ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

Using Text Mining Algorithm to Track Job Seeker Search Patterns in Ghana

Delali Kwasi Dake[1]

Lecturer, Department of Information and Communication Technology, University of Education, Winneba[1]

ABSTRACT: Data Mining using text analysis is an emerging research area which helps to discover patterns in unstructured text.Text mining can help derive valuable insightful relations from text-based content to a structured format for further analysis.In Ghana, one key focus of any Government is not just the creation of jobs but how to create jobs to cover a wide range of graduates from different subject areas. This job creation agenda in the 21st century needs an analytical approach through the discovery of hidden information. The relevance of this research is to deploy a Text Mining algorithm on job search data obtained from online job aggregators and to determine the most searched keywords from the unstructured text to better understand the graduate unemployment issue. This study uses the process of text summarization with five key text mining extensionsin RapidMinerto predict sectors that Government and Academia can focus on in the job creation strategy.

KEYWORDS –Data Mining;Text Mining;Employment; Prediction; Summarization

I. INTRODUCTION

Huge data is generated in the employment market in Ghana every day. The emergence of Job Search Aggregators has potentially escalating data regarding job search queries with relevant importance to categories of interest. There is a close relationship between the job search trend patterns of graduates and the availability of such jobs in the Ghanaian job market. With the advent of text classification, it is possible to discover relevant search patterns out of this huge data and advise Government on decaying job categories which mostly gives rise to higher searches, the academia on course policy direction and the graduates on employability predictions.

Text mining is the process of obtaining useful and interesting information from unstructured text [1] using statistical and machine learning algorithms. Text mining is a multi-disciplinary field based on information retrieval, data mining, machine learning, statistics and computational linguistics [2]. Text mining techniques are continuously applied in academia, web applications, internet, industry and other related fields [3]. In text mining, words and cluster of words in documents can be analysed with similarities or how they are related to other variables of interest in a data mining project.

Generic process of text mining performs the following steps [4]:

- Collection of unstructured data from different sources available in different file formats such as plain text, web pages, pdf files etc.
- Pre-processing and cleansing operations performed to detect and remove anomalies. Cleansing process captures the real essence of text available and is performed to remove stop words.
- Processing and controlling operations are applied to audit and further clean the data set by automatic processing
- Pattern analysis is implemented by Management Information System
- Information processed in the above steps are used to extract valuable and relevant information from effective and timely decision making and trend analysis [5].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

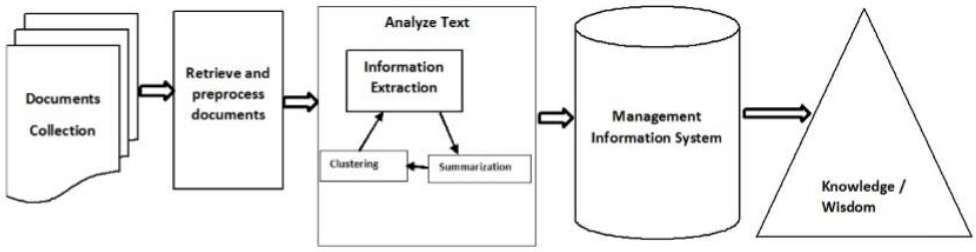


Fig. 1. The Text Mining Process

In this research, Text Mining classification algorithm will be deployed on data obtained from Job Aggregators in Ghana. Relevant job search keywords will be ranked by the algorithm to determine the most and the least searched keywords.

II. PROBLEM STATEMENT

The emergence of job search portals from print media advertisement has created the need for text classification to better understand the graduate unemployment issue. The Ghana Government and Tertiary Institutions have pushed the need for policy directions especially in revamping industrialisation by building factories. The centre of focus is the kind of jobs that can be created by these industries and whether the skilled market set of graduates can benefit. Text classification using tokenization on the already available huge data from Job Aggregators will help expose the need and influence policy direction on the kind of factories government can build in order to cater for graduates from all fields. Mostly, a graduate will only search for a job that is difficult to get in the print media or directly from a website.

III.RELATED WORK

IBM's Technology Watch, developed jointly with Synthema in Italy [6] is a text mining application in the scientific domain. It performs document clustering plus visualization in the form of maps for patent databases and technical publications.

In Customer Relationship Management [7], the most widespread applications are related to the management of the content of client's messages. This kind of analysis helps to reroute unique requests to the targeted service provider or at supplying immediate answers to the most frequently asked questions. Services research has emerged as a green field area for application of advances in computer science, engineering and IT.

One relevant application of text mining is in multilingual processing of natural languages where a good deal of attention must be paid to the languages to be recognised by the system [8]. In this application, a recognition system for Italian and German is built, thus the properties of both languages are important. Therefore, the recognition system must not only cope with dialectal variations, but also with a certain amount of accent by the speaker.

In QUASAR [9] system, user provides a question and this is handed over to the Question Analysis and Passage Retrieval modules. Next, the answer extraction obtains the answer from the expected type, constraints and passages returned by the Question Analysis and Passage retrieval modules. The main objective of the question analysis module is to derive the expected answer type from the question text.

One main application of biomedical text mining is in Named Entity Recognition (NER) systems [10]. The goal is to identify within a collection of text all the instances of a name for a specific type of thing: for example, all of the drug names within a collection of journal articles, or all the gene names and symbols within a collection of MEDLINE abstracts. Hansich and d Bruijn [11] believed that solving this problem allow more complex text-mining tasks to be addressed. The idea is that recognising biological entities in text allows for further extraction of relationships and other information by identifying the key concepts of interest and allowing those concepts to be represented in some consistent, normalised form.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

Choudhary[12], in their research used text mining to uncover patterns, associations and trends from Post Project Reviews (PPRs). A case study of this project was done on two constructions companies by analysing 50 PPR reports collected during the last three years. The results obtained after the text analysis identified problem areas, enhanced processes and improved customer relationships.

IV.METHODOLOGY

In this research, text summarization is the main text mining technique deployed on the job search data. Text summarization deals in collecting and producing concise representation of the original text documents [13]. In pre-processing the search data for summarization, the search text extracted is tokenized to break up the document into discrete bits or tokens. The simplest meaningful token is a word. After this process the tokenized text is filtered to remove stop words. These are common words such as propositions, conjunctions and adverbs. The next step is stemming of the words to their barest minimum in order to keep the core characteristics of words which convey effectively the same meaning. Then we performed n-gramming on the text to group pairs of words which mostly go together. Identifying such pairs will allow intelligent parsing through the text document [14].

For the job categorization section, sixteen (16) categories of job segments are deduced after comparing categories from Jobweb Ghana and Jobberman Ghana, two of the largest Job Aggregators in Ghana.

Accounting/Finance Jobs	Administrative Jobs	Healthcare Jobs	Oil and Gas Jobs	Sales Jobs
Banking Jobs	Customer Service Jobs	Agricultural Jobs	Engineering Jobs	Advertisement/Media Jobs
Logistics Jobs	Human Resource Jobs	Education/Teaching Jobs	NGO Jobs	IT/Telecom Jobs
Construction/Real Estate Jobs				

Tab. 1. Job categories considered for this research

These categories were selected based on the discretion that they have a generic distribution across the Ghanaian job market.

V. RESULTS ANALYSIS

To run the text classifier, a monthsample job search data (20th December 2017 – 17th January 2018) was taken from the Jobweb Ghana jetpack search tracker. The search data was then analysed using the RapidMiner text processing extension.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

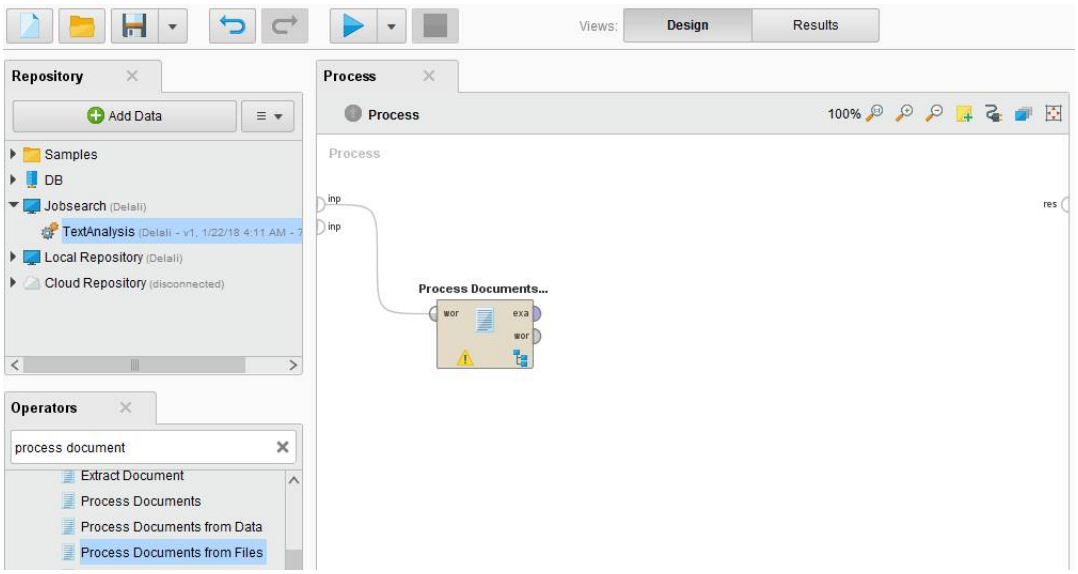


Fig.1.Interface of the Job Search TextAnalysis in RapidMiner

The text data is prepared for analysis as shown in Fig.2using theRapidMiner. A repository is created with a stored procedure. The text data path is added to the Process Documents from Files function from the text extension in RapidMiner.

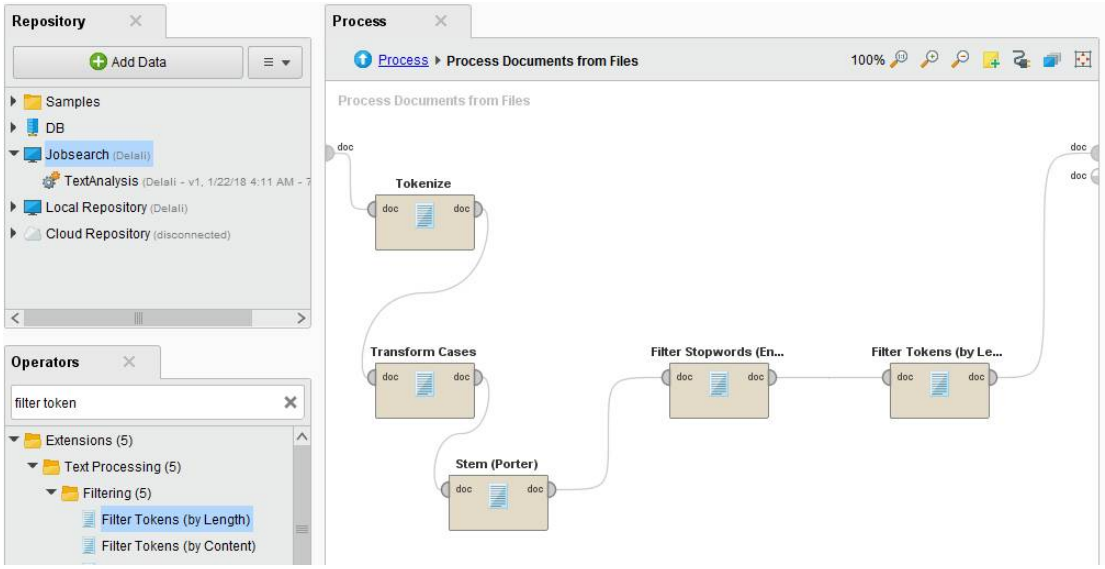


Fig. 3.The Job Search Text Analysis Process in RapidMiner

The operators used for the text analysis as shown in In Fig. 3, were based on the text extension algorithms in RapidMiner. The selected operators includes the following processes: Tokenize, Transform Cases, Stem Procedure, Filter Stopwords and Filter Tokens.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

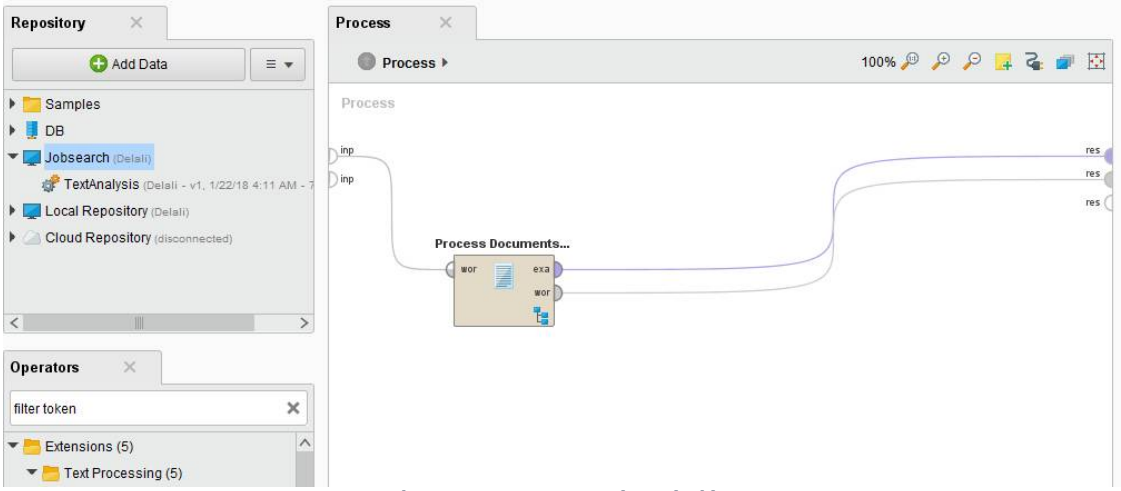
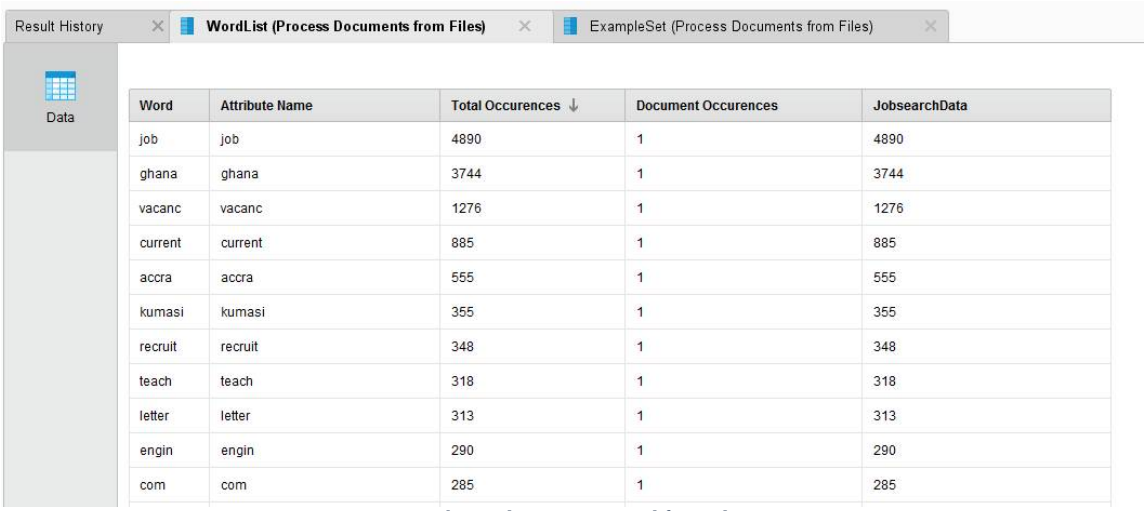


Fig.4. The Process now connected to a double output

The double output operator as shown in Fig. 4, caters for the class file used for the analysis and the job search data obtained from Jobweb Ghana. The two outputs are necessary to determine which class file a particular search data is derived.



Word	Attribute Name	Total Occurences	Document Occurences	JobsearchData
job	job	4890	1	4890
ghana	ghana	3744	1	3744
vacanc	vacanc	1276	1	1276
current	current	885	1	885
accra	accra	555	1	555
kumasi	kumasi	355	1	355
recruit	recruit	348	1	348
teach	teach	318	1	318
letter	letter	313	1	313
engin	engin	290	1	290
com	com	285	1	285

Fig. 5.The results now generated for analysis

The detailed results generated for analysis are shown in Fig. 5. The Attribute Name gives the total occurrences of each Word and the class file represents the document occurrence which is one for this research.

Job Category	**Frequency**	**Job Category**	**Frequency**
Accounting/Finance Jobs	138	Education/Teaching Jobs	473
Banking Jobs	361	IT/Telecom Jobs	2
Logistics Jobs	20	Oil and Gas Jobs	99
Construction/Real Estate	77	Engineering Jobs	411

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

Jobs			
Administrative Jobs	60	NGO Jobs	199
Customer Service Jobs	23	Sales Jobs	25
Human Resource Jobs	19	Advertisement/Media Jobs	26
Healthcare Jobs	147	Agricultural Jobs	62

Tab. 2. Job search categories frequency of distribution

The occurrence of each job search category of focus is tabulated in Tab. 2 and represents the extracted frequency after deploying the text mining algorithm on the class file. This tabulated frequency is converted to a graph format in Fig. 6 for easy analysis.

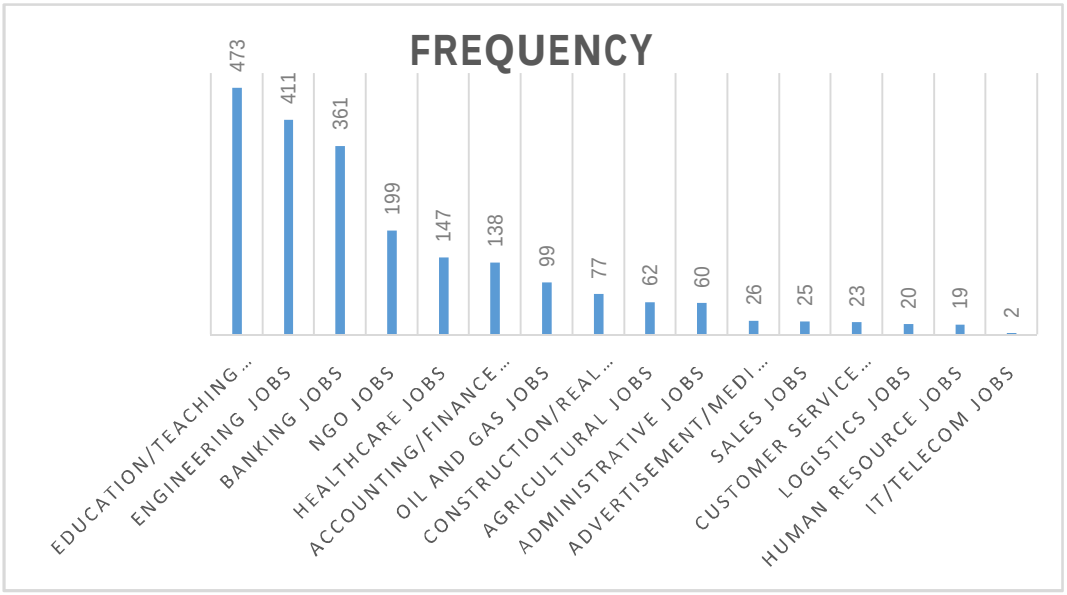


Fig.6.Job Search Frequency

After the text mining, the results from Tab. 2 and Fig.6 shows a significant search relevance for Education, Engineering, Banking, NGO, Healthcare and Accounting Jobs. Such high searches through search engines for these job categories implies scarcity when looking for jobs in those areas. Graduates mostly from such fields or who are really interested in such fields have to deploy search engines hopefully to land a job. More job opportunities should be created in such areas through industrialization or an HR policy in making jobs in such areas public.
The least of the job searches are in IT/Telecom, Human Resources, Logistics, Customer Service and Sales Jobs. It means there are more opportunities in terms of Jobs in those areas and Graduates don't need to use Search platforms to locate such jobs. In such cases, Graduates can use the print media or directly visit the relevant job websites.

VI. CONCLUSSION

In this paper, we deployed a text summarization algorithm on a job search data sample from Jobweb Ghana to determine the most and the least job searches. The text miner results shows a high search relevance in Education, Engineering, Banking, NGO, Healthcare and Accounting categories with least significance searches in IT/Telecom,

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 1, January 2018

Human Resources, Logistics, Customer Service and Sales categories. The most search categories as discussed in the analysis should be prioritized by the Government in revamping job categories in such areas. The text mining algorithm exposed the need to understand the job seeker and create meaningful job opportunities in all sectors.

REFERENCES

1. McCallum, Andrew, and Kamal Nigam, "A comparison of event models for naive bayes text classification",AAAI-98 workshop on learning for text categorization, Vol. 752, 1998.
2. W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining",*Communications of the ACM*, Vo1. 49, no. 9, pp. 76–82, 2006.
3. S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011", Expert Systems with Applications, Vo1. 39, no. 12, pp. 11303–11311, 2012.
4. S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," Expert Systems with Applications, Vo1. 39, no. 12, pp. 11303–11311, 2012.
5. N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining", IEEEtransactionsonknowledgeanddataengineering,vol.24, no. 1, pp. 30–44, 2012.
6. Tan, Ah-Hwee. "Text mining: The state of the art and the challenges",*Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, Vol. 8, 1999.
7. Gupta, Vishal, and Gurpreet S. Lehal. "A survey of text mining techniques and applications", Journal *of emerging technologies in web intelligence* 1.1, pp. 60-67, 2009.
8. U. Ackermann, B. Angelini, F. Brugnara, M. Federico ,D. Giuliani, R. Gretter, G. Lazzari and H. Niemann ,"SpeeData: Multilingual Spoken Data Entry", International Conference, IEEE, Trento, Italy, pp. 2211-2215, 2009.
9. Emilio Sanchis, DavideBuscaldi, Sergio Grau, Lluis Hurtado and David Griol, "SPOKEN QA BASED ON A PASSAGE RETRIEVAL ENGINE", Proceedings of IEEE international conference, pp. 62-65, 2006.
10. De Bruijn, B. and Martin, J., 'Getting to the (c)ore of knowledge: Mining biomedicalliterature', Int. J. Med. Inf., Vol. 67, pp.7–18, 2002.
11. Hanisch, D., Fluck, J., Mevissen, H. T. and Zimmer, R., 'Playing biology's name game: Identifying protein names in scientific text', in 'Proceedings of the 8th Pacific Symposium on Biocomputing Hawaii, pp. 403–414, 2003.
12. CHOUDHARY,A.Ket.al., 'The needs and benefits of Text Mining applications on Post-Project Reviews', Computers in Industry conference, pp.728-740, 2002.
13. B. A. Mukhedkar, D. Sakhare, and R. Kumar, "Pragmatic analysis based document summarization", International Journal of Computer Science and Information Security, Vol. 14, no. 4, pp. 145-146, 2016.
14. Bala Deshpande (2012, January 14). 3 ways to use text mining with RapidMiner to juice up your job search. Retrieved from http://www.simafore.com/blog/bid/111839/3-ways-to-use-text-mining-with-RapidMiner-to-juice-up-your-job-search