Identification Of Kols In Pharmaceutics Using Network Analysis

Priya Shelke, Sachin Gotecha, Manas Kulkarni, Sanul Raskar
Information Technology, Vishwakarma Institute of Information Technology
priya.shelke@viit.ac.in
sachin.gotecha@viit.ac.in
manas.kulkarni@viit.ac.in
sanul.raskar@viit.ac.in

Abstract

Key Opinion Leaders (KOLs) are the influential people who have done research in specific fields and are experts in that field. Our project aims to provide a platform for identification of KOLs in pharmaceutics using network analysis. The platforms constitute following modules - Module 1 is the user interface for entering keywords displaying list of KOLs and information database, Module 2 consists of libraries for formation of network based on keyword entry, Module 3 consists of algorithms for analysis of network formed. The paper includes I) Introduction, II) Related work and literature survey, III) Implementation, IV) Future Scope and V) Conclusion.

Keywords – influencers, webscraping, network, kol, pharmaceutics

I. INTRODUCTION

Key opinion leaders are the people who have done a lot of research in a particular field and are experts of that field [1]. When a pharmaceutical industry starts exploring new medicines or diseases it may take a very long time for it to get that medicine work the way it should or find a perfect medicine for a particular disease. But if the industry could hire a researcher who has been doing research upon that particular medicine or disease, then it could save a lot of its time, money and resources. There could be a lot of researchers who have been working around that medicine or disease and their experiences may vary. At times just based on one or two researches industry can't decide which researcher to hire. Here in this paper we propose a platform which will facilitate hiring right researcher for the job. We aim to use European PubMed Central's data for collecting information about researchers, their publications and citations. Using this data, we will create a network of authors and then analyze this network against network analysis algorithms [4].

II. RELATED WORK AND LITERATURE SURVEY

Hengmin Zhou and Daniel Zeng [1] have tried Finding Leaders from Opinion Networks. Their goal was to utilize results of opinion mining which in turn would facilitate social network analysis. They had collected data of e-commerce website on which visitors could read new and old reviews about variety of items to help them decide on purchase. They used this dataset to create social network. Their work was mainly based on opinion of people. Huanhuan Liu, Xiaoqing Yu, Jing Lu [2] have worked on identifying Top-N Opinion Leaders for finding the people who has important influence on message propagation on local social network. They apply different centrality measures at once on China's local social network ('sinamicroblog' which is called hybrid of Facebook and Twitter). They compare their result with base mark methods such as PageRank and HITS algorithm. They have found out that person who is identified as an opinion leader with their proposed algorithm is likely to be the opinion leader identified with PageRank and HITS algorithms. Zhixiao Wang, Changjiang Du, Jianping Fan, Yan Xing [3] had proposed a multi-attribute ranking method based on position of node and its neighborhood. This method utilizes iteration information of the K-shell decomposition to further distinguish the node position and also fully considers the neighborhood's effect upon the influence capability of a node.

III. Implementation

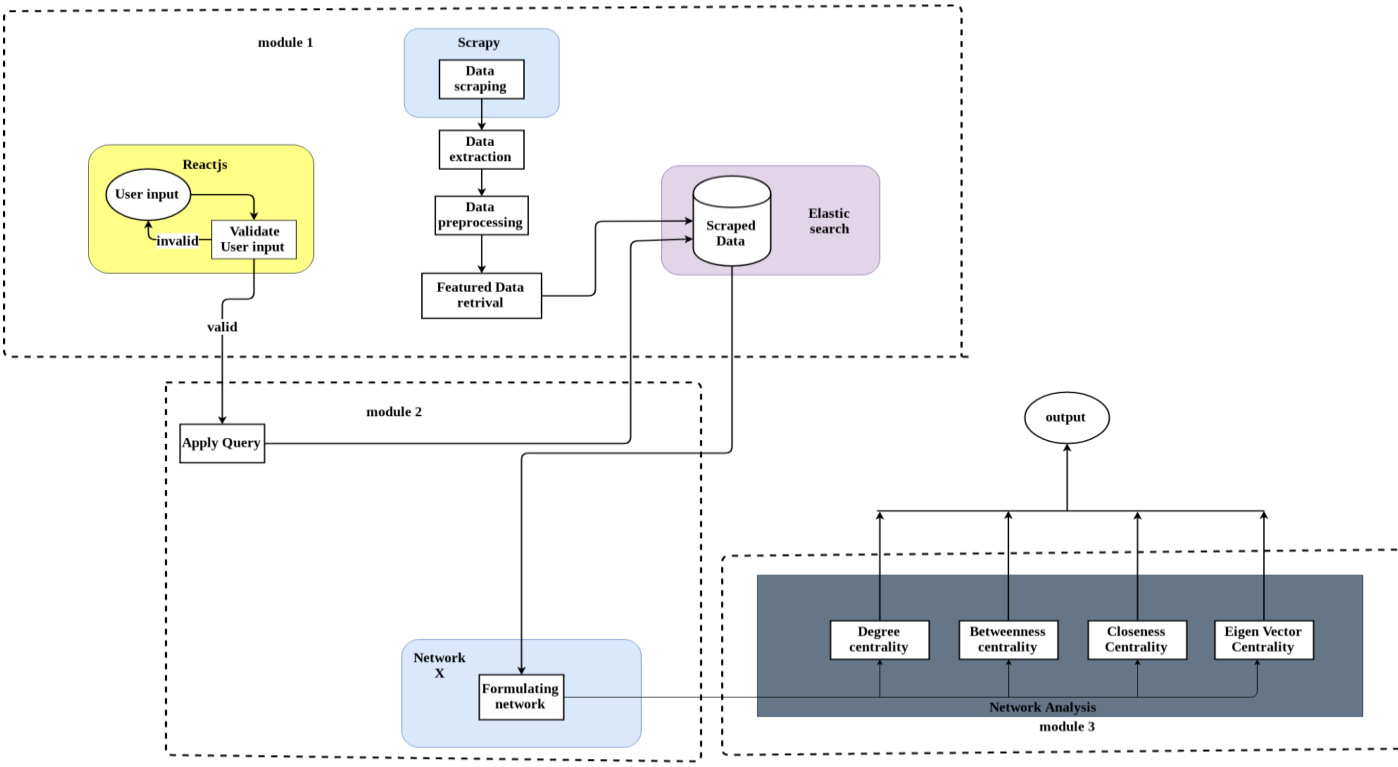


Fig. 1 Block Diagram

A. Frontend Technologies

The following are technologies used to build the client.

1) Reactjs: It is a JavaScript framework for building single-page or mobile applications and the advantage of fetching changing data rapidly that needs to be stored [10].

2) Redux: Redux is an open-source JavaScript library for managing application state. It is most commonly used with libraries such as React [11].

3) D3: D3js is a JavaScript library for generating dynamic and interactive data visualizations in web browsers. It mainly uses SVG, HTML, and CSS [12].

4) Bootstrap: Bootstrap is a CSS framework used for responsive and mobile-first frontend development. It contains CSS, JS based templates for buttons, forms, navigation, and other UI components [13].

B. Backend Technologies

The following are technologies used to build the server.

1) Django: Django is an open-source and free web-based framework, in Python. A web framework is defined as a set of components that helps you to develop websites more efficiently [14].

2) Scrapy: It is an open-source framework that is used for extracting the data websites [15].

3) NetworkX: NetworkX is a package written in python for creating, manipulating, and studying the structure, functions, and dynamics of complicated networks. Various features provided by NetworkX include Data structures for graphs, digraphs, and multigraphs, many traditional graph algorithms, measures for analyzing network structure, generators for classic graphs, random graphs, and synthetic networks [8].

4) Elastic: It's an open-source, RESTful search engine that is built on top of Apache Lucene and released under an Apache license. It is Java-based and used to search. and index document files in diverse formats [16].

C. Module 1

Module 1 is comprised of a user interface that prompts the user to enter the keyword-based on which researchers' data will be found. The elastic database is used to store the data scraped from the European PubMed website [7]. It includes information about researchers, their field of research and their publications/citations. This keyword is sent from the client to the server. If the data for a keyword is not present in the database, then web crawlers are activated by Scrapy framework and data is scraped from the website [7]. We scraped data with fields like research paper title, authors, citations, date of publication and paper ID. Data is cleaned and stored in an elastic database. Elasticsearch parses, normalizes, indexes the data given to it.

D. Module 2

This module consists of network formation. Stored scraped data is retrieved from the database to form the network. Before giving the data to the NetworkX, the data is preprocessed, and unwanted data is removed and remaining useful data is given as input to the next module. The data retrieved from the database is then given to Python's NetworkX library. This library converts this data into a network. In the author's collaboration network, every node represents an author while every edge represents at least one publication common between the two nodes (authors). The numbers of citations for that publication is represented by the weight.

E. Module 3

This module consists of algorithms used for analyzing the network. The algorithms include [5]:

1) Degree centrality: The degree of a node means the number of nodes that a particular node is connected to. Using the degree centrality algorithm, we calculate the degree of each node.

2) Closeness centrality: Using the closeness centrality algorithm we calculate the minimum distance of each node from all other nodes in the network and do their sum. This results in greater weight lesser the distance and vice versa.

3) Betweenness centrality: We calculate the sum of the minimum distance between two nodes if the path between those two nodes is passing through the node for which we are calculating betweenness centrality and divide it by the total number of shortest paths present between those two nodes [6].

4) Eigenvector Centrality Algorithm: In this algorithm, we first formulate adjacency matrix A (n*n) and then we consider vector v of having dimension n*1. Vector v contains values of each node in terms of their influence; greater value node will have greater influence in the network. And one damping factor is also used in order to control increasing vector v values after each iteration. We keep on doing iterations till it converges to some value.

F. Output

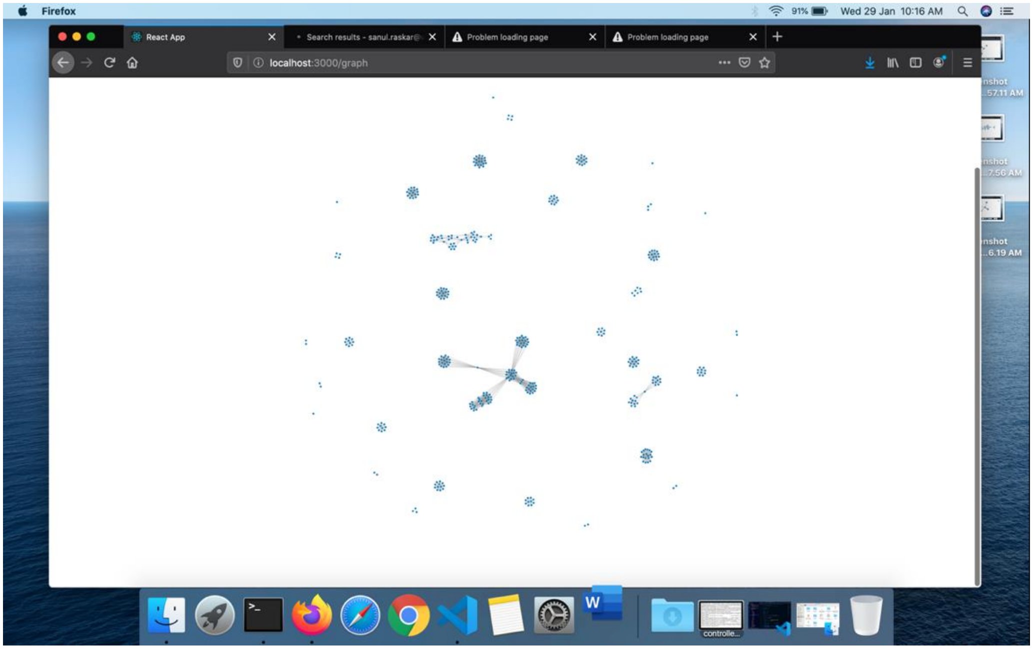
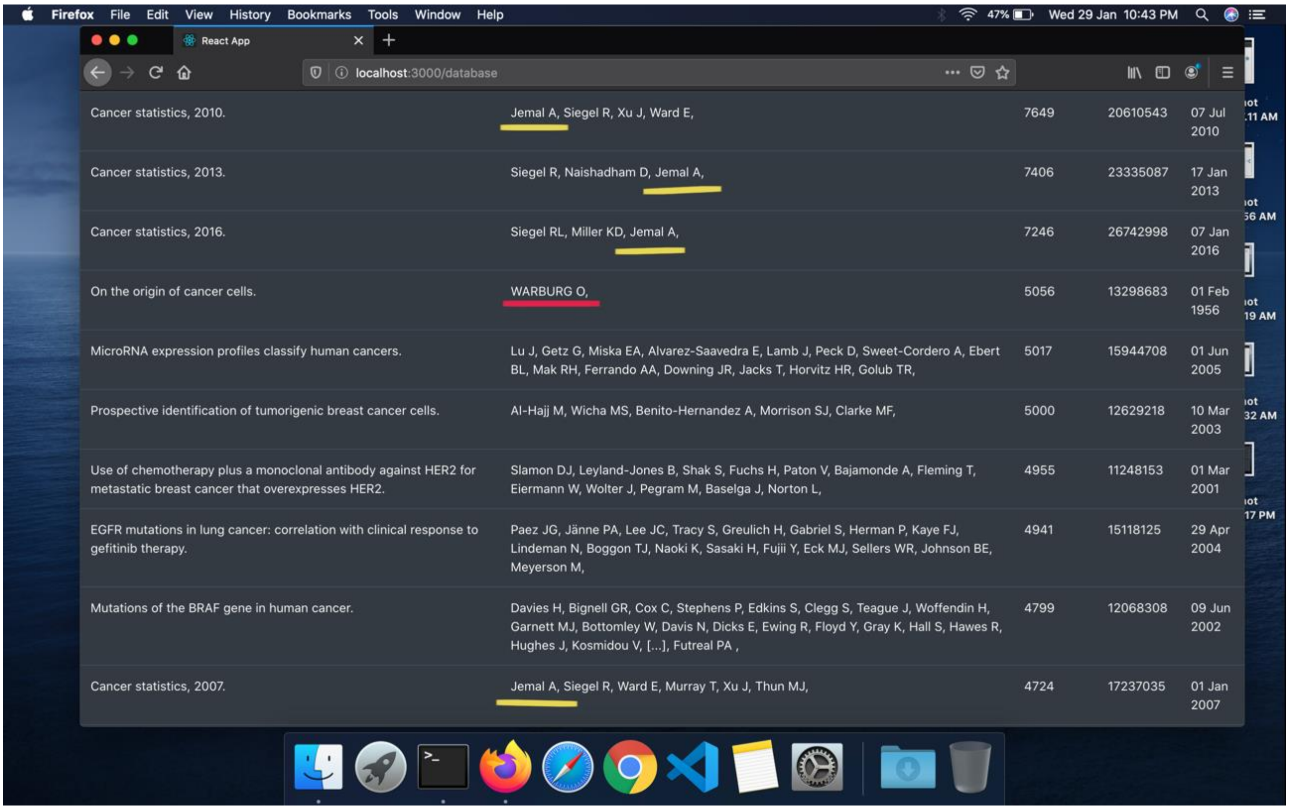


Fig. 2 Complete Network

As of now, we have implemented module 1 to scrape data and store cleaned data to the database. We have integrated the redux library in Reactjs for app state management. Scrapy crawlers can be initiated from the Django server. As our server uses restful API, clients can send GET and POST HTTPS request and get a response in JSON format.

The formed network is exported as a JSON document from the NetworkX library and stored in the database. The network is displayed on WebApp using the D3 visualization library.



Cancer statistics, 2010.	Jemal A, Siegel R, Xu J, Ward E,	7649	20610543	07 Jul 2010
Cancer statistics, 2013.	Siegel R, Naishadham D, Jemal A,	7406	23335087	17 Jan 2013
Cancer statistics, 2016.	Siegel RL, Miller KD, Jemal A,	7246	26742998	07 Jan 2016
On the origin of cancer cells.	WARBURG O,	5056	13298683	01 Feb 1956
MicroRNA expression profiles classify human cancers.	Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR,	5017	15944708	01 Jun 2005
Prospective identification of tumorigenic breast cancer cells.	Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF,	5000	12629218	10 Mar 2003
Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2.	Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, Baselga J, Norton L,	4955	11248153	01 Mar 2001
EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy.	Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, Naoki K, Sasaki H, Fujii Y, Eck MJ, Sellers WR, Johnson BE, Meyerson M,	4941	15118125	29 Apr 2004
Mutations of the BRAF gene in human cancer.	Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks E, Ewing R, Floyd Y, Gray K, Hall S, Hawes R, Hughes J, Kosmidou V, [...], Futreal PA ,	4799	12068308	09 Jun 2002
Cancer statistics, 2007.	Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ,	4724	17237035	01 Jan 2007

Fig. 3 Database for Cancer Keyword

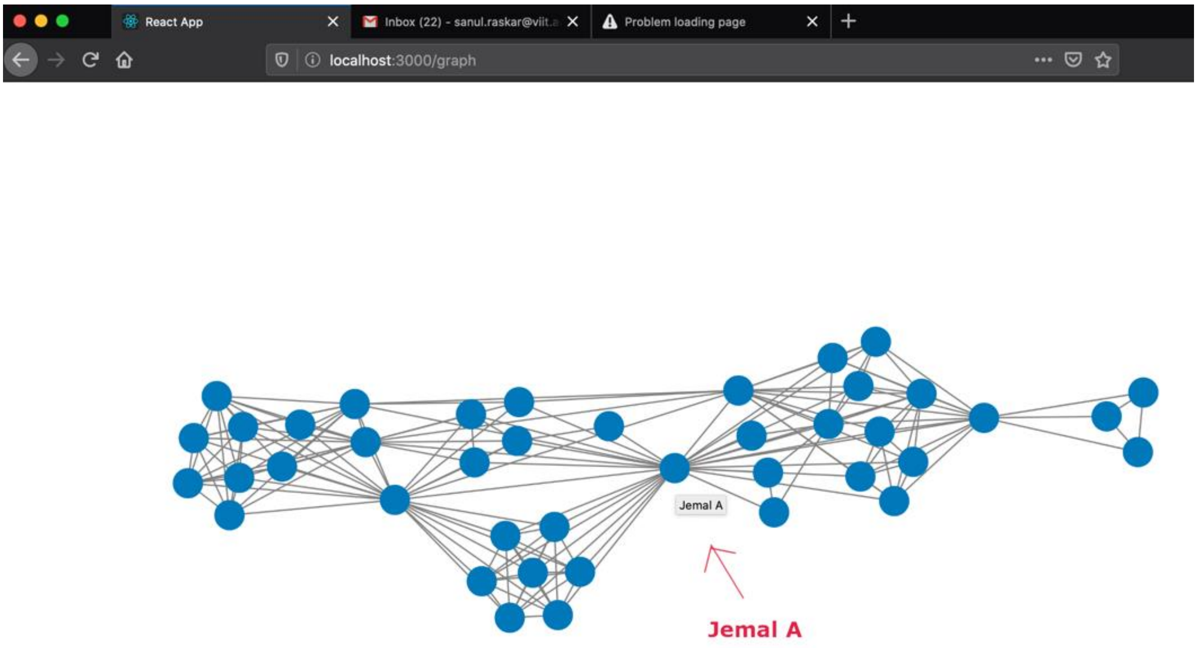


Fig. 4 Network with author Jemal

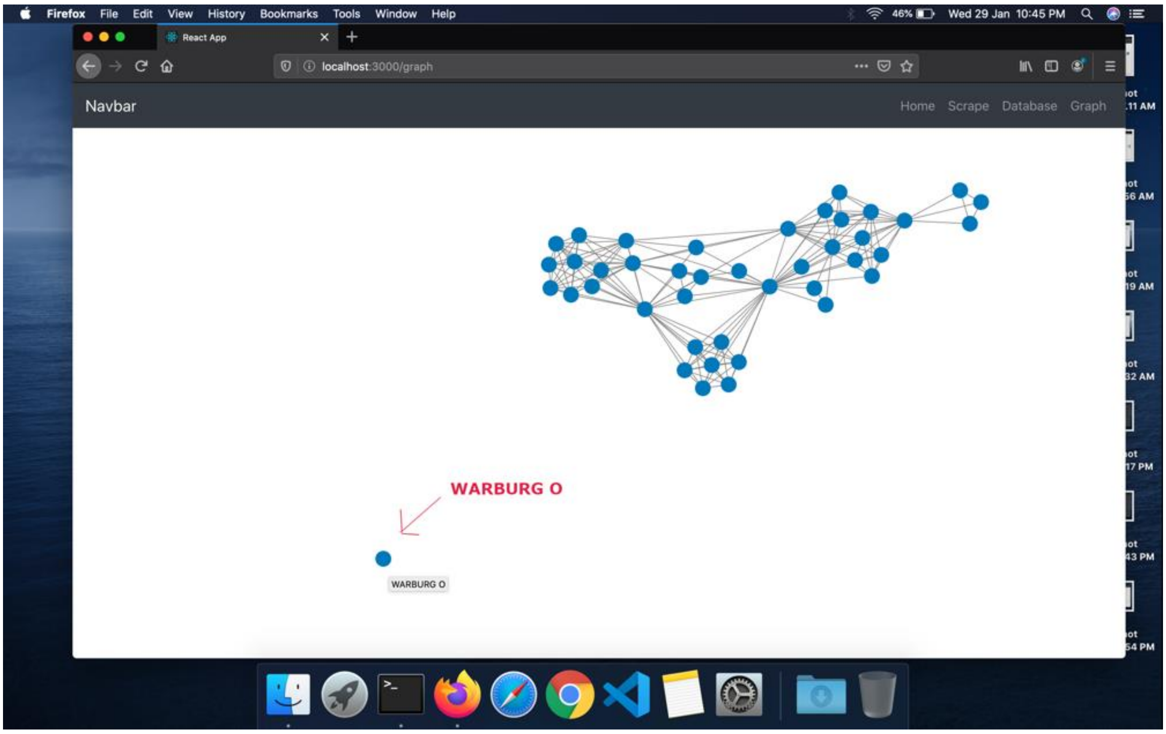

Firefox File Edit View History Bookmarks Tools Window Help
Wed 29 Jan 10:45 PM
React App
localhost:3000/graph
Navbar
Home Scrape Database Graph
WARBURG O
WARBURG O

Fig. 5 Network with author Warburg

Here, we have scraped 250 research papers for keyword cancer. Consider one author named Jemal A (underlined with yellow color in Figure 3). He has contributed to many research papers so in the network, he has many edges connected to other authors. Now consider author WARBURG O (underlined with red color in Figure 3), he has only contributed to one research paper so in the network we can find just the node of him without any edges associated with him.

IV. FUTURE SCOPE

To apply degree centrality, closeness centrality, betweenness centrality and eigenvector centrality algorithm on formed network and to find most influential key opinion leader.

V. CONCLUSION

In this phase of the project we have studied literature related to our project topic and this report has been produced as the result of studied literature. From the literature studied we have proposed system for identification for KOLs in pharmaceutics. It consists of three modules. First one is for data scaping and collection, second is for network formation from the data in which nodes represent authors and links between nodes represent co-authored publications. The final module consists of network analysis algorithms (degree, betweenness, closeness, eigen-vector) using which dominant nodes are identified as KOLs.

REFERENCES

[1] Zhou, H., Zeng, D. and Zhang, C., 2009, June.: Finding leaders from opinion networks In: *2009 IEEE International Conference on Intelligence and Security Informatics* (pp. 266-268). IEEE

[2] Liu, H., Yu, X., Lu, J.: Identifying Top-N Opinion Leaders on Local Social Network. In: Smart and Sustainable City 2013 (ICSSC 2013), IET International Conference on Date 19-20 Aug. 2013.

[3] Wang Z, Du C, Fan J, Xing Y. Ranking influential nodes in social network based on node position and neighborhood Neurocomputing. 2017 Oct 18;260:466-77.

[4] Zhou, H., Zeng, D., Zhang, C.: Finding Leaders from Opinion Networks. In: Intelligence and Security Informatics, 2009. ISI ’09. IEEE International Conference on Date 8-11 June 2009.

[5] Identifying Key opinion leaders using social network analysis. Cognizant System Private Limited’s White Paper on June 2015.

[6] Duan, J., Zeng, J., Luo, B.: Identification of Opinion Leaders Based on User Clustering and Sentiment Analysis. In: 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)

[7] Vojtech Draxl:Web Scraping Data Extraction from websites, Wien, 04.02.2018

[8] Python NetworkX -https://networkx.github.io/

[9] Europe PubMed Central - https://europepmc.org/

[10] React JS - https://reactjs.org/

[11] Redux JS - https://redux.js.org/

[12] D3 JS - https://d3js.org/

[13] Bootstrap - https://getbootstrap.com/

[14] Django-https://www.djangoproject.com/

[15] Scrapy - https://scrapy.org/

[16] Elastic - https://www.elastic.co/