Institute of Information Systems and Marketing (IISM)
Kaiserstraße 89-93
Kollegiengebäude am Kronenplatz (Geb. 05.20)
76133 Karlsruhe
http://iism.kit.edu

DEMYSTIFYING JOB ROLES IN DATA SCIENCE: A TEXT MINING APPROACH

Research Paper

Michalczyk, Sven, Institute of Information Systems and Marketing (IISM), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, sven.michalczyk@kit.edu

Dr. Nadj, Mario, Institute of Information Systems and Marketing (IISM), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, mario.nadj@kit.edu

Professor Dr. Mädche, Alexander, Institute of Information Systems and Marketing (IISM), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany,

Dr. Gröger, Christoph, Robert Bosch GmbH, Stuttgart, Germany, christoph.groeger@de.bosch.com

Abstract

The continuing proliferation of data science these days is causing organizations to reassess their workforce demands. Simultaneously, it is unclear what types of job roles, knowledge, skills, and abilities make up this field and how they differ. This ambiguity is generating a misleading myth around the Data Scientist's role. Against this background, this paper attempts to provide clarity about the heterogeneous nature of job roles required in the field of data science by processing 25,104 job advertisements published at the online job platforms Indeed, Monster, and Glassdoor. We propose a text mining approach combining topic modeling, clustering, and expert assessment. Therefore, we identify and characterize six job roles in data science that are in a request by organizations, described by topics classified in three major knowledge domains. An understanding of job roles in data science can help organizations in acquiring and cultivating job roles to leverage data science effectively.

Keywords: Roles in Data Science, Job Advertisement Analysis, Text Mining, Topic Modeling, Clustering

1 Introduction

Data science (DS) – "the study of the generalizable extraction of knowledge from data" (Dhar, 2013, p. 1) – has reached considerable attention in academia and practice in recent years (Sirje and Emmanouel, 2020). This interdisciplinary field requires a wide range of knowledge as it builds on different disciplines including "statistics, informatics, computing, communication, management, and sociology to study data and its environments […] by following a data-to-knowledge-to-wisdom thinking and methodology" (Cao, 2017, p. 8). Fueled by the continuing proliferation of DS these days, organizations are beginning to reassess their workforce demands. They are rapidly trying to secure the necessary knowledge in this field to remain competitive in the market (De Mauro *et al.*, 2018). However, the Data Scientist's sole job role is not able to cover the full spectrum of knowledge, skills, and abilities (KSAs) required in this field (Miller, 2014). Simultaneously, it is unclear what types of job roles make up this field and how they differ, generating a misleading myth around the Data Scientist's job role (De Mauro *et al.*, 2018). In particular, this job role is often nebulously defined and seen as capable of dealing with any kind of analytical, technical, or organizational problem throughout the process, from business understanding to solution deployment, leading to unrealistic expectations regarding the Data Scientist (De Mauro *et al.*,

2018; Miller, 2019). These considerations disregard the variety and specificity of the related job roles in DS that are necessary to collect, organize, and transform data into insights to create business value for organizations (De Mauro *et al.*, 2018).

In this regard, seminal work extracted KSAs from job advertisements (or short job ads) in related fields, such as big data (BD) (De Mauro *et al.*, 2018), business intelligence (BI) (Debortoli, Müller and Vom Brocke, 2014), analytics (Handali *et al.*, 2020), and Industry 4.0 (Pejic-Bach *et al.*, 2020). However, the data sets were mainly small and typically crawled from only one platform. For instance, De Mauro *et al.* (2018) crawled 2.786 job ads from Dice, while only Handali *et al.* (2020) crawled a higher amount of job ads (17.282) from Monster. While all studies commonly applied quantitative methods (e.g., topic modeling) to extract KSAs, only De Mauro *et al.* (2018) and Pejic-Bach *et al.* (2020) defined job roles. However, De Mauro (2018) used term frequencies in job titles to define job roles qualitatively, without the aid of quantitative methods and with a specific focus on BD. In contrast, Pejic-Bach *et al.* (2020) relied on hierarchical clustering to define job roles based on the previously extracted KSAs but explicitly focused on Industry 4.0. Thus, we see the need to systematically identify and characterize the most prominent job roles and KSAs in DS by combining topic modeling, clustering, and expert assessment.

Against this background, this paper attempts to provide clarity about the heterogeneous nature of job roles required in DS by analyzing and preprocessing 25,104 job ads published online on the job platforms Indeed, Monster, and Glassdoor. Hereby, we formulated the following research question: *What are the most prominent job roles in DS, and how are they characterized?* To address this research question, we propose a text mining approach based on a combination of topic modeling, clustering, and expert assessment. We use a topic model for topic extraction of job ads. Subsequently, we rely on these topics as input features for a partition-based clustering algorithm to define the underlying job roles. With this, topics consist of frequent word combinations from the job ads, and job roles consist of a distribution of topics. On this basis, we assign topics to three KSAs classes (i.e., business, analytical, and technical) as suggested by Cegielski and Jones-Farmer (2016), identify six job roles in DS that are requested by organizations (i.e., Data Analyst, Data Engineer, Data Scientist, Business User, Software Developer, and Software Architect), and characterize each job role. For instance, according to our analysis, Data Scientists have the highest analytical expertise of all identified job roles, as could be expected. They are specialists in the analytical domain with a deep understanding of analytical methods (e.g., machine learning, short ML) and underlying programming languages (e.g., Python). Besides, a strong understanding of the technical domain (e.g., BD technologies, cloud architectures) completes this job role. Remarkably, our results did not reveal any explicit KSAs in the business domain. Thus, a collaboration with domain experts seems to be necessary to perform the underlying analyses effectively. In turn, our results suggest that the Data Analyst requires business expertise, such as expertise in quality management, supply chain management, manufacturing, or customer relationship management. However, the business expertise's dominance comes at the expense of the analytical and technical KSAs of this job role. Specifically, Data Analysts make greater use of traditional analytical methods to perform their analyses and produce reports with query results for Business Users but lack dedicated technical and statistical expertise.

In summary, such job role understanding can (1) help organizations to define strategies and a common language for acquiring and cultivating the right KSAs to effectively use DS, (2) support employees to evaluate their KSAs in DS, and (3) supplement related training and education portfolios. Methodologically, our text mining approach is reproducible to define job roles in other domains. Theoretically, we deliver an empirically grounded, up-to-date conceptualization of job roles in DS. The remainder of the paper is structured as follows: First, we introduce related work and describe our method. Next, we illustrate and discuss the results. Lastly, we conclude our article.

2 Related Work

From an academic standpoint, DS boundaries have not been formally determined yet (De Mauro *et al.*, 2018). Consent exists that this interdisciplinary field covers a broad spectrum of KSAs and related job

roles (Cao, 2017, p. 8). Therefore, scholars have started to offer multiple descriptions of what DS should display (De Mauro *et al.*, 2018). For instance, Muller et al. (2019) moved from the term Data Scientist to DS workers because many different people are doing the work of DS with a broad range of tools in various contexts and across multiple job titles. From a process perspective, DS workers perform the following activities: "discovery", "capture", "curation", "design", and "creation". "Discovery" refers to locate the data without applying transformations. "Capture" involves filtering and deleting for subsequent analysis. "Curation" considers activities that transform the data into a format for visualization tools, databases, or statistical packages. If the format is tabular, this is known as tidying (Wickham, 2014). The next activity is "design" and includes the creation of new data by integrating existing data, where "data appear to be more produced than discovered or revealed" (Muller *et al.*, 2019, p. 3). The activities *curation* and *design* are equivalent with *data preparation* in the well-established CRISP-DM methodology (Chapman *et al.*, 2000). Lastly, "creation" is the activity that enables knowledge creation as "ground truth data" (Muller *et al.*, 2019, p. 11).

To classify KSAs in DS, several job ad studies relied on a coding schema developed by Todd et al. (1995). This scheme categorizes KSAs along with three knowledge domains (KDs), business, analytical, and technical, and is well-established and successfully applied in related studies (e.g., Debortoli, Müller and Vom Brocke, 2014; Cegielski and Jones-Farmer, 2016; Handali *et al.*, 2020). For instance, Debortoli *et al.* (2014) illustrated considerable differences between BI and BD KSAs along the analytical and technical KDs. In particular, the KSAs for BI were related to commercial products from prominent software vendors. However, no such reference could be established for the KSAs in the BD field. Besides, most BD job ads analyzed required advanced software development skills, statistical knowledge, and expertise in ML. In turn, these KSAs were requested to a much lesser degree in BI job ads. Additionally, the business KD was equally essential for both fields BI and BD. Furthermore, Gurcan and Cagility (2019) arrogated today's Software Developers and Engineers to be Data Scientists. This study reported that one-third of Software Developers and Engineers require KSAs that refer to data-related technologies (technical KDs) and analytical KDs. Thus, Software Developers and Engineers also seem to contribute to the DS field but rather on the technical end of the spectrum.

Practice does not seem to have clearly drawn the DS boundaries either. Saltz and Grady (2017) analyzed five case studies from two standard bodies (i.e., NIST, 2015; EDISON, 2017), two industry organizations (e.g., SAIC (Grady, 2016); Springboard, 2020), and the advisory firm Gartner (Idoine *et al.*, 2018). A set of six job roles occurred: Data Scientist, Data Engineer, Data Architect, Data Science Programmer, Data Science Researcher, and Data Analyst. Although these are among the top self-reported job roles in a sample of 19,717 DS professionals (Kaggle, 2019), only the Data Scientist and the Data Engineer appear in all cases analyzed.

Against this backdrop, text mining is a promising approach to demystify job roles in DS, as research in related fields such as BD (De Mauro *et al.*, 2018), BI (Debortoli, Müller and Vom Brocke, 2014), analytics (Handali *et al.*, 2020), and Industry 4.0 (Pejic-Bach *et al.*, 2020) has successfully extracted and described sets of KSAs based on job ads. Hereby, online job platforms such as Monster (Handali *et al.*, 2020), Dice (De Mauro *et al.*, 2018), and LinkedIn (Pejic-Bach *et al.*, 2020) were crawled. In the analysis step, term frequencies and statistical models such as topic modeling using Latent Dirichlet Allocation (LDA) (De Mauro *et al.*, 2018; Gurcan and Cagiltay, 2019; Handali *et al.*, 2020), Latent Semantic Analysis (Debortoli, Müller and Vom Brocke, 2014), or k-Nearest Neighbor (Wowczko, 2015) were applied. Topic modeling approaches were of particular interest, as frequent terms are non-exclusively clustered into topics based on the term's job ad affiliation, thus providing a dense description of the data. Beyond just describing the data in terms of KSAs, two studies went on to define job roles. De Mauro *et al.* (2018) used term frequencies in job titles to define job roles qualitatively, described by a fitted topic model. Pejic-Bach *et al.* (2020) applied a hierarchical clustering approach to previously extracted most frequent bi-grams.

In summary, the boundaries of DS are not yet formalized. Using the term DS worker, Muller *et al.* (2019) characterized this discipline by defining sets of activities without creating roles. Depending on the nature and volume of data, the job role's KD distributions vary (Debortoli, Müller and Vom Brocke,

2014). First attempts from practice and research to define roles in DS exist. In this context, text mining is a promising approach because (1) a large amount of data can be analyzed globally, (2) job ads represent the current demand of human resources in the industry, and (3) in conjunction with topic modeling, a dense description of job ads can be extracted as a basis for defining job roles in specific fields. In the following sections, we build on this related work and suggest a text mining approach that combines topic modeling, clustering, and expert assessment to demystify and characterize which job roles are in demand by organizations in DS.

3 Method

Our method relies on text mining based on seminal work in related fields such as BI (Debortoli, Müller and Vom Brocke, 2014) or Industry 4.0 (Pejic-Bach *et al.*, 2020). Figure 1 summarizes our methodological approach. First, we *crawled* job ads with a combination of different *search terms* from three major online job platforms *from March to June 2020* (see step 3.1 in Figure 1). Next, we preprocessed documents by *striping the HTML tags* and *keeping only English job ads*. In addition, we relied on the cosine similarity to identify and discard duplicates (see step 2). Afterward, we *tokenized* the documents into *n-grams* which we subsequently pruned in several iterations by applying common text mining practices (see list items in step 3.3) to *reduce the dimensionality*. The resulting n-grams with their respective *TF-IDF weighted frequencies* per job ad serve as features for the model (see step 3.3). Next, we fitted an *LDA* topic model with *Gibbs sampling* to create dense representation in the form of interpretable topics (see step 3.4). We used a *grid search for parameter tuning*. Lastly, we extended the seminal work by applying the *clustering algorithm K-Medoids* to the new data representation in form of *probability distributions of topics per job ad* to define job roles (see step 3.5). Because of this data representation, we used the Jensen Shannon Divergence *(JSD) as distance measure*. The results of steps four and five are assessed quantitatively by relying on the silhouette method, and qualitatively, by experts who labeled and pruned topics respectively clusters (see step 3.6).

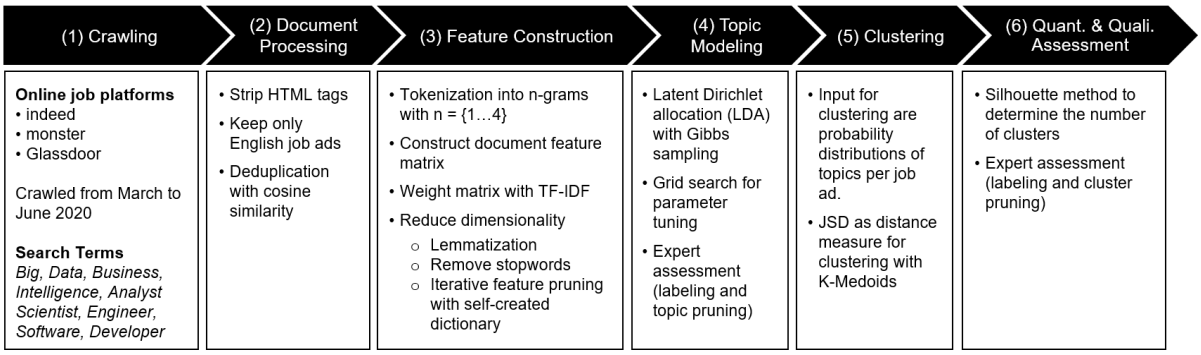

Figure 1. Methodological Approach

(1) Crawling. Based on the related work, we collected job ads with a combination of the following search terms: *big data* (Debortoli, Müller and Vom Brocke, 2014; De Mauro *et al.*, 2018; Gardiner *et al.*, 2018) or *business intelligence* (Debortoli, Müller and Vom Brocke, 2014) or *data analyst* (Lyon and Mattern, 2017) or *data scientist* (Schmid and Baars, 2016) or *data engineer* (Lyon and Mattern, 2017) *software engineer* or *software developer* (Gurcan and Cagiltay, 2019). We observed that online job platforms have a limited capability of using logical operators. Search terms are cross combined, which leads to the inclusion of relevant terms like *big data developer* or *business analyst*. To avoid getting a bias from one online job platform, we selected three platforms that are among the top ones according to job seekers and recruiters' visits worldwide (Jobboard Finder, 2021). In particular, we crawled Indeed (18,811), Monster (3,500), and Glassdoor (2,768). Because job ads get frequently published and deleted once a job is taken, we decided to run our crawler two times a week from March to June 2020 for four months. We developed our crawler in Python, relying on Beautiful Soup (2020) as a crawling framework, Requests for sending HTTP requests (2020), and Selenium, as a web driver (2020).

(2) Document Processing. Next, we investigated the structure of our 25,104 crawled job ads. We observed that a job ad usually has five building blocks: (a) a company section, (b) a description of tasks and responsibilities, (c) a section about the requirements for this job, (d) a section about the offering and benefits, and lastly (e) a specification about the application process, which is in line with Cegielski and Jones-Farmer's (2016) classification. To define roles, we are solely interested in sections (b) and (c) because they contain the company's expectation about the KSAs of a candidate respectively job role (Cegielski and Jones-Farmer, 2016). Unfortunately, the HTML tags did not reflect those building blocks on any platform. Ideally, HTML tags would have enabled the automated extraction of these sections using path languages like XPath (W3C, 2017). Consequently, we decided to use the full description and strip the HTML tags by relying on the R packages rvest (Wickham, 2021), stringr (Wickham, 2019), and rebus (Cotton, 2017). Next, we kept English job ads only by using Google's Compact Language Detector 3 (Ooms, 2021) because (1) they make up the major proportion of 20,390 job ads, and (2) we require a common language for the feature construction step. In general, the data preparation was assisted by dplyr (Wickham, 2020) and we visualized our results with ggplo2 (Wickham *et al.*, 2020).

The document processing is finished by removing duplicates and similar job ads. This step is necessary because we crawled platforms twice a week, which resulted in many duplicates. Therefore, we compared exact matches and additionally identified similar job ads by calculating the cosine similarity. Such a course of action is suitable for our application as it does not consider the length of a job ad but the most common features. For calculating, we already ran through a couple of feature construction iterations described in the next section because the cosine similarity is a feature-based similarity measure (Gomaa and Fahmy, 2013). It represents the similarity as a vector of inner product space measuring the pairwise angle between job ads (Han, Kamber and Pei, 2012). We had to set a threshold above which we consider job ads to be similar for the deduplication. We started with a threshold of 99%, and stepwise decreased the measure investigating in each step the similarity of a representative sample of the candidates. To judge the similarity, we compared job titles, companies hiring, and if necessary, the full text of the job ads. We finally defined a threshold of 90%, which identifies reliable changes of a job ad irrelevant to our analysis, such as changes of contact persons, locations, or spelling corrections made after republishing job ads.

(3) For feature construction, 11,402 job ads remained. We applied the vector space model (Salton, Wong and Yang, 1975), where each job ad represents a vector of frequencies of each term. Thus, the order of terms is not considered (i.e., bag of words model). Therefore, we decided to account beside unigrams (single terms) on bi-, tri-, and four-grams, which are combinations of two, three, and four terms in a sequence (Nigam et al., 2000). For instance, this also created the feature *problem_solving* instead of having only two separate terms, *problem* and *solving*. The general form is called n-gram (e.g., a unigram has an $n = 1$ and a bigram $n = 2$). Subsequently, we constructed a document feature (term) matrix (DFM) (Manning, Raghavan and Schütze, 2008). The rows are the job ads, and the columns refer to the features. Initially, our matrix consisted of six million (m) features. As this matrix was sparse, not many shared features among the job ads existed, but those were of particular interest. Next, we performed the following dimensionality reduction steps: First, we removed standard stop words (Snowball, 2020) such as *at* or *very* (5.4 m features left). Subsequently, we applied a lemmatization dictionary (Mechura, 2017), which replaces words such as *are* and *am* with the infinitive *be* (leading to 5.2 m features). After that, we consulted previous research and evaluated well-established thresholds to trim the matrix. We decided to prune features present (1) in more than 50% of the job ads and (2) in less than 1% of the job ads (Manning, Raghavan and Schütze, 2008). The upper limit of 50% discards widespread features like *data*, *team*, *experience*, *skill*, or *knowledge* that are not differentiating job ads. The lower limit of 1% was also applied by the related work of Debortoli *et al.* (2014) and successfully discarded particular features in our data set, such as *meteosat* and n-gram features like *validation_standard* or *services_time*. Next, we applied a custom lemmatization dictionary, leading to 7,462 features. Thereby, we grouped similar features like *analytics, analyse and analysis* into one lemma *analyt* as the standard lemmatization dictionary does not contain them. Alphabetic sorting of unigrams helped us to identify domain-specific lemmas.

Finally, we developed an iterative process to prune features that were not part of the DS field. First, we ordered all features decreasingly by the overall frequency in all job ads, taking DFM's row sums. The sums lead us to consider the features with higher significance first. Subsequently, we went through approximately 20% of the features and decided whether they can be part of the DS field. We would need to decide for prune-able unigram features if they had some explanatory power as part of an n-gram. If not, we could prune all features that contained this feature as n-gram. If that applied, we would mark them with "discard from n-gram construction" otherwise with "discard as unigram". We decided not to apply this logic for prune-able n-grams with $n > 1$ due to practicability, although it is reversely possible to apply this logic. After we marked approximately 20% of the data, we removed flagged features from the DFM. Once again, we started with the top features of the updated DFM. The distinction between two flags has the advantage that in the subsequent iteration, $n\text{-grams} > 1$ containing unigrams with the flag "discard from tokenization" are pruned and do not have to be labeled again, which makes the process more efficient. We ran through this process for approximately ten iterations leading to 3.724 highly relevant features. As a common text mining practice (Beel *et al*., 2016), we applied a TF-IDF weighting to the matrix. TF-IDF is a measure of a feature's originality by comparing the frequency of features in a job ad with the number of job ads the feature appears in. The feature constructing step resulted in a DFM, with job ads as rows, features as columns, and TF-IDF-weighted feature counts as values conducted with the assistance of quanteda (Benoit *et al.*, 2020).

(4) Topic Modelling – Extracting Knowledge Domains. Top features could be considered as first characteristics, but this information is still sparse, and the relation of features stays unrevealed. We followed the suggestions of related studies (De Mauro *et al.,* 2018; Gurcan and Cagiltay, 2019; Handali *et al.*, 2020) and applied a topic modeling relying on an efficient implementation of Blei *et al.* (2003) by using the R package topicmodels (Grün *et al.*, 2020). Although newer clustering approaches for textual data consider the semantic meaning of words in sentences (Moody, 2016), they are unsuitable for our study as they add distortions to our kind of data (Handali *et al.*, 2020). For instance, many technologies required for DS are polysemes, such as *Spark* or *Storm.* However, a topic model extracts latent information in the form of topics using the co-occurrence of features in job ads. For instance, *Java* and *programming* might be forming one topic. It is a mixed membership model allowing features to account for different topics (Airoldi *et al.*, 2008). As a result, two probabilities distributions are estimated: (1) the distribution of features per topic, (2) the distribution of topics per job ad.

To fit a topic model, we needed to set three primary parameters. A crucial parameter is k, the number of topics to obtain. We relied on a quantitative measure as suggested by Devauad *et al.* (2014) to select the number of topics in conjunction with a qualitative assessment. Thereby, we trained models with k from *3 to 100*, leading to the optimal number of topics at $k = 30$. We also had a consensus for this number of topics from our qualitative assessment because 20 topics described the dataset imprecisely, whereas 40 topics lead to many irrelevant topics. The following two parameters are priors, which serve as initial values for approximating the distributions. Alpha controls for the per job ad topic distribution. A high alpha means that a job ad is likely to contain a mixture of all topics and not any single topic specifically. For setting alpha, we followed Griffiths and Steyvers (2004) and set it to 50 divided by the number of topics. Beta, in turn, controls the prior for the topic-feature distributions. A high beta means that each topic is likely to contain a mixture of most of the features. We set beta slightly below the default value suggested by Grün and Hornik (2011), with beta equals 0.1 multiplied by 95% to control for a more distinct feature allocation per topic. We relied on Gibbs sampling instead of the VEM algorithm to approximate these distributions, as we achieved better results with lower computational costs.

Our final model consists of 30 topics explained by feature distributions per topic. For instance, *machine_learning, communication,* or *data_science* are features describing the topics. To get a holistic and foundational base for defining job roles in DS, we classified our labeled topics with three KDs: business, analytical, and technical. This classification was first developed by Todd *et al.* (1995) and successfully applied by related studies (Todd, McKeen and Gallupe, 1995; Debortoli, Müller and Vom Brocke, 2014; Cegielski and Jones-Farmer, 2016; Handali *et al.*, 2020). On this basis, we were able to dive deeper into the individual KD items. Thereby, we included essential functions of a business such

as manufacturing or finance, and general KSA unclassifiable into the analytical or technical direction (like language professions), into the business KD (Handali *et al.*, 2020). Furthermore, we relied on Debortoli *et al.* (2014) for the classification decisions regarding the technical KD. Accordingly, KSAs like BD technologies and architectural expertise were grouped into this KD. Additionally, we included KSAs like those related to statistics, ML, and visualization into the analytical KD, as suggest by Cegielski and Jones-Farmer (2016) and Handali *et al.* (2020). Two researchers and one professional did the classification and labeling independently to ensure intercoder reliability. A third researcher consolidated the results and resolved discrepancies. We discarded six topics because they were irrelevant for our context (Boyd-Graber, Mimno and Newman, 2014). Some topics are related to the company's description or benefits as they contain features like *lunch* or *cut_edge_technology*. This resulted in a total of 24 topics.

(5) Clustering Job Roles. As a result of our topic model, we reduced the dimensionality from 3,724 n-gram features to distributions of 24 topics classified in three KDs. To define job roles, we used the topic assignment by the topic model as input to a clustering algorithm. Thus, we required a measure of pairwise dissimilarities between job ads, as clustering algorithms typically use the distance between objects for grouping them. We relied on the JSD because it is a dissimilarity measure between two probabilities distribution (Endres and Schindelin, 2003). Our results fulfill this property because the topic model describes each job ad as a probability distribution of topics. The resulting dissimilarity matrix (with 11,402 columns and rows) served as input for the clustering. We tried out the density-based clustering algorithm DBSCAN (Ester *et al.*, 1996). Two parameters, the density threshold minPts, and neighborhood radius eps, needed to be specified. For minPts, we relied on an established heuristic and set it to the number of dimensions equal to 24 topics. To determine epsilon, we relied on a kNN-distance plot (Hahsler, Piekenbrock and Doran, 2019). However, this setting leads to one big cluster with some job ads labeled as noise. A 3D scatterplot revealed the ellipse-shaped sphere's structure, which indicates that the transitions between job roles are relatively smooth. Thus, we decided to apply a partition-based clustering approach to cut the sphere into partitions. We used K-Medoids (Kaufmann and Rousseeuw, 1990) as it is a robust alternative to k-means clustering and less sensitive to noise and outliers because it uses medoids as cluster-centers instead of means.

(6) Quantitative and Qualitative Assessment of Clustering Results. Critical in the application of K-Medoids is to determine the number of clusters that describe the underlying data structure best. To increase objectivity, we followed a two-step assessment of the clustering result, consisting of a quantitative and a qualitative assessment. First, we trained for $k = \{3 \ldots 13\}$ models and plotted the average width of clustering silhouettes. This is a well-established approach to evaluate a clusterings' validity because the silhouette "shows which objects lie well within their cluster" (Rousseeuw, 1987, p. 1). The optimal number of clusters is indicated by a maximum in the curve, in our case, at $k = 9$. Thus, 11,402 job ads were best separated into nine clusters. To explain these clusters regarding the 24 topics, we aggregated the topics from the job ads to the clusters' level.

In the next step, we qualitatively assessed the clustering results. Two researchers and one professional with working experience in DS labeled the clusters independently as potential job roles. We only considered the ten highest loading topics per cluster in our assessment to keep our analysis focused. In the labeling process, we discarded three clusters for the following reasons: We discarded the first cluster due to the heavy loading of the cloud architecture topic (T7). It is only supported by 154 job ads that is below the average of 1,262 job ads per cluster, and is therefore considered an outlier. Because the second cluster contains a weak uniform distribution of the topics "Communication & Problem-Solving Ability" (B2), "Big Data Technologies" (T3), "Marketing" (B4), "Consulting" (B10), we discarded this topic with the consensus of the labeling experts. The third cluster might describe a clinical or laboratory scientist. However, we would not like to highlight any specific context demanded at the time of crawling. Hence, we decided to remove this cluster because it was the only context-specific cluster. In summary, we identified a total of six job roles in the nine clusters. The job roles are described by a combination of KSAs along with the business, analytical, and technical KDs. The results are elaborated on in the following sections.

4 Results

4.1 Knowledge Domains

Topic	High-Loading Features (Probability %)
B1: Work Experience	amazon (7.33), +_year_experience (4.64), 3_+ (2.02), 3_+_year (2), 5_+ (1.93), 5_+_year (1.8), aws (1.79), connect (1.71), 2_+ (1.48), 2_+_year (1.41)
B2: Communication & Problem-Solving Ability	communicate (8.89), communicate_skill (5.79), analytic (4.6), skill_ability (3.21), verbal (2.45), interpersonal (2.07), problem_solve (1.68), write_communicate (1.63), attention (1.56), oral (1.49)
B3: Customer Relationship Management	relationship (3.12), strategy (2.39), plan (2.13), presentation (1.67), strategic (1.67), revenue (1.53), manage (1.48), lead (1.37), sell (1.37), executive (1.27)
B4: Marketing	brand (2.41), content (2.41), tech (1.97), mobile (1.82), consumer (1.82), web (1.81), passion (1.72), app (1.6), ad (1.45), search (1.36)
B5: Professional Language Skills	english (6.63), professional (2.98), fluent (2.73), dynamic (1.82), advantage (1.58), german (1.36), fluent_english (1.29), master (1.2), professional_experience (1.01), start (0.93)
B6: Finance & Risk	financial (7.06), risk (4.64), bank (3.27), finance (3.12), investment (2.08), asset (1.89), write_verbal (1.4), trade (1.26), credit (1.26), verbal_communicate (1.25)
B7: Project & Quality Management	report (3.59), plan (3.41), review (2.83), implementation (1.85), audit (1.76), control (1.72), documentation (1.69), maintain (1.67), manage (1.64), compliance (1.57)
B8: Supply Chain Management & Manufacturing	operation (6.12), automate (3.92), improvement (3.24), supply (2.53), operational (2.35), plan (2.2), chain (2.14), manufacture (1.99), supply_chain (1.78), optimization (1.76)
B9: Leadership & Management	lead (9.15), manage (4.76), leadership (4.35), senior (4.05), manager (3.74), delivery (2.52), leader (2.07), strategy (1.99), mentor (1.7), roadmap (1.28)
B10: Consulting	consult (5.05), professional (2.9), lead (2.54), transformation (2.37), consultant (2.15), train (2), strategy (1.79), expert (1.78), delivery (1.75), big_data (1.52)

Table 1. Topics in the business KD

Business. We identified ten business-oriented topics. The first topic (**B1**) highlights that between two and more than five years' "Work Experience" is required, denoted by the feature *+_year_experience* and *year* in associated with different numbers. We considered the features of *amazon* and *aws* in this topic as outliers because most features, namely seven of the top ten features, indicate a relation to work experience. Unfortunately, we could not generally prune *amazon* and *aws* in the previous feature construction step due to their relevance in other topics. The next topic (**B2**) has two complementary components. First, "Communication", with the top features *communicate*, *communicate_skill*, and second, features like *verbal*, *write_communicate*, and *oral,* which are consistent with the definition of communication as "verbal or written message" (Merriam-Webster, 2021). We derived the second component, "Problem-Solving Ability" from the third-highest loading feature *analytic* together with *problem_solve* as *interpersonal* KSA, which lead to the topic's label "Communication & Problem-Solving Ability". In topic **B3**, the features *revenue* and *sell* indicate a relationship to the customer, and in combination with *relationship* and *manage,* resulting in the label "Customer Relationship Management" (CRM). Other features like *strategy* and *plan* support this assumption, leading to topic **B3** "Customer Relationship Management", a subdiscipline of Marketing Strategy (Kumar, 2010). The unique feature combination in the topic **B4** of *brand*, *content*, *consumer*, *web*, *search*, and *ad* leads to the assumption that the business function "Marketing" is predominant in this topic because these features describe fundamental building blocks of marketing, as marketing "includes advertising, selling, and delivering products to consumers or other business" (Twin, 2020). The next topic (**B5**) describes the requirement of "Professional Language Skills" in *English* and *German*. The topic "Finance & Risk" (**B6**) was clearly identified by the same-named top features, *financial* and *risk*. Other features like *band, investment, asset, trade,* and *credit* support this assumption because all are strongly related to finance. Additionally, *written* and *verbal communication* skills seem to be important in this topic. The follwoing topic **B7** consists of two components. The first one, "Project Management", was featured by *plan* and *report*; the second one, "Q*uality Management*" (QM) was featured by *review*, *audit*, and *compliance*. In combination with *manage*, we labeled this topic "Project & Quality Management". The next topic (**B8**)

focuses on the *operations* of a company, particularly on the *optimization* of the *supply_chain* by *automating* and *improving* it. The field of *manufacturing* is related to that. Therefore, we labeled this topic "Supply Chain Management & Manufacturing". Next, the topics "Leadership & Management" (**B9**) and "Consulting" (**B10**) were derived by the highest loading features *leadership* and *consult,* respectively. In summary, the topic **B9** and **B10** can, like the topics **B3** "Customer Relationship Management", **B4** "Marketing", **B6** "Finance & Risk" and **B7** "Project & Quality Management", be considered as business functions in which DS, as our results reveal, plays an important role. The topic **B8** is a hybrid in which "Supply Chain Management" refers to the latter, and "Manufacturing" is a core function directly related to a company's value chain. **B1**, **B2,** and **B5** refer to *interpersonal* KSAs.

Topic	High-Loading Features (Probability %)
T1: Cloud Architecture (Amazon)	aws (8.15), cloud (3.62), architec (3.25), compute (2.91), implementation (1.81), amazon (1.52), database (1.5), professional (1.24), web (1.23), cloud_compute (1.22)
T2: Computer Science Degree	computer_science (6.51), c (4.29), degree_computer (3.27), degree_computer_science (3.1), bachelor (2.07), c_+ (2.04), c_+_+ (1.99), python (1.87), algorithm (1.84), software_engineer (1.62)
T3: Big Data Technologies	big_data (7.57), spark (4.36), hadoop (3.63), python (2.71), scala (2.15), distribute (2.01), java (2.01), pipeline (1.96), stream (1.92), hive (1.66)
T4: IT Operations	security (6.76), infrastructure (3.63), linux (2.57), automate (2.26), script (2.18), maintain (2.14), troubleshoot (2), server (1.84), database (1.71), deployment (1.54)
T5: Agile Software Engineering & Development	code (4.84), agile (3.54), developer (3.06), software_engineer (2.64), web (2.18), software_development (2.04), framework (2.03), integration (1.68), java (1.61), automate (1.61)
T6: Software Development	c (2.63), amazon (2.28), software_development (1.75), architec (1.47), object-oriented (1.31), java (1.3), architec_design (1.24), c_+_+ (1.13), java_c (1.12), c_+ (1.1)
T7: Cloud Architecture (Microsoft & Google)	cloud (12.02), architec (4.88), azure (3.4), google (2.47), infrastructure (1.68), microsoft (1.68), cloud_platform (1.35), big_data (1.25), google_cloud (1.25), implementation (1.16)

Table 2. Topics in the technical KD

Seven **technical** topics emerged from our analysis. The topics **T1** and **T7** are related to "Cloud Architecture" as they include *cloud* and *architecture* as top features. The difference lies in the expertise in specific cloud technologies, such as *amazon's aws* in **T1**, as well as the *google_cloud* and *microsoft azure* in **T7**. Therefore, we labeled the topics "Cloud Architecture" with the respective provider in parentheses. **T7** might be more provider agnostic as *cloud* is the top feature compared to **T1,** in which *aws* is the top feature. The next topic (**T2**) requires a degree in *computer_science* with experience in the programming languages *C*, *C++*, and *Python*. The next topic (**T3**) refers to "Big Data Technologies", such as *Spark*, *Hadoop,* and *Hive.* Thereby, the programming languages *Python*, *Scala,* and *Java* seem to be in demand to process data in *streams*. These features were also categorized in the BD field by Debortoli's *et al.* (2014) and differently labeled as "distributed programming" by Handali *et al.* (2020). We related the next topic, **T4,** to "IT operations", which objective is to monitor, control, and *maintain infrastructure* components and applications in a day-to-day routine as defined by the ITIL (2007). "IT operations" includes the administration of *servers* and *databases*, which requires *Linux*, a well-established Unix-like operating system. The next topic (**T5**) refers to "Agile Software Engineering & Development" because *agile, software_engineer*, and *software_development* are the top features. Other features like *code*, *framework,* and *integration* support this label. **T6** contains *object-oriented software_development* as features complemented by classical programming languages such as *C, C++,* and *Java,* underlining the *software_development* focus. The *design* of *architecture* is also crucial for this topic, representing a relevant developer KSA (Gurcan, 2019). Unlike the feature *agile* in T5, this topic leads to no assumption about a project management methodology. In summary, we named **T6** "Software Development".

We identified seven **analytical** topics from a total of 24 topics. The first topic **A1** is related to the ability to derive *actionable* (2.0%) insights from data because it is constructed from the top features *decision* and *analytic*. Therefore, we labeled A1 with "Analytical Decision-Making". We labeled the following

topic (**A2**) as a concatenation of the topic's top features, *statistics*, *model,* and *technique*. Our choice is supported by the features *R*, indicating the statistical environment and language *R*, the features *mathematics* and *quantitative,* and one of the objectives of *statistics*, that is, making *predictions*. Also, *data_mining* can be found in this topic, a predecessor of what we call ML today, grounded in multivariate *statistics* (Fayyad, Piatetsky-Shapiro and Smyth, 1996). Next, we labeled **A3** with "Business Intelligence" because it contains features like *report*, *sql*, *visualization*, and *dashboards* that are essential constituents in *BI* to do *data analytics*. With those terms, Howard Dresner introduced the field of BI in 1989 (Power, 2007). In the next topic (**A4**), *machine_learning* is by far the highest loading feature in topic **A4,** which we therefore used as a label for this topic. Features like *deep learning* and *natural* language processing support this choice, characterizing specializations of ML. We gave the topic **A5** the label "Data Science" as *data_science* is among the top features in this topic. Other assigned features foster the understanding of what is required to do DS, such as *experience* with *data, Python* and *R* as common programming languages, and the importance of *visualizations*. The next topic (**A6**) focuses on science, with the *scientist* doing *scientific experiments* in *clinical* or *lab* settings possibly holding a *Ph.D.* degree. In summary, we called topic **A6** "Scientific Experience". The topic "Data Engineering" (**A7**) explicitly contains the job role label *data_engineer* and is described with features like data *warehouses*, *sql*, and the implementation of *ETL pipelines*.

Topic	**High-Loading Features** (Probability %)
A1: Analytical Decision-Making	decision (5.43), analytic (3.89), recommendation (2.43), actionable (1.97), metric (1.69), decision_make (1.47), analyse (1.45), strategy (1.42), quantitative (1.41), translate (1.32)
A2: Statistical Modelling Techniques	statistic (10.37), model (9.72), technique (2.99), predict (2.8), r (2.55), mathematics (2.3), quantitative (1.93), mining (1.85), data_mining (1.39), predict_model (1.36)
A3: Business Intelligence	data_analyt (8.51), report (7.49), sql (3.83), visualization (3.29), bi (2.88), tableau (2.62), dashboard (2.01), business_intelligence (1.79), data_visualization (1.73), analyt_data (1.69)
A4: Machine Learning	machine_learning (15.76), model (3.9), algorithm (3.23), deep_learning (1.99), machine_learning_model (1.44), artificial (1.35), artificial_intelligence (1.31), scientist (1.28), learning_model (1.27), natural (1.19)
A5: Data Science	data_science (10.53), scientist (10.13), data_scientist (9.76), python (4.15), experience_data (2.27), r (2.09), model (1.45), visualization (1.18), senior_data (1.08), science_team (1.07)
A6: Scientific Experience	scientific (3.35), clinical (2.72), scientist (2.41), study (2.34), phd (1.83), university (1.77), patient (1.65), experiment (1.37), lab (1.36), r_have (1.24)
A7: Data Engineering	data_engineer (3.3), warehouse (2.75), etl (2.57), data_warehouse (2.45), sql (2.36), database (2.28), data_model (2.07), pipeline (2.01), model (1.89), implement (1.43)

Table 3. Topics in the analytical KD

4.2 Clustered Roles

We have identified a spectrum of six job roles in DS: The Business User, the Data Analyst, The Data Scientist, the Data Engineer, the Software Architect, and the Software Developer (Table 4). The first job role, the **Business User,** is purely constructed from business topics. It represents a customer-oriented job role as CRM (B3, most frequent), consulting (B10), and marketing (B4) are assigned. Thereby, this role must be able to communicate well to solve customer's problems (B2), takes a leadership position (B9), and is often accounted for the management of projects, ensuring the quality of *implementations* (B7). The next role, the **Data Analyst,** is responsible for making decisions (A1) with the assistance of BI systems (A3). In particular, data are loaded from data warehouses which this job role subsequently visualizes, analyzes, and puts into reports. In order to make decisions based on data, this job role already makes first use of statistical modeling techniques (A2) and is requiring business domain knowledge in "Supply Chain Management", "Manufacturing" (B8), CRM (B3) or "Finance & Risk" (B6). In turn, the **Data Scientist** has no specific business KD assigned. However, this job role relies strongly on statistics (A2), ML (A4), and DS (A5) for decision making (A1). Technologically, expertise with BI (A3) and BD (T3) is required with an understanding of the underlying computer science concepts (T2). For this job role, work experience is assumed (B1). The **Data Engineer** implements ETL pipelines to extract

data (A7) from source systems into other systems such as data warehouses (A3), which includes the definition of data models. This is fostered by understanding how data is subsequently analyzed by other DS job roles (A3, A5). Besides, expertise with BD technologies (T3) and cloud architecture is required (T1, T7). Working with BD technologies and the cloud assumes a computer science degree (T2) but also work experience (B1) with KSAs in "Agile Software Engineering & Development" (T5). The next role, the **Software Architect,** has substantial experience with cloud architecture (T1) and BD technologies (T3). The necessary background for this is a computer science degree and the ability to communicate well (B5). The latter is because architectural decisions, which definition is the primary responsibility of a software architect, need to be aligned with business functions such as "Marketing" (B4) or "Finance & Risk" (B6). The **Software Developer,** with the self-name topic "Software Development" (T6) firmly assigned, is characterized by having a computer science degree (T2) and by being a senior role with much work experience (B1) and leadership KSAs (B9). The working style of this job role is agile (T5). Also, technologies like the cloud (T1) and BD (T3) are essential for the software developer. Additionally, this role is involved in ML (A4), properly when models need to be deployed in production.

Role (Count Job Ads)	**High-Loading Topics** (Probability %)
Business User (596)	B3: Customer Relationship Management (14.44), B10: Consulting (6.99), B9: Leadership & Management (6.31), B4: Marketing (5.61), B7: Project & Quality Management (3.75), B2: Communication & Problem-Solving Ability (3.67)
Data Analyst (1531)	A3: Business Intelligence (9.87), A1: Analytical Decision-Making (7), B7: Project & Quality Management (6.32), B2: Communication & Problem-Solving Ability (4.69), B6: Finance & Risk (4.64), A2: Statistical Modelling Techniques (4.32), B8: Supply Chain Management & Manufacturing (3.98), B3: Customer Relationship Management (3.77), B9: Leadership & Management (3.52)
Data Scientist (1927)	A2: Statistical Modelling Techniques (12), A4: Machine Learning (10.67), A5: Data Science (6.79), A1: Analytical Decision-Making (4.36), A3: Business Intelligence (3.38), B1: Work Experience (3.27), T2: Computer Science Degree (3.27), T3: Big Data Technologies (3.11)
Data Engineer (809)	A7: Data Engineering (18.71), T3: Big Data Technologies (7.42), B1: Work Experience (4.86), T7: Cloud Architecture (Microsoft & Google) (4.68), T2: Computer Science Degree (4.01), A3: Business Intelligence (3.55), A5: Data Science (3.14), T1: Cloud Architecture (Amazon) (3.05), T5: Agile Software Engineering & Development (3.04)
Software Architect (1467)	T5: Agile Software Engineering & Development (13.64), T4: IT Operations (6.59), T3: Big Data Technologies (5.72), T7: Cloud Architecture (Microsoft & Google) (4.12), B5: Professional Language Skills (3.66), T2: Computer Science Degree (3.51), B4: Marketing (3.42), B6: Finance & Risk (3.17)
Software Developer (447)	T6: Software Development (38.88), B1: Work Experience (11.98), T2: Computer Science Degree (4.22), T5: Agile Software Engineering & Development (3.2), T1: Cloud Architecture (Amazon) (2.89), B4: Marketing (2.83), A4: Machine Learning (2.31), B9: Leadership & Management (2.13), T3: Big Data Technologies (2.05)

Table 4. Job Roles in DS

5 Discussion

Theoretical Contribution. Looking from a bird's eye view on the job roles (Figure 2), we can distinguish highly technical roles (Data Engineer, Software Architect, and Software Developer) from job roles that are more business and analytics affine (Business User, Data Analyst, Data Scientist). In the latter, a trade-off between analytical and business KDs is present. On the one side of the spectrum, we identified an exclusive business role (Business User), through a job role with still a lot of business understanding but also substantial analytical knowledge (Data Analyst) to a job role with the highest level of analytical expertise (Data Scientist). The Data Scientist is still the most demanded job role in DS, according to the highest number of job ads in this cluster (see the first column in Table 4). Applying statistical modeling techniques (A2) and building ML models (A4) are in focus but without having dedicated business understanding like in "Finance & Risk" (B6). In this direction, at least two analytical flavors of the Data Scientist are suggested by literature. First, the **DS Researcher** that can build advanced mathematical models and create new ML algorithms by following a scientific approach, including hypothesis testing (Saltz and Grady, 2017). Second, the **Applied Data Scientist** that focuses on applying the vast amount of already established algorithms in DS with wide-ranging DS knowledge

(Spruit and Jagesar, 2017). Our analysis did not reveal with which flavor we are dealing with, nor which one is currently more in demand on the market. There is neither a clear focus on the development of new algorithms nor the direct application of DS in business domains through extracted KSAs noticeable. In contrast, the Data Analyst explicitly requires knowledge in business domains (i.e., B7, B3, B6) to make decisions analytically (A1) with the help of traditional BI systems (A3). However, the Data Analyst only makes first use of statistical modeling techniques (A2), lacking the wide-ranging DS knowledge of the "Applied Data Scientist". Against this backdrop, the Data Analyst could be a role that serves as a linkage between Data Scientists and Business Users because this job role has concurrent KSAs in the analytical and business KD. In this regard, literature suggest the **Business Analyst** (IIBA, 2021) that translates business problems into DS projects (Saltz and Grady, 2017). However, we did not find any support for this job role in our data. Hence, the Business Analyst is currently not demanded on the market but presumably present in companies. As other job roles are in high demand, we see an opportunity to develop employees, probably represented as the Business Users in our data, with no previous analytical expertise, into Business Analysts. Their business expertise provides an advantage in consuming complex ML models. Training employees would relieve the high-demanded DS job roles in companies such as the Data Scientist. Additionally, innovative Business Intelligence & Analytics (BI&A) systems such as Self-Service BI&A systems can enable these Analysts (Michalczyk *et al.*, 2020) with minimum training efforts.

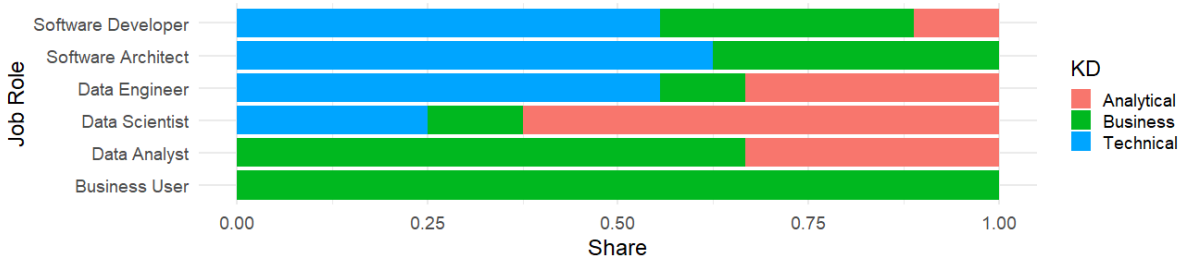

Figure 2. Distribution of KDs

The second group comprises highly technical job roles, which require at least 50% technical KSAs. They differ mainly in the distribution of business and analytical KSAs. The Data Engineer is located at the analytical top end of this second group and focuses on making data available (A7) and on BD technologies (T3). The latter also plays an essential aspect for Software Architect, which led to the assumption that BD pipelines are usually prepared by Data Engineers with the support of Software Architects and are consumed for analytical purposes by Data Scientists, for which we could not identify the "Data Engineering" (A7) KSA. Thereby, Software Architects streamline with their knowledge in "IT operations" (T4), "Cloud Architecture" (T7), and "Big Data Technologies" (T3) the required infrastructure by defining architectures (Gröger, 2018). In literature, we identified a DS specialization of the Software Architect, the **Data Architect**, responsible for synthesizing requirements across DS projects (Childs, 1993) to derive architectural patterns. Thereby, this job role selects appropriate technologies and designs analytical platforms (Fattah, 2015; Gröger, 2018). Also, non-functional attributes of platforms such as security, usability, and stability are evaluated (Fattah, 2015). Finally, the Software Developer is the most technical and most code-savvy role because expertise in object-oriented programming in C, C+++, and Java is required. In this regard, Gurcan and Cagilty (2019) argue that today's software developers can be seen as Data Scientists because knowledge in data processing, data frameworks, and data analytics make up one-third of their study of software developers. However, our results do not support these findings because the Software Architect and Software Developer require only 11% analytical KSAs. However, we could draw a relation of Gurcan and Cagility's (2019) findings to the Data Engineer because it is a third technical role in between, having one-third of analytical knowledge assigned. We see a benefit in the code focus of Software Developers, which helps Data Analysts and Data Scientists bridging the gap to transfer DS artifacts such as classifiers into production (Venturebeat, 2019), making DS initiatives successful.

Practical Contribution. To this end, we believe that our work can help organizations define strategies and a common language for the acquisition and cultivation of the right KSAs for the effective use of DS. Companies can capture the status quo of DS KSAs in their organization to determine current demands at the strategic level. On this basis, existing employees can be trained or new ones hired. Not all KSAs can be developed on the job, but our analysis reveals that especially analytics-affine employees could be trained to Business Analysts or even Data Analysts. In this regard, our work can complement related training portfolios. In order to reduce training efforts, companies could establish mentoring programs. For example, a Business Analyst might be coached by a Senior Data Analyst in order to grow to the level of a Data Analyst. The development of Data Scientists and Data Engineers, on the other hand, is a challenge, since a solid background in statistics or computer science is required, which is difficult to provide through on-the-job training. In this direction, our work can facilitate hiring those job roles because it is based on the most frequent term in DS job ads which can assist the composition of relevant job ads and the effective targeting of promising candidates in hiring campaigns. On the operational level, our work can help in stuffing teams for DS projects through an understanding of job role's KSAs and how the roles complement each other in a team. For individuals, our work (1) is useful to assess their expertise in DS to develop specific KSA in order to accelerate career path systematically, and (2) serves as a reference that novices, scholars, government representatives, and human resources manager can use for further examinations. For example, scholars in universities can align their education portfolio considering the current DS demand.

6 Conclusion

To effectively leverage DS, organizations must develop the right KSAs among their workforces. However, this goes far beyond the recruitment of Data Scientists alone and creates the need to address the job roles' diversity and specificity (De Mauro *et al.*, 2018). Against this background, we have offered clarity about the heterogeneous nature of job roles needed. Hereby, we analyzed and preprocessed 25,104 real job ads published on the online job platforms Indeed, Monster, and Glassdoor. Our text mining approach relied on topic modeling, clustering, and expert assessment. We suggested 24 topics along three KDs (i.e., business, analytical, and technical), identified six job roles in DS that are in demand by organizations (i.e., Data Analyst, Data Engineer, Data Scientist, Business User, Software Developer, and Software Architect), and characterized each job role. Methodologically, our text mining approach could also be reproduced to study new topics such as job roles in other disciplines like Industry 4.0 (Pejic-Bach *et al.*, 2020). We are also aware that this article has limitations. In particular, any bias in the algorithms' parameter settings could distort the results of the topic modeling and clustering approach (Föll, Hauser and Thiesse, 2018). To reduce this possibility, our proposed approach was based on established methodological recommendations. All threshold and parameter settings are defined clearly with a prior quantitative assessment. Furthermore, we run through a qualitative expert assessment to select the parameters and suitable labels for our topics and clusters.

For future work, we plan to train a seeded topic model (Lu *et al.*, 2011) by using the results of a series of interviews we conducted with thirteen professionals of a large automotive supplier as a basis for optimizing our model. We will rely on data triangulation (i.e., text mining, interviews, literature), to understand additional but important roles we can only identify solely through literature (e.g., Data Architect, DS Researcher). Furthermore, we could also investigate how the identified job roles change over time, for instance, by applying an LDA-based topic model over time (Wang and Mccallum, 2006). Such an approach seems important because roles are living artifacts that need to be reassessed regularly (Tyler, 2013). For example, one could investigate whether the Business User is increasingly expected to fulfill a Business Analyst's role.

Acknowledgment. The authors would like to thank Dr. Harald Beier for his valuable comments.

References

Airoldi, E. M. *et al.* (2008) 'Mixed Membership Stochastic Blockmodels', *Journal of Machine Learning Research*, 9, pp. 1981–2014.

Beautiful Soup (2020) *Beautiful Soup 4.9.0 documentation*. Available at: https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (Accessed: 30 October 2020).

Beel, J. *et al.* (2016) 'Research-paper recommender systems: a literature survey', *International Journal on Digital Libraries*. Springer Verlag, 17(4), pp. 305–338. doi: 10.1007/s00799-015-0156-0.

Benoit, K. *et al.* (2020) 'quanteda: Quantitative Analysis of Textual Data'. Comprehensive R Archive Network (CRAN). Available at: https://cran.r-project.org/package=quanteda (Accessed: 21 March 2021).

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) 'Latent Dirichlet allocation', *Journal of Machine Learning Research*, 3, pp. 993–1022. doi: 10.1016/b978-0-12-411519-4.00006-9.

Boyd-Graber, J., Mimno, D. and Newman, D. (2014) 'Care and feeding of topic models: Problems, diagnostics, and improvements', in *Handbook of Mixed Membership Models and Their Applications*. Chapman and Hall CRC, pp. 259–288. doi: 10.1201/b17520.

Cao, L. (2017) 'Data science: A comprehensive overview', *ACM Computing Surveys*, 50(3), p. 42. doi: 10.1145/3076253.

Cegielski, C. G. and Jones-Farmer, L. A. (2016) 'Knowledge, Skills, and Abilities for Entry-Level Business Analytics Positions: A Multi-Method Study', *Decision Sciences Journal of Innovative Education*. Wiley-Blackwell, 14(1), pp. 91–118. doi: 10.1111/dsji.12086.

Chapman, P. *et al.* (2000) 'CRISP-DM 1.0: Step-by-step data mining guide'. Available at: https://www.semanticscholar.org/paper/CRISP-DM-1.0:-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72 TS - www.semanticscholar.org M4 - Citavi.

Childs, D. B. (1993) 'Data architecture and the data architect', in *AIP Conference Proceeding*, pp. 550–564. doi: 10.1063/1.44465.

Cotton, R. (2017) 'Build Regular Expressions in a Human Readable Way'. Comprehensive R Archive Network (CRAN). Available at: https://cran.r-project.org/package=rebus (Accessed: 21 March 2021).

Debortoli, S., Müller, O. and Vom Brocke, J. (2014) 'Comparing business intelligence and big data skills: A text mining study using job advertisements', *Business and Information Systems Engineering*, 6(5), pp. 289–300. doi: 10.1007/s12599-014-0344-2.

Deveaud, R. *et al.* (2014) 'Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval', *Document Numérique, Lavoisier*, pp. 61–84. doi: 10.3166/DN.17.1.61.

Dhar, A. (2013) 'Data Science and Prediction', *Communications of the ACM*, 56(12), pp. 64–73. doi: 10.1145/2500499.

EDISON (2017) *EDISON Data Science Framework (EDSF) | Edison Project*. Available at: https://edison-project.eu/edison/edison-data-science-framework-edsf/ (Accessed: 15 November 2020).

Endres, D. M. and Schindelin, J. E. (2003) 'A new metric for probability distributions', *IEEE Transactions on Information Theory*. Institute of Electrical and Electronics Engineers Inc., 49(7), pp. 1858–1860. doi: 10.1109/TIT.2003.813506.

Ester, M. *et al.* (1996) 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise', in *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*.

Fattah, A. (2015) *The Role of the Analytics Architect -- as analytics moves to a mission-critical role, architecture becomes essential*, *Linkedin*. Available at: https://www.linkedin.com/pulse/role-analytics-architect-moves-mission-critical-becomes-fattah/ (Accessed: 7 March 2021).

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From Data Mining to Knowledge Discovery in Databases', *AI Magazine*, 17(3), p. 37. doi: 10.1609/aimag.v17i3.1230 M4 - Citavi.

Föll, P., Hauser, M. and Thiesse, F. (2018) 'Identifying the skills expected of IS graduates by industry: A text mining approach', in *International Conference on Information Systems 2018, ICIS 2018*, pp.

1–17.
Gardiner, A. *et al.* (2018) 'Skill Requirements in Big Data: A Content Analysis of Job Advertisements', *Journal of Computer Information Systems*. doi: 10.1080/08874417.2017.1289354.
Gomaa, W. H. and Fahmy, A. A. (2013) *A Survey of Text Similarity Approaches*, *International Journal of Computer Applications*.
Grady, N. W. (2016) 'KDD meets Big Data', in *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*. doi: 10.1109/BigData.2016.7840770.
Griffiths, T. L. and Steyvers, M. (2004) 'Finding scientific topics', in *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, pp. 5228–5235. doi: 10.1073/pnas.0307752101.
Gröger, C. (2018) 'Building an Industry 4.0 Analytics Platform', *Datenbank-Spektrum*. Springer Science and Business Media LLC, 18(1), pp. 5–14. doi: 10.1007/s13222-018-0273-1.
Grün, B. *et al.* (2020) 'Topic Models'. Comprehensive R Archive Network (CRAN). Available at: https://cran.r-project.org/web/packages/topicmodels/index.html.
Grün, B. and Hornik, K. (2011) 'Topicmodels: An r package for fitting topic models', *Journal of Statistical Software*, 40(13), pp. 1–30. doi: 10.18637/jss.v040.i13.
Gurcan, F. (2019) 'Extraction of core competencies for big data: Implications for competency-based engineering education', *International Journal of Engineering Education*, 35(4), pp. 1110–1115.
Gurcan, F. and Cagiltay, N. E. (2019) 'Big Data Software Engineering: Analysis of Knowledge Domains and Skill Sets Using LDA-Based Topic Modeling', *IEEE Access*, 7, pp. 82541–82552. doi: 10.1109/ACCESS.2019.2924075.
Hahsler, M., Piekenbrock, M. and Doran, D. (2019) 'Dbscan: Fast density-based clustering with R', *Journal of Statistical Software*. American Statistical Association, 91(1), pp. 1–30. doi: 10.18637/jss.v091.i01.
Han, J., Kamber, M. and Pei, J. (2012) *Data Mining: Concepts and Techniques*, *San Francisco, CA, itd: Morgan Kaufmann*. doi: 10.1016/B978-0-12-381479-1.00001-0.
Handali, J. P. *et al.* (2020) 'Industry Demand For Analytics: A Longitudinal Study', in *Proceedings of the 28th European Conference on Information Systems*. Available at: https://aisel.aisnet.org/ecis2020_rp.
Idoine, C. *et al.* (2018) 'Staffing Data Science Teams: Map Capabilities to Key Roles'. Available at: https://www.gartner.com/en/documents/3888468/staffing-data-science-teams-map-capabilities-to-key-role (Accessed: 25 February 2020).
IIBA (2021) *A Guide to the Business Analysis Body of Knowledge (BABOK-Guide)*, *International Institute of Business Analysis*. Available at: https://www.iiba.org/career-resources/a-business-analysis-professionals-foundation-for-success/babok/ (Accessed: 24 March 2021).
ITIL (2007) *IT Operations Management | IT Process Wiki*, *Service Operations, ITIL V3*. Available at: https://wiki.en.it-processmaps.com/index.php/IT_Operations_Management (Accessed: 25 March 2021).
Jobboard Finder (2021) *10 Best job sites worldwide*. Available at: https://www.jobboardfinder.com/en/best-job-sites (Accessed: 22 March 2021).
Kaggle (2019) *2019 Kaggle ML & DS Survey - The most comprehensive dataset available on the state of ML and data science*. Available at: https://www.kaggle.com/c/kaggle-survey-2019/overview/description (Accessed: 26 October 2020).
Kaufmann, L. and Rousseeuw, P. J. (1990) 'Partitioning Around Medoids (Program PAM)', in *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Ltd, pp. 68–125. doi: 10.1002/9780470316801.ch2.
Kumar, V. (2010) 'Customer Relationship Management', *Wiley International Encyclopedia of Marketing*. Chichester, UK: John Wiley & Sons, Ltd. doi: 10.1002/9781444316568.wiem01015.
Lu, B. *et al.* (2011) 'Multi-aspect Sentiment Analysis with Topic Models | Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops', in *ICDMW '11: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 81–88. Available at: https://dl.acm.org/doi/10.5555/2117693.2119585 (Accessed: 5 November 2020).
Lyon, L. and Mattern, E. (2017) 'Education for Real-World Data Science Roles (Part 2): A Translational

Approach to Curriculum Development', *International Journal of Digital Curation*. Edinburgh University Library, 11(2), pp. 13–26. doi: 10.2218/ijdc.v11i2.417.

Manning, C. D., Raghavan, P. and Schütze, H. (2008) *Introduction to Information Retrieval*. Cambridge University Press. Available at: https://nlp.stanford.edu/IR-book/information-retrieval-book.html (Accessed: 30 October 2020).

De Mauro, A. *et al.* (2018) 'Human resources for Big Data professions: A systematic classification of job roles and required skill sets', *Information Processing and Management*, 54(5), pp. 807–817. doi: 10.1016/j.ipm.2017.05.004.

Mechura, M. (2017) *Machine-readable lists of lemma-token pairs in 23 languages.* Available at: https://github.com/michmech/lemmatization-lists (Accessed: 30 October 2020).

Merriam-Webster (2021) 'Definition of Communication', *by Merriam-Webster*. Available at: https://www.merriam-webster.com/dictionary/communication#synonyms (Accessed: 24 March 2021).

Michalczyk, S. *et al.* (2020) 'A State-of-the-Art Overview and Future Research Avenues of Self-Service Business Intelligence Analytics', *ECIS 2020 Research Papers*, pp. 1–18. Available at: https://aisel.aisnet.org/ecis2020_rp/46 (Accessed: 17 October 2020).

Miller, G. J. (2019) 'The influence of big data competencies, team structures, and data scientists on project success', in *2019 IEEE Technology and Engineering Management Conference, TEMSCON)*. Institute of Electrical and Electronics Engineers Inc., pp. 1–8. doi: 10.1109/TEMSCON.2019.8813604.

Miller, S. (2014) 'Collaborative Approaches Needed to Close the Big Data Skills Gap', *Journal of Organization Design*, pp. 26–30. doi: 10.7146/jod.9823.

Moody, C. E. (2016) 'Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec'. Available at: http://arxiv.org/abs/1605.02019 (Accessed: 24 October 2020).

Muller, M. *et al.* (2019) 'How data science workers work with data', in *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–14. doi: 10.1145/3290605.3300356.

Nigam, K. *et al.* (2000) 'Text classification from labeled and unlabeled documents using EM', *Machine Learning*. Kluwer Academic Publishers, 39(2), pp. 103–134. doi: 10.1023/a:1007692713085.

NIST, B. D. P. W. G. (2015) 'NIST Big Data Interoperability Framework: Volume 2, Big Data Taxonomies', *NIST Special Publication*, 2(1500), p. 31. doi: 10.6028/NIST.SP.1500-2 M4 - Citavi.

Ooms, J. (2021) 'cld3: Google's Compact Language Detector 3'. Comprehensive R Archive Network (CRAN). Available at: https://cran.r-project.org/web/packages/cld3/index.html.

Pejic-Bach, M. *et al.* (2020) 'Text mining of industry 4.0 job advertisements', *International Journal of Information Management*. Elsevier, 50(December 2018), pp. 416–431. doi: 10.1016/j.ijinfomgt.2019.07.014.

Power, D. J. (2007) *A Brief History of Decision Support Systems*, *DSSResources.COM*. Available at: https://dssresources.com/history/dsshistory.html (Accessed: 25 March 2021).

Requests (2020) *Requests: HTTP for Humans™ — Requests 2.24.0 documentation*. Available at: https://requests.readthedocs.io/en/master/ (Accessed: 30 October 2020).

Rousseeuw, P. J. (1987) 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*. North-Holland, 20(C), pp. 53–65. doi: 10.1016/0377-0427(87)90125-7.

Salton, G., Wong, A. and Yang, C. S. (1975) 'A Vector Space Model for Automatic Indexing', *Communications of the ACM*, 18(11), pp. 613–620. doi: 10.1145/361219.361220.

Saltz, J. and Grady, N. W. (2017) 'The ambiguity of data science team roles and the need for a data science workforce framework', in *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*. Institute of Electrical and Electronics Engineers Inc., pp. 2355–2361. doi: 10.1109/BigData.2017.8258190.

Schmid, B. and Baars, H. (2016) 'Die Rollen des Data Scientist und des Data Analyst in der BI'.

Selenium (2020) *Selenium Browser Automation*. Available at: https://www.selenium.dev/ (Accessed: 30 October 2020).

Sirje, V. and Emmanouel, G. (2020) 'Data science and its relationship to library and information science: a content analysis', *Data Technologies and Applications*. Emerald Publishing Limited, 54(5), pp.

643–663. doi: 10.1108/DTA-07-2020-0167.
Snowball (2020) *Snowball*. Available at: https://snowballstem.org/ (Accessed: 30 October 2020).
Springboard (2020) *Springboard: Online Courses to Future Proof Your Career*. Available at: https://www.springboard.com/ (Accessed: 15 November 2020).
Spruit, M. and Jagesar, R. (2017) 'Power to the People ! - Meta-Algorithmic Modelling in Applied Data Science Power to the People', in *Proceedings ofthe 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K2016)*, pp. 400–406. doi: 10.5220/0006081604000406.
Todd, P. A., McKeen, J. D. and Gallupe, R. B. (1995) 'The evolution of IS job skills: A content analysis of IS job advertisements from 1970 to 1990', *MIS Quarterly: Management Information Systems*, 19(1). doi: 10.2307/249709.
Twin, A. (2020) *What Is Marketing?*, *Investopedia*. Available at: https://www.investopedia.com/terms/m/marketing.asp (Accessed: 25 March 2021).
Tyler, K. (2013) *Job Worth Doing: Update Descriptions*, *HR Magazine*. Available at: https://www.shrm.org/hr-today/news/hr-magazine/pages/0113-job-descriptions.aspx (Accessed: 16 November 2020).
Venturebeat (2019) *Why do 87% of data science projects never make it into production? | VentureBeat*, *Venturebeat*. Available at: https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/ (Accessed: 30 October 2020).
W3C (2017) *Xpath Standard W3C*. Available at: https://www.w3.org/TR/xpath/#axes (Accessed: 2 November 2020).
Wang, X. and Mccallum, A. (2006) 'Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends', in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*. New York, New York, USA: ACM Press, pp. 424–433.
Wickham, H. (2014) 'Tidy Data', *Journal of Statistical Software*. American Statistical Association, 59(10), pp. 1–23. doi: 10.18637/jss.v059.i10.
Wickham, H. (2019) 'stringr: Simple, Consistent Wrappers for Common String Operations'. Comprehensive R Archive Network (CRAN). Available at: https://cran.r-project.org/package=stringr (Accessed: 21 March 2021).
Wickham, H. (2020) 'dplyr: A Grammar of Data Manipulation'. Comprehensive R Archive Network (CRAN). Available at: https://cran.r-project.org/package=dplyr (Accessed: 21 March 2021).
Wickham, H. *et al.* (2020) 'ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics'. Comprehensive R Archive Network (CRAN). Available at: https://cran.r-project.org/package=ggplot2 (Accessed: 21 March 2021).
Wickham, H. (2021) 'rvest: Easily Harvest (Scrape) Web Pages'. Comprehensive R Archive Network (CRAN). Available at: https://cran.r-project.org/package=rvest (Accessed: 21 March 2021).
Wowczko, I. (2015) 'Skills and Vacancy Analysis with Data Mining Techniques', *Informatics*. MDPI AG, 2(4), pp. 31–49. doi: 10.3390/informatics2040031.