MA615 Final Project: Job Board Text Analysis

Anna Cook

12/13/2020

Introduction

The purpose of this project is to analyze words frequently used in data science job postings. This is important because, as a graduate student in statistical practice, myself and many of my peers will soon begin applying for jobs. There are many job sites to search and apply for data science jobs, so much so that it can be overwhelming to narrow down the search to find exactly the right jobs. In this analysis, I will determine and compare common words that are used in job descriptions on several popular job sites. This will allow me to identify important key words to use in my online job search in the future.

Data Collection/Description/Processing

The data was collected from three different job board sites: Adzuna, Github Jobs, and The Muse. For each website, a job search was conducting use "data science" as a key phrase, United States as the location, and the results were filtered to include only full-time jobs. The search on Adzuna returned 50 results, Github Jobs returned 8, and The Muse returned 20. The results were then downloaded using the jsonlite R package.

From the downloaded text files, I extracted only the job descriptions, and excluded all other information including job titles, salary, and location. I then manipulated the data frame to turn it into a tidy text format, with individual words being used as tokens. From these new tidy text data frames, I created three tables, one for each website, showing each word and the number of times it was used in a job description on that site (see Table 1). These tables will be visualized and further analyzed below.

Word Frequency Visualizations

For this project, I created two types of visualizations for the word frequencies in job descriptions on each job board site: histograms and wordclouds.

Table 1: Most frequent words for each job board: Adzuna, Github Jobs, and The Muse, respectively

word	n	word	n	word	n
strong	490	experience	37	span	524
data	149	strong	32	data	180
scientist	64	data	30	strong	123
science	56	team	30	experience	102
experience	19	development	27	business	73
description	15	software	25	technical	45

Histograms

The first way I chose to visualize the word frequencies was through a series of histograms. First, I created histograms for the top 20 most frequently used words from each of the three websites. Then, I combined the three data frames and created a histogram for only the words which were common across all three websites. Each of these histograms are shown below.

Adzuna

From the histogram for the Adzuna website, we can see that "strong" is the most frequently used word by a long shot, followed by "data" and "scientist" (see Figure 1) This finding leads me to speculate that the job postings on Adzuna are likely to appeal to more experienced candidates with a lot of very strong skills. "Data" and "scientist" are unsurprising words to be used frequently, since the job search included "data science" as a key phrase. The other words are used much less frequently than these three, but are important to take note of nonetheless.

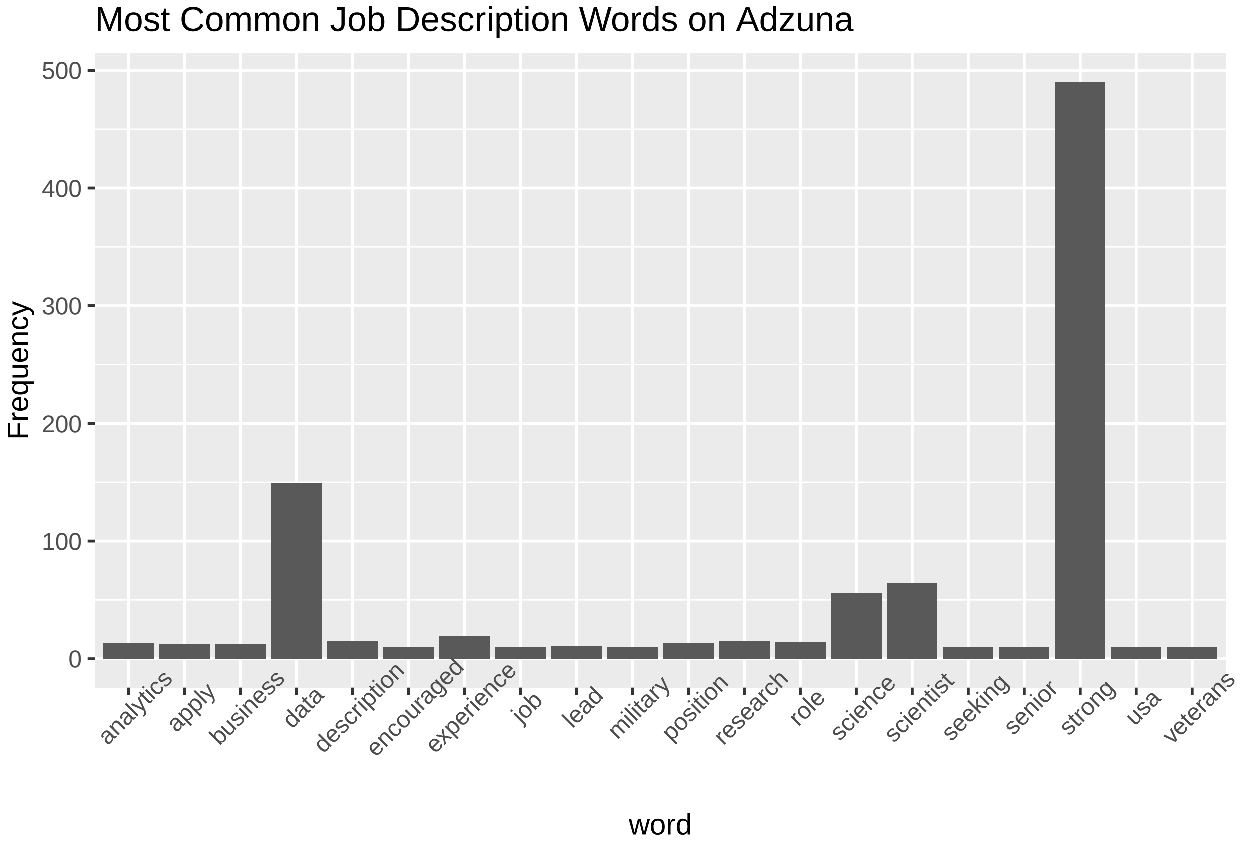

Figure 1: 20 most frequent words used in job description on Adzuna job board website.

Github Jobs

From the histogram for Github Jobs (see Figure 2), we can see that the frequency counts for these words are lower, which makes sense since the initial job search turned up fewer results than that of the other websites. The most commonly used word is "data," followed by "experience," "strong," and "team." This suggests that the job postings on github tend to appeal to candidates who have lots of experience and strong skills working with data, and are looking for candidates who work well in teams. The word "teams" is particuarly a good one to make note of, because as a candidate, I could incorporate into an application or resume that I'm a good team player and collaborate well with others.

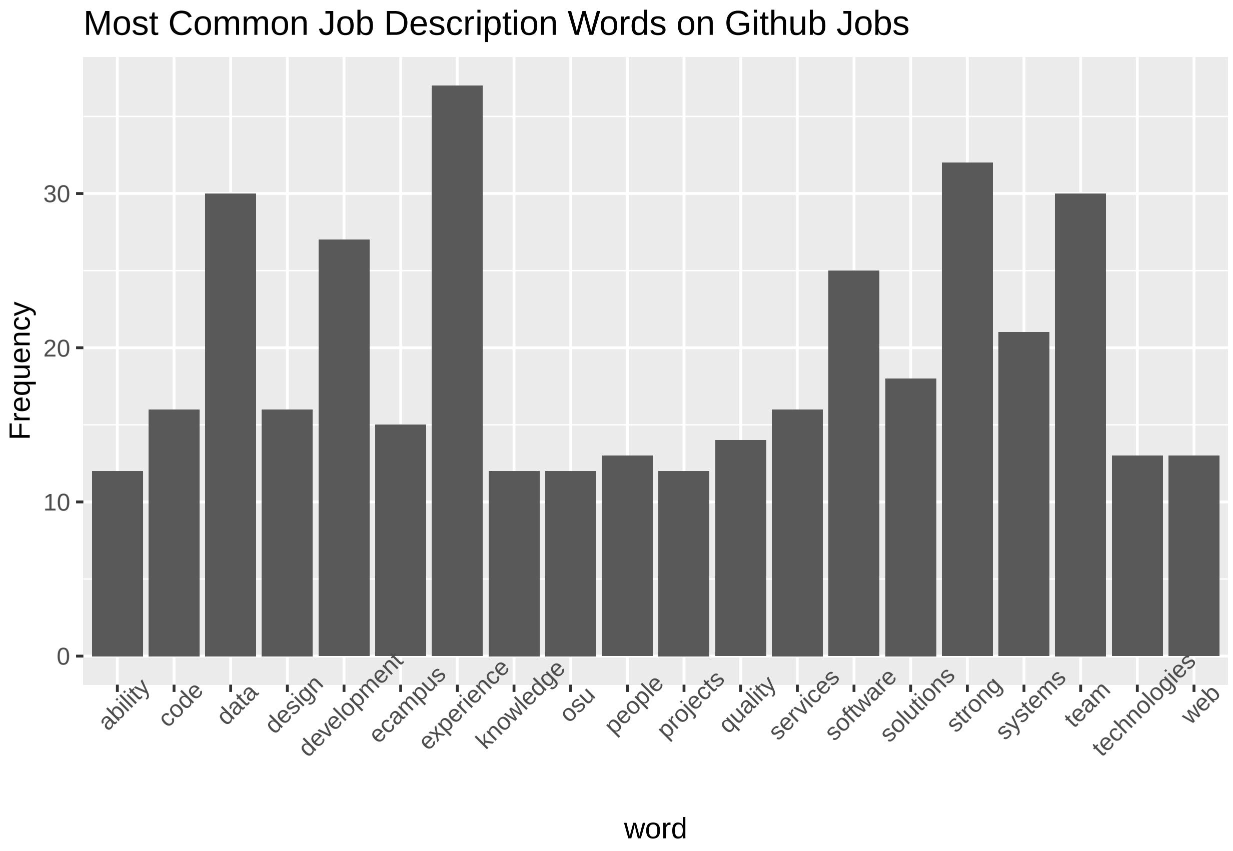

Figure 2: 20 most frequent words used in job description on Github Jobs website.

The Muse

Like Adzuna, the histogram for The Muse shows very high frequencies for the word counts, with the top word being "span" (see Figure 3). This finding is a bit difficult to interpret, and it makes me curious as to why it is used so frequency relative to other words. Other words that appear frequently on this website are "data," "strong," "experience," and "business." Business is a unique and interesting word, and it makes me curious if there are a lot of companies that prefer candidates who have experience in business settings. The other words are similar to the other websites and are unsurprising.

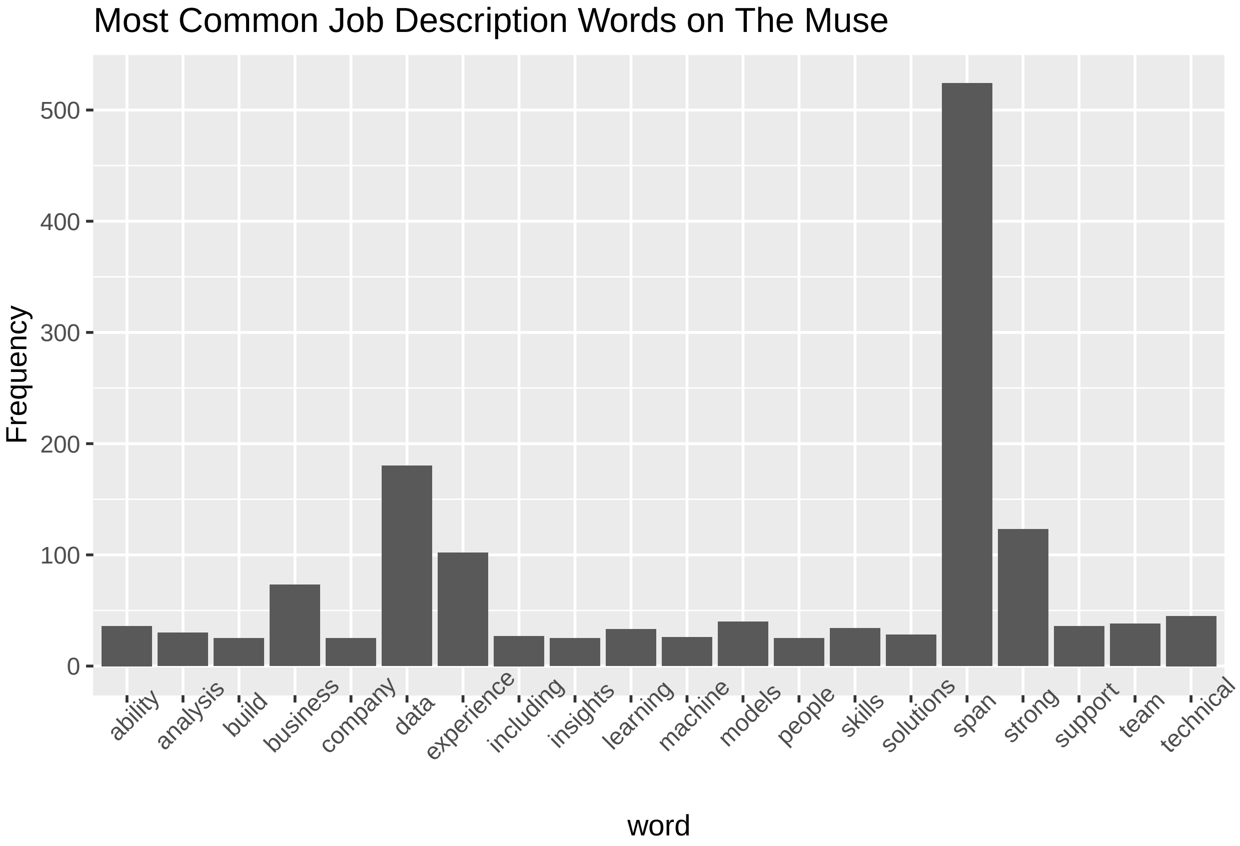

Figure 3: 20 most frequent words used in job description on The Muse job board website.

All Three Sites

The last histogram shows only the words that were in the top 20 of all three websites' most frequently used words (see Figure 4). This way, I am able to directly compare the counts of these words across websites. Github Jobs tended to have lower frequencies, but as mentioned previous this is likely explained by this website having fewer results for the job search that was conducted in order to collect the data. Muse and Adzuna show similar counts for most of the words. However, Adzuna had much higher counts for the word "strong," and The Muse had higher counts for the word "experience." It is also important to note that the word "learning" appears in the top 20 for all three websites. This is an important word because it could be applied to familiarity and skills surrounding machine learning, but it could also be in the context of a candidate's ability to learn new skills. Either way, these are both strengths that could be included in a resume or application.

Wordclouds

Although less informative than histograms, wordclouds are a unique way of visualizing text data, and can come in handy for things like presentations. I created word clouds for common words for each of the three

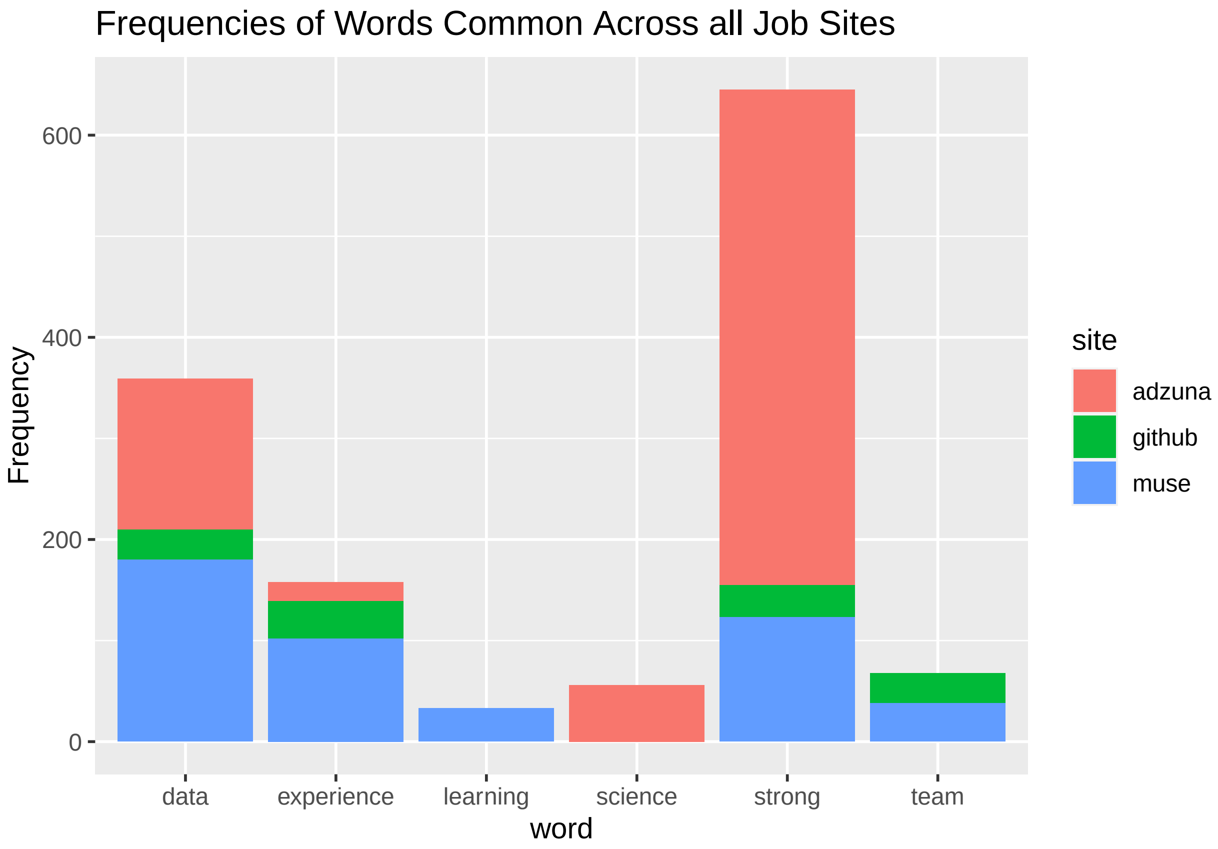

Figure 4: Histogram of words common to all three job board sites' top 20 words used in job descriptions.

websites. I won't discuss each of these in detail, but it is important to note that the same words that appeared in the combined histogram above, appear in each wordcloud, showing commonality between the two types of visualizations. These wordclouds can also be useful when writing resumes and applications as inspiration for important "buzz words" to include in order to appeal to more employers.

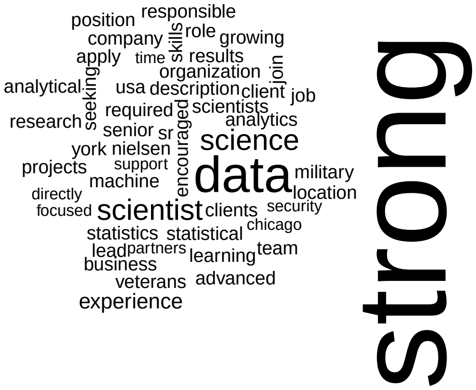

Figure 5: Wordcloud for frequently used words on Adzuma

Discussion

As mentioned previously, the purpose of this project was to explore word frequencies in job descriptions in order to determine key words that can be used when searching job postings, writing resumes, and filling out applications for jobs in data science. The three job boards that I used had several frequently used words in common. These included "strong," "data," "science," "experience," "learning," and "team."

Aside from the terms found most frequently, it is also important to make note of the number of results that each job board search returned. Adzuna returned the most results on a search for key phrase "data science," with United States as the location and a filter for full-time jobs only. This is good to know because Adzuna may be a more worthwhile site to search for job postings in the future compared with Github Jobs or The Muse.

Despite these interesting findings, there are some limitations to consider. First, there are many, many job board sites out there, and here I only explored three options. It would also be useful to explore sites like Indeed, Ziprecruiter, and LinkedIn, which seem to be very popular sites in the U.S. Second, here I have only explored the job descriptions. It may also be useful to explore other categories such as job titles or requirements.

References

Hadley Wickham (2020). tidyr: Tidy Messy Data. R package version 1.1.2. https://CRAN.R-project.org/package=tidyr

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.1. https://CRAN.R-project.org/package=kableExtra

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Ian Fellows (2018). wordcloud: Word Clouds. R package version 2.6. https://CRAN.R-project.org/package=wordcloud

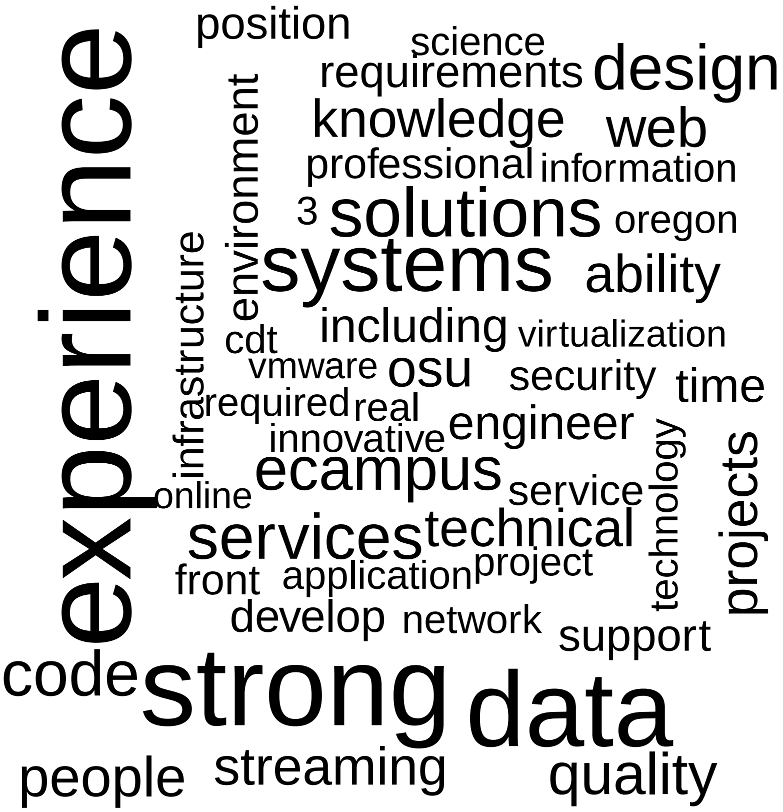

Figure 6: Wordcloud for frequently used words on Github Jobs

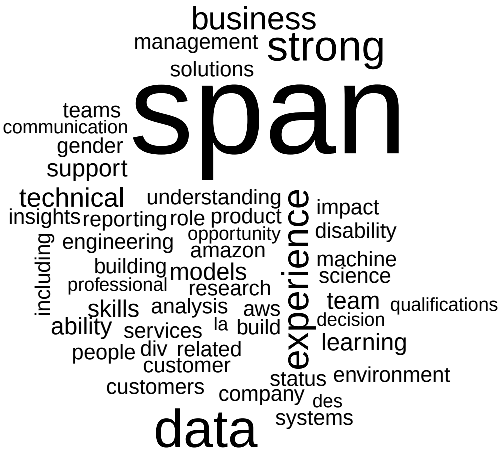

Figure 7: Wordcloud for frequently used words on The Muse

Jeroen Ooms (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] URL https://arxiv.org/abs/1403.2805.

Silge J, Robinson D (2016). “tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” *JOSS*, *1*(3). doi: 10.21105/joss.00037 (URL: https://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.