Jobs for Data Scientists: Skills, Locations and Industries

Yue Cao

Project Description:

Perform an analysis of “data scientist” jobs listed on job boards and on the employment pages of major companies. What are the most common skills that employers look for? What are the most unique skills that employers look for? Where are the types of companies that employ the most data scientists?

1. Introduction

As the modern society becomes more driven by digital activities and networks, data has become increasingly important in many industries. Besides software and Internet companies, even traditional industries, such as manufacturing and retail, are more dependent on high-quality data analysis. The demand for data scientists is rising accordingly.

Firstly, how to define a data scientist job? In general, Data Science is a field which studies how Information is gathered, what it conveys, and how that information can be converted into a valuable resource to generate meaningful insights for the betterment of a business as a whole[1]. Thus data scientists are the people who conduct these kinds of tasks in various industries.

To learn more about data scientist jobs, we perform a keyword-based skill search for data scientist jobs on three main online job boards. By extracting data scientist job information, such the job title, company name, skills, industry, etc., we can find out what are the most popular skills that employers preferred for data scientists and other unique characteristics.

2. Methods

2.1 Data Collection

In general, we collect data from online job boards using web scraping tools. This procedure contains a few key steps: job boards selection, web scraping and data cleaning. There are many popular online job boards which we can search for data scientists jobs. We first tried indeed.com and monster.com. However, these two websites seem to have instable html nodes and varying structure of subpage when following the link corresponding to a job title, which causes much trouble when we do web scraping to extract data science job information. We notice that Stackoverflow.com, Dice.com and Glassdoor.com have nice and stable web structures for web scraping. Especially, Dice returns more than 38,000 jobs when we search for data science jobs, and Glassdoor contains pre-classified industry information. Thus, we use these three jobs boards as the source of data science job data.

We mainly use `rvest` package by Hadley Wickham to scrape information from webpages in R. To know the HTML nodes related to specific parts of a webpage, such as job title, company name or job description, we need to use a Chrome extension, SelectorGadget. It is a web tool which returns the corresponding name of CSS or XPath when we click a specific part on a webpage. The main steps are as follows:

- According to the pattern of page number in the URL, obtain the URLs for all the pages (or the first 1 ~ 40 pages if there are more than 40 pages) of the search result for data scientist jobs.
- Use SelectorGadget to obtain the URLs of subpages corresponding to each job title in the search result page. For each subpage, use SelectorGadget to get CSS or XPath of job information, and apply html_node() function to extract job title, company name, industry, company size, location and job description.
- Use grep() function to search for the occurrence of theses skill keywords: Hadoop, Spark, R, SAS, Stata, Java, Perl, Python, SQL, NoSQL, Tableau, Excel, Machine Learning, Amazon Web Service, which are skills commonly asked by employers looking for data scientists. If the search result of a keyword is `TRUE` for a job, it means this keyword is mentioned in the job description and we regard it as a required skill for the job.
- Use readLine() function to extract industry information from the HTML source file of Glassdoor. Since there is no such nicely listed information on Stackoverflow and Dice, in this study, we only use industry information from Glassdoor.
- Create a data frame to combine all the job information together, including job title, company name, industry, location and search result of skill keywords.

2.2 Data Cleaning

In the data cleaning procedure, we first remove all the duplicates in the data sets and remove the rows which contain missing values. The readLine() function may produce some HTML language which needs to be transferred to plain English. For example, we substitute the string "&" in the industry fields to "&". In each data set of the three job boards, we summarize the frequency (percent) of each skill by counting the numbers of `TRUE` divided by the total number of jobs.

After data cleaning, we obtained data of 101 jobs from Stackoverflow, 498 jobs from Dice, and 444 jobs from Glassdoor.

3. Exploratory Data Analysis

3.1 Skill Frequency for Data Scientists Jobs on Stackoverflow, Dice and Glassdoor

We first add a "source" variable to each data sets of the three jobs boards, indicating the data source website. R package `ggplot2` is then applied to create a barplot to compare skill frequency for data scientists jobs on stackoverflow, dice and glassdoor.

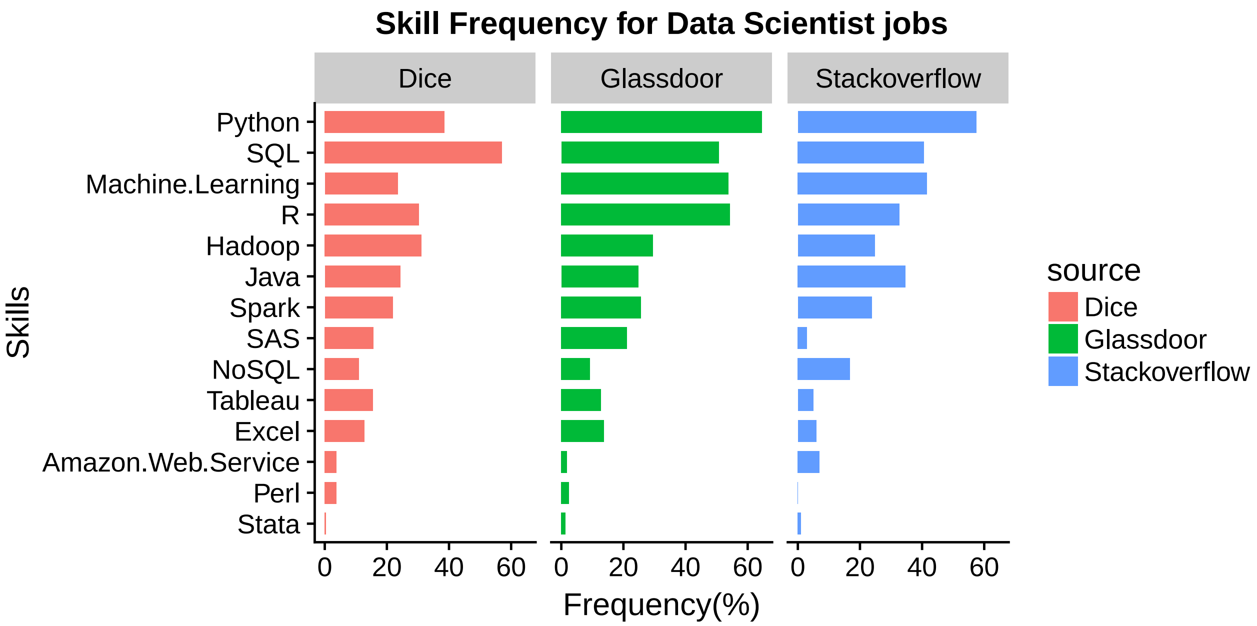

Figure 1. Compare Skill Frequency for Data Scientists Jobs on Dice, Glassdoor and Stackoverflow

In Figure 1, we can conclude that Python is the most popular skill required by employers on Glassdoor and Stackoverflow, while SQL is the most mentioned skill on Dice. More specifically, 64.64% of the data scientists jobs on Glassdoor required Python, and 57.43%, 57.03% of jobs on Stackoverflow and Dice require Python and SQL, respectively. Machine Learning, R, Java and Hadoop are also popular requirements for data scientists jobs. If we combine the data from the three job boards together, as we can see in Table 1, the most common skills that employers look for are (in the descending order of) SQL, Python, R, Machine Learning and Hadoop. Among these popular skills, SQL and Python are needed for more than half of the jobs in the combined data from the three job boards.

	skill	count	percent
1	SQL	550	52.73
2	Python	537	51.49
3	R	425	40.75
4	Machine.Learning	398	38.16
5	Hadoop	311	29.82

Table 1. Top Five Required skills in Combined Data Sets from Stackoverflow, Dice and Glassdoor (N=1043)

3.2 Number of Data Scientists Jobs Varies by Location and Industry

We use the combined data set to find out which locations have the most data scientists jobs. In Figure 2 (A), Among all the 1043 jobs, New York, NY has the most abundant data scientists jobs. San Francisco, CA and Chicago, IL also have many data scientists positions currently open. To be detailed, there are 81, 69 and 56 data scientists jobs in the location of New York, NY, San Francisco, CA and Chicago, IL, respectively.

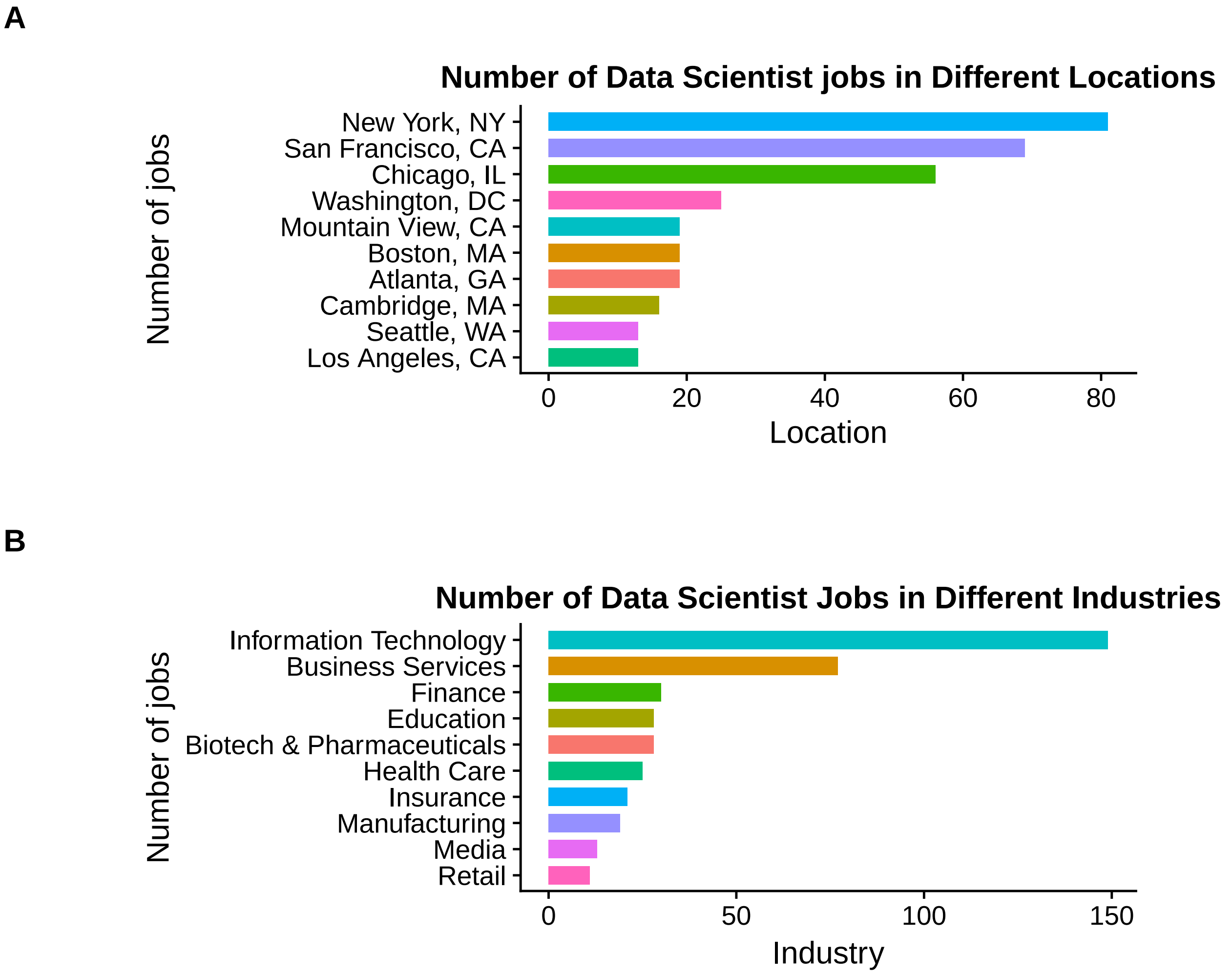

Figure 2. (A) Top 10 Locations Employing Most Data Scientists (N=1043); (B) Top 10 Industries Employing Most Data Scientists (Glassdoor)

We also draw a barplot to explore the relationship between the number of data scientists jobs and various types of industries. As we only have industry information from Glassdoor, Figure 2 (B) is created from the data set of Glassdoor with 444 observations and show the top 10 industries employing most data scientists. Companies in Information Technology employ much more data scientists than other industries. Business Services industry is also in a high demand for data scientists. There are 149 data scientists jobs from Information Technology and 77 jobs from Business Services in total.

3.3 Hierarchical Clustering on Industries

Can industries be clustered into groups based on their skill needs? To solve this question, we first summarize the number of each skill grouped by industry, and then conduct a hierarchical clustering on the distance matrix of the scaled skill data set. Here we use Euclidean distance and complete linkage for hclust() function. In Figure 3, we set the number of clusters to be 4, and use different colors to indicate the clustering results. Base on skill needs for data scientists, Information Technology and Business Services form two one-element clusters;

Finance, Biotech & Pharmaceuticals, Media, Health Care and Insurance are in the same cluster; and the other industries in the dendrogram constitute the last cluster.

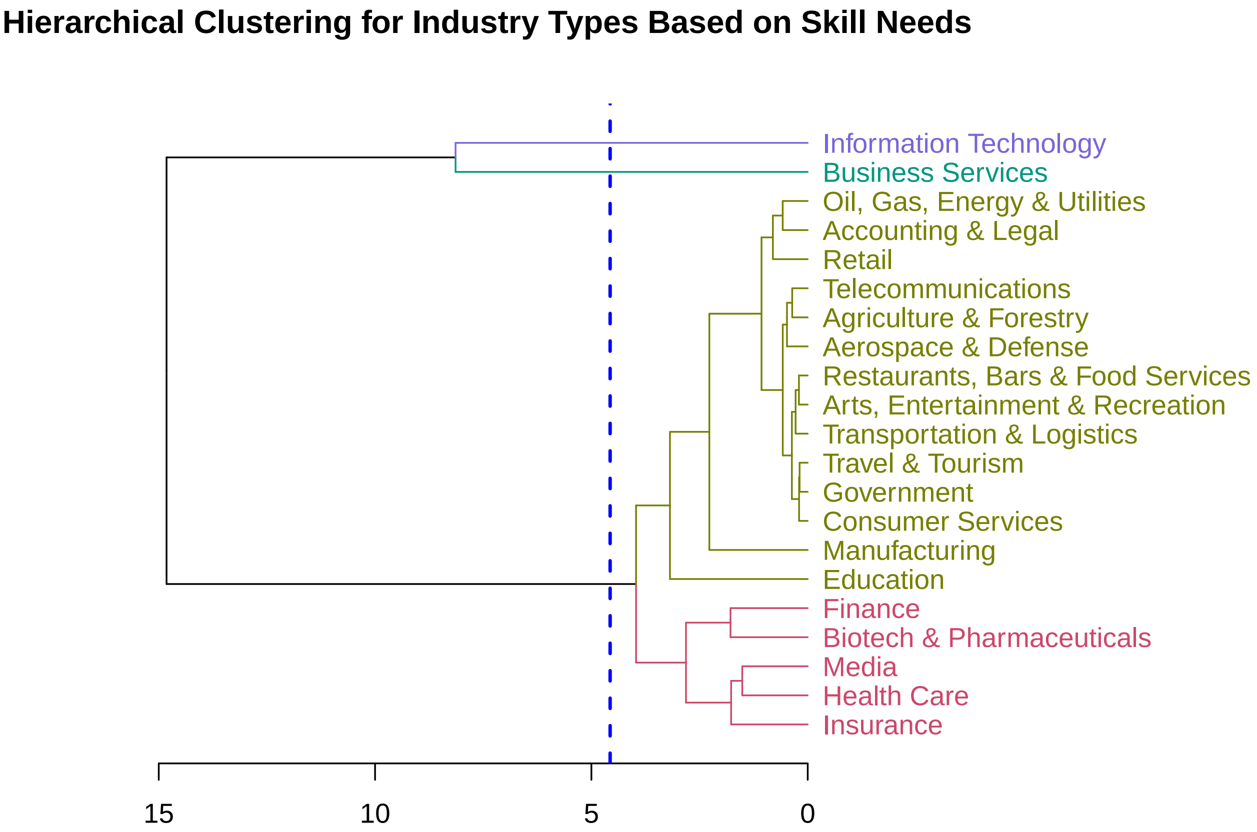

Figure 3. Hierarchical Clustering for Industry Types Based on Skill Needs (Glassdoor)

Discussion

This study has several limitations. First, the jobs from three job boards may have overlapping parts. Although we have removed duplications, there is still a chance of two slightly different job descriptions are actually referring to the same position. Second, in the data collection procedure, we consider a skill mentioned in a job description as a required skill by the employer. However, if the jobs description says “no need for Python experience”, for example, it would cause an opposite conclusion for the skill requirement. Third, the numbers of jobs in each industry are quite different, which may cause a negative influence on the result of hierarchical clustering, due to the imbalance of information. For example, as Information Technology and Business Services have the largest number of jobs, they tend to form individual clusters since they are different from other industries in quantity.

Reference

[1] Shiv Shet (Apr. 29, 2016). An Introduction to Data Science. (https://dzone.com/articles/an-introduction-to-data-science, Accessed Oct. 9, 2017)

[2] Hadley Wickham (2016). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.2. https://CRAN.R-project.org/package=rvest

[3] Hadley Wickham (2016). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.1.0. https://CRAN.R-project.org/package=stringr

[4] Hadley Wickham and Romain Francois (2016). dplyr: A Grammar of Data Manipulation. R package version 0.5.0. https://CRAN.R-project.org/package=dplyr

[5] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

[6] Yihui Xie (2016). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.14.

[7] Tal Galili (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics. DOI: 10.1093/bioinformatics/btv428

[8] David B. Dahl (2016). xtable: Export Tables to LaTeX or HTML. R package version 1.8-2. https://CRAN.R-project.org/package=xtable

[9] Claus O. Wilke (2017). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 0.8.0. https://CRAN.R-project.org/package=cowplot

[10] Stephen Cristiano, jobs_scrape.R, (https://slack-files.com/T6TTWE3G8-F75420VQD-7f4696035e, Accessed Sep. 23 2017)