5-2019

Topic Clustering of Data Science and Analytics Job Descriptions

Kayako Yamakoshi
Trinity University, kayamakos@gmail.com

Topic Clustering of Data Science and Analytics Job Descriptions

A THESIS PRESENTED

BY

KAYAKO YAMAKOSHI

TO

THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF REQUIREMENTS

FOR THE DEGREE OF

BACHELOR OF SCIENCE

IN THE SUBJECT OF

COMPUTER SCIENCE

TRINITY UNIVERSITY

SAN ANTONIO, TX

APRIL 2019

TOPIC CLUSTERING OF DATA SCIENCE
AND ANALYTICS JOB DESCRIPTIONS
KAYAKO YAMAKOSHI

A DEPARTMENT HONORS THESIS SUBMITTED TO THE
DEPARTMENT OF COMPUTER SCIENCE AT TRINITY UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
GRADUATION WITH DEPARTMENTAL HONORS

DATE APRIL 21, 2019

MATTHEW A. HIBBS	YU ZHANG
THESIS ADVISOR	DEPARTMENT CHAIR

Michael Soto, AVPAA

Student Agreement

I grant Trinity University (“Institution”), my academic department (“Department”), and the Texas Digital Library ("TDL") the non-exclusive rights to copy, display, perform, distribute and publish the content I submit to this repository (hereafter called "Work") and to make the Work available in any format in perpetuity as part of a TDL, Digital Preservation Network (“DPN”), Institution or Department repository communication or distribution effort.

I understand that once the Work is submitted, a bibliographic citation to the Work can remain visible in perpetuity, even if the Work is updated or removed.

I understand that the Work's copyright owner(s) will continue to own copyright outside these non-exclusive granted rights.

I warrant that:

1) I am the copyright owner of the Work, or
2) I am one of the copyright owners and have permission from the other owners to submit the Work, or
3) My Institution or Department is the copyright owner and I have permission to submit the Work, or
4) Another party is the copyright owner and I have permission to submit the Work.

Based on this, I further warrant to my knowledge:

1) The Work does not infringe any copyright, patent, or trade secrets of any third party,
2) The Work does not contain any libelous matter, nor invade the privacy of any person or third party, and
3) That no right in the Work has been sold, mortgaged, or otherwise disposed of, and is free from all claims.

I agree to hold TDL, DPN, Institution, Department, and their agents harmless for any liability arising from any breach of the above warranties or any claim of intellectual property infringement arising from the exercise of these non-exclusive granted rights.”

I choose the following option for sharing my thesis (required):

[X] Open Access (full-text discoverable via search engines)
[ ] Restricted to campus viewing only (allow access only on the Trinity University campus via digitalcommons.trinity.edu)

I choose to append the following Creative Commons license (optional):

N/A

Topic Clustering of Data Science and Analytics Job Descriptions

ABSTRACT

Since the rise of data science and analytics, the definitions of data scientists and analysts have been obscure. Various perceptions of those positions originate from companies, which have been attempting to gain competitive advantages over their competitors by using data in their business. However, the data science and analytics job markets are suffering from a high turnover rate as well as long times to fill job openings. This research conducts a hierarchical clustering and topic modeling in order to demonstrate how data scientist positions and data analyst positions can be distinguished from each other by finding hidden correlations among words used in job descriptions.

Contents

Acknowledgments

First, I would like to express my sincere gratitude to my thesis advisors, Dr. Hibbs and Dr. Young for their continuous support. Dr. Hibbs guided me through the research process and kindly allowed me to utilize his clustering program, which played a significant role in this research. Dr. Young provided tons of advice and insight from business perspectives when I was deciding the research topic. I cannot imagine having better advisors for my undergraduate thesis.

I am also profoundly grateful to Dr. Myers for his mentorship since the first day I got to Trinity University. Not only he has been my academic adviser, he has also been supporting me as a faculty international student liaison. I sincerely appreciate his service on the thesis committee.

I also thank my peers in the Department of Computer Science. It was a great pleasure working with them day and night.

Last but not least, I would like to thank my family from my heart. They have been always supportive and encouraging me to

keep working toward my goal. I am also thankful for the inspirations I have received from my brother, Hiro Yamakoshi, who has been my role model and brought me into the field of Computer Science.

My research would have been impossible without aid from everyone above. I would like to thank everyone again for the great guidance and support.

Chapter 1: Introduction

The data science and analytics fields have been growing considerably fast since they emerged quite recently. Being a relatively new field in Computer Science, the definition of "data science" and "data analytics" have been evolving in their history. This chapter explores how "data science" and "data analytics" have been established to become what they are today, and how people perceive those fields these days. Then it will show why those current perceptions of the fields could be problematic.

1. The History of Data Science and Analytics

The study of data was recognized as one independent science rather than a subcategory in statistics or mathematics in around 1962. John Tukey stated in "The Future of Data Analysis" that "I thought I was a statistician, interested in

inferences from the particular to the general. But . . . I have come to feel that my central interest is in *data analysis . . .* Data analysis . . . must take on the characteristics of science rather than those of mathematics… data analysis is intrinsically an empirical science (Tukey, 1962).” In this paper, he defined data analytics as procedures for analyzing data, techniques for interpreting the results of such procedures, and ways of planning the gathering of data make its analysis easier. In 1974, the term “data science” was formally established by Peter Naur in “Concise Survey of Computer Methods”. Data science was described as “the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences (Naur, 1974).” The connection between data science and analytics as well as business started to gather attention in 1994. The article regarding database marketing made the front page of Bloomberg Businessweek, explaining how companies are using consumer data to

predict their preferences and behaviors (Berry, 1994). The competitive advantages between companies by the use of data science and analytics became profound around 2005. The concept of data science and analytics pervaded the field and a few journals specializing in Data Science were initiated around this time (Press, 2013). The research focused on the new type of competition based on the extensive application of data in business was conducted in 2005, and it proved that the organizational culture at many companies is changing to be analytically oriented. It also indicated that companies would require changes in process, behavior, and skills for multiple employees (Davenport, Cohen, & Jacobson, 2006). As many companies aimed to excel at data utilization in business, data science and analytics jobs were becoming to be recognized as high growth job markets. The confusions caused by the various usages of the terms “data science” and “data analytics” became prevalent at this time while but many companies were desperately attempting to figure out how to

incorporate data to stay in business. A tremendous amount of articles have been written on what exactly data science is since 2009 (Press, 2013). Even today, the definition of "data science" and "data analytics" and the difference between the two seem to confuse people. Codeup had to post on their blog explaining the difference between the two since they have received too many inquiries from their prospective students, who were trying to pick which course better suits them (Codeup, 2018).

2. Significance of Accurate Definition of Data Science and Analytics

As discussed in the previous section, the popularity of data science and analytics positions have been increasing these past ten years. However, professionals in the field express many frustrations. *Financial Times* reported that most people in the field say that they spend 1-2 hours every week looking

for a new job (Waters, 2018). The article also states that data science has the second highest percentage of professionals in the field who are actively looking for a new job compared to other Computer Science related professions. The high turnover rate is not only troubling for employees, but also could be fatal for companies. According to the well-known recruiter Jörgen Sundberg, the cost of hiring a new employee is $240,000 (Fatemi, 2016). The U.S. Department of Labor also says that the price of a bad hire is at least 30 percent of the employee's first-year earnings, which corresponds with Sundberg's number if a position pays six-figures. Especially for companies hiring data scientists, the opportunity cost loss is immense as the position is generally one of the highest paying. According to Glassdoor, the average base pay for a data scientist is $117,345 per year, and $67,377 for a data analyst (Glassdoor, 2019).

While there are several possible reasons for the high turnover rate for data scientists and analysts, the wrong

expectation is one of the major causes of the issue. Edward Chenard discusses three top reasons why data scientists are leaving their jobs, and raised the wrong perceptions of responsibilities and tasks as one factor (Chenard, 2018). An attempt to get rid of inaccurate perceptions and definitions of data scientists and analysts is essential to promote the growth of the job market that is very influential in business.

Chapter 2: Background

The previous chapter covered how the study of data has evolved and how much impact it has in society today. Data science and analytics also brought urgent issues in their job markets. This chapter discusses the related researches in this subject and introduces an overview of the intended project.

1. Related Researches

IBM predicts in its "Quant Crunch Report" that the number of data science and analytics job listings will grow from nearly 364,000 job listings to about 2,720,000 openings by 2020. This is a 15% growth (IBM & Burning Glass Technology, 2016). Assuming that there are 2.8 million professionals with deep data analytics or science skill sets following the estimation provided McKinsey, this growth indicates that most of the

current employees need to change their jobs every year to fill these new job openings. However, the job market is already struggling to fill the job openings in the fields today and is not keeping up with the increasing amount of new job opportunities that become available every day. The average time to fill a job opening is 45 days for all the data science and analytics related jobs and the data systems developer takes the longest, which is 50 days. Data scientist position is second highest on the list, which scores 46 days, and data analyst positions take the shortest, which is 38 days. IBM infers this is because of “a lack of a common framework and vernacular for DSA jobs and skills.” They also point out that the Bureau of Labor Statistics has neither a clear definition of data science and analytics jobs nor key metrics of these jobs. This leads to inconsistencies in job titles across many types of data science and analytics jobs. The IBM paper provides a set of unique skills required for each type of data science and

analytics job as a criterion to identify one from another.

DSA Framework Category	Occupation	Analytical Score (2015)
Analytics Managers	Financial Analysis SQL SAS Data Analysis Business Intelligence	Budgeting Project Management Risk Management Accounting Financial Planning
Data Analysts	Data Analysis SQL Business Intelligence Data Warehousing SAS	Project Management Microsoft Access Business Process SAP Business Analysis
Data Systems Developers	SQL Database Administration Extraction, Transformation, and Loading Data Warehousing Apache Hadoop	Project Management LINUX Software Development UNIX JAVA
Data Scientists & Advanced Analysts	Apache Hadoop Machine Learning Big Data R Data Science	Python JAVA Economics C++ Project Management
Data-Driven Decision Makers	SQL Financial Analysis Data Analysis Data Management Data Validation	Budgeting Project Management Accounting Supervision Product Management
Functional Analysts	Financial Analysis SQL Data Analysis Data Management SAS	Budgeting Accounting Business Analysis Business Process Economics

Figure 1. 1. 1: Top Analytical and specialized skills for each type of data science and analytics jobs. Adapted from (IBM & Burning Glass Technology, 2016).

However, these skill group criteria are not completely reliable as people's perception of required skills for each type of data science and analytics jobs vary under the inconsistent

definitions of data science and analytics. The University of California Berkeley, which is recognized as one of the strongest science school in the United States, explains the combinations of skills required for data scientist and data analyst. Data scientists need “Programming skills (SAS, R, Python), statistical and mathematical skills, storytelling and data visualization, Hadoop, SQL, machine learning”, while data analysts need “Programming skills (SAS, R, Python), statistical and mathematical skills, data wrangling, data visualization (University Of California Berkeley, 2018).” While the IBM paper categorizes SQL only under data analyst, UC Berkeley categorizes SQL only under data scientist. While this comparison is but one example, the discrepancy between the two indicates that the unique combination of skills needed for each kind of data science and analytics jobs is subject to change.

2. Current Understanding and Hypothesis

So far there are no clear criteria that enable us to identify the category in which each data science and analytics job belongs. The attributes of each job, such as a job title and skills required, are influenced heavily by the writer of each job description and reflect the writer's own definition of data science and analytics. As such, this present research established a hypothesis that there might exist stronger correlations between the categories and more hidden attributes of each job description, such as words used in the description. This is based on the theory that a writer tends to be more conscious when naming a job title or specifying the required skills, but not as much when writing the rest of the description.

3. Project Overview

This research aims to find a hidden correlation between the category and words used in a job description, and evaluate its accuracy. In order to achieve this goal, the job descriptions for data scientist and data analyst have been retrieved from the job posting website, Indeed. After cleaning the job descriptions, Java TreeView and Hidra were used to conduct a hierarchical cluster analysis. Then, a topic model was introduced to find latent relationships between data science and analytics job descriptions. In this research, Latent Dirichlet Allocation was used for a topic model. The next chapter lists the methods and steps taken in the research in details.

Chapter 3: Methods

This research aims to find hidden correlations between the data science and analytics job categories and words used in job descriptions. This chapter introduces the software used in order to conduct the research.

1. Open-Source Software Used

Web scraping was done by using Beautiful soup, which is a Python library for pulling data from HTML and XML files (Richardson, 2019). In addition to a Python standard library, urllib3 was also used to fetch URLs (Petrov, 2019). Hierarchical clustering was done by scikit-learn (Cournapeau, 2019) and Java Treeview (Saldanha, 2003). Additionally, Hidra, the modified Java Treeview was applied to compare the data from two files. While analyzing a hierarchical clustering, G*Power was used to calculate the

value of effective correlation (Buchner, 2007). Latent Dirichlet Allocation was performed by gensim, a Python library for topic modeling. The use of each software tool is discussed in more detail in the following sections.

2. Obtaining Job Descriptions

The job description for data scientists and analysts were retrieved from Indeed by traversing the returned search results using Beautiful Soup. The search for full-time entry-level “data scientist” returned 1024 results after all duplicates were removed. Similarly, the search for full-time entry-level “data analyst” returned 1094 results.

3. Hierarchical Clustering

After the job descriptions from Indeed were cleaned, the word occurrences were calculated using count-vectorizer in

scikit-learn library. Then the separate program clustered the data and generated the tab-delimited text file called clustered data file (cdt file). Java Treeview then read in this cdt file to visualize the data. The main output of Java Treeview is a dendrogram, which shows the similarities of words in a same small cluster. In this research, Hidra, the modified version of Java Treeview developed by Dr. Matthew Hibbs, was also performed. While Java Treeview can only handle one cdt file, Hidra can read two files at the same time to compare the dendrograms generated from two separate cdt files. It shows if the words, which clustered in the same group for one dendrogram, are also clustered in the other dendrogram. If not, it also tells to which cluster that specific word belongs in the other dendrogram. As stated by Alok Saldanha, "a hierarchical clustering is a sequence of partitions in which each partition is nested into the next partition in the sequence (Dubes & Jane 1988)." In this research, a hierarchical clustering method is more appropriate than a partitional

clustering such as K-means, because it allows us to see partitions at a different level of correlations. The words from job descriptions can be estimated to positively correlate even if they belong to a different cluster. Visualizing partitions in a tree structure allows us to view correlations at different levels.

4. Topic Modeling

Topics modeling is a statistical model to find topics that are observed in a collection of documents. In this research, Latent Dirichlet Allocation (LDA) was applied. LDA consists of three-level hierarchical Bayesian models. It supposes that each document in a collection is a finite mixture of topics. Then each topic is modeled as an infinite mixture of topic probabilities (Blei, Ng, & Jordan, 2003). In order to generate a model, the dictionary was created from the job descriptions. Stop words and words that are commonly used in job

descriptions for any kinds of jobs were removed at this point. The bag of words containing a frequency of each word was created, and plugged in to the LDA model from gensim, a Python library for topic modeling.

Chapter 4: Results and Conclusion

This chapter shows the results of the research conducted by using the methodologies that are discussed in the previous chapter. First, we will go through the results from the hierarchical clustering and topic model. Then the implications from those results as well as the problem difficulty of this research will be addressed.

1. Hierarchical Cluster Analysis

Figure 4.1.1 is the dendrogram of full-time entry-level data scientist job descriptions that are created by Java Treeview. In this dendrogram, the column represents each job description and the raw represents each word. The red spot is shown in the intersection of a word and a job description in which this specific word was used. Although it is hard to scrutinize each small cluster due to the large size of data, it clearly shows that

there are some evident correlations between the words that are used together.

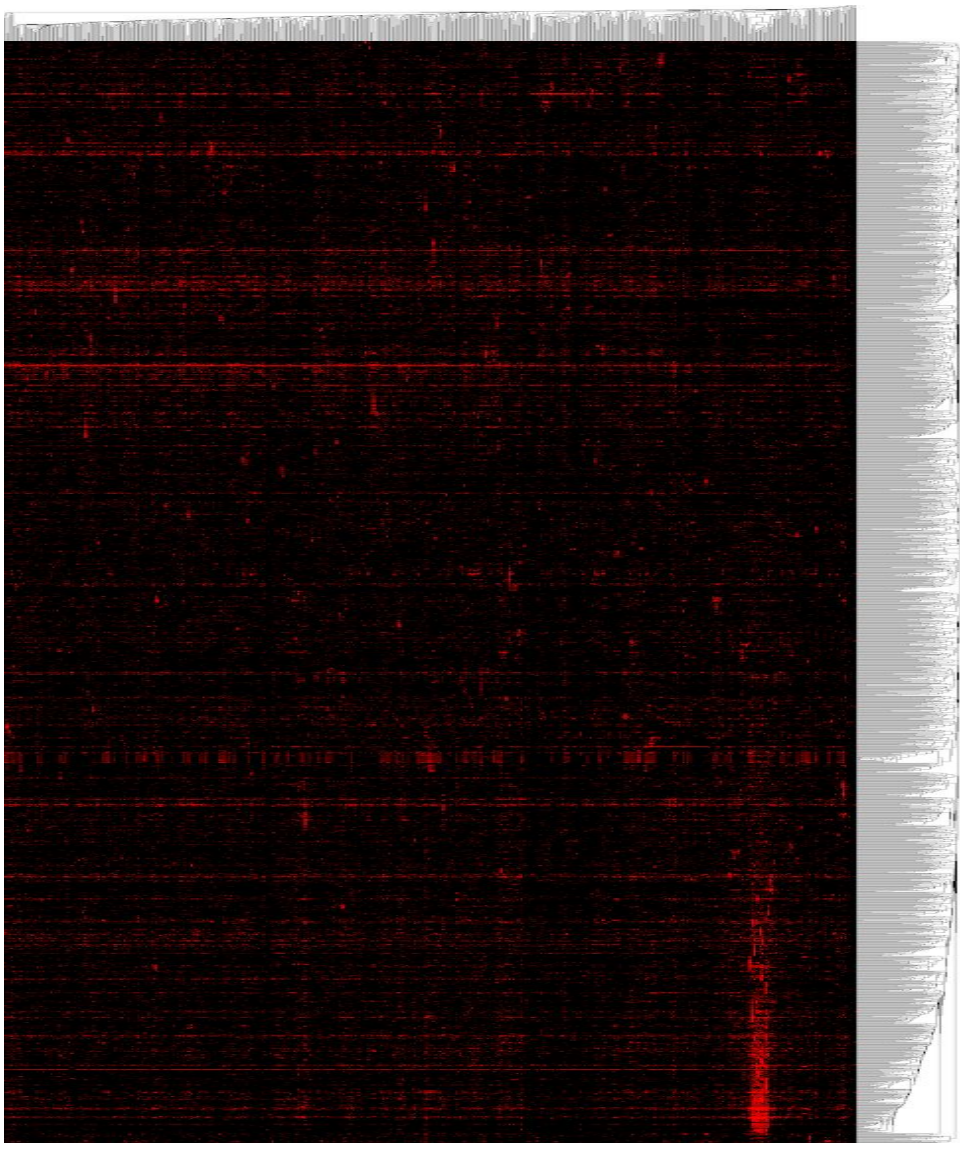

Figure 4. 1. 1: Hierarchical Clustering of full-time entry-level data scientist job descriptions.

For instance, positions at governmental institutions such as U.S. Department of the Air Force and U.S. Federal Government are grouped together. They have a correlation of 0.579.

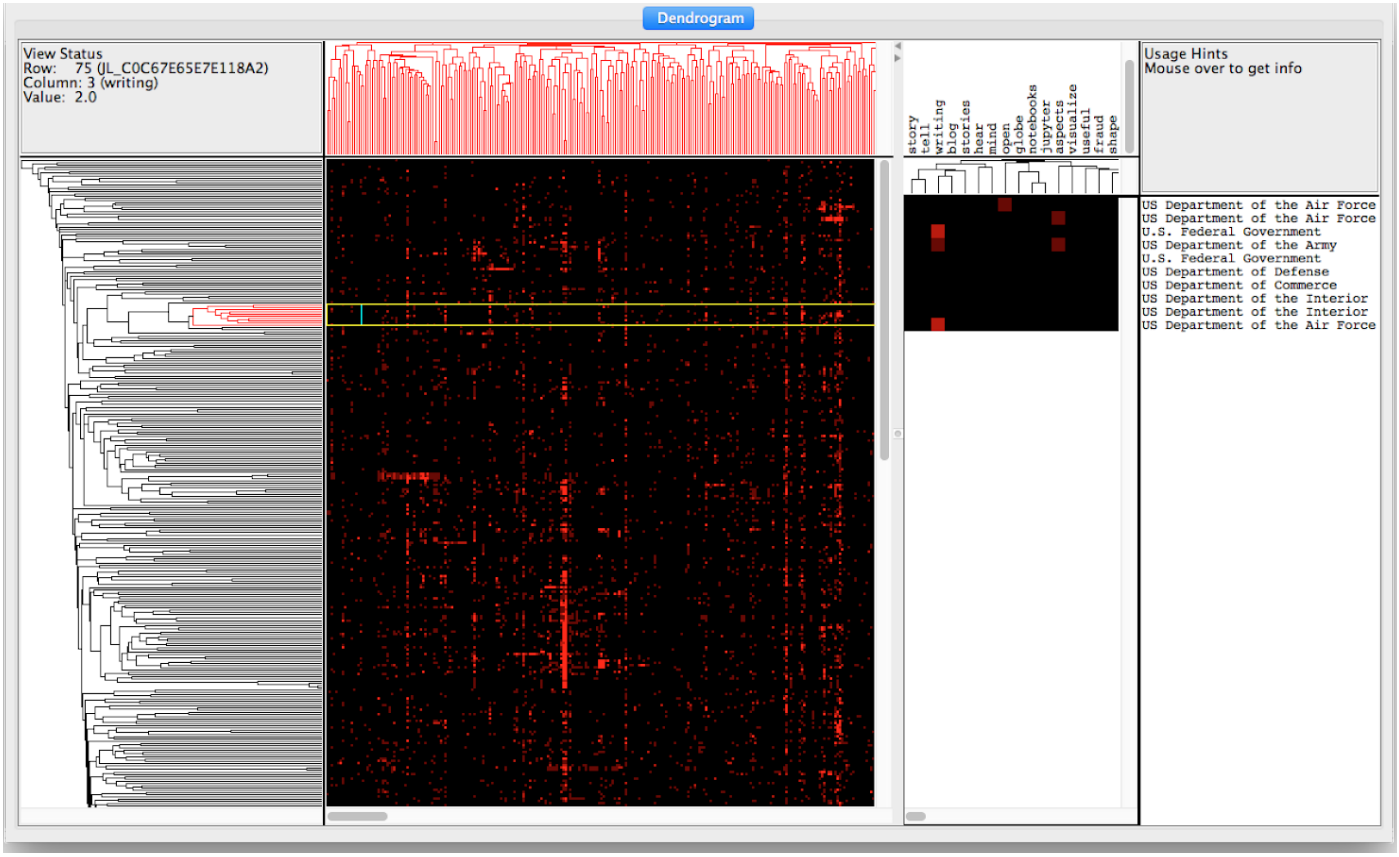

Figure 4.1.2: The small cluster of job positions at the governmental institution.

The obvious set of words that are usually used together were also clustered together. Figure 4. 1. 3 shows that words related to equal employment opportunity belong to the same cluster. This cluster includes words such as equal, age, identity, color, and gender. They have 0.425 correlations.

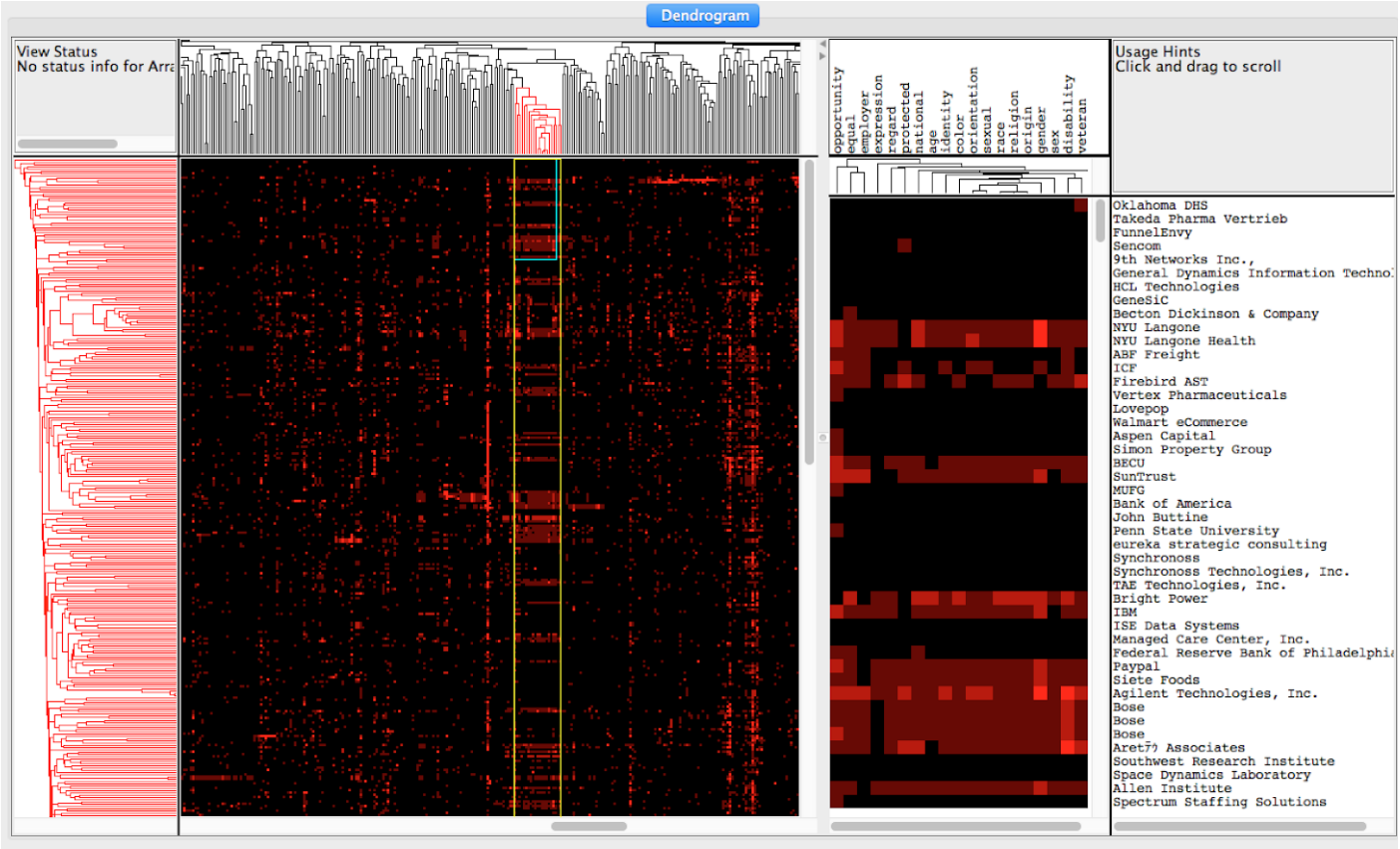

Figure 4. 1. 3: The cluster for words related to equal employment opportunity.

In order to look into details, words that are too general and used in any kind of job descriptions were removed for the next step. The words that did not occur much and occurred too often were removed as well. The list of those general words was obtained by gathering the frequent words used in job descriptions for non-technical occupations: writer, receptionist, and stylist. After removing those words, the data scientist file contained 506 words, and the data analyst file contained 398 words. Figure 4. 1. 4 is the dendrogram of the data scientist job descriptions with limited words. In this dendrogram, the column represents each word and the raw represents each job description.

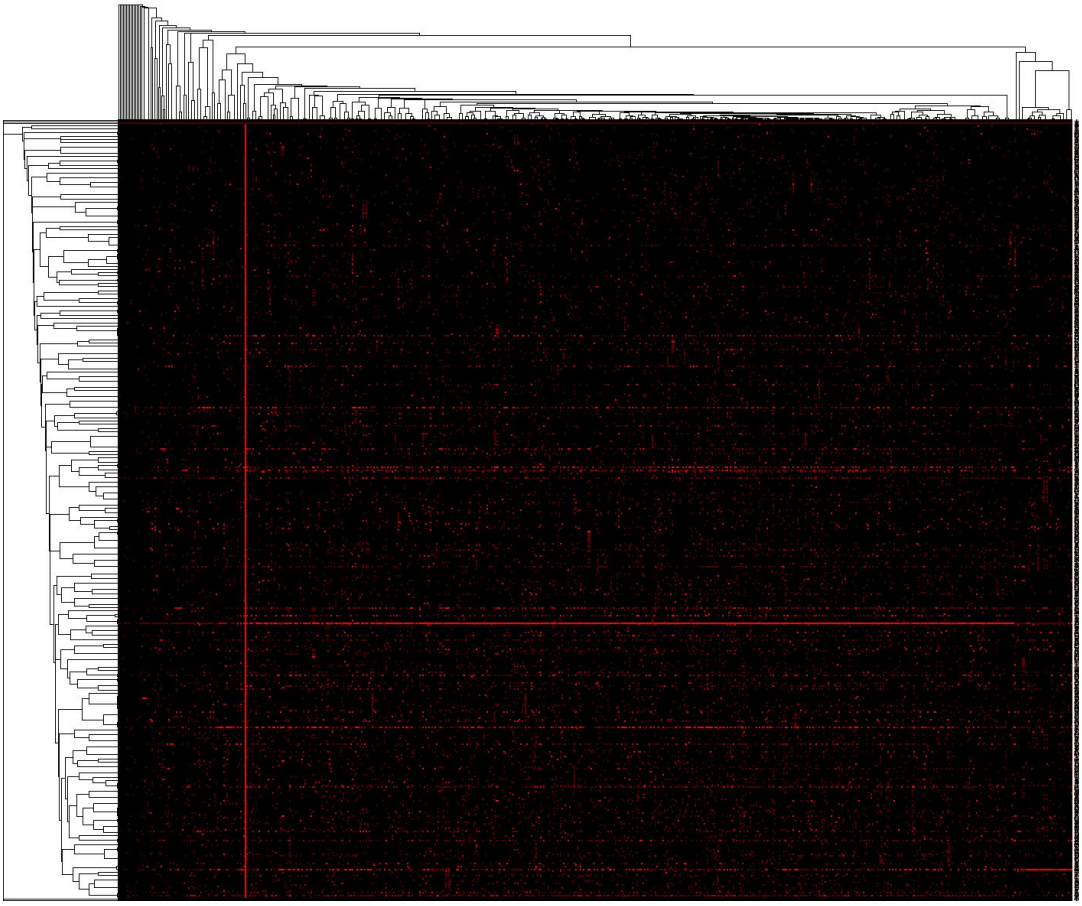

Figure 4. 1. 4: The full-time entry-level data scientist dendrogram with limited words.

From this dendrogram, the clusters that have a minimum of 0.3 correlation have been extracted. This threshold was determined by G*Power, which calculates an effective correlation given sample size and required extent of accuracy. The correlation of 0.3 has a 95% prediction interval

on 1000 samples. Below is the list of some clusters that have correlations higher than 0.3.

Universities Utilities Substantial Logic Notice Assessments Linear Theory Probability Rules	Air Autonomy Wrangling Pharmaceutical Raw Modeling Builds Insightful Derive
Functionality Validation Automation Enables Deploying Scala Regression Frameworks Experiment Analysts Computational	Dynamics Literature Suitable Approaches Proposed
Epidemiology Uncover Cognitive Analyst	Costs Analyzes Trees Tables Dashboards

Simulations Numpy Algebra Framework Matplotlib Hidden Cloud	Spectrum Prototypes Formal Laboratories Code
Pipeline Computation Biomedical Collaboratively Deployment NLP Extraction Biology Natural Cell	Reinforcement Torch Tensorflow Architecture NIPS Mathematics ICML Cambridge Advances
Computing Stata PHP POOL AWS Sciences CSS Decision DS	Implementation Querying Strongly Onsite Analyzing Multidisciplinary Analytical DC

Table 4. 1. 1: Example word clusters that have correlations higher than 0.3.

Some of the above clusters are easy to comprehend the reason why the words were clustered together in the same group. The clusters also include the words that are not the name of the skills, i.e., the words that would be listed under the required skills section in job descriptions. The examples include but are not limited to programming languages, related courses, and degrees. As we discussed in Chapter 2, clear attributes of job descriptions such as job titles and technical skills needed are more subject to a writer's own definition of "data science" or "data analytics". Those words, which were used in job descriptions in order to explain the position in detail, could be some criteria to distinguish data scientist positions from data analyst positions even when the same skill sets are required for both positions.

In order to investigate if the words co-occurred with highly technical words in the data scientist job descriptions differ from the ones in the data analyst, Hidra was performed on these two datasets. Hidra allows us to import two cdt files,

and see if words clustered in a same group were also clustered in a same group in the data from the other file. Figure 4. 1. 5 focuses on the cluster from data scientist job descriptions that have multiple words related to machine learning. The right side is the dendrogram created by the data scientist file and the left is the one by the data analyst file. While the word “mathematics” also showed up in this cluster for data scientist, the technical words related to machine learning and “mathematics” were not used together in the data analyst descriptions.

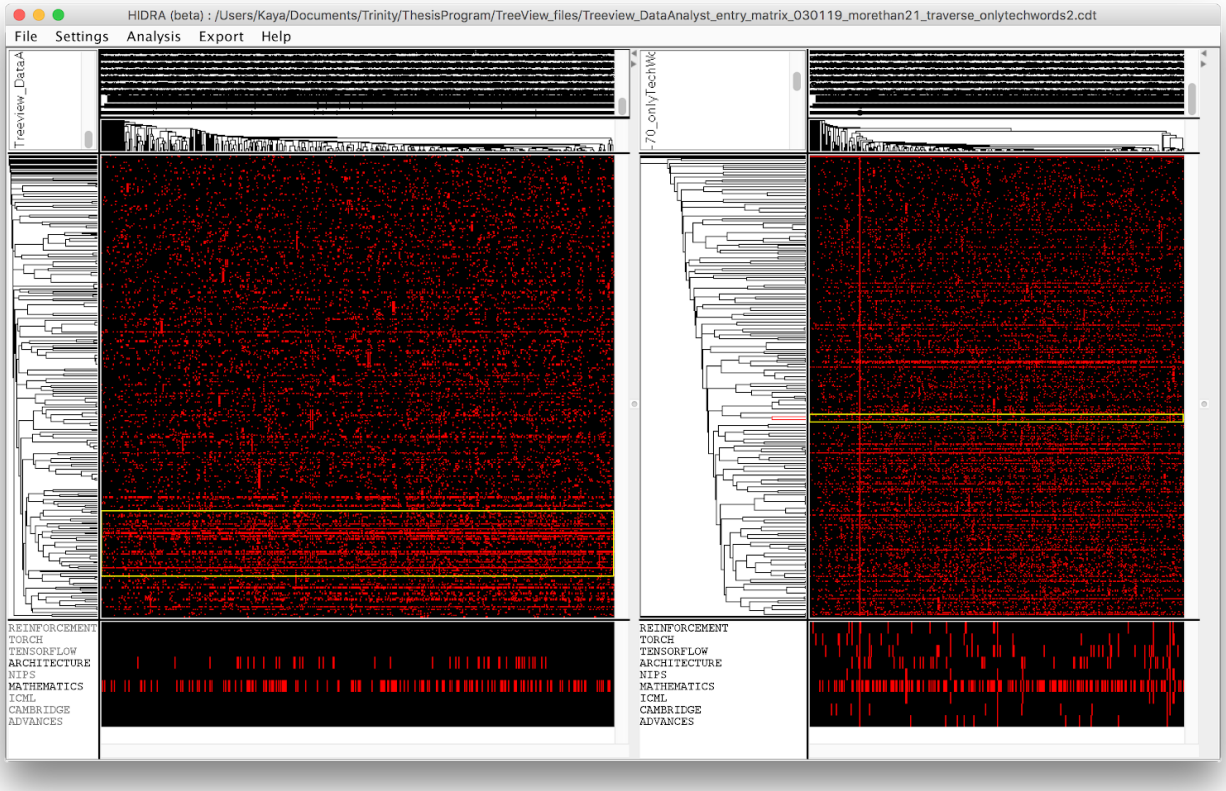

Figure 4. 1. 5: A dendrogram comparison using Hidra. The word “mathematics” was correlated with machine learning words in the data science descriptions, but not in the data analyst descriptions.

Likewise, in Figure 4. 1. 6, the words “analytical” and “analyzing” have a strong correlation in the data scientist descriptions as well as in the data analyst descriptions.

However, the other words in this cluster such as

"multidisciplinary" and "onsite" were not correlated well with words related to analysis in the data analyst descriptions.

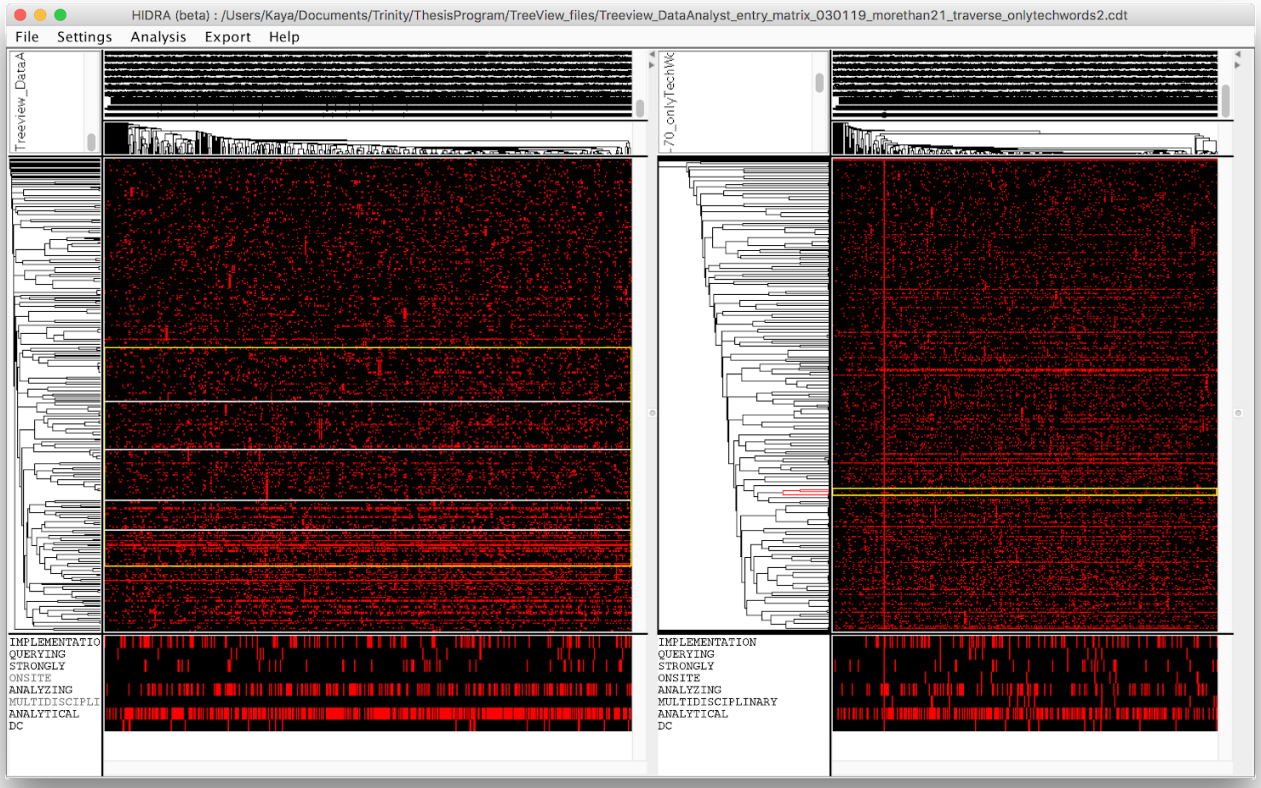

Figure 4. 1. 6: A cluster comparison using Hidra. While many words co-occurred with words related to "analysis" in the data science job descriptions also have some correlations in the data analyst job descriptions, the two words, "onsite" and "multidisciplinary" were not clustered in the same cluster.

These patterns agree with the hypothesis of setting apart data scientist positions from data analyst positions by the difference of words that are used with a specific technical word such as a programing language. However, it is hard to identify which one of those words surrounding the very technical word can be an accurate criterion to distinguish between the two jobs.

2. Topic Modeling

50 data scientist job descriptions and 50 data analyst job descriptions were fed into the topic model, and the LDA were tasked to divide all the job descriptions based on the hidden correlations between the words. As LDA is an unsupervised learning algorithm, it is asked to come up with the definition of each category rather than being given the definition. In this research, we coerced the model into two topics as the purpose is to distinguish data scientist job descriptions from data analyst job descriptions.

In order to validate the accuracy of this LDA model first, we input randomly chosen 50 data scientist and 50 receptionist job descriptions to the model.

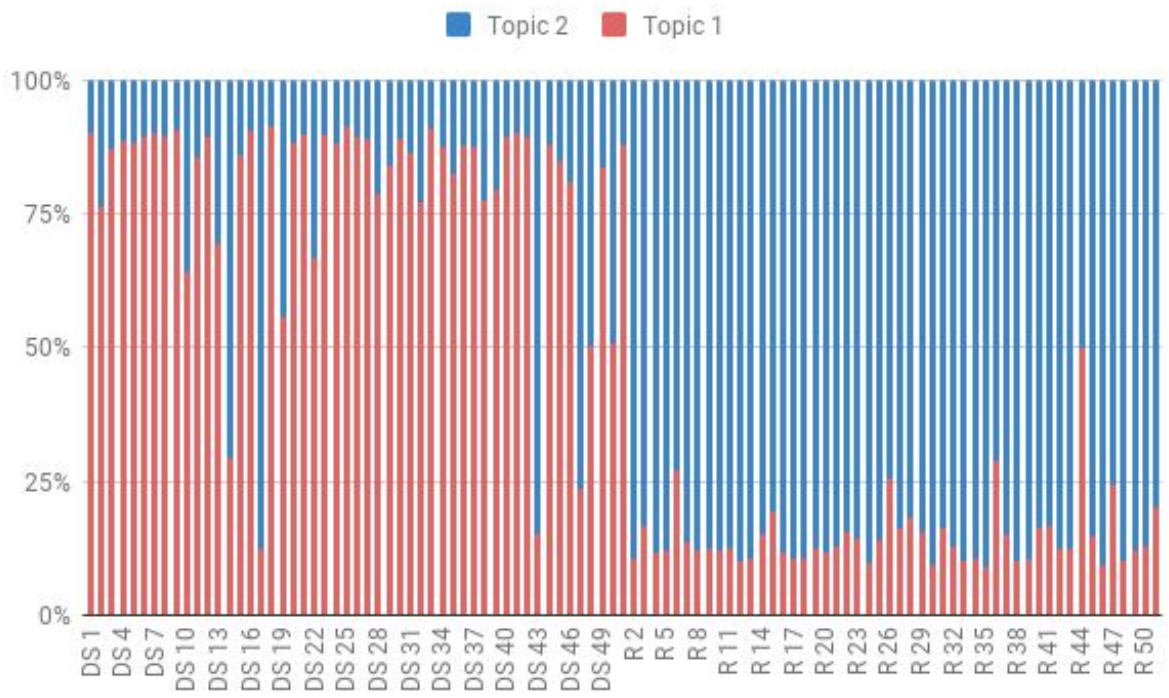

Figure 4. 2. 1: The LDA analysis on 50 data scientist and 50 receptionist job descriptions.

The y-axis is the probability that each job description belongs to either topic 1 or topic 2, which are LDA generated topic clusters based on its learning. “DS” stands for data scientist

and “R” stands for receptionist. While there are a few outliers in the graph, most data scientist jobs allocated to the topic 1 and most receptionist jobs allocated to the topic 2.

Since we have validated the LDA model, we now input randomly chosen 50 data scientist and 50 data analyst job descriptions. Figure 4. 2. 2 shows the result of this LDA analysis. “DS” stands for data scientist and “DA” stands for data analyst.

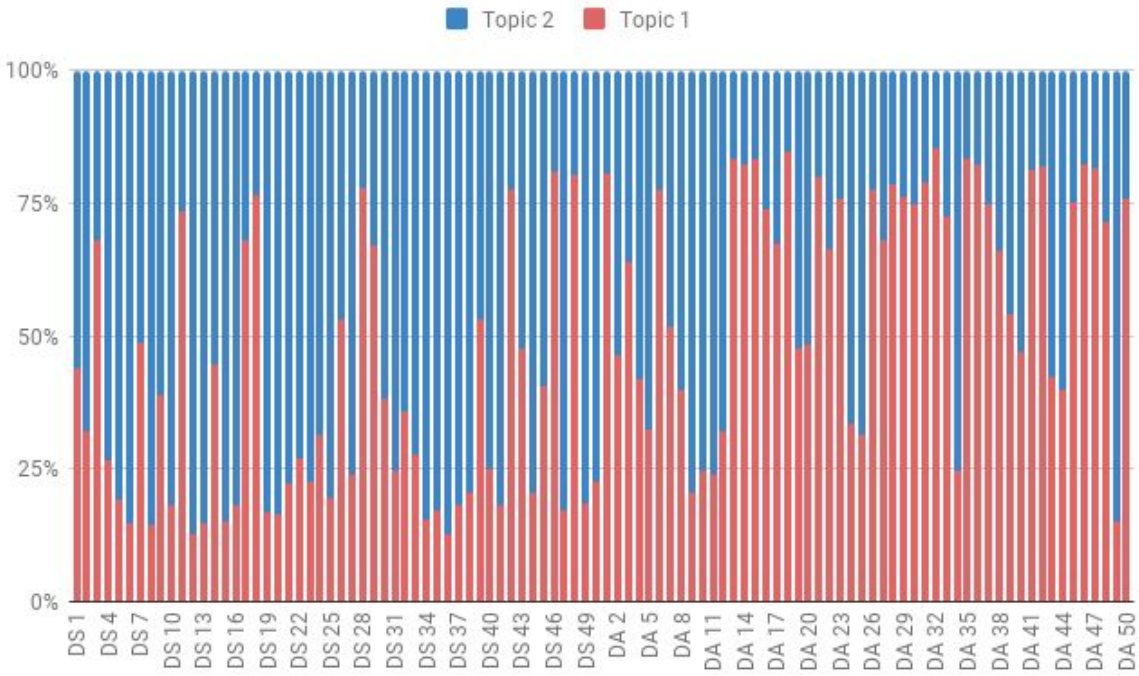

Figure 4. 2. 2: The LDA result on 50 data scientist and 50 data analyst job descriptions. It shows how likely each job description belongs to Topic 1 or 2, which are defined by LDA.

Although there are more outliers in this graph compared to Figure 4. 2. 1, more data scientist jobs are allocated to the topic 1 and more data analyst jobs are allocated to the topic 2. Table 4. 2. 1 shows how LDA defined the two topics for this

dataset. It is the list of words that have highest probabilities to be observed in each topic.

	Topic 1	**Topic 2**
1	research	r
2	analyst	statistics
3	r	modeling
4	statistics	algorithms
5	system	scientist
6	engineering	solve
7	modeling	research
8	quantitative	strategic
9	sources	analyst
10	trends	engineering

Table 4. 2. 1: Ten words that have highest probabilities to be used in each of topic 1 and 2.

6 out of 10 words that have highest probabilities in each topic are same. However, topic 1 is allocated more data scientist jobs and topic 2 is allocated more data analyst jobs. This

suggests that the hypothesis we have established earlier is correct. There exist hidden correlations between not so technical words used in descriptions that differentiate data science job descriptions from data analyst descriptions, and vice versa.

3. Problem Difficulty

The hierarchical cluster analysis and topic modeling provided some insights of a possible way to distinguish data science and analytics jobs from the co-occurrences of the words. However, those analyses could not identify exactly what set of words can become accurate criteria and differentiate two job fields.

The major cause of this problem was the variety of words people used when writing job descriptions. The number of vocabularies used in total was much broader than we assumed. The other issue that we encountered was the

formatting differences in job descriptions. A non negligible number of job descriptions were missing detailed descriptions of the positions because the detailed descriptions were placed under the wrong section of the job posting form that was provided by Indeed. This hindered us from getting the sufficient amount of information when traversing the job postings by web scraping.

Chapter 5: Closing Remarks

At the end of the previous chapter, the possible cause of the issues encountered in the research were discussed. In this chapter, a potential solution to those issues will be addressed. Furthermore, how this research can lead to the next step to solve the problem of unclear definitions of data science and analytics will be shown.

1. Future Work with Clustering

The issues of excessive variety of vocabularies may be solved or alleviated by introducing Latent Semantics Indexing (LSI). While the number of dimensions was equal to the number of words in LDA, LSI allows us to cut down the number of dimensions and obtain a refined outcome.

2. Applications of the Research

While this research requires improvements, it is possible to create a software tool that evaluates the quality of data science and analytics job description-writing as a future work. Based on the topic model, it would tell whether a job description is well written or poorly written so that employers can avoid producing inaccurate job postings and hiring an employee that is unsuitable for the position.

Furthermore, the research would be able to provide some insights to establish clear definitions of data science and analytics once the model becomes capable of clustering each job at higher accuracy. This would be a direct solution to the disrupting data science and analytics job markets as it would eradicate confusing job titles or descriptions. It would also help higher education institutions to build curriculum that better prepare students who wish to go into the data science and analytics fields upon

graduation. As a result, a high turnover rate as well as long times to fill job openings in the fields would be reduced.

References

Berry, J. (1994, September 4). Database Marketing. *BusinessWeek.* Retrieved from https://www.bloomberg.com/news/articles/1994-09-04/database-marketing

Blei, M., D., Ng, Y., A., & Jordan, I., M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022. Retrieved from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Buchner, A. (2007). G*Power: Statistical Power Analyses for Windows and Mac. Retrieved from http://www.gpower.hhu.de/

Chenard, E. (2018, January 18). Why Are Data Science Leaders Running for the Exit? *Oracle + Datascience.com.* Retrieved from https://www.datascience.com/blog/why-data-science-leaders-fail

Codeup. (2018, October 17). Data Science vs Data Analytics: What's the Difference? Retrieved from https://codeup.com/data-science-vs-data-analytics-whats-the-difference/

Cournapeau, D. (2019). Scikit-learn. Retrieved from https://github.com/scikit-learn/scikit-learn

Davenport, T., Cohen, D., & Jacobson, A. (2005). Competing on Analytics. *Working Knowledge Research Report.* Retrieved from http://www.babsonknowledge.org/analytics.pdf

Dubes, C., R., & Jain, K. A. (1988). *Algorithms for Clustering Data.* New Jersey, Upper Saddle River, Prentice-Hall Inc.

Fatemi, F. (2016, September 28). The True Cost Of A Bad Hire -- It's More Than You Think. *Forbes.* Retrieved from https://www.forbes.com/sites/falonfatemi/2016/09/28/the-true-cost-of-a-bad-hire-its-more-than-you-think/#7e2963be4aa4

Glassdoor. (2019). [Graph illustration of Salary Distribution]. Data Scientist Salaries. Retrieved from https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm

IBM & Burning Glass Technology. (2016). The Quant Crunch: How the Demand for Data Science Skills Is Disrupting the Job Market. Retrieved from https://www.ibm.com/downloads/cas/3RL3VXGA

Naur, P. (1974). Concise Survey of Computer Methods. *Journal of the Association for Information Science and Technology, 27, no. 2,* 125-126. https://doi.org/10.1002/asi.4630270213

Petrov, A. (2019). Urllib3. Retrieved from https://urllib3.readthedocs.io/en/latest/

Press, G. (2013, May 28). A Very Short History Of Data Science. *Forbes.* Retrieved from https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#7e28c70c55cf

Richardson, L. (2019). Beautiful Soup. Retrieved from https://www.crummy.com/software/BeautifulSoup/

Saldanha, A. (2003). Java Treeview. Retrieved from https://sourceforge.net/projects/jtreeview/files/

Tukey, John W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics, 3, no. 1,* 1-67. doi:10.1214/aoms/1177704711.

University of California Berkeley. (2018). What is Data Science? Retrieved from https://datascience.berkeley.edu/about/what-is-data-science/

Waters, R. (2017, November 29). How machine learning creates new professions — and problems. Financial Times. Retrieved from https://www.ft.com/content/49e81ebe-cbc3-11e7-8536-d321d0d897a3