A Data Science Approach to Defining a Data Scientist

Andy Ho[1], An Nguyen[1], Jodi Pafford[1], and Dr. Robert Slater[1]

Master of Science in Data Science, Southern Methodist University, Dallas TX 75275 USA {atho,angyuen2, jpafford, rslater}@smu.edu

Abstract. In this paper, we present a novel approach for the problem of not having a common definition and list of skillsets for a Data Scientist. As it relates to the field of data science, adoption has spanned nearly all industries and disciplines. This has increased the necessity for a wide range of technical skills, soft skills, experiences, and education to work on applications within data science (i.e. Artificial Intelligence, Machine Learning, Big Data Analytics, Data Visualization, etc.). The result is an overlap and ambiguity of various roles such as data scientist, data engineer, data analyst, software engineer, database administrator, and statistician. To solve the problem, we collected over 8,000 job postings from Indeed.com that for the six job titles to form our data set. Natural Language Processing (NLP) techniques, advanced Machine Learning (ML), and Universal Sentence Encoder (USE) was performed to cluster and classify the corpuses that contains job qualifications, skills, responsibilities, educational preferences, and requirements to arrive at our key findings on the definition and list of skillsets of a Data Scientist. A secondary finding verified various industries (and even within the same industry) experience this same problem. Our conclusion is a Data Scientist is an individual who discovers valuable insights from large amounts of data possessing programming skills, well versed in statistics and mathematics, and can adequately visualize and communicate the findings.

1 Introduction

Over 2.5 quintillion bytes of data is created every single day and the pace is expected to accelerate with the growth of the Internet of Things (IoT), a society more dependent on data, and more businesses making data-driven decisions. On any given day, 500 million tweets are sent, 294 billion emails delivered, 4 petabytes of data created on Facebook, and IoT products such as driverless cars, wearables, and smart cities will push the daily amount collected to 463 exabytes. This expansion of data has created the need for individuals (i.e. Data Scientists) who are trained, skilled, and educated in the field (and application) of Data Science.

There is no common definition and a list of skillsets for a Data Scientist. This becomes evident when the question of What is a Data Scientist or What does a Data Scientist do is presented or when a job search is ran on multiple job

sites (i.e. Indeed, Monster, ZipRecruiter). According to Dictionary.com, a data scientist is a person employed to analyze and interpret complex digital data, such as the usage of statistics of a website, especially in order to assist a business in its decision-making. But we postulate that there is more to a data scientist, and a common definition along with a list of skills can be reached using data science techniques.

To apply data science to this problem, our solution was to gather data, clean the data, model and analyze to find patterns and/or classifications within the data. NLP and ML was utilized to arrive at our solution. The end result is visualizations and explanations of our findings.

Since there is no public data sets available for this problem, we web scraped job postings listed under Data Scientist and other similar roles from multiple job sites to form our training data set. Features such as job title, job description/summary, experience and skill requirements, and location was collected. NLP was performed to read, decipher, understand, and make sense of each job posting. Unsupervised ML within NLP was performed to find document and text similarities between each job posting. Multiple comparative analytics provided additional results and insights to arrive at a common definition and list of skillsets of a data scientist.

The main conclusions are the definition and list of skills of data scientist varies from one job posting to another. We observed that multiple job sites have different algorithms presenting roles that seem like a data scientist but in fact an entirely different role. In applying data science techniques to this problem, the following are a few key findings:

1. Data Scientists at minimum have programming expertise in Python or R
2. Data Scientists have a background or knowledge of mathematics or statistics
3. Data Scientists have different domain expertise
4. Data Scientists, in general, possess degree(s) in computer science, mathematics, or some related science field

In addition to the Abstract and Introduction section, this paper is organized in the following manner: Tutorial Sections to educate readers on the general principles and techniques used in NLP, ML, USE for this research; a Data Set section explaining how data was collected, features selected and visualization from the data exploration analysis; a Methods section on the data science techniques and algorithms performed; a Related Work section discussion other studies already completed; a Results section listing and explaining our experiment; an Analysis section on the results; an Ethics section on the possible ethical ramification of our findings; and a Conclusions section listings multiple conclusions from our analysis.

2 Artificial Intelligence Explained

To understand the data science techniques applied to our research, we want to first provide a baseline understanding for Artificial Intelligence (AI). This

term has been associated to robots functioning on their own or a global network of computers rising up against mankind. Though that may be the case in the distant future, present AI is the development of advanced software programs to perform tasks that would normally require human intelligence. Alan Turings paper established the fundamental goal and vision for AI [7]. Much has developed since Turings endeavor to simulate human intelligence in machines, Table 1 shows a few examples of applied AI today:

Table 1.

1. Translation between languages (i.e. Google Translate)
2. Facial Recognition (i.e. Auto Tag on Facebook)
3. Virtual Assistant (i.e. Amazon Alexa)

The data science techniques applied will be in the realm of Narrow AI where machine learning and natural language processing is used to perform tasks in data collection, analysis, and modeling.

2.1 The Importance of Machine Learning to AI

In short, ML is when computers learn from the data it is provided. Without ML, AI could not provide insights, decisions, and/or actions for the examples of applied AI. There are many ML algorithms, all of which can be stratified into 3 classifications, described in Table 2.

Table 2. Machine Learning Classifications.

Classification	Description
Supervised Learning (SL)	Labeled input/output data is fed into an algorithm multiple times to arrive at a pattern for prediction. Algorithm examples could be Linear/Logistic Regression, K-Nearest Neighbor, Nave Bayes, Decision Tree.
Unsupervised Learning (UL)	Labeled input is fed into an algorithm multiple times to form clusters for the unlabeled output data. A new input is then added to predict which cluster it is associated with. Algorithm examples could be K-means clustering, Principal Component Analysis.
Reinforcement Learning (RL)	A reward base learning where feedback is provided on the output to improve the prediction accuracy.

In this paper, we will apply both SL and UL algorithms to find clusters and patterns for the data collected from natural language processing.

2.2 The Application of Natural Language Processing

NLP is a subfield of AI that focuses on how to program computers to process and analyze large amounts of human/natural language data. Both speech and written language is included in NLP. For this paper, only written text will be used to solve the problem. Using NLP, we were able to web scrape job postings from multiple sources to create our data set to perform lexical, syntactic, and semantic analysis. Throughout this paper, Python packages specializing in NLP will be leveraged to get text and document similarity scores of the job postings for each job title searched.

2.3 Feature Vector Creation using Universal Sentence Encoder

Googles TensorFlow is an open source library for advanced Machine Learning and for numerical computation. In addition, the library has an arsenal of algorithms for deep learning for digit classification, image recognition, word/sentence embeddings, recurrent neural networks, and for this paper natural language processing. The Universal Sentence Encoder (USE), which uses TensorFlow library, encodes text into feature vectors for the purpose of text classification, semantic similarity, and other natural language tasks. At the core, USE produces sentence embeddings for transfer learning [10] and is made publicly available on TF Hub. On a high level, a corpus of text is fed into the encoder and a 512-dimensional vector is outputted. The vectors are then used for clustering and classification for the purpose of this paper. An example is below in Fig. 1.

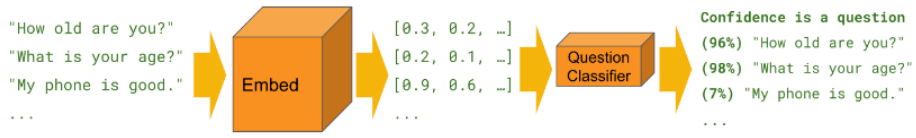

Fig. 1. Classification using a simple binary text classifier

3 Data Set Creation

Data is collected from Indeed.com. Indeed is the number one job site in the world with over 250 million unique visitors every month [6]. Indeed gives users free access to complete job-seeking tasks such as searching for jobs, posting resumes, and researching companies. Globally, Indeed.com has 9.8 jobs adds posted every second. Data is not readily available for download from Indeed.com, so it must be scraped. Web scraping, also referred to as web harvesting or web data extraction, is a process commonly used by Data Scientist to get data from the World Wide Web directly from the respective website (NEED REFERENCE). Web

Scraping was done using the python library, Beautiful Soup (beautifulsoup4). Beautiful Soup allowed us to scrape Indeed.com for information on several job titles and extract the information Beautiful Soup sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree'. (reference: https://pypi.org/project/beautifulsoup4/). Table 3 illustrates the data that was gather during the scraping process. Table 3 illustrates the data that was gather during the scraping process.

Table 3. Data Set Description.

Data Set	Variables
Job Search Page	Individual Job Posting Page Link
Job Search Page	Job Title
Job Search Page	Location
Separated from Location	City
Separated from Location	State
Separated from Location	Zip
Separated from Location	Country
Job Posting Page	Qualifications
Job Posting Page	Skills
Job Posting Page	Responsibilities
Job Posting Page	Education
Job Posting Page	Requirements
Job Posting Page	Full Description
Job Search Page	Salary (if available)

3.1 Data Set Contents

Data was scraped from Ineeded.com for the following 6 job titles: Data Scientist, Data Analyst, Data Engineer, Database Administrator, Software Engineer, and Statistician and from the following 16 cities: Atlanta, GA, Austin, TX, Bellevue, WA, Boston, MA, Chicago, IL, Cupertino, CA, Dallas, TX, Denver, CO, Houston, TX, Los Angeles, CA, Mountain View, CA, New York, NY, Pittsburgh, PA, Seattle, WA, San Francisco, CA, and Washington, DC. The goal was to initially scrape a maximum of 300 job postings per city and job title. There was a potential of having over 28,000 job posts. However, many cities did not have 300 that pulled and some of our cities were so close to each other that duplicates were pulled. Those duplicates were eliminated. A distribution of all the job postings pulled can be seen in Fig. 2, 3),and 4) below.

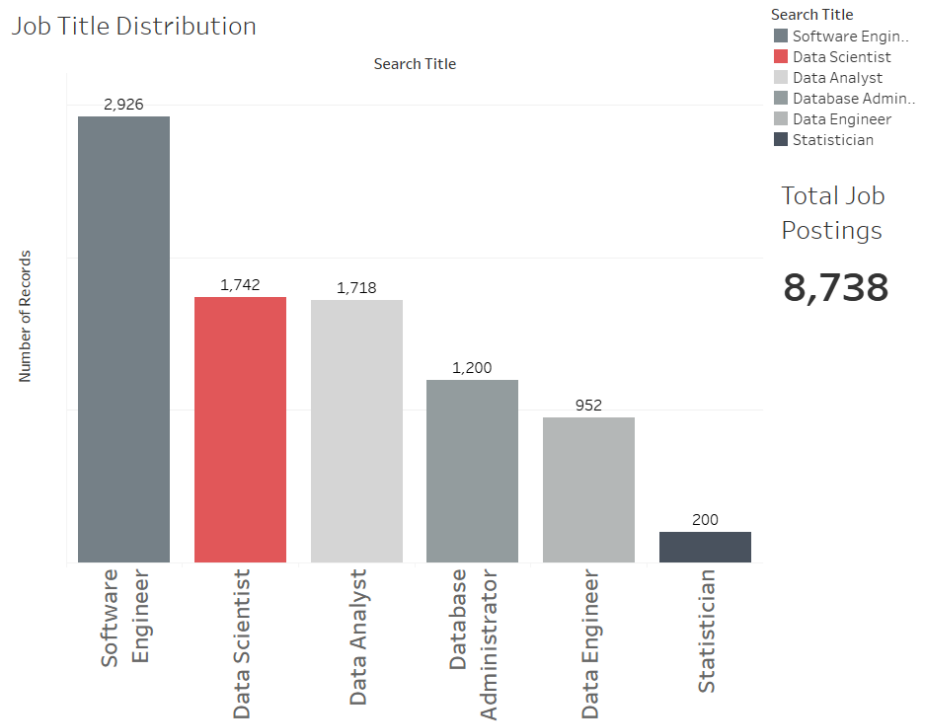

Fig. 2. Count of Job Postings

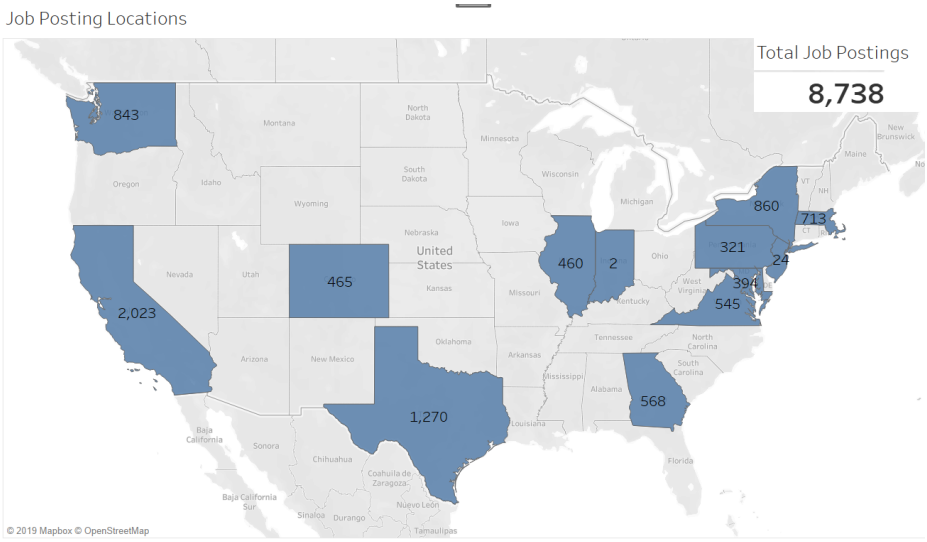

Fig. 3. Distribution of job postings across the United States

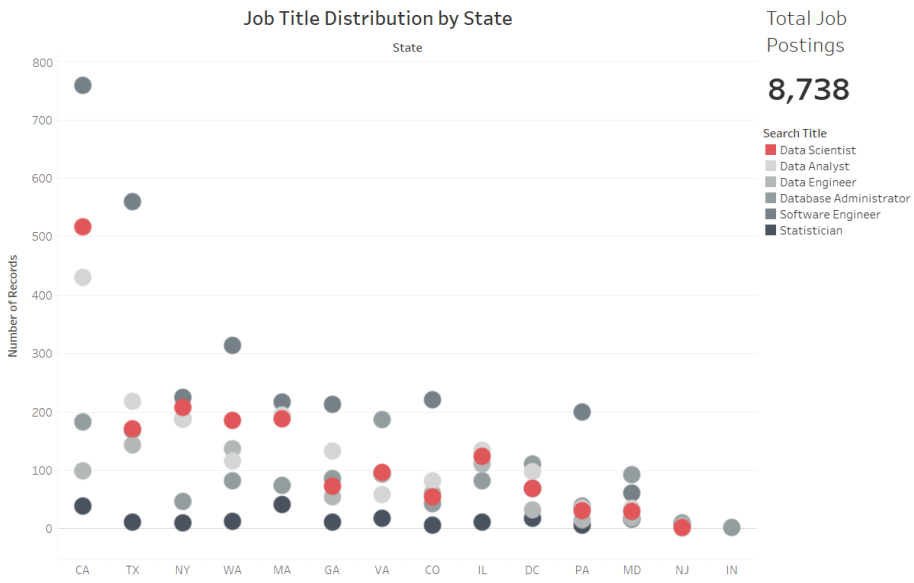

Fig. 4. Count of Job Titles by State

3.2 Exploratory Data Analysis

Indeed.com allows companies to post jobs with their own HTML code beyond the generic required information. This means that some companies included salary information, company logo, and/or company rating while others did not. This also means that most of the Job Description Summary sections are different base on how the company posted the job (bold with bullets, one single paragraph, and a myriad of other variation in between). Additionally, if we ran the entire job description summary through our analysis, we would end up capturing a lot of non-discrimination and company information. After examination of the data, we found that enough of the job postings had bullet points posted that this was a good way to focus our data set on applicable information. After we removed job postings who did not utilize bullet points, we ended with 4,156 job postings. The final count of jobs used can be found in Fig. 5). It is interesting to note that all of our final counts ended up with a total of two different digits within each total (i.e.-1,311 uses the digits 1 and 3). This does not have any affect on the outcome, but we found it interesting. Fig. 6 below shows the breakdown of those jobs throughout the United States. Table (see Fig. 7) shows the breakdown by job title searched.

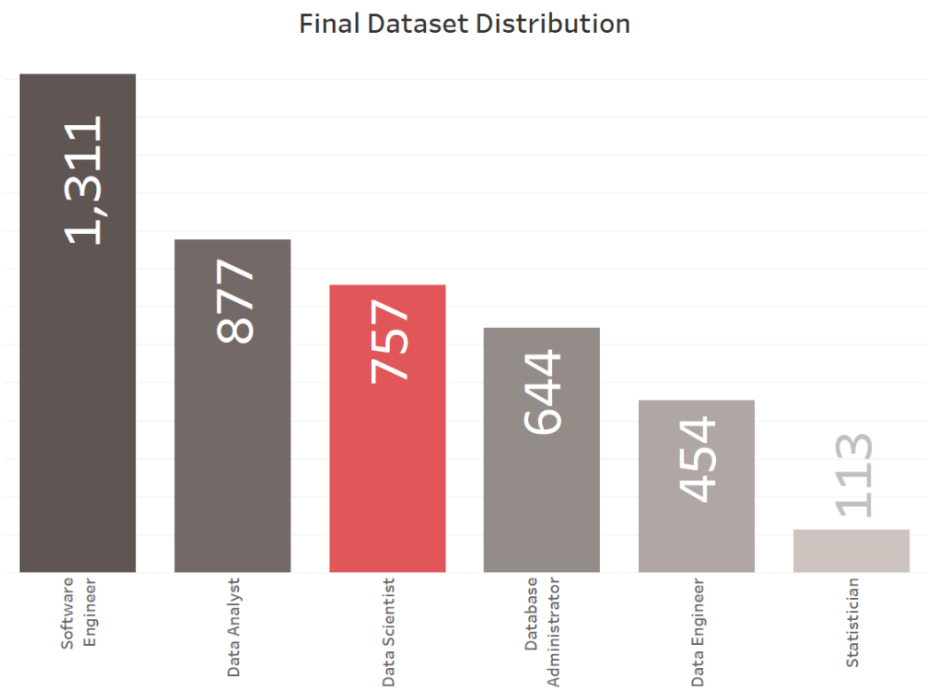

Fig. 5. Count of Jobs in the Final Dataset

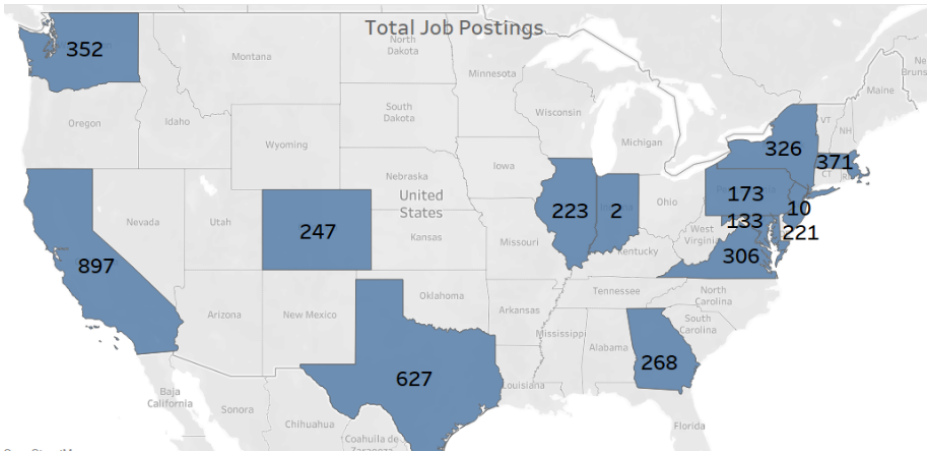

Fig. 6. Final Data Set Counts by State

Job Posting Used in Analysis

Search Title	State														
	CA	CO	DC	GA	IL	IN	MA	MD	NJ	NY	PA	TX	VA	WA	Grand Total
Data Scientist	205	28	43	42	51		83	8	1	74	14	92	54	62	**757**
Data Analyst	207	47	52	67	73		99	21		78	18	122	31	62	**877**
Data Engineer	54	31	22	31	50			7	3	78	7	79	44	48	**454**
Database Administrator	90	18	68	32	42	2	54	60	2	20	20	72	126	38	**644**
Software Engineer	326	121	25	93			108	26		70	110	255	43	134	**1,311**
Statistician	15	2	11	3	7		27	11	4	6	4	7	8	8	**113**
Grand Total	**897**	**247**	**221**	**268**	**223**	**2**	**371**	**133**	**10**	**326**	**173**	**627**	**306**	**352**	**4,156**

Fig. 7. Distribution of Job Titles Throughout the States

4 Methodology

4.1 Document Sentiment Comparitive Analysis

Document sentiment is being achieved by NLP.

4.2 Cluster Analysis on Key Terms

K Nearest Neighbor (K-NN) was used as the baseline model for the comparison of the neural network. Seven models were generated from the scraped data. First each job title was clustered individually then all six titles was concatenated into one data set and clustered. The number of clusters for each analysis was determined by locating the elbow from plots of the sum of the squared distances of each sample to their closest cluster center (Fig. 8). Figure 9)Figure 8 is the combination of all six job titles.

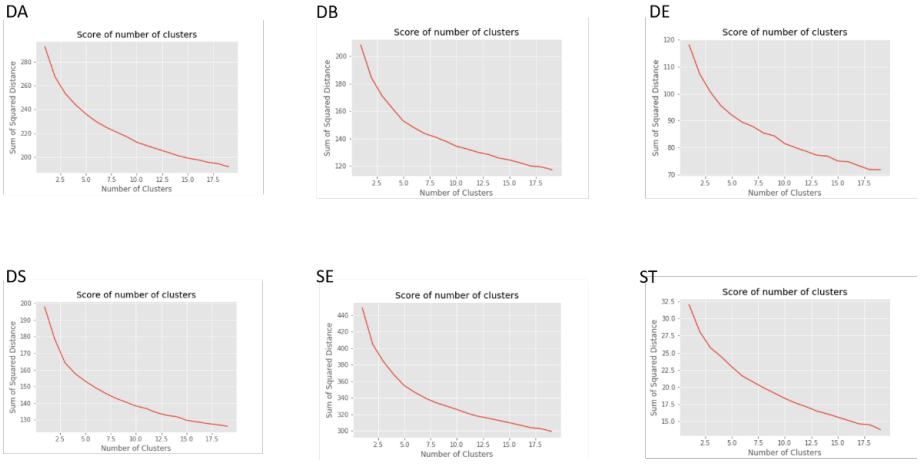

Fig. 8. Sum of Squared Distance for each job titles with different number of clusters.

Fig. 9. Sum of Squared Distance for all the job titles combined.

The optimum number of clusters suggested by this method was 3 clusters for the individual job titles and 4 clusters when all job titles are combined into one. In order to visualize the job descriptions, the data are transformed with Principal Component Analysis, figure 10 is the visualization of each K-NN analysis with three clusters. With the knowledge that each analysis consists of only one cluster figure 11 visualizes the data as one cluster.

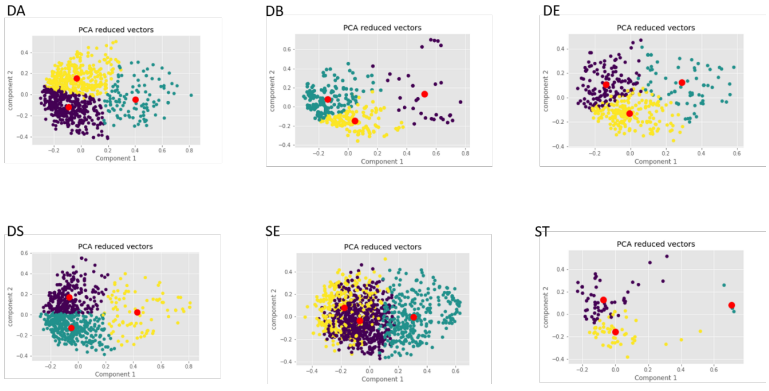

Fig. 10. PCA visualization of each job description as three clusters

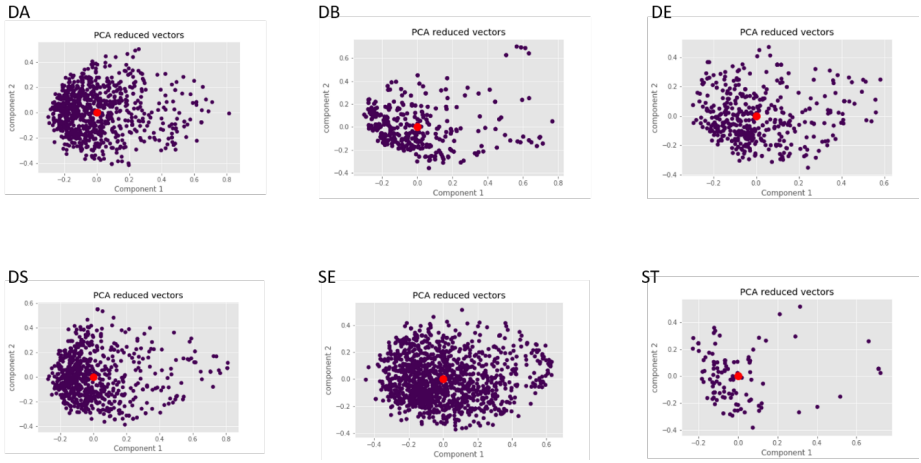

Fig. 11. PCA visualization of each job description as one cluster.

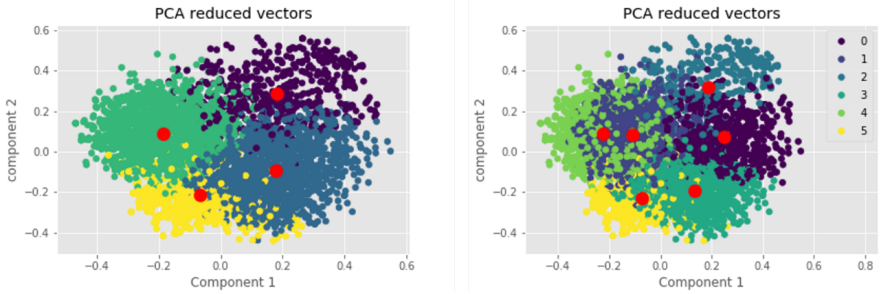

Fig. 12. PCA visualization of the combination of all six job titles with 4 and 6 clusters

Because this was an unsupervised clustering each label were auto generated (05). These labels are then mapped to each individual job description by taking the ratio between the count of each job per auto generated group and the number of job description scraped and taking the highest number per group. The Data Scientist job is mapped to group 1, dark blue. Data Analyst job is mapped to group 2, dark green. Statistician is mapped to group 4, light green. Database Administrator is mapped to group 5, yellow. Software Engineer is mapped to both group 0 and 3, purple and green. Data Engineer was not detected by the K-NN algorithm. Data Scientist and Statistician has a large overlap. Data Analyst group is separated from the Data Scientist/Statistician group. Database Administrator is also well separated from Data Scientist/Statistician group. Software Engineer is well separated from the Data Scientist/Statistician group and the

Data Analyst group but overlap with the Database Administrator group. The algorithm could not detect the Data Engineer group. A possible explanation is that Data Engineer is too similar to the Software Engineer group, observe figure 10. The accuracy of the model is 53.51%.

5 Related Work

In a paper from the University of Rome, Mauro et al. presented a classification of job roles and skills in the area of Big Data Analytics. The researchers used web scraping to retrieve job postings from many prominent websites. Natural language processing was then applied to this dataset to discover four essential job groups, most frequent bigrams appearing in the job titles: Business Analysts, Data Scientists, Developers and System Mangers. Then using the Latent Dirichlet Allocation, LDA, classification techniques the authors clusters skills into 9 topics that were generated by human interpretation of the skills. The 9 topics are: Cloud, Coding, Database management, Architecture, Project Management, Systems Management, Distributed Computing, Analytics, Business impact. Finally, the job skill sets are mapped to job roles by a measure of the extent at which each skill set is represented within each job post description[1]. (see Fig. 13)

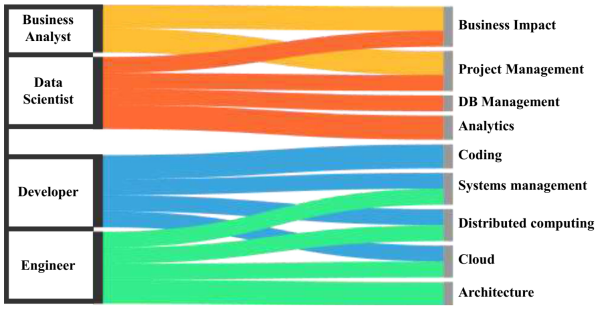

Fig. 13. Job skill sets are mapped to job roles by a measure of the extent at which each skill set is represented within each job post description.

A second group out of California State University focused entirely on the difference between Business data analytics, DBA and data science, DS. Radovilski et al. manually collected job descriptions of DBA and DS jobs from job boards. Using the job description they identify skill sets associated with Business, Analytical, Technical and Communication knowledge domains. Using text mining approaches, Document Data Matrix, Term Cloud, Singular Vector Decomposition, VARIMAX rotation and Latent Class Analysis the authors found the most frequent BDA and DS terms used, (see Fig. 14).[8].

Some published papers to add to the paper at a later time. [4, 5, 2, 6, 7]

#	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%

#	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%

Fig. 14. Job skill sets are mapped to job roles by a measure of the extent at which each skill set is represented within each job post description.

6 Results

6.1 What is a Data Scientist?

7 Analysis

7.1 Discoveries

8 Ethical Considerations

Ethics plays a role in the entire job search and interviewing process. There are many laws and regulations that oversee the process once the interviewing begins, however, there are not many laws and regulations when it comes to the job search process.

8.1 Website Usage

Scraping data from the website must be done with extreme caution. Each website is required to publish a robots.txt file that describes sections of the website that is not allowed to be scraped. Additionally, a websites terms and conditions may prevent someone from scraping. Falling outside the guidelines and/or company policies can be bad. There are criminal implications such as identity theft and

hacking if information is scraped from a website without following the proper protocol. For the novice programmer, it can be easy to make this mistake. According to the 2016 lawsuit, LinkedIn V. Doe Defendants, LinkedIn sued 100 people who scraped their website anonymously. The lawsuit was stopped in the U.S. District Court where Judge Edward Chen ruled that LinkedIn couldn't block companies from deploying bots to scrape data from a public website. Though this was not held up in court, it speaks to the breadth of the dangers and risks of web scraping.

8.2 Job Search Ramifications

Many ethical issues related to job searches revolve around the truthful representations of jobs. Employers may try to entice more applicants by displaying the role as more desirable than it is. According to the Society for Human Resource Management (SHRM), creating fake job descriptions is a common way to get more applicants in a pool even though the role does not exist and is not advertised as a pool. SHRM has a code of ethics for the overall human resource profession which addresses recruiting [3]. Inversely, applicants could utilize algorithm or apply data science to falsify or embellish their resumes. The goal would be to trick resume tracking software into classifying the applicant has qualified, competent, and/or a fit for the open position. Not only is this misrepresentation but it prevents other qualified applicants from being interviewed.

8.3 Model Used for Profiling Candidates - aka AI bias

One shared concerned regarding Artificial Intelligence is its ability to be fair and neutral. Although there are many advantages of AI (i.e. speed and capacity to process), the software/programs are still written by humans whom are bias and judgmental. The application of NLP and ML to identify features or patterns in job postings could also be used by employers to profile applicants. The risk is knowingly removing applicants based on gender, race, creed, religion, and/or sexual orientation.

9 Conclusion and Future Work

9.1

References

1. Andrea DeMauro, Marco Greco, M.G.P.R.: Human resources for big data professions: A systematic classification of job roles and required skill sets. Information Processing & Management **54**(5) (9 2018)

2. Ankita Srivastava, Yogesh Tiwari, H.K.: Attrition and retention of employees in bpo sector. International Journal of Computer Technology and Applications **2**(6), 3056–3065 (2011)
3. Bates, S.: Do recruiters need a code of ethics (2019), https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/do-recruiters-need-code-of-ethics.aspx
4. Brijesh Kishore Goswami, S.J.: Attrition issues and retention challenges of employees. International Journal of Scientific & Engineering Research **3**(3) (4 2012)
5. Collins Marfo Agyeman, P.V.P.P.: Employee demographic characteristics and their effects on turnover and retention in msmes. International Journal of Recent Advances in Organizational Behaviour and Decision Sciences **1**(1) (2014)
6. Kaplan, J.: Artificial Intelligence: What Everyone Needs to Know. What Everyone Needs To Know®, Oxford University Press (2016), https://books.google.com/books?id=wPvmDAAAQBAJ
7. Touring: Alan Turing: His Work and Impact. Computing Machinery and Intelligence®, Gale Virtual Reference Library (GVRL) (2013)
8. Zinovy Radovilsky, Vishwanath Hegde, A.A.U.U.: Skills requirements of business data analytics and data science jobs: A comparative analysis. Journal of Supply Chain and Operations Management **16**(1) (3 2018)