# Skills Requirements of Business Data Analytics and Data Science Jobs: A Comparative Analysis

Zinovy Radovilsky*
*California State University, East Bay, Hayward, California, USA*

Vishwanath Hegde
*California State University, East Bay, Hayward, California, USA*

Anuja Acharya
*California State University, East Bay, Hayward, California, USA*

Uma Uma
*California State University, East Bay, Hayward, California, USA*

This research paper identifies and compares knowledge domains and skills that characterize Business Data Analytics (BDA) and Data Science (DS) professions. By reviewing the literature sources, we develop a model to systemize the expected knowledge domains and skills for these two fields. Then we collected primary BDA and DS job posting data from online job-related websites and analyzed this data using text mining methods. The results of the text mining analysis reveal some similarities as well as important differences between the required knowledge domains and groups of skills for BDA and DS. These results provide insights that are vital in designing curriculum and training in the evolving business analytics and data sciences areas, and also enable professionals to sharpen their skills that are aligned with job market requirements.

* Corresponding Author. E-mail address: zinovy.radovilsky@csueastbay.edu

## I. INTRODUCTION

Big data analytics is increasingly shaping the way organizations manage their decision-making processes, create new services and products, and gain competitive advantage. The demand for analytics and its applications is driving the expansion of information technology and associated analytical methods and tools (Kennedy et al., 2015). As a result, a number of analytics professions emerged, among which business analytics, business data analytics (in this paper, we use the term "business data analytics" or BDA) and data science (DS) professions are the most prominent ones. According to a Forbes article (Columbus, 2015), jobs related to big data analytics represent one of the fastest growing segments of the overall job market. This article also stated that the number of people using the analytics titles in LinkedIn has

doubled in 2011-2015. Based on a MarketWatch.com survey, the data scientist job was ranked in 2017 as #1 Best Job in America, and the analytics manager job was ranked as #4 (Fottrell, 2017).

Due to the fact that the big data fields like BDA and DS are still evolving over recent time, there is an ongoing discussion on how to define each of these areas and what they really represent. As Miller (2014) stated "Exactly what a data scientist is and what skill set is required is a subject of intense debate." A number of academic and practitioner papers define these terms in a variety of ways. In some papers, the authors do not distinguish between the terms "business data analytics" and "data science" as separate entities (Henry and Venkatraman, 2015; Miller, 2014; Power, 2014; Kemper and Mathews, 2017; Priestley, 2013). For example, Henry and Venkatraman (2015), state that "Big data infrastructure comprises of a big data repository, data analytics software, and the data scientists. Data scientists use their knowledge and expertise in the growing field of data sciences, which need strong mathematical, statistical, and information technology." This explanation does not contemplate a business data analytics profession to be an entity separate from the data scientist profession. In a similar way, Kemper and Mathews (2017) point out that all work in data analytics is done by "the Data Scientist." They define the "data science" term in conjunction with the "data analytics" term using the definition provided by the National Institute of Standards and Technology (NIST, 2015): "data science is the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing; and the analytics process is the synthesis of knowledge from information."

In many academic and practitioner literature sources, "business data analytics" or simply "business analytics," and "data science" terms are defined separately (Phelps and Szabat, 2017; Bichler et al., 2017; Kambatla et al., 2014; Aasheim et al., 2015). This separation between BDA and DS is also evident by the relevant job postings in the market. Summarizing various definitions of "data analytics," specifically in conjunction with "business analytics," Phelps and Szabat (2017) concluded that the BDA definitions highlight four key aspects: data, technology, statistical and quantitative analysis, and decision-making support. They also stated that data science definitions include four key aspects: data, databases, computer systems (taking a set of inputs, processing them to produce outputs), and advanced analysis (e.g., machine learning, natural language processing, artificial intelligence) including statistics. This summarization of BDA and DS definitions provides a clear difference between the two terms. DS emphasizes computer systems, algorithms, and computer programming skills, whereas BDA has a substantial focus on statistical and quantitative analysis of data, and decision-making support.

In addition to defining the BDA and DS terms, there is also an ongoing discussion in the academic and practitioner literature on the knowledge and skills that are required for the two professions (Phelps and Szabat, 2017; Kemper and Mathews, 2017; Cegieelski, and Jones-Farmer, 2016; Kambatla et al., 2014). This discussion is specifically important at the time when the job market for these professions is rapidly growing, and, at the same time, an extensive number of academic and professional programs in BDA and DS have been recently introduced by a large number of universities and colleges. A clear and in-depth understanding of relevant knowledge and skills for BDA and DS fields may

tremendously benefits several important stakeholders: (a) business organizations creating job requirements and qualifications for big data professionals; (b) current and future specialists that would like to develop and/or improve their skills in the BDA and DS fields, and (c) academic and professional institutions that are creating or already teaching analytics and data science programs.

In this paper, our goal is to identify, systemize and compare groups of knowledge domains and skills that characterize BDA and DS professions. For identifying these groups, we analyze existing in the market business and data analysts as well as data scientist job offerings using methods of text mining analytics. The objectives of this research study include the following:

- Based on literature sources, identify the expected knowledge domains and skills for BDA and DS fields.
- Systemize the groups of skills required for BDA and DS fields by analyzing primary job posting data.
- Compare the BDA and DS knowledge domains and groups of skills to identify their similarities and differences.
- Provide important insights for companies and education of BDA and DS professionals.

The structure of this paper is as follows. After the Introduction section I, we present in section II, a literature review of existing research on BDA and DS knowledge and skills. In section III, we describe research methodology, data collection, and analytical approaches applied in this research. In section IV, we provide, using text mining methods, analysis of BDA and DS job requirements and present the results of this analysis. We finish the paper with the conclusions in section V.

## II. LITERATURE REVIEW

As the entire field of big data analytics is relatively new and developing, the existing research pertaining to the knowledge domains and skills required in big data analytics professions is also quite recent and evolving. In this section we analyze existing literature sources pertaining to: (a) methodology and actual professional skills essential to the BDA and DS fields, and (b) comparison of the knowledge domains and skills of these two fields.

As a professional field that involves utilization of technology, quantitative/statistical methods, and business-related functionality, BDA and DS involve an extensive variety of knowledge and skills traits. This necessitates developing a methodology for systemizing those skills and present them in coherent and understandable ways. We identified a number of papers that describe such methodologies and associated skills in BDA (Dubey and Gunasekaran, 2015; Gorman and Klimberg, 2014; Cegieelski, and Jones-Farmer, 2016; De Mauro et al., 2017; Ellway et al., 2014; Wymbs, 2016) and DS (Power, 2016; Mills et al, 2016; Schoenherr and Speier-Pero, 2015; Song and Zhu, 2016; Tang and Sae-Lim, 2016; Kemper and Mathews, 2017; Englmeier and Murtagh, 2017).

For BDA, Dubey and Gunasekaran (2015), based on a survey of data and business analytics executives, develop a framework for education and training in BDA by clustering all skills into two major groups: hard skills and soft skills. The hard skills include some statistical/quantitative methods like statistics, forecasting and optimization, and also functional areas such as finance, financial accounting, and marketing. The soft skills contain some managerial and communication attributes like leadership abilities, team skills,

listening, interpersonal skills, etc., which can be relevant to any business profession. Overall, this research, being a good initial approach in systemizing the BDA skills, does not address all of them, e.g., technology-related skills. From the latter standpoint, a more detailed framework, which is based on benchmarking of BDA academic programs (Gorman and Klimberg, 2014), includes not only statistics and quantitative groups of skills, but also information systems and business intelligence group. This group of skills incorporates skills in big data technologies, enterprise systems, data marts, databases, online analytical processing (OLAP) and programming, which are overlapped with quantitative methods, forecasting and data mining. Being more detailed, the Gorman and Klimberg (2014), however, do not address managerial and communication skills associated with BDA.

Cegieelski, and Jones-Farmer (2016) present an even more comprehensive research of the BDA knowledge, skills, and abilities by analyzing a large number of undergraduate business analytics programs. This research identifies, using expert ranking of those programs, the important components of knowledge, skills, and ability in three main knowledge domains required for the BDA profession, i.e.:

- Business domain – includes communication skills, project management, client orientation, time management, etc.
- Analytical domain – incorporates problem definition and problems-solving skills, predictive analysis, integrative analysis, and some others.
- Technical domain – is divided into three subgroups:
  - Applications – Excel, SAS, Tableau, etc.
  - Languages – R, Python, SQL, and Java
  - Infrastructure – Hadoop, Casandra, Oracle, Linux, MapReduce, Hive, and Pig.

We now continue this literature review with a discussion of methodologies related to DS knowledge and skills. In analyzing management information systems programs, Mills et al. (2016) adopted four pillars of skills in educating data scientists including:

- Data preprocessing, storage, and retrieval – e.g., NoSQL, data modeling, and data warehousing skills.
- Data exploration – statistical analysis and visualization.
- Analytical models and algorithms – machine learning, data mining, and natural language processing.
- Data product – data and information organization, knowledge representation, and application development.

This framework is somewhat similar to those for BDA, but distinguishes itself by introducing the "data product" pillar, which includes, among other things, application development. Conversely, the framework provided by these authors (Mills et al., 2016) does not incorporate a business domain, which is quite common among BDA groups of skills.

Schoenherr and Speier-Pero (2015), in their research on big data analytics in supply chain management, establish three groups of skill sets for DS, which include: (a) enterprise business processes and decision making, (b) analytical and modeling tools, and (c) data management. Although these three groups describe DS skills, they also come close and can be easily adopted for a BDA skills framework.

Another research paper on identifying DS skills (Song and Zhu, 2016) presents the DS framework as a combination of four components:

- Big Data Infrastructure – includes big data technologies such as Hadoop ecosystems, NoSQL databases, in-memory computing and cloud computing.
- Big Data Analytics Lifecycle – covers all stages of data analysis, data understanding, data preparation, and integration, model-building, evaluation, deployment, and monitoring.
- Data Management Skills – incorporate traditional data modeling and relational database knowledge.
- Behavioral disciplines – contain soft skills related to people and business such as abilities to think critically, to communicate with domain experts, and to make project outcomes relevant to business.

This DS framework also comes close to the BDA groups of skills, specifically, in terms of data management skills and behavioral disciplines. With respect to the question on distinguishing between BDA and DS skills, we identify only a handful number of research papers, which consider the differences between the groups of skills in these fields (Aasheim et al., 2015; Debortolli et al., 2014; Liberatore and Luo, 2013). For example, Aasheim et al. (2015) discusses the similarities and differences between the BDA and DS undergraduate programs based on the course descriptions. The authors identified that these two types of programs have a number of similarities, specifically, in terms of the topics taught. However, they also recognized several important differences in the program curriculum: "First, DS programs require additional mathematics courses – at least through linear algebra, which is typically after Calculus II, and most require discrete math. Second, they all require at least nine hours of programming courses and at least two statistics courses" (Aasheim et al., 2015). The limitation of these results with respect to our study is that it does not cover all potential BDA and DS programs, and also do not consider, as has been described before, the real job market requirements for these professions.

We identified, however, several practitioner papers that describe the most common DS skills based on analyzing DS job positions from various websites. For example, Junco (2017) used text mining to extract DS skills from Glassdoor.com, a popular website with various job positions (see Figure 1).

Gonzalez (2017) provided a list of most common DS skills based on the 2015 job data from LinkedIn. Based on web scabbing tools, Steinweg-Woods (2015) identified DS skill lists for the job markets in various cities (San Francisco, New York, Chicago, etc.) and nationwide. All of these lists contain just the technical skills, and do not have any other relevant DS skills, e.g., analytical and business skills pertaining to the DS profession.

Overall, the analysis of the described academic literature sources shows that they: (a) do not clearly distinguish between the group of skills required for BDA and DS; and (b) those papers are predominately based on analysis of existing academic programs or on surveys of big data professionals, and do not take into consideration real job market requirements for BDA and DS knowledge domains and skills.

**The Ten Most Common Data Science Skills in Job Postings**

| Skill | Percentage of Job Listings |
|---|---|
| Python | 72% |
| R | 64% |
| SQL | 51% |
| Hadoop | 39% |
| Java | 33% |
| SAS | 30% |
| Spark | 27% |
| Matlab | 20% |
| Hive | 17% |
| Tableau | 14% |

glassdoor

**FIGURE 1. TEN MOST COMMON DATA SCIENCE SKILLS IN JOB POSTINGS (JUNCO, 2017)**

## III. METHODOLOGY

In this paper, we intent to answer the following research questions related to the BDA and DS fields:

- What are groups of knowledge and skills required for BDA and DS professions?
- What are common and distinguishing groups of skills between the two professions?
- What are important insights that can be revealed to business organizations and educational programs in terms of the BDA and DS knowledge and skills?

Our research model is based on identifying knowledge domains and associated skill sets for the BDA and DS fields. In addition to the three knowledge domains – *Business*, *Analytical*, and *Technical* – that are used in the Cegieelski, and Jones-Farmer (2016) paper discussed in the Literature Review section, we add one more knowledge domain, *Communication*. The latter is related to the written, communication, and presentation skills required today for the BDA and DS specialists (Song and Zhu, 2016; Dubey and Gunasekaran, 2015). Figure 2 depicts our research model, which connects the BDA and DS with 4 described knowledge domains and groups of skills (one or several groups) connected with each domain. In general, some groups of skills may also contribute simultaneously to several knowledge domains, e.g., analytical and technical.

As described in the Literature Review section, most of the academic research that considers the BDA and DS knowledge and skills is based on analyzing existing academic and professional programs without proper emphasis on and attention to the existing job market requirements for these professions, which actually defines the real knowledge and skills for these big data professions. Therefore, to answer our research questions, we analyze the primary data related to the posted online job requirements, qualifications, and skills for the BDA and DS professions. For these purposes, we collected around 1050 unique records of job requirements in BDA and DS for a period of 5 months from July through November of 2017. The job records were approximately evenly split between the two positions (close to 550 records for BDA and 500 records for DS positions). A specific number of 500 records for each data set represents a substantial and sufficient array of records to be able to perform a qualified text mining

analysis and receive statistically significant results of this analysis. The collection of these records was done manually without utilizing any web-scaping (web-crawling) tools, because with the latter tools we would not be able to structure the scrapped data into specific data/column structure we used in our research. They were derived from various online job posting sites including LinkedIn.com, Monster.com, Dice.com, Indeed.com, glassdoor.com, and careerbuilder.com.

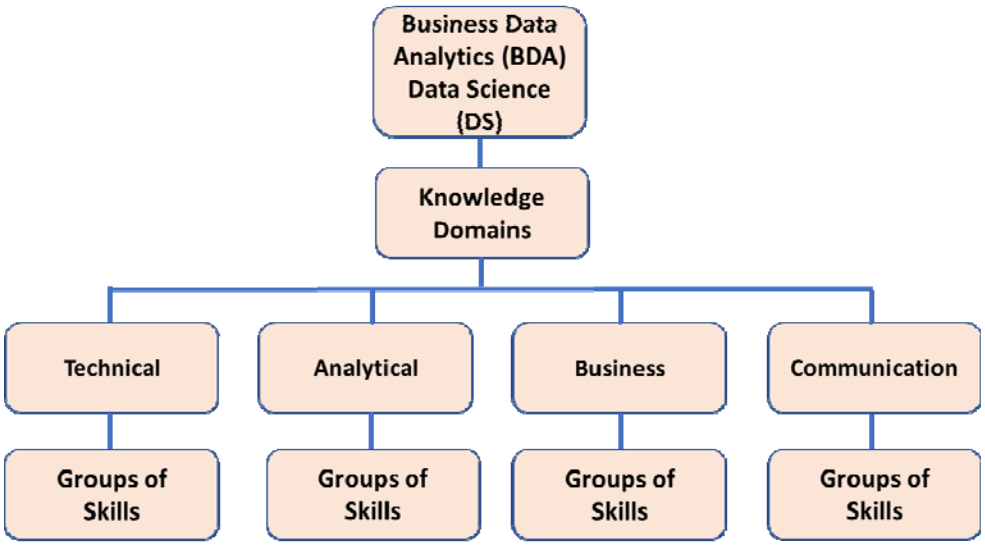


**FIGURE 2. RESEARCH MODEL.**

We develop two separate data sets, one included positions for DS jobs, and another data set – for BDA jobs. The titles of these job positions included in each data set were strictly related to either "Data Scientist" or "Data Analyst" jobs titles, respectively. We did not consider titles that may entail both DS and BDA in them, or other titles that were unclear in terms of their inclusion in one of the two data sets.

The primary job posting data was initially collected in a separate spreadsheet data set for each profession (BDA or DS) with the following column structure:

- Job Title – the title of a job position by which it is advertised in the website.
- Industry – a specific industry to which this job belongs, e.g., Information Technology, Insurance, Biotechnology, Oil and Gas, etc.
- Position Level – position level is derived by the following two methods:
  - The first option is a position level (e.g., Entry level, Associate, Director) as advertised for the job position in the website.
  - The second option, in case of the position level is not advertised in the website is to use the following assumptions: entry level (0-3 years of experience), mid-senior level (3-5 years of experience), senior level (5+ years of experience)
- Education – education level required for the job.
- Experience – the number of years of professional experience required for the job.

- Responsibilities – the job description which entails what a candidate is supposed to do in the given role.
- Requirements/Skills/Qualification – the skillset required for the job title. This includes, for example, technical skills like scripting languages (R, Python etc.), statistical and quantitative analytical models (classification, regression models, forecasting, optimization model etc.), domain experience (Insurance, Media etc.), business-related skills, and others.

In this research, we mostly concentrate on the last column with the information on the jobs' requirements, skills, and qualification. If these information is not provided in a job posting, we consider responsibilities as a primary source of information to identify required job skills. This information is central to our research agenda of identifying and comparing the knowledge domains and groups of skills for BDA and DS professions.

The data we analyze – requirements, skills, and qualifications of BDA and DS job positions – are mostly qualitative textual data that demonstrate a sufficient complexity derived from its unstructured nature and amount of records used. This analysis cannot be done by simple data visualization or manual counting, and therefore requires more sophisticated approaches of big data analytics. In our case, we utilize text mining methods to identify the existing frequencies of various skills associated with the BDA and DS professions, group (cluster) those skills, and compare between each other. For each of these professions, we identify the following:

- *Document Data Matrix (DTM)* with useful words and phrases, defined as "terms," describing the job requirements, skills and qualifications, and also their count and relative frequency.
- *Term Cloud* to visualize the relative frequencies of terms in DTM.
- *Singular Vector Decomposition (SVD)* of the data to reduce noise and redundancy in the analyzed DTM that improves the ability to capture the essence of existing relationships between the terms.
- *VARIMAX Rotation* to develop, based on the relationships between terms, several important groups of topic terms (words and phrases) that can characterize the groups of skills required for each position.
- *Latent Class Analysis (LCA)* to identify clusters of terms (similar to topic words) based on the closeness of their relations.

The described quantitative approaches are commonly used for text mining analytics, and are well described in a variety of literature sources (Berry and Kogan, 2010; Kwartler, 2017; Klimberg and McCollough, 2016, Weiss et al., 2015).

To utilize all these text mining methods, we employ the JMP software, which is one of the SAS software products. The current version of JMP software has an extensive built-in statistical and predictive analytics capabilities, and, in particular, text mining analytics. It is relatively easy to use, provides good visualization of the derived results, and can make output scripts with the methods applied to any major data analytics language like Python, R, and SAS. The results of our analysis based on the specified text mining methods, and comparison of the BDA and DS groups of skills are presented in the next section.

## IV. TEXT MINING ANALYSIS AND RESUTS

As discussed in the Methodology section, we utilize the text data describing requirements, skills, and qualifications of the accumulated BDA and DS job postings. Based on this text data, we develop a *Document Term Matrix (DTM)* for each profession, which contains a combination of terms, i.e., individual words and simple phrases, specific to a particular profession (BDA or DS). The DTM development was using *Stemming* of words and *Regex* option for tokenizing those words. We also use a *Stop Add Word* utility to remove unnecessary words, e.g., "a," "the," "and," etc., and remove words or their combinations that are general and not specific to BDA or DS, e.g., "strong," "easy," "experience," and some others. In addition, we use a *Recode* utility to change the words with common names, for example, "ml" was recoded as "machine learning," or recode misspelled words.

Based on these DTM improvements and cleaning, they finally contained 985 terms for BDA and 1,350 terms for DS. The 20 most frequent BDA and DS terms and their percentages in the total number of postings are presented in Tables 1 and 2, respectively. As can be seen from the tables, the most frequent terms are quite different between the two sets of job postings. For BDA jobs, the most frequent terms start with "sql," "tools," "reports," "business," and "environment," which describe a more business-oriented nature of those jobs. At the same time, the most frequent words in the DS jobs contain "machine learning," "analytical," "python," "big data," "analytics," and "algorithms," which characterizes more technical (programming and software) orientation of those jobs. The development of DTM in JMP does allow to use, besides individual terms, the phrase terms combined of two or more words up to a certain limit specified by a DTM developer (we limited it to 4 words in a phrase term). Therefore, two-word terms like "machine learning," "big data," and "business intelligence" appear in DTM (see Tables 1 and 2).

The term clouds in Figures 3 and 4 also support the same conclusion with respect to more business-oriented BDA terms versus technical orientation of the DS terms.

**TABLE 1. 20 MOST FREQUENT BDA TERMS.**

| # | Term | Proportion of total, % |
|---|---|---|
| 1 | sql | 74.5% |
| 2 | tools | 56.9% |
| 3 | reports | 54.9% |
| 4 | business | 53.9% |
| 5 | environment | 52.9% |
| 6 | analytics | 51.0% |
| 7 | understand. | 47.1% |
| 8 | excel | 47.1% |
| 9 | database | 44.1% |
| 10 | develop | 44.1% |
| 11 | proficient | 44.1% |
| 12 | python | 36.3% |
| 13 | required | 36.3% |
| 14 | r | 35.3% |
| 15 | business intelligence | 35.3% |
| 16 | technical | 31.4% |
| 17 | tableau | 30.4% |
| 18 | team | 28.4% |
| 19 | query | 27.5% |
| 20 | analysis | 27.5% |

**TABLE 2. 20 MOST FREQUENT DS TERMS.**

| # | Term | Proportion of total, % |
|---|---|---|
| 1 | machine learning | 71.6% |
| 2 | analytical | 47.0% |
| 3 | python | 42.0% |
| 4 | big data | 28.2% |
| 5 | analysis | 26.8% |
| 6 | algorithms | 26.4% |
| 7 | written communication | 24.0% |
| 8 | sql | 22.6% |
| 9 | r | 20.0% |
| 10 | techniques | 18.6% |
| 11 | tools | 15.8% |
| 12 | statistics | 15.4% |
| 13 | data processing | 14.8% |
| 14 | natural language | 14.2% |
| 15 | hadoop | 13.4% |
| 16 | models | 12.6% |
| 17 | environment | 12.0% |
| 18 | data mining | 11.8% |
| 19 | technologies | 11.6% |
| 20 | collaborate | 11.4% |


**FIGURE 3. TERM CLOUD FOR BDA JOBS.**


**FIGURE 4. TERM CLOUD FOR DS JOBS.**

We next apply *Latent Semantic Analysis (LCA)* to extract useful information from DTM concerning knowledge and skills of the BDA and DS professions. Due to a large size of the DTMs for both job positions, we utilize, as a part of LCA, *Singular Value Decomposition (SVD)* to reduce the DTM dimensionality to a manageable size that more clearly depicts the underlying structure of the data (Klimberg and McCullough, 2016, Weiss et al., 2015). SVD reduces noise and redundancy in the data, and creates new dimensions that capture the essence of existing relationships. To develop the SVD output, we utilize the minimum term frequency of 10 (a term is mentioned at least 10 times in the job postings), and *term frequency—inverse document frequency (TF-IDF)* weighting. TF-IDF considers the tradeoff between the frequency of a term throughout the data set and the frequency of the term in DTM, which makes results more interpretable (Klimberg and McCullough, 2016, Weiss et al., 2015).

To develop SVD, we apply 20 singular vectors. The number of singular vectors can be assigned by a developer in the SVD specifications. For the visualization purposes, we utilize the first two vectors only, which capture around 14% and 10% of the variance in the BDA and DS data sets, respectively. The SVD outputs for the first two term singular vectors (Vector 1 and

Vector 2) are shown in Figures 5 and 6. In each graph, the singular vectors have three branches that are shown with different colors and markers (red squares, blue circles, and green "plus sign" points). The branches describe some common themes existing in the text data. For example, in the SVD plot for BDA (Figure 5), the branch with blue circles describes the data analytics technologies and languages (Python, R, Hadoop, Hive, Spark, Java, and similar related items). The branch with plus sign points represents "machine learning," "statistics," and "techniques." The branch with a substantial number of red squares characterizes a theme associated with business requirements in business data analytics.

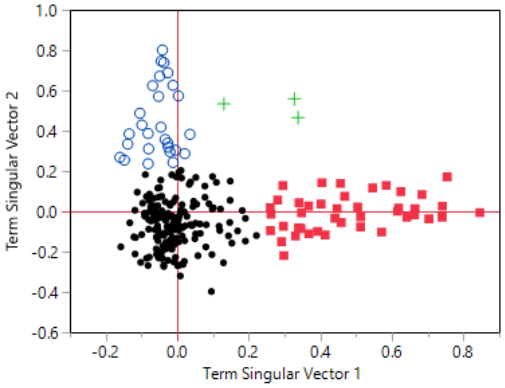


**FIGURE 5. SVD PLOT FOR BDA TERMS.**

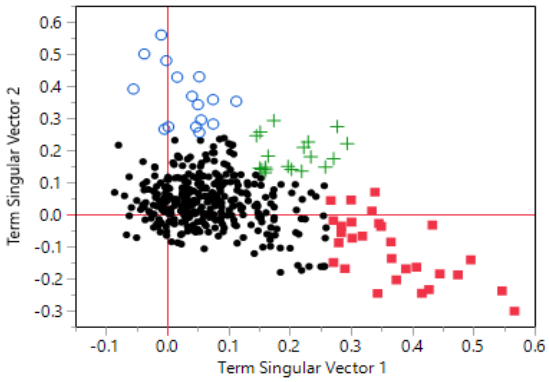


**FIGURE 6. SVD PLOT FOR DS TERMS.**

The three branches in the SVD plot for DS terms (Figure 6) represent different groups of terms from those in the plot for BDA terms. For example, the blue circles in the DS plot (Figure 6) are associated with analytical tools like "machine learning "svm," "naïve bayes," and some others. The branch with green plus sign points represents "algorithms," "architecture," "models," and other similar terms. The red square dots describe a variety of software development terms like "software development," "open source," "distributed computing," and others. Finally, the black dots in both SVD plots (Figures 5 and 6) denote BDA and DS terms that are not grouped by the first two singular vectors, Vectors 1 and 2.

To develop more coherent groups of terms for BDA and DS, we use the SVD data for topic extraction with *VARIMAX Rotation*, which is, essentially, the factor analysis of terms. This approach allows extracting groups of terms that, according to squared loadings or squared correlations between the terms, are closely related to each other (Kwartler, 2017; Klimberg and McCollough, 2016, Weiss et al., 2015). In our case, we apply the *Topic Analysis* with VARIMAX Rotation for both DTMs. The number of selected topics could vary, typically, from 2 to 10, and can be used to identify the best fit of the term groupings. In our research, we identify that the best fit for the BDA terms happens with the selected number of topics equal to 10, and for the DS terms the best number of topic appears to be 8 (see *Appendix A*). Each topic in *Appendix A* contains a list of skills that belong to each topic. The specific skills included in each topic are closely related to each other according to squared loadings between the skills. These skills are also listed in descending order of their respective loadings, from top-rated (high loadings) down to lower-rated (lower loadings).

In essence, each topic identifies a group of terms, which may be described as a

group of skills generally required for a BDA or DS job. Based on the combination of terms included in the topics in *Appendix A*, we were able to provide a possible definition, with only certain degree of certainty, for each of those topics, and associate it with four knowledge domains, i.e., Technical, Analytical, Business, and Communication described in the Methodology section of this paper. The topic or group of skills definitions, example of those skills, and their allocation to specific knowledge domains are presented in Tables 3 and 4. In these tables, all terms in the *Examples of Skills* column are shown in lowercase, which is done to make them consistent with the DTM terms (Tables 1 and 2) and data mining results in *Appendix A*.

**TABLE 3. TOPICS DEFINITIONS FOR BDA.**

| *Topic* | *Group of Skills* | *Examples of Skills* | *Domain* |
|---|---|---|---|
| Topic 1 | Database Design and Management | xml, database design, packages, collect organize analyze and disseminate | Technical |
| Topic 2 | Data Mining/Machine Learning | clustering, regression, factoring, algorithms, neural network | Analytical |
| Topic 3 | General Business | business, requirements, lead, mapping, documents | Business |
| Topic 4 | Office Programs | programs, office, results, outlook, powerpoint | Business |
| Topic 5 | Effective Data Management | scheduling, media, data visualization, effective, solutions | Business |
| Topic 6 | Big Data Technology and Languages | hadoop, spark, hive, c/c++, big data technologies, languages | Technical |
| Topic 7 | Data Presentation | tables, pivot/table, sharepoint, microsoft office, vba | Analytical |
| Topic 8 | Data Warehousing | cube, kimball, ssas, sql server, data warehouse | Technical |
| Topic 9 | Data Analytics Software | r, tableau, qlikview, python, stata | Analytical |
| Topic 10 | Presentation and Communication | reports, database, web, crystal reports, interpersonal skills | Communication |

**TABLE 4. TOPICS DEFINITIONS FOR DS.**

| *Topic* | *Group of Skills* | *Examples of Skills* | *Domain* |
|---|---|---|---|
| Topic 1 | Analytical Tools | naïve bayes, knn, svm, machine learning | Analytical |
| Topic 2 | Information Management and Communication | information, process, manage, written communication, verbal communication | Business and Communication |
| Topic 3 | Implementation of Analytics and Big Data Technologies | implementation, service, customer, data mining, hadoop, statistical modeling | Technical |
| Topic 4 | Software Development | software development, distributed computing, hive, architecture, apache | Technical |
| Topic 5 | Software and Data Management | software, open source, testing, apache, data management | Technical |
| Topic 6 | Analytical Methods | clustering, modeling, quantitative, analytical skills, logistic regression, sas | Analytical |
| Topic 7 | Algorithms and Languages | bayesian, frameworks, algorithms, architecture, models | Analytical |
| Topic 8 | Communication and Presentation | communicate complex, real world, problems, team, solve | Business and Communication |

Tables 3 and 4 show several similar topics for both BDA and DS, for example, *Topic 2 Data Mining/Machine Learning* and *Topic 10 Learning and Communication* in BDA are comparable with *Topic 1 Analytical Tools* and *Topic 8 Communication and Presentation* in DS, respectively. However, we also noticed substantive differences in the groups of skills in those profession. As can be seen from Table 4, the BDA groups of skills have a major emphasis on business-related skills like *General Business* skills and *Effective Data Management*, and also technical skills such as *Database Design and Management*, and *Data Warehousing*. Contrary to that, the DS job skills focus on *Software Development*, *Algorithms and Languages*, and *Software and Data Management*, whereas the business skills appeared to be substantially less prominent in DS.

We continue our text mining research with cluster analysis of the BDA and DS' terms, which is referred to as *Latent Class Analysis (LCA).* It is used in text mining to group together DTM's common terms into a designated number of clusters (Berry and Kogan, 2010; Klimberg and McCollough, 2016; Weiss et al., 2015). As previously with the SVD and VARIMAX Rotation, we used for consistency 10 clusters for BDA and 8 clusters for DS. The clustering results including the DTM terms in each cluster are presented in Appendix B. Employing the combination of terms allocated to the clusters in *Appendix B*, we also provide a possible definition for each of those clusters, and associate each cluster with one or several of the four knowledge domains. The topic definitions and their allocation to specific domains for each profession is presented in Tables 5 and 6.

**TABLE 5. DEFINITIONS of CLUSTERS FOR BDA.**

| *Cluster* | *Group of Skills* | *Examples of Skills* | *Domain* |
|---|---|---|---|
| Cluster 1 | Business Problem Solving and Communication | communications skills, problem solving, creativity, collaboration, problem | Business and Communication |
| Cluster 2 | Big Data Technology and Languages | python, develop, languages, machine learning, hadoop, r, hive, statistics, programming | Technical and Analytical |
| Cluster 3 | Office Programs | excel, access, proficient, organizational, powerpoint | Business |
| Cluster 4 | General Management | manage, problem solving, fast paced, excel | Business |
| Cluster 5 | Effective Data Management | sql, result, proficient, communication skills | Business |
| Cluster 6 | Database Design and Management | advanced, sql server, build, etl | Technical |
| Cluster 7 | Database Design and Management | database, sql, develop, analytical skills, design | Technical |
| Cluster 8 | Presentation and Communication | complex, develop, business, reports, presentation, communication skills | Communication |
| Cluster 9 | Data Analytics | analytics, proficient, results, business intelligence | Analytical |
| Cluster 10 | General Management | environment, develop, management, work independently, product | Business |

Comparing the allocation of terms into BDA clusters and BDA topics (Tables 3 and 5), we can see a high degree of similarity between them. Although not in the same sequence, the common topics and clusters for BDA are: *General Business/General Management*, *Office Programs*, *Effective Data Management*, *Big Data Technologies and Languages*, and some others to the lesser level of similarity. In the same way, there are a number of common themes in the topics and clusters related to DS (Tables 4 and 6), e.g., *Analytical Methods*, *Algorithms*, *Communication and Presentation*, and *Big Data Technologies*. All this supports the point that several different text mining methods identify a number of similar groups of skills within BDA and DS professions. In addition, the cluster analysis also confirms, as previously identified through the Topics Analysis, that there are a number of different groups of skills between BDA and DS. Overall, the LCA cluster analysis validates the results of VARIMAX Rotation of terms for the BDA and DS groups of skills.

**TABLE 6. DEFINITIONS OF CLUSTERS FOR DS.**

| *Cluster* | *Group of Skills* | *Examples of Skills* | *Domain* |
|---|---|---|---|
| Cluster 1 | Technologies and Programming Languages | technologies, programming languages | Technical |
| Cluster 2 | Analytical Methods | machine learning, natural language, statistics, analytical tools, optimization | Analytical |
| Cluster 3 | Analytical Methods | machine learning, tools, analysis, python, | Analytical |
| Cluster 4 | Communication and Presentation | written communication, non technical, analytical, results, environment | Communication |
| Cluster 5 | Algorithms | techniques, algorithms, machine learning, big data, svm | Technical |
| Cluster 6 | Algorithms | deep learning, machine learning, algorithms, python, problem solving | Technical |
| Cluster 7 | Business and Communication | written communication, business, analytical, projects, research | Business and Communication |
| Cluster 8 | Big Data Technologies | spark, big data, algorithms, hadoop, modeling | Technical |

## V. CONCLUSIONS

Creating a detailed understanding of relevant knowledge and skills for business data analytics (BDA) and data science (DS) fields is extremely important for various organizations looking for big data professionals. It is also critical for current and future professionals that would like to develop and/or improve their skills in these areas. In addition, the clear recognition of the BDA and DS knowledge and skills may be vital for an increasing number of academic and professional programs that teach business data analytics and data science.

The described in this paper literature review of academic and practitioner papers show that they do not clearly distinguish between the groups of skills required for BDA and DS. Those papers are predominately based on analysis of existing business analytics and data science educational programs, or on surveys of big data professionals, and do not take into

consideration real job market requirements for BDA and DS knowledge and skills.

In this research study, we apply the primary data of job requirements, qualifications, and skills for business/data analyst and data scientist from various online job-related websites in order to identify and systemize the groups of knowledge and skills required for BDA and DS professions. We also compare the two sets of data to distinguish between the two professions, and provide important insights for companies and education of BDA and DS professionals.

Our research model is based on identifying knowledge domains and skill sets for the BDA and DS fields. According to this model, we consider four knowledge domains – Technical, Analytical, Business, and Communication – and associated groups of skills (one or several groups) connected with each domain. To identify these groups of skills we utilize a variety of methods of text mining analytics, i.e., developing a Document Data Matrix and text terms cloud, performing Singular Value Decomposition of the Document Data Matrix, and using the latter to make VARIMAX Rotation and Latent Class Analysis for grouping and clustering the text terms, respectively.

The text mining analysis, in particular, VARIMAX Rotation's topics and LCA clusters, reveal some strong similarities as well as substantial differences between the knowledge domains and groups of skills for BDA and DS (see Tables 4-6). The similarities are demonstrated by the groups of skills related to *Analytical Methods*, *Data Mining and Machine Learning*, and *Communication and Presentation* skills. We also identify that the most notable difference in BDA vs. DS groups of skills is that the former has a major emphasis on business-related skills including *General Business* and *Effective Data Management*, and also technical skills such as *Database Design and Management*, and *Data Warehousing* (Table 3). Contrary to that, the DS groups of skills focus on the technical skills related to *Software Development*, *Algorithms and Languages*, *Software and Data Management* (Table 4). Applying the text mining analysis, we were also able to identify the top ranked skills that belong to each group of skills.

Overall, this paper presents a unique research study, where the knowledge domains and groups of skills relevant to BDA and DS are compared and contrasted based on the job-related skill requirements and qualifications for these professions. This research also produced a number of contributions to the theory and practice of big data analytics. First, we extended the list of knowledge domains for BDA and DS to four main domains including Technical, Analytical, Business, and Communication, and also develop a new research model that incorporates these domains along with the groups of skills. Second, we identified and titled those groups of skills (topics and clusters) with their respective top-ranked skills examples for BDA and DS professions. Third, from the practical standpoint, the systemized knowledge domains and associated groups of skills for BDA and DS can be extremely helpful to: (a) employers developing job requirements for bid data job positions; (b) employees and specialists for better understanding and for sharpening their skills in BDA or DS; and (c) academic and professional education for developing new or improving existing analytics and data science programs and certificates.

This research can be extended in the future in several ways. We may analyze the BDA and DS skills in conjunction with several important job-related characteristics and attributes, i.e., years of experience (e.g., 1-3 years, 3-5 years, 5+years), position level (entry, middle, and senior level), education

level (bachelor, masters, Ph.D.), job location, and some others. In addition, we may compare and contrast the groups of BDA and DS job-related skills with those imbedded in the existing academic and professional programs in these fields.

## REFERENCES

Aasheim, C.L. and Williams, S., Rutner, P. and Gardiner, A., "Data Analytics vs. Data Science: A Study of Similarities and Differences in Undergraduate Programs Based on Course Descriptions", *Journal of Information Systems Education*, 26(2), 2015, 103-115.

Berry, M. and Kogan, J. *Text Mining: Applications and Theory*, John Wiley and Sons, 2010.

Bichler, M., Heinzl, A. and Van der Aalst, W.M., "Business Analytics and Data Science: Once Again?", *Business & Information Systems Engineering*, 59(2), 2017, 77-79.

Cegieelski, C.G. and Jones-Farmer, L.A., "Knowledge, Skills, and Abilities for Entry-Level Business Analytics Positions: A Multi-Method Study", *Decision Science Journal of Innovative Education*, 14(1), 2016, 91-118.

Columbus, L., "Where Big Data Jobs Will Be in 2016", *Forbes Magazine*, November 15, 2015, http://www.forbes.com (accessed October 5, 2017).

Daniel, B., "Big Data and analytics in higher education: Opportunities and challenges", *British Journal of Educational Technology*, 46(5), 2015, 904-920.

Debortolli, S., Muller, O. and Von Brocke, J., "Comparing Business Intelligence and Big Data Skills: A Text Mining Study Using Job Advertisements", *Business and Information Systems Engineering*, 5, 2014, 289-300.

De Mauro, A., Greco, M., Grimaldi, M. and Ritala, P., "Human resources for Big Data professionals: A systematic classification of job roles and required skill sets," *Information Processing & Management*, June, 2017, 1-22.

Dubey R. and Gunasekaran, A., "Education and training for successful career in Big data and Business Analytics", *Industrial and Commercial Training*, 47(4), 2015, 174-181.

Ellaway, R.H., Puci, M.V., Galbraith, R.M. and Cameron, T., "Developing the role of big data and analytics in health professional education", *Medical Teacher*, 36, 2014, 216-222.

Englmeier, K. and Murtagh, F., "Editorial: What Can We Expect from Data Scientists?", *Journal of Theoretical and Applied Electronic Commerce Research*, 12(1), 2017, 1-5.

Fottrell, Q., "The most sought-after job in America pays $110,000 a year – if you have this skill", *MarketWatch: Stock Market News*, November 19, 2017, http://www.MarketWatch.com (accessed December 1, 2017).

Gonzalez, C., "Data Analysis and Computing: The Future Job Wave," *Machine Design*, April 14, 2017, http://www.machinedesign.com/industrial/data-analysis-and-computing-future-job-wave (accessed November 28, 2017).

Gordon, M.E, "Are Business ready to Take Advantage of the Growing Number of Data Scientists?", *Big Data Consulting*, February 6, 2016, http://www.experfy.com (accessed December 1, 2017).

Gorman, M. F. and Klimberg, R.K., "Benchmarking Academic Programs in Business Analytics", *Interfaces*, 44(3), 2014, 329-341.

Henry, R. and Venkatraman, S., "Big Data Analytics the Next Big Learning Opportunity", *Journal of Management Information and Decision Sciences,* 18(2), 2015, 17-29.

Holden, G., "Big Data and R&D Management", *Research-Technology Management*, September-October, 2016, 22-26.

Junco, P.R., "Data Scientist Programs: What Skills Do They Have and How Much Do They Make?", *Glassdoor Job Search*, September 21, 2017, http://www.glassdoor.com (accessed November 10, 2017).

Kambatla, K., Kollias, G., Kumar, V. and Grama, A, "Trends in big data analytics", *Journal of Parallel and Distributed Computing*, 74, 2014, 2561-2573.

Kemper, S. and Mathews, T., "Earth Science Data Analytics: Definitions, Techniques and Skills, *Data Science Journal*, 16(6), 2017, 1-8.

Kennedy D.T., Crossen, D.M. and Szabat, K.A., "Business Analytics and Big Data: Driving Organizational Change", in *Handbook of Research on Organizational Transformations through Big Data Analytics*, IGI Global, 2015.

Klimberg, R. and McCullough, B.D., *Fundamentals of Predictive Analytics with JMP*, Second Edition, SAS Institute, Inc., 2016.

Kwartler, T., *Text Mining in Practice with R*, John Wiley & Sons Ltd, 2017.

Liberatore, M. and Luo, W., "ASP, The Art and Science of Practice: A Comparison of Technical and Soft Skill Requirements for Analytics and OR Professionals", *Interfaces*, 43(92), 2013, 194-197.

Miller, S., "Collaborative Approaches Needed to Close the Big Data Skills Gap", *Journal of Organization Design*, 3(1), 2014, 26-30.

Mills, R., Chudoba, K.M. and Olsen, D.H., "IS Programs Responding to Industry Demands for Data Scientists: A Comparison between 2011 – 2016", *Journal of Information Systems Education*, 27(2), 2016, 131-140.

NIST Big Data Interoperability: 2015 NIST Big Data Public Working Group Definitions and Taxonomies Subgroup. *Framework: Volume 1, Definitions, NIST Special Publication 1500-1*, 2015.

Phelps, A.L. and Szabat, K.A., "The Current Landscape of Teaching Analytics to Business Students at Institutions of Higher Education: Who is Teaching What?", *The American Statistician*, 71(2), 2017.

Power, D.J., "Data science: supporting decision-making", *Journal of Decision Systems*, 25(4), 2016, 345-356.

Priestley, J., "Big Data Career Switch: 4 Key Points", *InformationWeek Connects to Business Technology Community*, July 17, 2013, http://www.informationweek.com (accessed November 10, 2017).

Ransbotham, S., Kiron, D. and Prentice, P.K., "The Talent Dividend", *MIT Sloan Management Review*, 56(3), Spring 2015, 63-68.

Schoenherr, T. and Speier-Pero, C., "Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential", *Journal of Business Logistics*, 36(1), 2015, 120-132.

Song, I.Y. and Zhu, Y., "Big data and data science: what should we teach", *Expert Systems*, 33(4), 2016, 364-373.

Steinweg-Woods, J., "Web Scrapping for Key Data Science Job Skills," *Jesse Steinweg-Woods, Ph.D.*, March 17, 2015, http://jessesw.com/Data-Science-Skills (accessed November 10, 2017).

Tang, R. and Sae-Lim, W., "Data science programs in U.S. higher education: An

exploratory content analysis of program description, curriculum structure, and course focus", *Education for Information*, 32, 2016, 269-290.

Weiss, S.M., Indurkhya, N. and Zhang, T., *Fundamentals of Predictive Text Mining*, Second Edition, Springer-Verag London, 2015.

Wymbs, C., "Managing the Innovation Process: Infusing Data Analytics into the Undergraduate Business Curriculum (Lessons Learned and Next Steps)", *Journal of Information Systems Education*, 27(1), 2016, 61-74.

# APPENDIX A
# RESULTS OF VARIMAX ROTATION FOR BDA

| Topic 1 | |
|---|---|
| **Term** | **Loading** |
| xml | 0.88063 |
| database design | 0.78871 |
| packages | 0.75177 |
| collect organ· analyze and disseminate | 0.74700 |
| javascript | 0.73582 |
| datasets | 0.69896 |
| segmentation | 0.67588 |
| ai | 0.66750 |
| object· | 0.66731 |
| data analytics | 0.66304 |
| data models | 0.65943 |
| etl | 0.65908 |
| spss | 0.62745 |
| understand· | 0.55994 |

| Topic 2 | |
|---|---|
| **Term** | **Loading** |
| clustering | 0.92141 |
| regression | 0.83244 |
| factoring | 0.80845 |
| algorithms | 0.78392 |
| neural network | 0.77214 |
| decision trees | 0.73692 |
| probability | 0.67449 |
| deep learning | 0.65676 |
| scenario· | 0.64068 |
| statistics | 0.52094 |
| methodolog· | 0.51635 |
| techniques | 0.48832 |
| number· | 0.45855 |

| Topic 3 | |
|---|---|
| **Term** | **Loading** |
| business | 0.72035 |
| requirements | 0.70143 |
| lead· | 0.61084 |
| mapping | 0.60608 |
| document· | 0.59245 |
| sources | 0.57522 |
| present· | 0.57382 |
| quantitative | 0.52530 |
| develop | 0.49399 |
| decision | 0.49341 |
| complex | 0.43243 |
| analytics skills | 0.42633 |
| analytics | 0.41026 |
| quick· | 0.40749 |
| oriented | 0.40245 |

| Topic 4 | |
|---|---|
| **Term** | **Loading** |
| programs | 0.72612 |
| office | 0.71524 |
| result· | 0.65822 |
| outlook | 0.58220 |
| terminolog· | 0.52815 |
| powerpoint | 0.48317 |
| healthcare | 0.47723 |
| reason· | 0.46306 |
| supervis· | 0.45005 |
| word | 0.44892 |
| logic· | 0.44124 |
| industri· | 0.39038 |
| strong problem solving | 0.38805 |
| work independ· | 0.38688 |
| oper· | 0.38114 |

| Topic 5 | |
|---|---|
| **Term** | **Loading** |
| schedul· | 0.70564 |
| media | 0.62606 |
| data visualization | 0.62509 |
| train· | 0.58535 |
| effect· | 0.56831 |
| solutions | 0.51221 |
| teradata | 0.50165 |
| fast pace· | 0.48969 |
| managing | 0.48428 |
| tools | 0.48410 |
| agile | 0.47759 |
| product· | 0.47183 |
| recommend· | 0.45953 |
| environment | 0.44282 |
| depth | 0.43770 |
| oracle | 0.43491 |

| Topic 6 | |
|---|---|
| **Term** | **Loading** |
| hadoop | 0.73192 |
| spark | 0.64426 |
| hive | 0.57655 |
| c/c++ | 0.54148 |
| big data technologies | 0.49627 |
| languages | 0.48913 |
| java | 0.46880 |
| technologies | 0.44101 |
| integration | 0.43847 |
| nosql | 0.38636 |
| cloud | 0.38244 |
| linux | 0.36212 |
| azure | 0.36137 |
| programming | 0.35441 |

| Topic 7 | |
|---|---|
| **Term** | **Loading** |
| tables | 0.6022 |
| pivot/table | 0.5640 |
| sharepoint | 0.5364 |
| suite | 0.5220 |
| microsoft office | 0.4890 |
| reason· | 0.4257 |
| vba | 0.3921 |
| view· | 0.3874 |
| group | 0.3838 |
| function· | 0.3756 |
| macros | 0.3742 |
| review· | 0.3499 |
| passion· | -0.3402 |
| analysis | -0.3329 |
| build | 0.3229 |

| Topic 8 | |
|---|---|
| **Term** | **Loading** |
| cube· | 0.78972 |
| kimball | 0.78972 |
| ssas | 0.73856 |
| sql server | 0.70620 |
| data warehous· | 0.62529 |
| schema· | 0.60503 |
| ssis | 0.51059 |
| design | 0.45140 |
| oral and written communication skills | 0.44004 |
| develop | 0.42574 |
| quick· | 0.37991 |
| methodolog· | 0.37930 |

| Topic 9 | |
|---|---|
| **Term** | **Loading** |
| r | 0.5941 |
| tableau | 0.5338 |
| qlikview | 0.4947 |
| python | 0.4740 |
| stata | 0.4702 |
| consult· | 0.4230 |
| spotfire | 0.4180 |
| modeling | 0.3983 |
| required | 0.3611 |
| statistics | 0.3543 |
| statistical analysis | 0.3419 |
| oper· | -0.3283 |
| looker | 0.3282 |
| visualization | 0.3229 |

| Topic 10 | |
|---|---|
| **Term** | **Loading** |
| reports | 0.5470 |
| database | 0.5379 |
| web | 0.4803 |
| problem | -0.4577 |
| crystal reports | 0.4553 |
| queri· | 0.4376 |
| tools | 0.3611 |
| interpersonal skills | 0.3467 |
| manager | 0.3431 |
| solving | -0.3388 |
| analyze | 0.3335 |
| oracle | 0.3300 |
| data modeling | 0.3278 |
| concepts | 0.3112 |
| sql | 0.3089 |

# RESULTS OF VARIMAX ROTATION FOR DS

| Topic 1 | |
|---|---|
| **Term** | **Loading** |
| naive bayes | 0.82145 |
| k nn | 0.80145 |
| svm | 0.71277 |
| nosql databases | 0.48640 |
| financial services | 0.39420 |
| techniques | 0.36149 |
| critical thinking | 0.34746 |
| aws | 0.34153 |
| big data | 0.33929 |
| statistical software | 0.31589 |
| algorithms | 0.30851 |
| machine learning | 0.29667 |

| Topic 2 | |
|---|---|
| **Term** | **Loading** |
| information | 0.50269 |
| process | 0.42999 |
| manage | 0.42842 |
| management | 0.38766 |
| written communication | 0.38138 |
| business | 0.36963 |
| innovative | 0.30888 |
| verbal communication | 0.30591 |
| problem solving | 0.28942 |
| team | 0.28714 |
| minimum | 0.25008 |
| develop | 0.24920 |
| developing | 0.24723 |
| technical | 0.24181 |

| Topic 3 | |
|---|---|
| **Term** | **Loading** |
| implementation | 0.54393 |
| services | 0.49413 |
| customer | 0.39938 |
| data mining | 0.38411 |
| project | 0.38041 |
| design | 0.34461 |
| leadership | 0.34267 |
| redshift | 0.34147 |
| hadoop | 0.32914 |
| methodologies | 0.31388 |
| google | 0.30194 |
| statistical modeling | 0.29093 |
| automation | 0.28893 |
| development | 0.26783 |

| Topic 4 | |
|---|---|
| **Term** | **Loading** |
| software development | 0.54253 |
| effectively communicate | 0.51235 |
| real world | 0.51092 |
| distributed computing | 0.50685 |
| presentation | 0.50479 |
| teamwork | 0.49363 |
| class | 0.48297 |
| hive | 0.48247 |
| architecture | 0.38747 |
| apache | 0.36215 |
| scikit learn | 0.33838 |
| methods | 0.33047 |
| technologies | 0.32825 |
| non technical | 0.31969 |

| Topic 5 | |
|---|---|
| **Term** | **Loading** |
| software | 0.52477 |
| open source | 0.45519 |
| testing | 0.41633 |
| best practices | 0.39711 |
| apache | 0.39115 |
| data management | 0.38502 |
| platform | 0.38255 |
| information retrieval | 0.33230 |
| development | 0.32694 |
| applications | 0.31815 |
| shell scripting | 0.31575 |
| data warehousing | 0.29820 |
| processes | 0.29236 |
| research | 0.25566 |

| Topic 6 | |
|---|---|
| **Term** | **Loading** |
| clustering | 0.4739 |
| concepts | 0.3606 |
| modeling | 0.3566 |
| quantitative | 0.3382 |
| databases | 0.3344 |
| oracle | 0.3246 |
| analytical skills | 0.3122 |
| communicate | 0.2812 |
| logistic regression | 0.2776 |
| sas | 0.2693 |
| data processing | -0.2689 |
| analytical | 0.2607 |
| supervised | 0.2579 |
| neural networks | 0.2539 |
| complex | 0.2471 |

| Topic 7 | |
|---|---|
| **Term** | **Loading** |
| bayesian | 0.44110 |
| frameworks | 0.43065 |
| algorithms | 0.41595 |
| architecture | 0.36904 |
| models | 0.36625 |
| c | 0.35786 |
| developing | 0.34125 |
| systems | 0.31484 |
| logistic regression | 0.29022 |
| web | 0.26386 |
| deep learning | 0.26033 |
| java | 0.23938 |
| production | 0.23896 |
| principles | 0.23301 |

| Topic 8 | |
|---|---|
| **Term** | **Loading** |
| communicate complex | 0.51279 |
| class | 0.49010 |
| real world | 0.46620 |
| quantitative analysis | 0.44393 |
| problems | 0.36538 |
| team | 0.34035 |
| problem | 0.32250 |
| solve | 0.31324 |
| solving | 0.31108 |
| nosql | 0.30334 |
| fast paced | 0.29283 |
| analyze | 0.28697 |
| model | 0.23751 |

# APPENDIX B
# RESULTS OF CLUSTERING ANALYSIS OF BDA

| Cluster1 | | Cluster2 | | Cluster3 | | Cluster4 | | Cluster5 | | Cluster6 | | Cluster7 | | Cluster8 | | Cluster9 | | Cluster10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Term | Score | Term | Score | Term | Score | Term | Score | Term | Score | Term | Score | Term | Score | Term | Score | Term | Score | Term | Score |
| communication skills | 5.9259 | python | 10.083 | excel | 13.63 | manage | 9.3575 | excel | 9.6497 | advanced | 9.2379 | database | 11.739 | complex | 12.487 | analytics | 14.174 | environment | 11.681 |
| problem solving | 3.297 | develop | 8.0135 | proficient | 9.0428 | problem solving | 8.2806 | sql | 9.3606 | organ· | 7.1091 | information | 11.338 | develop | 11.915 | proficient | 11.362 | develop | 11.559 |
| creativ· | 3.2094 | languages | 7.1578 | access· | 5.8114 | fast pace· | 8.2343 | result· | 7.362 | sql server | 5.4463 | sql | 9.2079 | business | 9.6203 | packages | 9.398 | management | 11.398 |
| oper· | 2.9312 | machine learning | 7.0742 | organizational | 2.9989 | excel | 7.7351 | proficient | 7.0524 | function· | 5.3284 | develop | 8.9724 | reports | 8.8175 | result· | 9.3117 | work independ· | 10.857 |
| collabor· | 2.5751 | hadoop | 6.342 | powerpoint | 2.9531 | statistics | 6.6401 | communication skills | 7.0315 | detailed oriented | 5.2493 | attention | 8.2119 | environment | 8.6945 | required | 9.1887 | product· | 10.237 |
| problem | 2.562 | r | 6.2729 | management | 2.8646 | creativ· | 6.2669 | understand· | 6.9697 | services | 5.1868 | statistical | 8.1431 | analyze | 8.3008 | business intelligence | 9.0986 | schedul· | 6.8714 |
| chang· | 2.5093 | hive | 5.3891 | visualization | 2.346 | machine learning | 5.0114 | tools | 6.9592 | build | 4.6836 | analytics skills | 7.9907 | present· | 8.2963 | tableau | 8.6053 | sql | 6.8252 |
| environment | 2.3778 | statistics | 5.2105 | methods | 2.2406 | accuracy | 4.9476 | database | 6.5283 | etl | 3.6841 | accuracy | 7.6102 | communication skills | 8.2881 | word | 8.1118 | relationships | 6.0608 |
| aws | 2.0385 | process· | 5.0289 | statistics | 2.2008 | reports | 4.9036 | applications | 5.5783 | quick· | 3.6767 | design | 7.5994 | requirements | 8.282 | tools | 7.7513 | fast pace· | 6.0136 |
| team· player | 2.0314 | programming | 4.7534 | vba | 2.1712 | data mining | 4.8185 | powerpoint | 5.5191 | healthcare | 3.6692 | hadoop | 6.2884 | effect· | 7.971 | reports | 7.522 | concepts | 6.0096 |

# RESULTS OF CLUSTERING ANALYSIS OF DS

| Cluster1 | | Cluster2 | | Cluster3 | | Cluster4 | | Cluster5 | | Cluster6 | | Cluster7 | | Cluster8 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Term | Score | Term | Score | Term | Score | Term | Score | Term | Score | Term | Score | Term | Score | Term | Score |
| technologies | 1.5765 | machine learning | 3.8583 | machine learning | 8.4171 | written communication | 6.8176 | techniques | 9.6783 | deep learning | 5.3887 | written communication | 6.3994 | analysis | 8.8 |
| excellent communication | 1.2223 | natural language | 2.7834 | tools | 5.4713 | non technical | 4.206 | algorithms | 8.9651 | machine learning | 5.1707 | business | 5.824 | machine learning | 8.5232 |
| programming languages | 1.1592 | statistics | 2.606 | analysis | 4.9993 | analytical | 3.8091 | machine learning | 6.0935 | algorithms | 5.1226 | analytical | 5.464 | spark | 7.8958 |
| industry experience | 0.6685 | data processing | 1.5109 | python | 4.8812 | results | 3.7169 | big data | 4.4252 | written communication | 4.9362 | projects | 5.1345 | big data | 7.2161 |
| apache | 0.5621 | analytical tools | 1.4062 | technical | 4.2532 | environment | 3.3002 | svm | 4.0049 | python | 4.9064 | research | 4.6251 | algorithms | 7.2066 |
| programming language | 0.4988 | programming language | 1.2344 | scikit learn | 4.204 | technical | 3.2803 | models | 3.3292 | problem solving | 4.6914 | development | 4.6121 | hadoop | 7.086 |
| b testing | 0.4301 | optimization | 1.1286 | methods | 4.1942 | collaborate | 3.0376 | aws | 3.0928 | natural language | 4.2917 | techniques | 4.3219 | modeling | 6.7509 |
| solving skills | 0.4088 | computational | 0.9281 | non technical | 4.1583 | tableau | 3.0268 | programming | 3.0852 | java | 4.0068 | verbal communication | 4.217 | sql | 6.3434 |
| sql experience | 0.3925 | scala | 0.8784 | technologies | 3.3679 | fast paced | 3.0087 | naive bayes | 3.0645 | c | 3.8192 | hadoop | 3.9311 | statistical modeling | 6.0605 |
| powerpoint | 0.3841 | collaborate | 0.8549 | natural language | 3.2894 | programming | 2.9832 | technologies | 2.976 | frameworks | 3.7389 | problem solving | 3.915 | bayesian | 5.7367 |