# A Semi-Automated Analytical Methodology for Classification of Professionals for Big Data Job Roles

Abhiraj P
Department of Information Technology
Government Engineering College Barton hill
Trivandrum

Vijayanand K S
Department of Information Technology
Government Engineering College Barton hill
Trivandrum

***Abstract*— Nowadays the Big data is playing a major role in the IT industry. Apparently the number of job opportunities in this area is rapidly increasing. The job market is unaware of the required skills to find the niche for this professions. The HR recruiters of the organization find difficulty for identifying suitable professionals for various roles related to big data .In this paper we propose a semi-automated analytical methodology for the classification of big data related job roles. By analyzing a large amount of real world job post published online, we classify the big data jobs into four families. Then recognize the nine groups of big data skill set by using topic modelling algorithm, then mapping the skill set to the job families according to the demand by the industry. From this structured classification of job families and skill set we assembled a semi-automated analytical methodology using machine learning algorithm and expert judgment for finding the suitable professionals to fit the big data job roles .The proposed method is applied over the resume of the professionals and the results obtained are confirmed with the manual classification by various HR experts.**

***Keywords:- Predictive system, machine learning, big data, data analytics, topic modelling,***

## INTRODUCTION

Big data has become one of the prominent player in the IT industry. Big data phenomenon can be defined as information assets characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value[1].The challenges in handling Big Data because of the enlarged data size include capturing, storage, analysis, search, transfer, querying, visualization updating and maintaining information privacy. Now it is also being used to refer user behavior analytic, predictive analytic, and certain other types of advanced data analytic methods through which value is extracted from data[2].Big data analysis helps in help finding modern trends of business trends, disease prevention and also fight crime. Data analytic is now being effectively used by medical practitioners, entrepreneurs and scientists alike for meeting daily challenges with humongous data sets in the world of Internet search, urban informatics, and business informatics, e-Science work, like meteorology, genomics, physics simulations, biology connectives, and environmental research. The multifaceted nature of Big Data facilitate numerous job opportunities that any aspiring individual can grab for a successful future and bright prospect. Big Data provides tailor made solutions for business problems across multiple sectors from health care to investment banking, gas to IT and oil, and insurance to education.

Big Data jobs have been created in almost all industries like marketing, banking/investment banking, finance, games etc. The number Big Data jobs have increased by about 212% in the last five years. But there is a lack of skilled personnel who can actually and competently fill these positions. Employers find it difficult to fill in these expert categories and has an effect on their salaries. It has been noticed that a median salary is that of $55000. When compared to other IT-related roles, it is indeed very high. Jobs in Big Data are forecast to be rising by 160% between 2017 and 2023.According to IDC, the Big Data market would be worth $46.34 billion by 2018, its a technology and associated services market is likely to grow at a compound annual growth rate (CAGR)[16] of 23.1% from 2013 to 2020. The annual spends might reach $48 billion in 2020. Hadoop is predicted to grow at 58.2% CAGR between 2013 and 2020. The IDC predictions are positive.

The rapid expansion of big data analytics is forcing companies to rethink their human resource need. Employers find it difficult to fill in these expert categories with suitable professionals. It is unclear which types of job roles and skills constitute this area. This confusion makes the HR recruiters to find niche profiles for this posts. As a solution for this we propose a semi-automated analytical methodology for big data professionals. This will help the HR recruiters and managers to solve this crisis. The paper is organized as follows; Literature review, methodology we have used, result analysis and summarizes our conclusions and suggests future extensions to the current work

## 2. LITERATURE SURVEY

Steven Miller[4] states that the big data analytics talent discussion has largely focused on data scientists. As its importance is going to the peak of networking, every professional occupation and universities should focus on employees and students with skills to cope with the era of big data. Focusing on big data jobs is more important now which includes information strategists, information system professionals and data governance and ethics professionals

. The information strategist: There is an increased need for graduates with both business and industry acumen. University data management courses concern themselves with technical issues and hence business-focused data skills are largely developed on the job

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

Big data information systems professionals: Information systems curricular need to evolve quickly to better prepare students for the emerging jobs of data professionals and roles

Data governance and ethics professionals: Data governance and data ethics have their less importance like they are offered to students as elective courses or are not been fully developed by universities.

Provost, F., Fawcett,[3] in their work states that what exactly data science is as it is umbrella term comprising all data science jobs . The importance of data science and data scientists is rapidly increasing and hence its principles should be discussed explicitly in order to realize the real potential of data science

Brynjolfsson, E., Hitt, L, M., Kim, [5] H.H in their work states that how the data driven decision making impact on firms aspect. The effect of data driven decision making on the productivity do not appear to be due to reverse causality which is found through instrumental variable Some of the most interesting Big Data roles according to Digividya[2] are s methods and our results provide some of the first large scale data on the direct connection between data-driven decision making and firm performance

Some of the most interesting Big Data roles according to Digividya [2] are

Data analyst: They are typical problem solvers and they analyze various data systems, create programmed systems for information retrieval from a database and compile reports

Database administrator: They manage the daily functioning of a database which includes controlling modifications and updates, maintaining backups and making sure that the database remains stable.

Data scientist: They needs to dive into raw data, in-depth analysis and presentation of findings to the business leaders for the latter to make smart decisions for achieving business goals.

Data architect: They create data workflows; design and test new database prototype sand also makes database solutions from business directives base manager: They are able to lead data items, maintain an entire database environment and also the standard management duties of managing people and departmental budget. Big data engineer: They need to stalwart communicators who not only understand the major goals of the company but also use data to achieve these goals. The mediate between data scientists and business executives so that they help the engineering team in processing data to meet business goals, evaluate new data sources and handle copious amounts of raw data. Baojan Ma, Nan Zhang, Guannan Liu, Liangqiang Li, Hua Yuan[6] stated in their work Semantic search for public opinions on urban affairs, a probabilistic topic modeling based approach used similar classification criteria that needed for this work. In order to search for relevant public opinions among unwanted ones, the office can conduct keywords rather than manual read or summarizing retrieved results through Latant Dirichlet Allocation (LDA) where prepossessing subject clustering for the comment data is done based on a probabilistic topic modeling approach while a semantic search tool is not yet routinely employed in various departments, the users are more satisfied with this SS-LDA method. Sergio Moro, Paulo Cortez, Paulo Rita[7] have used a text mining approach using LDA in their works, which resulted in several topics grouping articles in which each of those topics are characterized by three most relevant terms. The intrinsic limitations of clustering algorithms such as LDA have lead efforts toward validating the hypotheses for relations between the several terms and corresponding trends

SeanGerrish,ChongWang,DavidM,Blei[8]state that Probabilistic topic models are a popular tool for the unsupervised analysis of text, providing both a predictive model of future text and a latent topic representation of the corpus where the latent space is used to check models, summarize corpus and guide exploration of its contents. New quantitative methods for measuring semantic meaning in inferred topics are presented in this paper. The measures we develop here have a possibility of being incorporate with human judgments into the model-learning framework or creating a computational proxy that simulates human judgments

Ingo Feinerer, Kurt Hornik, David Meyer[9], is analyzing the R infrastructure and its usage in text mining which is widely used discipline utilizing statistical and machine learning methods. We presents the tm package for this text mining within R techniques for count-based analysis method, text clustering, text classification and string kernels and also interfacing with other open source toolkits like Weka or open NLP into the available technology in R, offering further methods for tokenization, stemming, sentence detection and part of speech tagging. Thinking on integrating tm with lsa package, we are working on memory-efficient clustering techniques in R to handle highly dimensional sparse matrices as found in larger text mining case studies. tm will be among the first to take advantage of new technology as researches are going on in analyzing large data sets by using sparse data structures.

## 3 METHODOLOGY

Predictive system for getting suitable human resource for big data professionals is done through the following steps by combining a series of existing analytical practices .First we have to gather the data set of substantial amount of related on line job posts, by means of web scraping techniques. Second, we will define the job families by expert judgment. Third, by applying suitable topic modeling algorithm relevant skill set is identified Fourth, mapping of skill set to job families and a presence matrix is formed. Finally resume of the professionals are analyzed with the presence matrix and suitable professionals are predicted.

### 3.1 Data set gathering

By analyzing a various job site across worldwide web, from https://www.dice.com an American based job portal retrieved the job post and the skillsets. Unlike from other web portals dice organize the job post with the rear roles are organized in alphabetic order .It will also have better With the help of the Spider a web scraping tool retrieved the data from the website. WebCrawlers can automatically retrieve information and store in the desired locations [10].22000+ online post where retrieved by web crawlers. The data pre-processing has to be done in this data set. Tokenization in which punctuation are removed and text are split in simple sentences then to words. All are been converted to lowercase. Stop words are removed. Unlike from other data preprocessing the words with less than 3 letters are not removed

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

## 3.2 JOB FAMILY IDENTIFICATION

Analyzing the data set with the help of literature discussed in previous section classified the job post into 4 families. Defining the job families in the way that they are non-ambiguous in nature. Job post descriptions are retrieved from the data and classified by expert judgment .The four job families are;

1. Business Analyst (Project Manager, Business Analyst, Product Manager, Program Manager)

2. Data Scientist (Data Engineer Data Scientist, Data Analyst, Data Consultant)

3. Developer (Software Engineer, Java Developer, Hadoop Developer, Software Developer)

4. Engineer (Data Architect, DevOps Engineer, Solution Architect, Systems Engineer

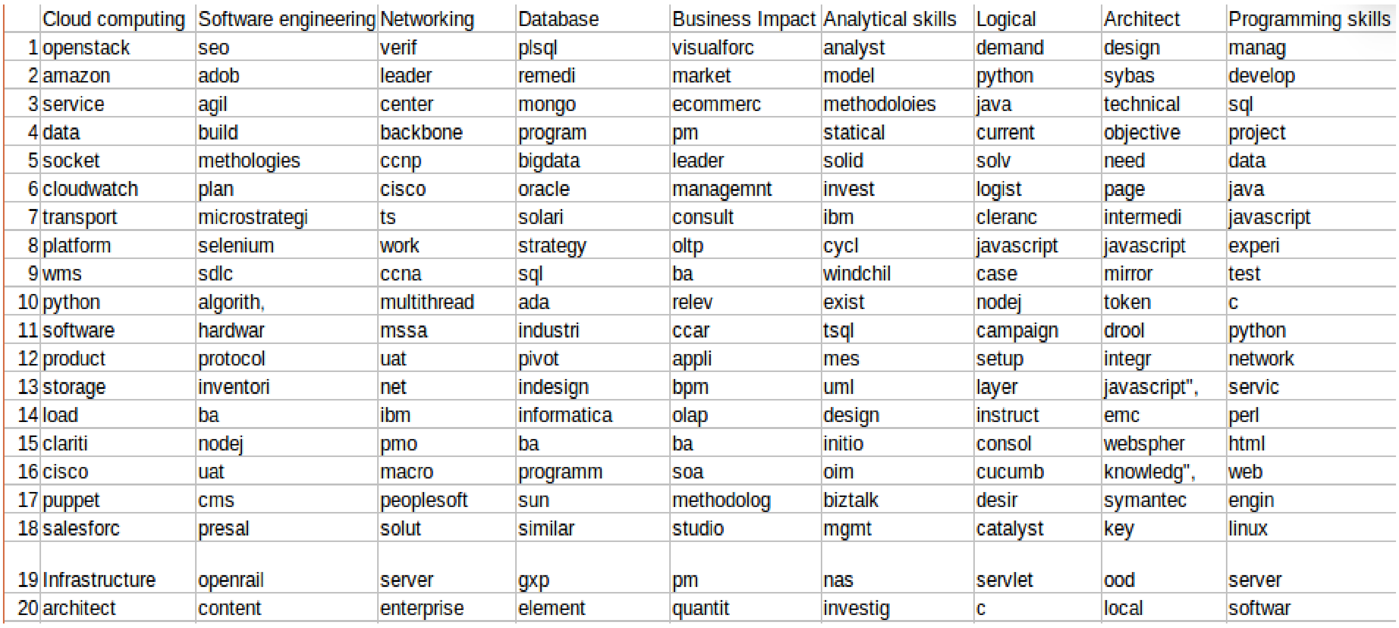
With the help of expert analyzing the job post from the data and are included in the four families and job descriptions are retrieved

## 3.3 SKILL SET CLASSIFICATION

From the data set the job skills are gathered. The skills are to be clustered for the homogeneous families' .Forming a number of skill set that compress of a number of skills. The skills are to be clustered into skill sets in a way that multiple skillset are to be required for a single job role and the skill set comprises of different skills in different proportion. Also, one skill may or may not be in multiple skill set .Hierarchical clustering algorithms and traditional algorithms like K means cannot be applied.[11]With help of literature in chapter 2.2 and 2.3 algorithm is confirmed. Topic modelling algorithm can be used in this context which uses mixed membership models. To identify various skill sets within job posts, we decided to adopt the mixed-membership model Latent Dirichlet Allocation, LDA (12)

### 3.3.1 LDA

LDA is a statistical model of document collections that tries to capture the intuition that documents exhibit multiple topics and it is most easily described by its generative process, the imaginary random process by which the model assumes the documents arose.[13]



| | Cloud computing | Software engineering | Networking | Database | Business Impact | Analytical skills | Logical | Architect | Programming skills |
|---|---|---|---|---|---|---|---|---|---|
| 1 | openstack | seo | verif | plsql | visualforc | analyst | demand | design | manag |
| 2 | amazon | adob | leader | remedi | market | model | python | sybas | develop |
| 3 | service | agil | center | mongo | ecommerc | methodoloies | java | technical | sql |
| 4 | data | build | backbone | program | pm | statical | current | objective | project |
| 5 | socket | methologies | ccnp | bigdata | leader | solid | solv | need | data |
| 6 | cloudwatch | plan | cisco | oracle | managemnt | invest | logist | page | java |
| 7 | transport | microstrategi | ts | solari | consult | ibm | cleranc | intermedi | javascript |
| 8 | platform | selenium | work | strategy | oltp | cycl | javascript | javascript | experi |
| 9 | wms | sdlc | ccna | sql | ba | windchil | case | mirror | test |
| 10 | python | algorith, | multithread | ada | relev | exist | nodej | token | c |
| 11 | software | hardwar | mssa | industri | ccar | tsql | campaign | drool | python |
| 12 | product | protocol | uat | pivot | appli | mes | setup | integr | network |
| 13 | storage | inventori | net | indesign | bpm | uml | layer | javascript", | servic |
| 14 | load | ba | ibm | informatica | olap | design | instruct | emc | perl |
| 15 | clariti | nodej | pmo | ba | ba | initio | consol | webspher | html |
| 16 | cisco | uat | macro | programm | soa | oim | cucumb | knowledg", | web |
| 17 | puppet | cms | peoplesoft | sun | methodolog | biztalk | desir | symantec | engin |
| 18 | salesforc | presal | solut | similar | studio | mgmt | catalyst | key | linux |
| 19 | Infrastructure | openrail | server | gxp | pm | nas | servlet | ood | server |
| 20 | architect | content | enterprise | element | quantit | investig | c | local | softwar |

Figure 1: Classification of skill set

The two-stage process in which the topic model is generated are:

- p(t|d) is the probability distribution of topics in documents
- p(w|t) is the probability distribution of words in topics
- Probability of a word given document [14]

$$p(w|d)= \sum_{tT} p(w \vee d)p(t \vee d) \qquad (1)$$

LDA is uploaded over the data set, where k value is taken as 9 and number of skill per skill set as 20.

## 3.4 PRESCENCE MATRIX FORMATION BY MAPPING OF SKILL SETS BY JOB FAMILY.

A matrix shows the presence of elements in the relevant rows is formed by the analytical mapping between the descriptions about the job families discussed in the above Sections withe the output of LDA Fig 1.Using the analytical functions [15] in R the matrix is formed and table 1 shows the matrix. It provide a structural classification for the big data job roles

## 3.5 RESUME MAPPING OF PROFESSIONALS

Resumes of the professionals are then mapped with the structured classification of the job roles and skills using the analytical functions in R. The skill set of the professionals are needed and it is mapped with the LDA output fig.1 and results are formed. The mapping result is taken for analysis. The result

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

of the mapping can take and to be analyzed niche persons for the post can be found.

## 4 RESULT ANALYSIS

The result will give an idea about the skills needed for the job and how much the skills in the profile matches with the required skills .Human resource recruiters will find easy with the structured classification to recruit the suitable one.

About 300 resumes of various professionals are used in the system and result are analyzed .For each resume the system will generate the presence matrix. Each matrix will show about the skill sets that the professionals possess. Table 2 shows the result of four candidates resume mapping result .As per the expert judgment candidate 1 is suitable for business analyze Figure 2 shows the graphical representation of business analyst and candidate 1 to the skill sets .Both look similar so it can be predicted that candidate 1 is business analyst. Similarly the other resumes are mapped and the result with the expert judgment is analyzed

Precision and recall values of the system is analyzed and is recorded as.

Precision = 0.91.

- Recall = 0.91
- F1 Score = 0.9
- Accuracy = 0.93

with result in table 1 .Comparing the mapping result of the both Result are analyzed separately for each job families and accuracy, precision ,f1 score and recall is recorded in table 2

Table 1: Job Families with Skill Set

| | Cloud | Software | Network | DB | Business Impact | Analytic | Logic | Architecture | Programing |
|---|---|---|---|---|---|---|---|---|---|
| Business Analyst | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 2 |
| Data Scientist | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| Data Developer | 4 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 2 |
| Data Engineer | 4 | 1 | 3 | 3 | 1 | 1 | 1 | 2 | 3 |

Table 2: Result of each job family

| | Precision | Recall | F1 score | Accuracy |
|---|---|---|---|---|
| Business Analyst | 1 | 1 | 1 | 1 |
| Data Scientist | 0.8 | 1 | .88 | .92 |
| Data Developer | 0.75 | 0.75 | 0.75 | 0.84 |
| Data Engineer | 0.8 | 0.8 | 0.8 | 0.85 |

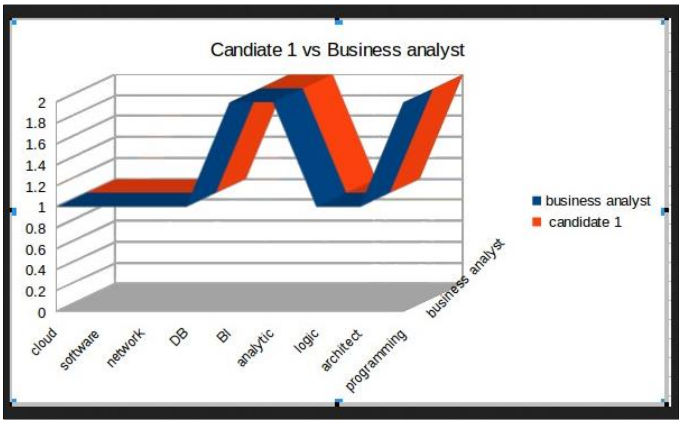


Figure 2: candidate 1 and business analyst

## 5 CONCLUSION AND FUTURE SCOPE

In this paper we assembled a semi-automated analytical process for classification of big data professionals .It will provide answer to managers on which new talent they need for the firm .It will also help to provide a knowledge regarding the upgrading of the skills of their current human resource .Functional managers can use our results to build more meaningful and structured job descriptions for hiring Provide useful guidance to educational institutions. This system will provide clarity upon the features of job roles. Regardless of intuition managers and HR recruiters can obtain experts for the profession. Job seekers will get clear idea regarding their skills upon the market needs. By changing the data set it can be implemented in various Fields such as syllabus framing for universities, training for professional with in the rm, prediction for other job posts. Accuracy of the system can be improved by enlarging data set. Instead of LDA deep learning algorithms may be implemented to improve precision values

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

# REFERENCE

[1] De Mauro, A., Greco, M., Grimaldi, M. (2016).” A formal definition of Big Data based on its essential features. ‘Library Review,” 65, 122–135. http://www. emeraldinsight.com/doi/abs/10.1108/LR-06-2015-0061. (accessed April 1, 2016). .

[2] https://www.digitalvidya.com/blog/big-data-and-big-data-jobs-thejobs-of-the-future-are-here/ .

[3] Provost, F., Fawcett, T. (2013).” Data science and its relationship to Big Data and data-driven decision making. Data Science Big Data, 1” , 5159. doi: 10.1089/big.2013.1508.

[4] Miller, S. (2014). “Collaborative approaches needed to close the Big Data skills gap. Journal of Organization Design, “3 , 2630. doi: 10.7146/jod.3.1.9823

[5] Brynjolfsson, E., Hitt, L. M., Kim, H. H. (2011). “Strength in numbers: how does data-driven decisionmaking affect firm performance? SSRN ElectronicJournal ,” 128. doi: 10.2139/ssrn.1819486

[6] Ma, B., Zhang, N., Liu, G., Li, L., Yuan, H. (2016).Semantic search for public opinions on urban affairs: A probabilis tic topic modeling based approach. Information Processing Management, 52 , 430â445. doi: 10.1016/j.ipm.2015.10.004 .

[7] Moro, S. M. C., Cortez, P. A. R., Rita, P. M. R. F. (2014). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Systems with Applications, 42 , 1314â1324. doi: 10.1016/j.eswa.2014.09.024 .

[8] Blei, D. M. (2012). Introduction to probabilistic topic models. Communi- cations of the ACM, 55 , 77â84. doi: 10.1145/2133806.2133826 .

[9] Feinerer, I., Hornik, K., Meyer, D. (2008). Text min- ing infrastructure in R. Journal of Statistical Software, 25

[10] Kobayashi, M., Takeda, K. (20 0 0). Information retrieval on the web. ACM Computing Surveys, 32 , 144–173. doi: 10.1145/358923.358934 . Ma, B., Zhang, N., Liu, G., Li, L., Yuan, H. (2016). Semantic search for public opinions on urban affairs: A probabilis tic topic modeling-based approach. Information Processing Management, 52 , 430–445. doi: 10.1016/j.ipm.2015.10.004 .

[11] Airoldi, E. M., Blei, D. M., Fienberg, S. E., Xing, E. P. (2008). Mixed membership stochastic blockmodels. Journal of Machine Learning Research, 9 , 1981–2014. doi: 10.1016/j.bbi.2008.05.010

[12] Blei, D. M. (2012). Introduction to probabilistic topic models. Communi- cations of the ACM, 55 , 77â84. doi: 10.1145/2133806.2133826 .

[13] https://medium.com/@tomar.ankur287/topic-modeling-using-lda-and-gibbs-sampling-explained49d49b3d1045

[14] “Surveying a suite of algorithms that offer a solution to managing large document archives””.By DaviD m.

[15] https://link.springer.com/article/10.1007/s00153-016-0483-x

[16] https://www.globenewswire.com/news-release/2019/07/12/1881968/0/en/High-Availability-Server-Marketto-Expand-at-13-5-CAGR-Rising-Demand-for-Big-Data-Analytics-Fosters-Growth-says-Fortune-BusinessInsights .html

(This work is licensed under a Creative Commons Attribution 4.0 International License.)