy clustering tools and methods enhance user's experiences by influencing outcomes as smooth as the company or designation name and title. A clustered content of data is represented in the Figure 5.10 which shows a controlling another option or point of the data sets using a simple tree widget approach. This approach is very useful to manage or filter data contents into sub parts as per various analyzing criterion.

Fig. 5.10. Displaying Intelligently Clustered Data enhances User's Experience

The analysis of various clustering algorithms used for different criteria are shown below in the Table 5.1.

Table 5.1 Comparative Analysis of Clustering Algorithms used for Different Criteria

Clustering Techniques	Time Complexity	Similarity Criteria	Similarity Metric	Clustered Instances
Greedy	$O(n^2)$	Job Title	Jaccard Distance	Instances 255 Clusters 9
Hierarchical	$O(n^3)$	Job Title	Jaccard Distance	Instances 255 Clusters 7
K-Means	O(kn)	Geographic Coordinates	Jaccard Distance	Instances 255 Clusters 3

5.2 Analysis and Visualization of Twitter Data

Here we extracted tweets about "AAM AADMI PARTY" (a recently grown political party) and built an environment using python to induce analytical insights from these tweets using frequency analysis, retweet analysis and sentiment analysis.

5.2.1 Frequency Analysis

Tweets are extracted in JSON format by using Twitter API and stored in a document oriented database system MongoDB which is suitable for storing JSON data. After that tweets are processed for frequency analysis and retweet analysis. Tweets are tokenized into words and some frequency analysis is carried out as shown in Figure 5.11, 5.12 and 5.13. Only top 10 items in each list are shown in Figure. Figure 5.11 shows the various words present in the tweets with their count information. Figure 5.12 shows various screen names present in the extracted tweets with their frequency and Figure 5.13 shows the various hashtags present in the extracted tweets with their frequency.

Word	Count
RT	130
Aam	101
Aadmi	95
Party	74
in	71
à¤•à¥	62
the	58
of	51
à¤•à¥€	47
aam	46

Fig. 5.11. Frequency Count of Words present in Tweets

Screen Name	Count
ArvindKejriwal	32
acrossUpdate	31
AamAmdaniParty	30
RW_HITMAN	15
boneexpert	10
AAPInNews	8
riya_sharma9280	7
sane_voices	7
hindu_blood	4
indiantweeter	3

Fig. 5.12. Frequency Count of Screen Names present in Tweets

Hashtag	Count
AAPKaSting	30
स्वराज	30
आपियो	30
democracy	19
AAPWAR	19
WarInAAP	17
BiasedTwitterI	16
100DaysOfMufflerMan	14
Arvi	9
hypocrisy	9

Fig. 5.13. Frequency Count of Hashtags present in Tweets

Now this calculated frequency data is visualized in the form of histogram. Figure 5.14 shows a histogram which displays a sorted directory of tuples where every tuple is a (token/word, frequency) pair; the x-axis on the plot corresponds to the index of the particular tuple, and the y-axis corresponds to the frequency of the token in that tuple. Now data values are grouped together in bins and each bin corresponds to a particular range of frequency. These bins can be easily displayed in a histogram. Histograms corresponding to frequency distribution computed as shown in Figure 5.11, 5.12 and 5.13 are shown in Figure 5.15, 5.16 and 5.17 respectively after grouping words, hashtags and screen names into bins.

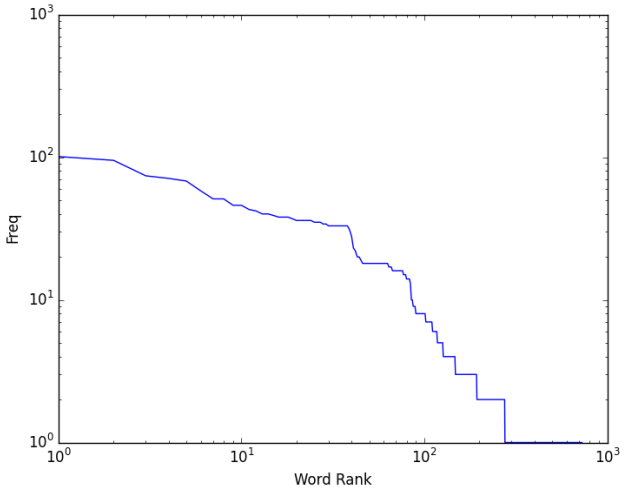


Fig. 5.14. Plot of Sorted Frequencies for words present in Tweets

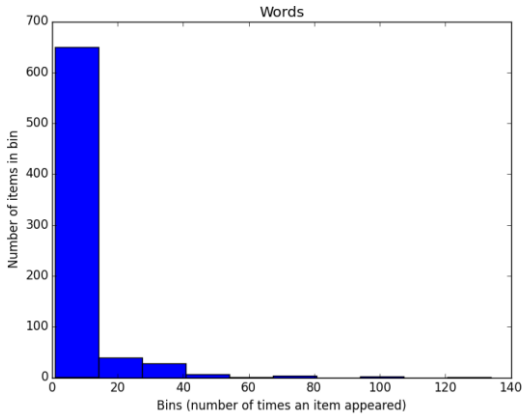


Fig. 5.15. Histogram of Computed Frequency Data for Words present in Tweets

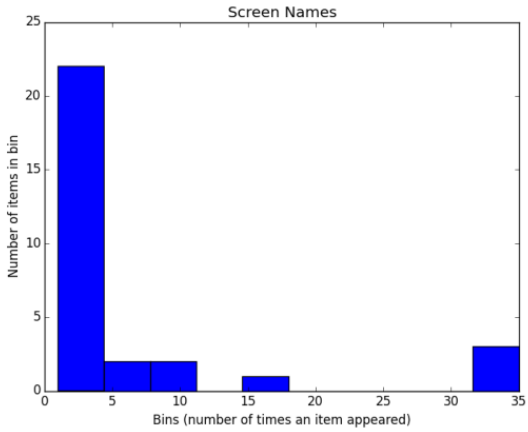


Fig. 5.16. Histogram of Computed Frequency Data for Screen Names

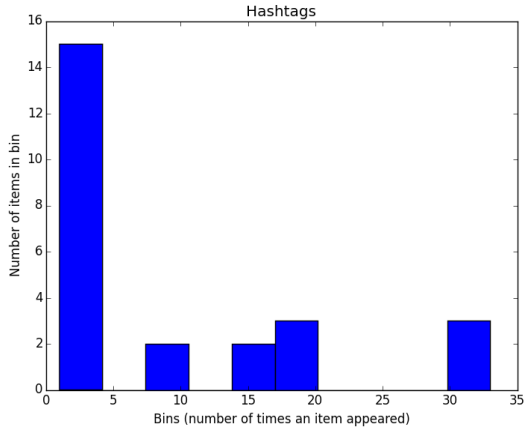


Fig. 5.17. Histograms of Computed Frequency Data for Hashtags

5.2.2 Computing Lexical Diversity of Tweets

Now Lexical Diversity of the tweets and average number of words present per tweet are calculated and shown in Table 5.2. Lexical diversity can be defined as number of distinct tokens in the tweets divided by the overall number of tokens present in the tweets. Lexical diversity is remarkable concept in the field of inter personal communications as it gives a quantitative assess for the diversity of a group's or individual's vocabulary.

Table 5.2 Lexical Diversity and average number of words from extracted Tweets

Lexical Diversity of words present in tweets	0.194711538462
Lexical Diversity of screen names	0.16393442623
Lexical Diversity of hashtags	0.113636363636
Average number of words per status text	18.72

Lexical Diversity of tokens present in the tweets is about 0.20, which means that about 2 out of 10 tokens are unique or every status update on Twitter carries about 20 percent unique information. Similarly the lexical diversity calculated on screen names is around 0.16, which means that around 2 out of 11 screen names mentioned in tweets are unique. The lexical diversity of hashtags is very low, around 0.11, means that very low values in the hashtags presents multiple times in the extracted tweets. The average number of words present per tweet is 18.72.

5.2.3 Retweet Analysis

Now Twitter Retweet API is used to further analyze the tweets dataset to determine how many times a particular tweet is retweeted or not. Figure 5.18 shows the various tweets and their corresponding retweet count. The histogram corresponding to retweet information is shown in Figure 5.19. The retweet analysis helps in finding out if there was a particular tweet that was extremely retweeted or if there were just plenty of "one-off" retweets.

Count	Screen Name	Text
490	DelhiTweeter	RT @DelhiTweeter: Even official page of "Jammat ud Dawa (JUD)" follows the Handle of "Aam Aadmi Party". Enough to Said. http://t.co/8cNvM4tâ€¦
178	rameshsrivats	RT @rameshsrivats: Maybe the basic problem with the Aam Aadmi Party is that it has too many aam aadmis partying.
178	rameshsrivats	RT @rameshsrivats: Maybe the basic problem with the Aam Aadmi Party is that it has too many aam aadmis partying.
69	AAPInNews	RT @AAPInNews: Aam Aadmi Party government: No FDI in multi-brand retail in Delhi - http://t.co/9O9daljnCN - http://t.co/f7XbGltybv
69	AAPInNews	RT @AAPInNews: Aam Aadmi Party government: No FDI in multi-brand retail in Delhi - http://t.co/9O9daljnCN - http://t.co/f7XbGltybv

Fig. 5.18. Tweets with their Retweet Count

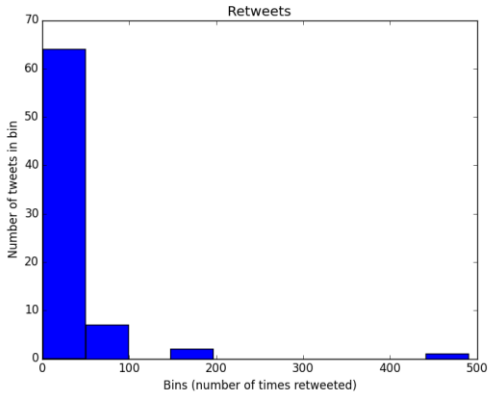


Fig. 5.19. Histogram of Retweet Frequencies

5.2.4 Sentiment Analysis

Sentiment analysis is carried out on more then 1000 collected tweets about "Aam Aadmi Party" using python, MongoDB, R and Alchemy API. The results obtained are shown in Figure 5.20. The number of positive tweets comes out is 377 and one more interesting point, the number of negative tweets comes out to be 758 and rest 574 tweets are classified as neutral. About 44.35% of tweets are having negative sentiment about Aam Aadmi Party. Figure 5.20 also shows the average positive tweet score and average negative tweet score with most negative and most positive tweet.

```
################
#    Stats     #
################


SENTIMENT BREAKDOWN
Number (%) of positive tweets: 377 (22.06%)
Number (%) of negative tweets: 758 (44.35%)
Number (%) of neutral tweets: 574 (33.59%)

AVERAGE POSITIVE TWEET SCORE: -0.432560
AVERAGE NEGATIVE TWEET SCORE: 0.000000

MOST POSITIVE TWEET
Text: @abhipintu26 @neha_aks if you are really generous feel free to ring AAP, t
hey just got their new helpline for accepting funds ;);)
User: ashmanthani
Time: Sun Apr 05 17:29:33 +0000 2015
Score: 0.947811

MOST NEGATIVE TWEET
Text: @OfficeOfRG never trust th Aam Aadmi Party. Traitors, back stabbers and tw
o faced hypocritical dirty @ArvindKejriwal @shaziailmi
User: neilhaslam90
Time: Thu May 07 17:50:54 +0000 2015
Score: -0.860547
```

Fig. 5.20. Stats obtained from Sentiment Analysis

R framework is used to visualize this data and 3 histograms are obtained as shown in Figure 5.21-5.23. Figure 5.21 represents a raw histogram that shows the sentiment scores for various positive or negative tweets in the dataset. Figure 5.22 is the kernel density estimation of the above histogram. The last is a bar chart that shows volume of Tweets (which are broken down by opinion) binned by hour.

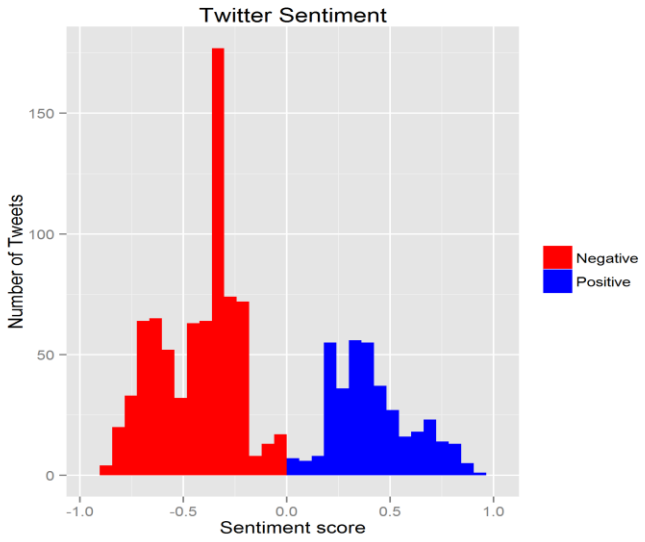


Fig. 5.21. Sentiment Score Histogram

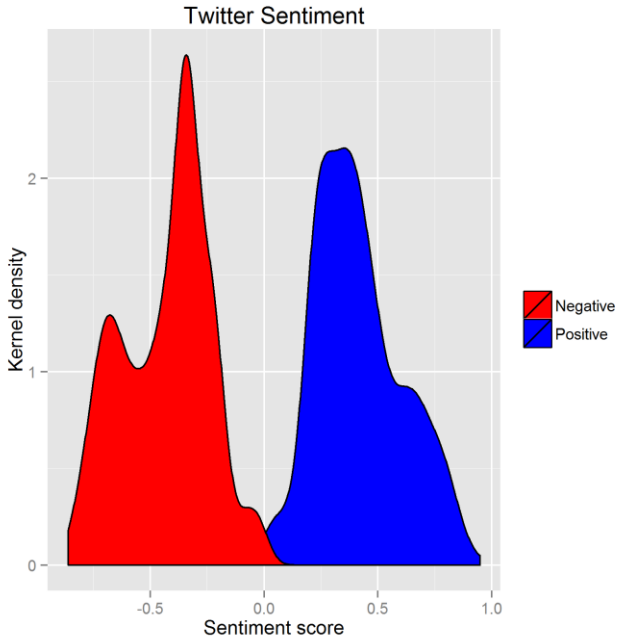


Fig. 5.22. Kernel Density Estimate of Sentiment Score

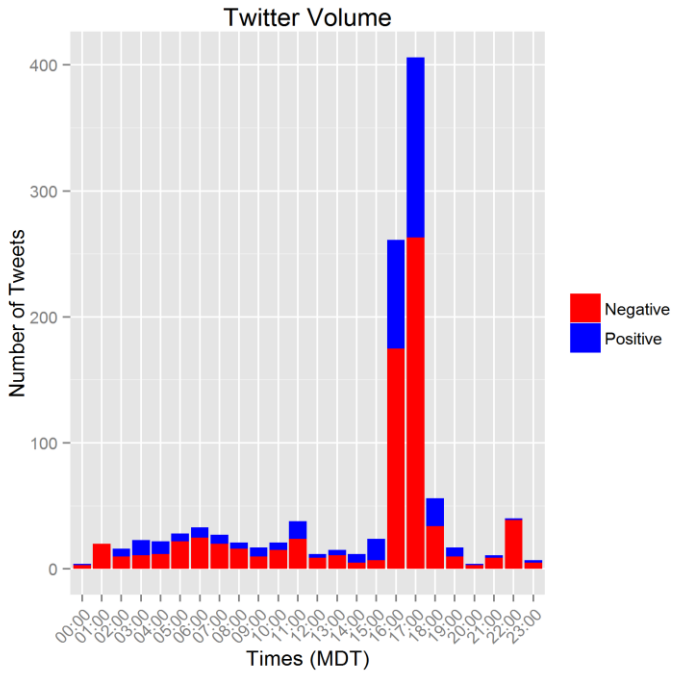


Fig. 5.23. Bar Chart of Volume of Tweets binned by hour

In this way by analyzing combined view of all the persons about a political party helps in building an overall image of the party out of these individual opinions. Analyzing these huge social datasets and predicting the opinions of individuals also plays an important role in business and academics.

Chapter 6 Conclusion and Future Scope

6.1 Conclusion

In this thesis, we explore the various research issues associated with social web analysis. Then we proceed for mining, analyzing and visualizing the two famous social networking sites that is LinkedIn and Twitter by using professional community detection, geo-clustering, frequency analysis and opinion mining.

We understand the fundamental concept of clustering and demonstrating a variety of ways to apply it to our professional network data on LinkedIn and also address common problems such as normalization of messy data, similarity computation on normalized data, and concerns related to the computational efficiency of approaches for a common data mining technique. By clustering the LinkedIn contacts various professional online communities according to job titles are detected and the clustered LinkedIn network is visualized to have a better insight into this professional dataset. A lot of information regarding our professional life can be visualized and analyzed using mining the social web such as LinkedIn which can not be seen otherwise so easily.

The spatial data present in LinkedIn dataset is also utilized in geo-clustering. We cluster the LinkedIn connection according to their location and compute the centroid of each cluster. Then the geo clusters thus obtained is visualized on Google Earth. Just visualizing our network can provide previously unknown insight, but computing the geographic centroids of our professional network can also open up some intriguing possibilities. For example, we might want to compute candidate locations for a series of regional workshops or conferences. Alternatively, if we are in the consulting business and have a hectic travel schedule, we might want to plot out some good locations for renting a little home away from home, or maybe we want to map out professionals in our network according to their job duties, or the socioeconomic bracket they are likely to fit into based on their job titles and experience. Beyond the numerous options opened up by visualizing our professional network’s location data, geographic clustering lends itself to many other possibilities, such as supply chain management and travelling salesman types of problems in which it is necessary to minimize the expenses involved in travelling or moving goods from point to point.

Secondly, in this research we extract tweets about “Aam Aadmi Party” from twitter REST APIs. The extracted tweets are analyzed by calculating some simple statistics and by performing frequency analysis, retweet analysis and sentiment analysis. Sentiment analysis depicts the opinion of a community towards a product, services or a political party.

6.2 Future Work

- We can explore the LinkedIn extended profile information by using various others LinkedIn API, and try to correlate where people work versus where they went to school and to analyze whether people tend to relocate into and out of certain areas.
- The LinkedIn API provides a means of retrieving a connection’s Twitter handle. We can check how many of our LinkedIn connections have Twitter accounts associated with their professional profiles? How active are their accounts? How *professional* are their online Twitter personalities from the perspective of a potential employer?
- We can also fetch various trending topics in different professional communities by connecting LinkedIn API with Twitter API.
- We can also classify the tweets of various connections in our LinkedIn profile into positive, negative and neutral from a professional point of view or from a different perspective whatever we want and can perform sentiment analysis and opinion mining on the same.
- By analyzing data from LinkedIn, we can also design a predictive system which recommends the preferred job for any person on the basis of his skill set and past experiences.
- We can explore twitter Streaming APIs to do these analysis at a streaming rate with real time. This continuous analysis of tweets can play a quite impressive role for a government to get an overall picture of its public’s opinion towards its rule.

References

[1] E. Ferrara, P. D. Meo, G. Fiumara and R. Baumgartner, "*Web Data Extraction, Applications and Techniques: A Survey*", Knowledge Based Systems, Vol. 70, No. 3, pp. 301-323, 2014.

[2] D. F. Nettleton, "*Data Mining of Social Networks Represented as Graphs*", Computer Science Review, Vol. 7, No. 2, pp. 1-34, 2014.

[3] J. Gantz and D. Reinsel, "*Extracting Value from Chaos*", IDC Interactive View, pp. 1–12, 2011.

[4] V. M. Schonberger and K. Cukier, "*Big Data: a Revolution that will transform how we live, work, and think*", Houghton Mifflin Harcourt.

[5] M. Chen, S. Mao and Y. Liu, "*Big Data A Survey*", Mobile Networks and Application, Vol. 19, No. 2, pp. 171-209, 2014.

[6] T. Miki, T. Nomura and T. Ishida, "*Semantic Web Link Analysis to discover Social Relationship in Academic Communities*", In Proceedings of IEEE Symposium on Application and the Internet, pp. 38-45, 2005.

[7] M. Jamali and H. Abolhassani, "*Different Aspects of Social Network Analysis*", In Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, pp. 66-72, 2006.

[8] D. Laney, "*3-D Data Management: controlling Data Volume, Velocity and Variety*", META Group Research Note, 2001.

[9] J. Chen, Y. Chen, D.U. Xiaoyong, L.I. Cuiping, L.U. Jiaheng, Z. Suyun and X. Zhao, "*Big Data Challenge: a Data Management Perspective*", Frontiers of Computer Science, Vol. 7, No. 2, pp. 157-164, 2013.

[10] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. H. Byers, "*Big Data: the next frontier for Innovation, Competition, and Productivity*", McKinsey Global Institute, 2011.

[11] J. Heer and S. Kandel, "*Interactive Analysis of Big Data*", The ACM Magazine for Students - Big Data, Vol. 19, No. 1, pp. 50-54, 2012.

[12] A. Jacobs, "*The Pathologies of Big Data*", Communications of the ACM, Vol. 52, No. 8, pp. 36-44, 2009.

[13] I. Kopanas, N. Avouris and S. Daskalaki, "*The Role of Domain Knowledge in a Large Scale Data Mining Project*", In Proceedings of Second Hellenic Conference Artificial Intelligence: Methods and Applications of Artificial Intelligence, pp. 288-299, 2002.

[14] J. I. Maletic and A. Marcus, "*Data Cleansing: beyond Integrity Analysis*", In Proceedings of Information Quality Citeseer, pp. 200–209, 2000.

[15] G. Nandi and A. Das, "*Online Social Network Mining: current Trends and Research Issues*", International Journal of Research in Engineering and Technology, Vol. 3, No.4, pp. 346-350, 2014.

[16] G. Nandi and A. Das, "*A Survey on Using Data Mining Techniques for Online Social Network Analysis*", International Journal of Computer Science Issues, Vol. 10, No. 6, pp. 162–167, 2013.

[17] R. Prabowo and M. Thelwall , "*Sentiment Analysis: A Combined Approach",* Journal of Informatics, Vol. 3, No.2, pp. 143-157, 2009.

[18] A. Nagpal, A. Jatain and D. Gaur, "*Review based on Data Clustering Algorithm*", In Proceedings of IEEE Conference on Information and Communication Technologies, Vol. 13, pp. 298-303, 2013.

[19] S. Bird, E. Klein and E. Loper, "*Natural Language Processing with Python–Analyzing Text with the Natural Language Toolkit*".

[20] S. Revathi and T. Nalini, "*Performance Comparison of Various Clustering Algorithms*", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, No. 2, pp. 67-72, 2013.

[21] S. Arora and I. Chana, "*A Survey of Clustering Techniques for Big Data Analysis*", In Proceedings of IEEE 5th International Conference on Confluence the Next Generation Information Technology Summit, Vol. 14, pp. 59-65, 2014.

[22] M.A. Russell, "*Mining the Social Web*", O'Reilly Media.

[23] M. Srinivas and C. K. Mohan, "*Efficient Clustering Approach using Incremental and Hierarchical Clustering Methods*", In Proceedings of IEEE International Joint Conference On Neural Networks, pp. 1-7, July 2010.

[24] C. Boutsidis, A. Zouzias, M. W. Mahoney and P. Drineas, "*Randomized Dimensionality Reduction for K-Means Clustering*", IEEE Transactions On Information Theory, Vol. 61, No. 2, pp. 1045-1062, 2015.

[25] G. A. Wilkin and X. Huang, "*K-Means Clustering Algorithms: Implementation and Comparison*", In Proceedings of Second IEEE International Multi Symposium on Computer and Computational Sciences, pp. 133-136, 2007.

[26] R. Kumar and R. Verma, "*Classification Algorithms for Data Mining: A Survey*", International Journal of Innovations in Engineering and Technology, Vol. 1, No. 2, pp. 7-14, 2012.

[27] C. Y. Christopher and D. N. Tobun, "*Analyzing Content Development and Visualizing Social Interactions in Web Forum*", In Proceedings of IEEE International Conference on Intelligence and Security Informatics, pp. 25-30, 2008.

[28] A. Younus, M. A. Qureshi, F. F. Asar, M. Azam, M. Saeed and N. Touheadd, "*What do the average Twitterers say: a Twitter model for public Opinion Analysis in the face of major Political Events*", In Proceedings of International Conference on Advances in Social Networks Analysis and Mining, pp. 618-623, 2011.

[29] E. Kouloumpis, T. Wilson and J. Moore, "*Twitter Sentiment Analysis: The good, the bad and the OMG!*", In proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, pp. 538-541, 2011.

[30] K. Kaur and R. Rani, "*Managing Data in Healthcare Information Systems: Many Models, One Solution*", Computer, Vol. 48, No. 3, pp. 52-59, 2015.

[31] M. Cha , H. Haddadi and P. K. Gummadi, " *Measuring User Influence in Twitter: the million follower fallacy*," In Proceedings of International Conference on Web and Social Media, pp.10-17, 2010.

[32] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis", *Foundations and Trends in Information Retrieval*, Vol. 2, No. 2, pp. 1-135, 2008.

[33] V. Hatzivassiloglou and K. McKeown, "*Predicting the Semantic Orientation of Adjectives*", In Proceedings of Eighth Conference on European Chapter of the Association for Computational Linguistics, pp. 174-181, 1997.

[34] A. Esuli and F. Sebastiani, "*SentiWordNet: A publicly available Lexical Resource for Opinion Mining*", In Proceedings of the 5th Conference on Language Resources and Evaluation, pp. 417-422, 2006.

[35] H. Yu and V. Hatzivassiloglou, "*Towards Answering Opinion Questions: Separating facts from Opinions and Identifying the Polarity of Opinion Sentences*", In Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 129-136, 2003.

[36] A. Bifet and E. Frank, "*Sentiment Knowledge Discovery in Twitter Streaming Data*", In Proceedings of 13th International Conference on Discovery Science, pp. 1-15, 2010.

[37] A. Pak and P. Paroubek, "*Twitter as a Corpus for Sentiment Analysis and Opinion Mining*", In Proceedings of the Conference on Language Resources and Evaluation, pp. 1320-1326, 2010.

[38] L. Barbosa, and J. Feng, "*Robust Sentiment Detection on Twitter from biased and noisy Data*", In Proceedings of the 23rd International Conference on Computational Linguistics, pp. 36-44, 2010.

[39] D. Davidov, O. Tsur and A. Rappoport, "*Enhanced Sentiment learning using Twitter Hashtags and Smileys*", In Proceedings of the 23rd International Conference on Computational Linguistics, pp. 241-249, 2010.

[40] Y. Choi, C. Cardie, E. Riloff and S. Patwardhan, "*Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns*" In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 355–362, Oct. 2008.

[41] K. Hiroshi, N. Tetsuya and W. Hideo, "*Deeper Sentiment Analysis using Machine Translation Technology*", In Proceedings of the 20th International Conference on Computational Linguistics, pp. 494–500, Aug. 2004.

[42] Y. J. Nasukawa, T. W. Niblack and R. Bunescu, "*Sentiment Analyzer: Extracting Sentiments about a given topic using Natural Language Processing*

Techniques", In Proceedings of the 3rd IEEE International Conference on Data Mining, pp. 427–434, Nov. 2003.

[43] A. Go, R. Bhayani and L. Huang, "*Twitter Sentiment Classification using Distant Supervision*", Technical report, Stanford, 2009.

[44] P. Mika, "*Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks***",** In Proceedings of 3rd. International Semantic Web Conference, pp. 211-223, 2005.

[45] F. Fu, L. Liu and L. Wang, "*Empirical Analysis of Online Social Networks in the age of Web 2.0*", Physica A: Statistical Mechanics and its Applications, Vol. 387, No. 2, pp. 675-684, 2008.

[46] A. Mislove, M. Macron, K. P. Gummadi, P. Druschel and B. Bhattacharjee, "*Measurement and Analysis of Online Social Networks*", In Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, pp. 29-42, 2007.

[47] F. E. John and D. Abbott, "*A Comparison of Leading Data Mining Tools*" In Proceedings of Fourth International Conference on Knowledge Discovery and Data Mining, pp. 19-25, Aug. 1998.

[48] H. V. Jagadish, B. C. Ooi and Q. H. Vu, "*BATON: a Balanced Tree structure for peer-to-peer Networks*", In Proceedings of the 31st International Conference on very large Data Bases, pp. 661-672, 2005.

[49] J. Chen, Y. Chen, D.U. Xiaoyong, L.I. Cuiping, L.U. Jiaheng, Z. Suyun and X. Zhao, "*Big Data Challenge: a Data Management Perspective*", Frontiers of Computer Science, Vol. 7, No. 2, pp. 157-164, 2013.

[50] S. Das, Y. Sismanis, K. S. Beyer, R. Gemulla, P. J. Haas and J. R. McPherson, "*Integrating R and Hadoop*", In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 987-998, 2010.

[51] Y. Zhao and G. Karypis, "*Evaluation of Hierarchical Clustering Algorithms for Document Datasets*", In Proceedings of Conference on Information and Knowledge Management, pp. 515-524, 2002.

[52] B. H. Park and H. Kargupta, "*Distributed Data Mining: Algorithms, Systems, and Applications*", In Proceedings of International Conference on Distributed Data Mining, pp. 341-358, 2002.

[53] J. A. Hartigan and M. A. Wong, "*A K-Means Clustering Algorithm*", Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1, pp. 100-108, 1979.

[54] C. Ordonez, "*Clustering Binary Data streams with K-Means*", In Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 12-19, 2003.

[55] J. Han, K. Koperski and N. Stefanovic, "*GeoMiner: a system prototype for Spatial Data Mining*", In Proceedings of the ACM SIGMOD International Conference on Management of data, pp. 553-556, 1997.

[56] C. A. Lu, C. H. Chen and P. J. Cheng, "*Clustering and Visualizing Geographic Data Using Geo-tree*", In Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Vol. 11, pp. 479-482, 2011.

List of Publications

[1] P. Garg, R. Rani and S. Miglani, "*Analysis and Visualization of Professional's LinkedIn Data*", In Proceedings of Springer 3rd International Conference on Emerging Research in Computing, Information, Communication and Application (ERCICA-15), NMIT Bangalore, July 31 – Aug 01, 2015. [**Accepted**]

[2] P. Garg, R. Rani and S. Miglani, "*Mining Professional's Data from LinkedIn*", In Proceedings of IEEE 5th International Conference on Advances in Computing and Communication (ICACC-2015), RSET Kochi, Sept. 2-4, 2015. [**Accepted**]

[3] P. Garg, R. Rani and S. Miglani, "*Analysis and Visualization of Twitter Data using Python, R and MongoDB*", IEEE 8th International Conference on Contemporary Computing (IC3), JIIT Noida, Aug 20-22, 2015. [**Communicated**]