DATA ANALYTICS JOB POSTINGS REPORT - CANADA

- AMAR VIVEK

SYNOPSIS

Through recognizing the importance of a qualified workforce, skills research has become one of the focal points in economics, sociology, and education. Great effort is dedicated to analyzing labor demand and supply, and actions are taken at many levels to match one with the other. In this work we concentrate on skills needs, a dynamic variable dependent on many aspects such as geography, time, or the type of industry. Historically, skills in demand were easy to evaluate since transitions in that area were fairly slow, gradual, and easy to adjust to. In contrast, current changes are occurring rapidly and might take an unexpected turn. Thus, it is very important to stay on top of such changes and be aware of them for continuous growth. This study is also useful for individuals who are looking to make a career transition as it helps them to understand some of the key skills in demand for a particular job type.

Wherever humans are concerned, jobs and careers are a significant part of the equation. This is always a matter of how one is making a living, which is directly dependent on what competencies one possesses. However, whether employed, freelancing, or a business owner, there is no doubt that professional choices are heavily constrained by a major player—labor market demands. As is apparent, minimalizing dissimilarity between skills needs and supply is the primary concern of policy makers around the world. Significant time, careful investigation, and financial resources are required to balance labor markets. Therefore employers, professional associations, educational bodies, governments, global agencies, and many more actively cooperate to develop tools that allow for assessing and forecasting skill needs.

There is no widely accepted and available coding scheme for job skills requirements across countries comparable to the International Standard Classification of Occupations (ISCO). Occupation is merely a concept and, as such, might take different meanings and interpretations. Furthermore, as a measure of skills, occupational title alone is insufficient because it is a nominal variable that offers relatively few broad categories—usually between two and ten highly aggregated groups. While occupational title is the essential starting point, more detail is necessary to better understand skills requirements. Therefore, several bodies have proposed an approach that helps with understanding skills needs by utilizing machine learning and digital vacancy data. The method, based on data mining techniques, has the advantage of being more flexible by retrieving detailed knowledge about competency requirements in a way that bypasses rigid occupational schemes.

SCOPE

This study scrapes data in an unstructured format from a job search website. Then, the data is structured and cleaned so that analysis can progress. Lastly, through text mining key skills are identified from those postings. The data is being collected for all the Job Postings present for the job of "Data-Analyst" on www.monster.ca. A total of 132 jobs were selected for this study. These jobs were selected randomly from the available job postings.

DATA COLLECTION

The first step is to collect data. We have used www.monster.ca as the data source. The data is being obtained through web scraping using Python. There are many open source tools available in the market that ease the process of web scraping but since the layout of all the job postings was not the same, we had to write a code in Python to do the scraping.

Python code was executed using Jupyter notebook. The web scraping code along with the output csv generated has been attached along with this document. The method defined within the script has been parameterized so that it can filter the results for a particular City. The data points collected for this study are:

- Company Name
- Location
- Industry
- Job Description

The main libraries used here were:

- Pandas
- BeautifulSoup
- urllib2.

DATA PREPARATION

The first step after the data has been scraped and structured in the defined format is to look at the data within the columns and clean it. Pre-processing of data is an important step since it lays the foundation to accurate insights.

At this stage, feature engineering also plays an important role. Based on the use case being looked at, it is pertinent that based on the domain knowledge features should be engineered. Feature Engineering plays an important role to uncover insights. For the purpose of this study, only one feature was engineered which was for the length of the text of job description.

CLEANING DATA:

- The location column consisted of the city, state and in some cases even the zip code. Since the Zip code is not significant for this study it can be removed and new columns could be created namely, Ciy and State.
- For this study, the most important column is the job description column which contains the textual description of the job requirement. Data needs to be cleaned of the following nuances:
  - Punctuations
  - Numbers
  - Web urls
  - Stop words

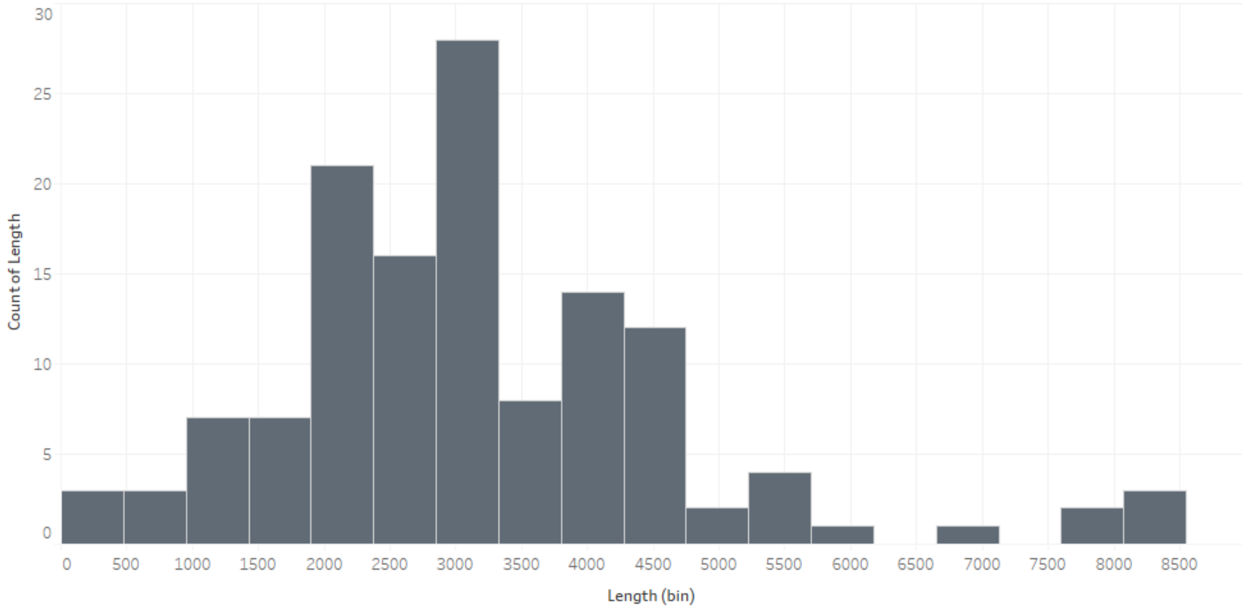
EXPLORATORY DATA ANALYSIS

After data cleaning, the data is now ready for some exploratory analysis. The visualization along with the dashboard has been prepared in Tableau and has been uploaded to Tableau Public server on the following link:

https://public.tableau.com/profile/amar.vivek#!/vizhome/DataAnalystJobPostingAnalysis/DataAnalystJobs?publish=yes

TEXT LENGTH

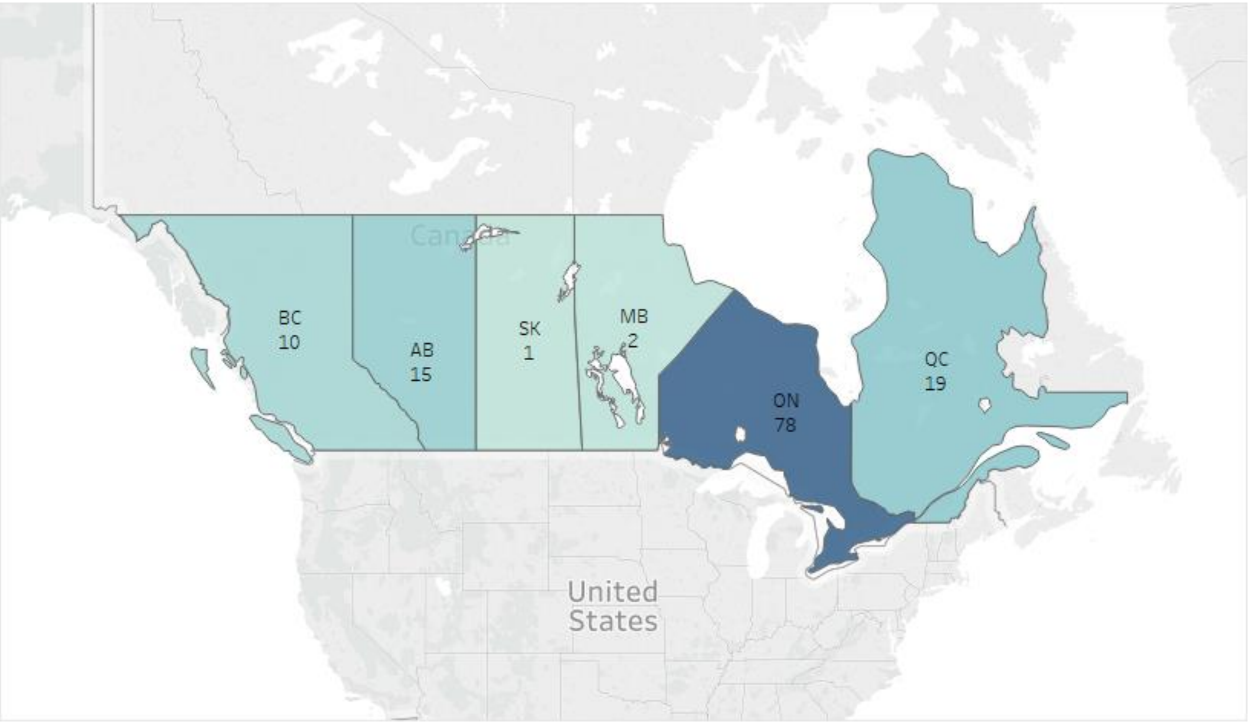
The following graph shows the bins for text lengths of the job description. It is evident that most of the job description lies in the range of 2000 – 3000 words.



There are also some outliers where the job description has close to 9000 words. On further investigation it was found that those jobs also had the Company Description within the Job Description.

Provincial Distribution

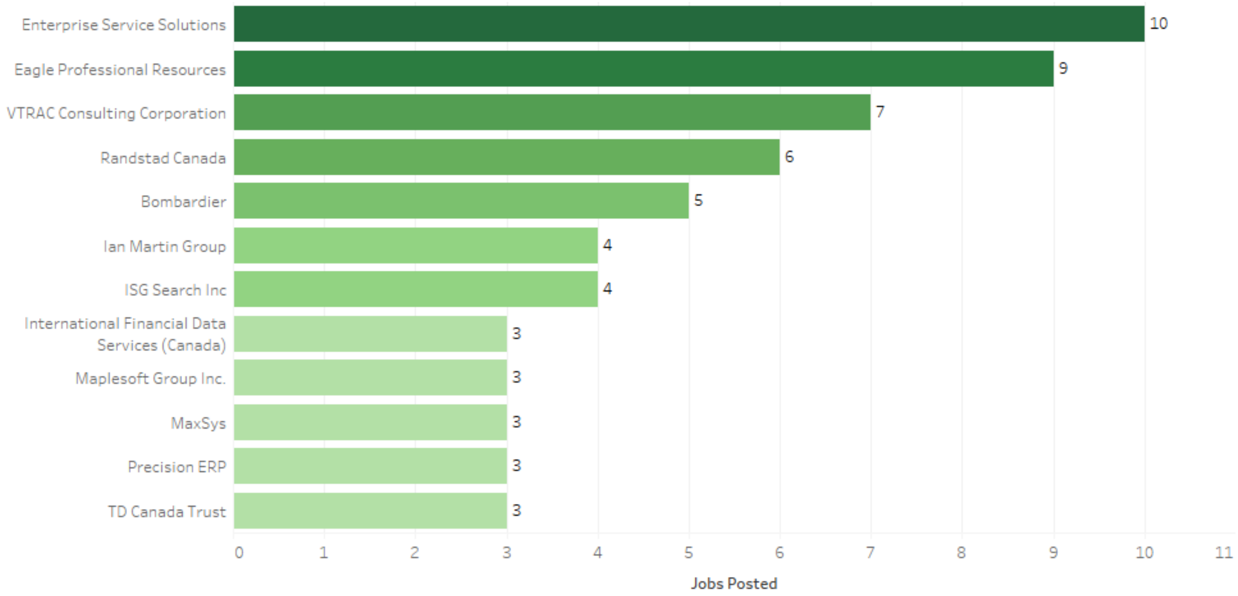
Out of the 132 job postings, the maximum were posted in Ontario followed by Quebec and Alberta.



In terms of Cities, Toronto had the highest job postings at 43 followed by Montreal and Ottawa at 10 each.

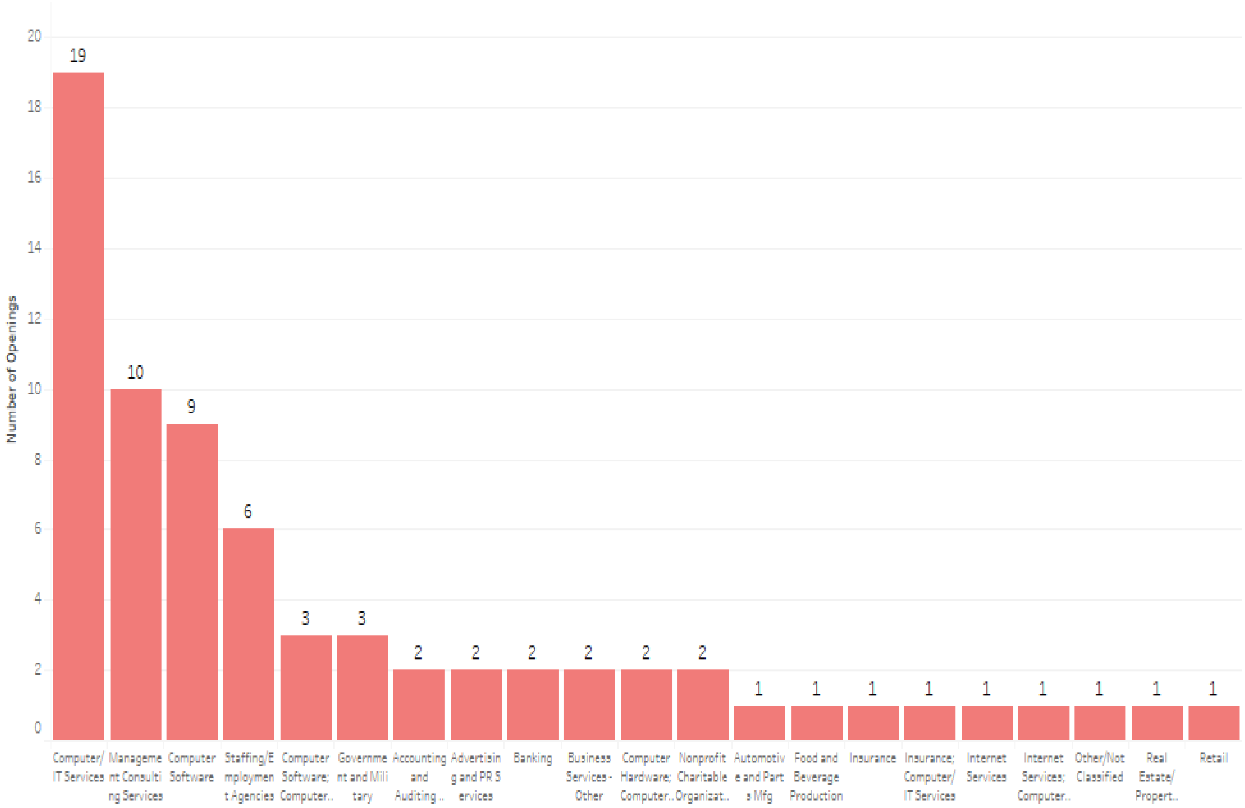
Hiring Companies

The following graph shows the companies and the number of job postings for "Data Analyst" job.



Industry Trends

On doing analysis with respect to Industries, it was found that “Computer/IT services” sector has the most number of openings followed by “Management Consulting Services” sector.



Text Mining

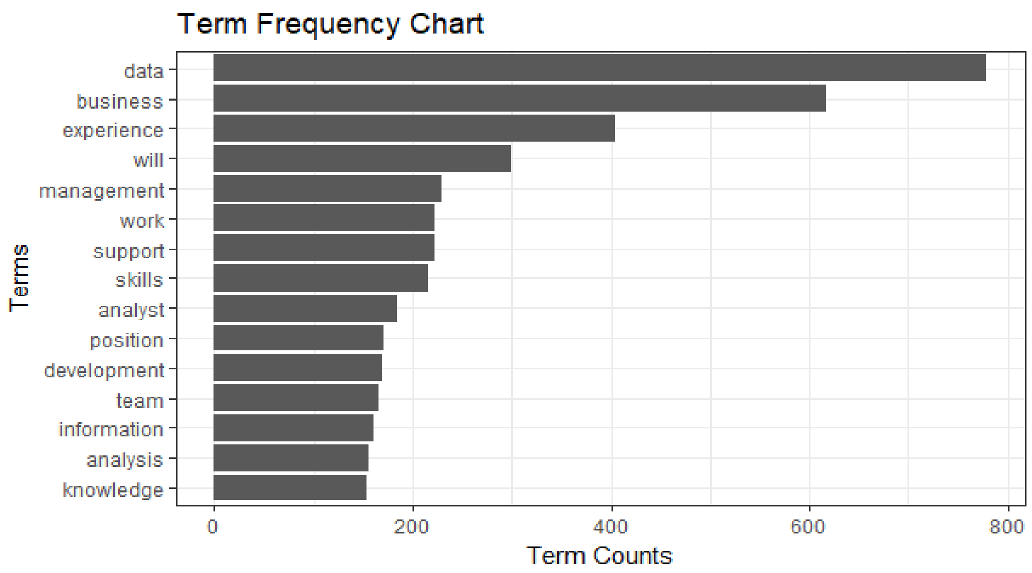
Now that we have looked at the data and are familiar with its content, we can proceed with creation of the Corpus and the term document matrix. Once that has been created, we can look at the word frequencies, word clouds and topic models. Text mining has been performed in R and utilizing packages like dplyr, ggplot2, SnowballC, tidytext etc.

Single Word Analysis

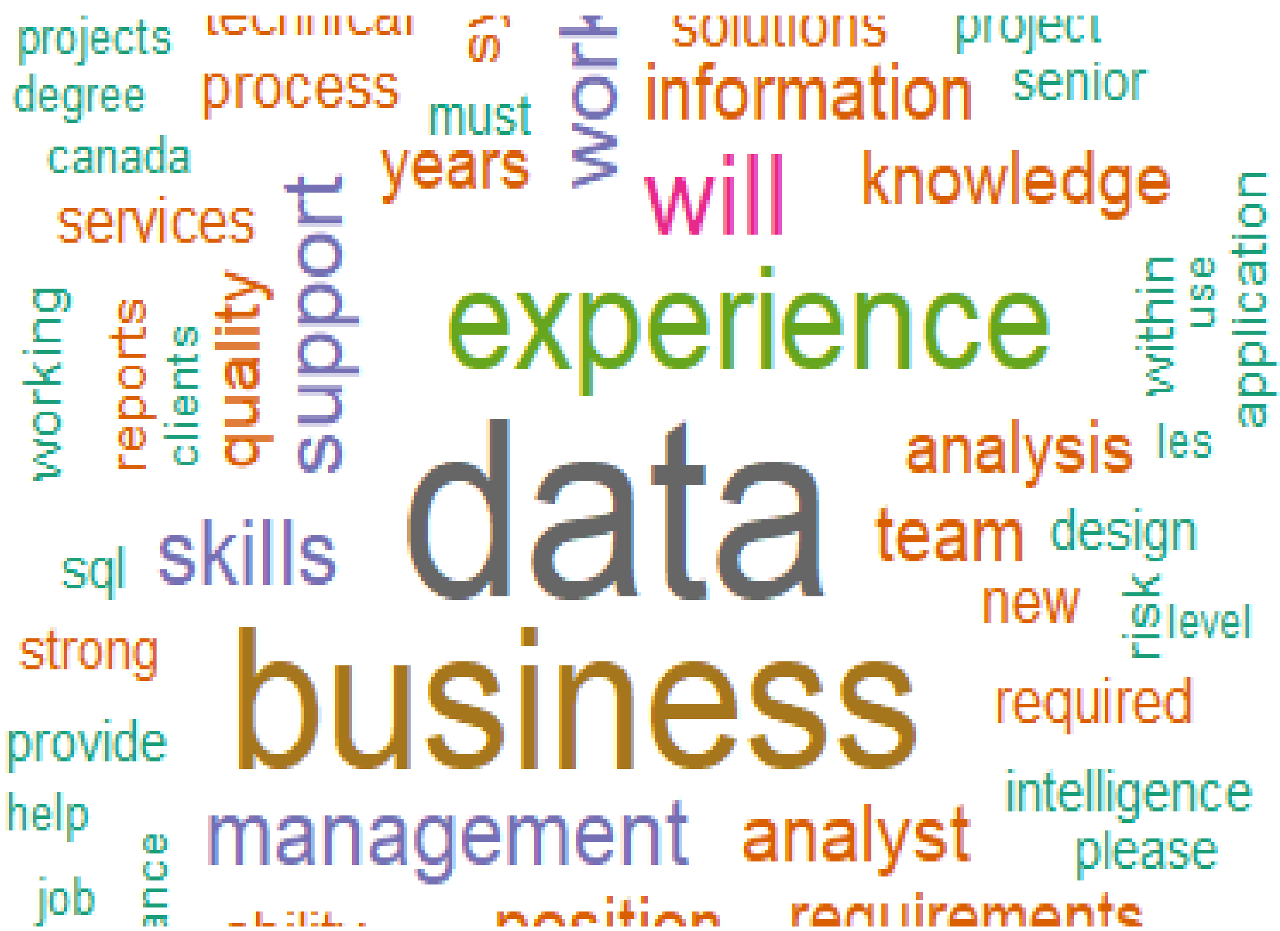
Firstly, we begin with one word analysis. The R-code for the same has been attached along with the assignment.

Term Frequency Chart

The term frequency chart shows that data is the most commonly used word. This is expected since the job posting is for Data Analyst, thus the mention of data is ought to be found. Other keywords that pop-out are business, experience, management and analysis. It seems that single word analysis is not very effective here since a lot of these words do not make much sense on their own. We would later also perform a bigram analysis on the job descriptions.


Term Frequency Chart
Terms
data
business
experience
will
management
work
support
skills
analyst
position
development
team
information
analysis
knowledge
0
200
400
600
800
Term Counts

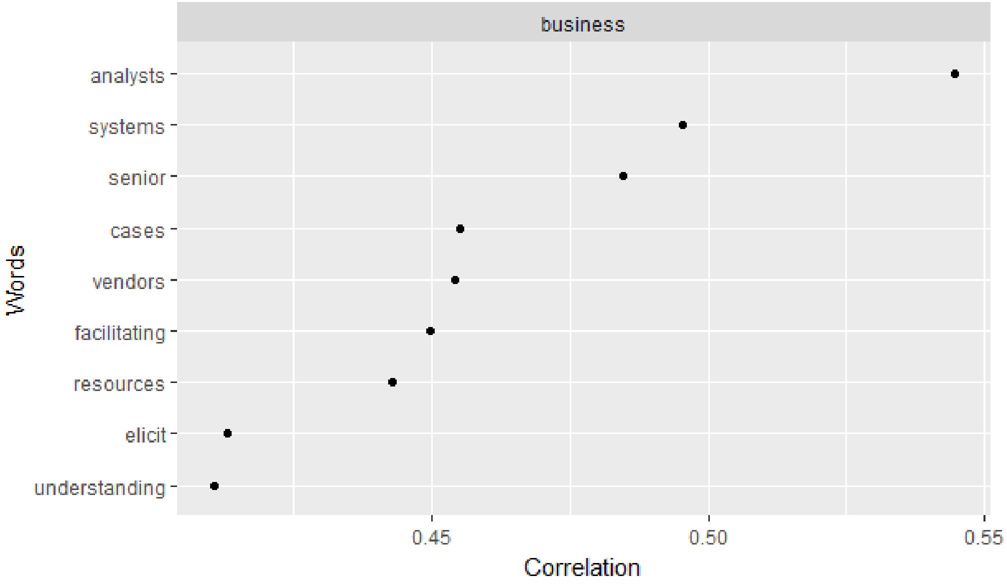
Word Cloud


projects
technical
solutions
project
degree
process
must
work
information
senior
canada
years
will
knowledge
services
support
quality
working
reports
clients
experience
within
use
application
data
analysis
les
skills
sql
team
design
new
risk
level
strong
business
required
provide
help
management
analyst
intelligence
please
job
position
requirements

Word Correlations

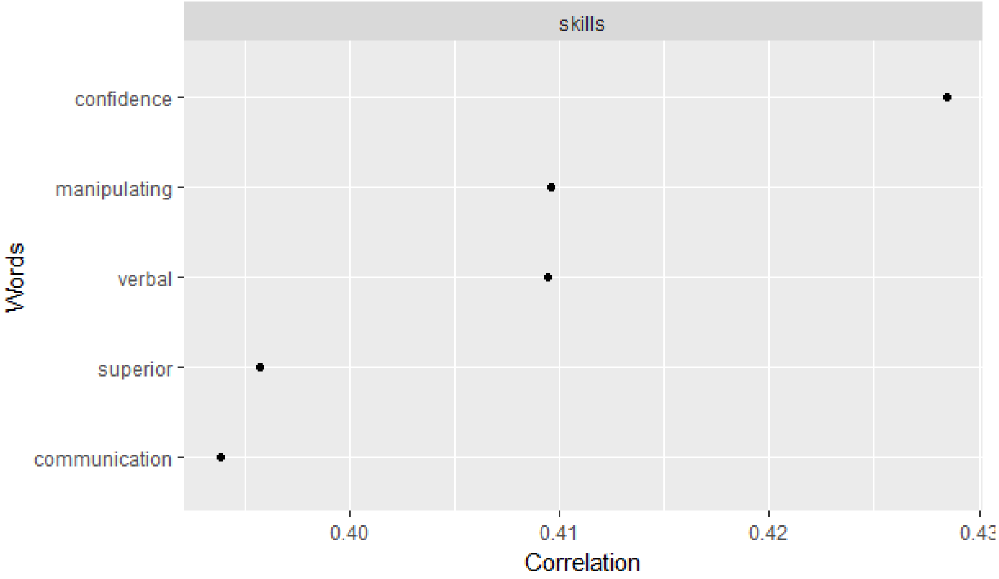
Business

The word correlation plot of business shows the correlation words with business like 'understanding', 'facilitating', 'systems' etc.



Skills

The word correlation plot of 'skills' shows some of the skills that are being asked off in the postings. Some of them are 'confidence', 'communication', 'verbal' etc.



Topic Modeling

After performing LDA (Latent Dirichlet Allocation) and setting the parameters to find 6 topics, following are the topics that are retrieved. These topics can be further utilized to assist with classification steps.

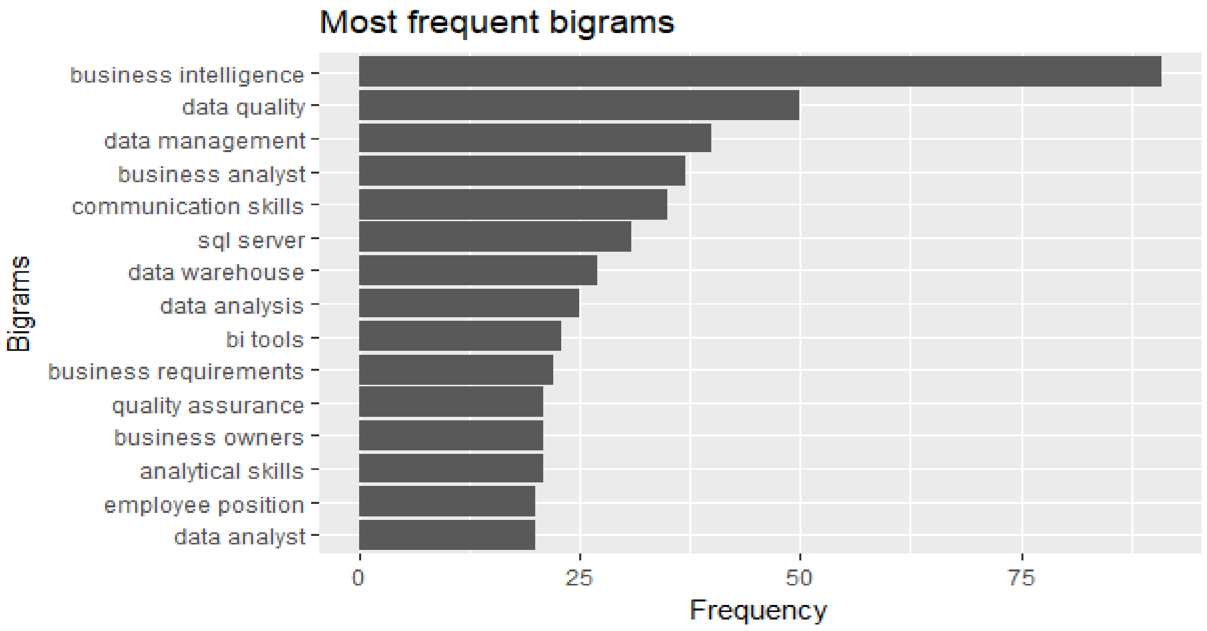
```
                                                                Topic 1
"business, data, experience, information, skills, marketing, including"
                                                                Topic 2
             "data, will, quality, experience, work, database, support"
                                                                Topic 3
    "data, business, experience, will, work, intelligence, information"
                                                                Topic 4
              "business, experience, security, data, will, years, work"
                                                                Topic 5
           "data, support, risk, management, business, new, experience"
                                                                Topic 6
      "business, experience, position, will, management, analyst, data"
```

Bigram analysis

Since most of the key skills are not single words but a combination of words, like business intelligence or sql server. Thus it becomes important that we also have a look at the analysis of combination of words that occur together in the text. Here, we would be performing analysis of a combination of 2 words, hence the term bi-gram.

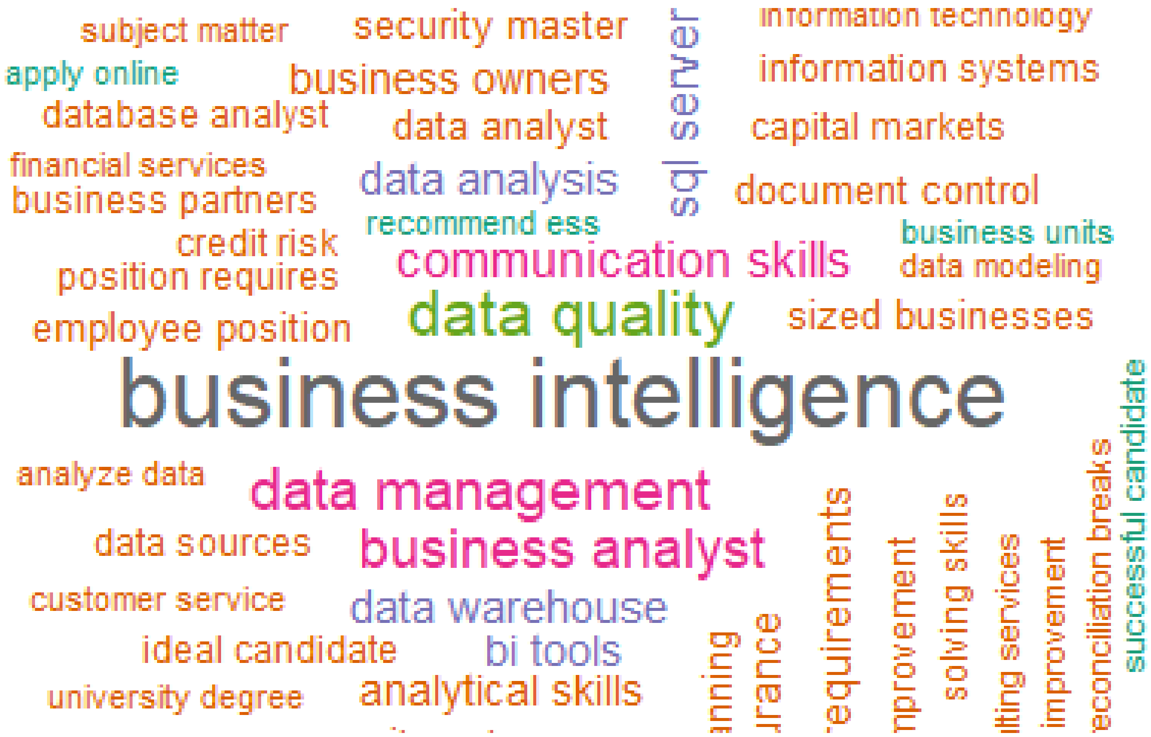
Bigram Frequency Chart

Below chart gives us a better picture of the key skills that are being requested in the job postings. Things like data quality, data management are some of the most sought after skills. Also, communication skills is amongst the top 5 skills being asked for.



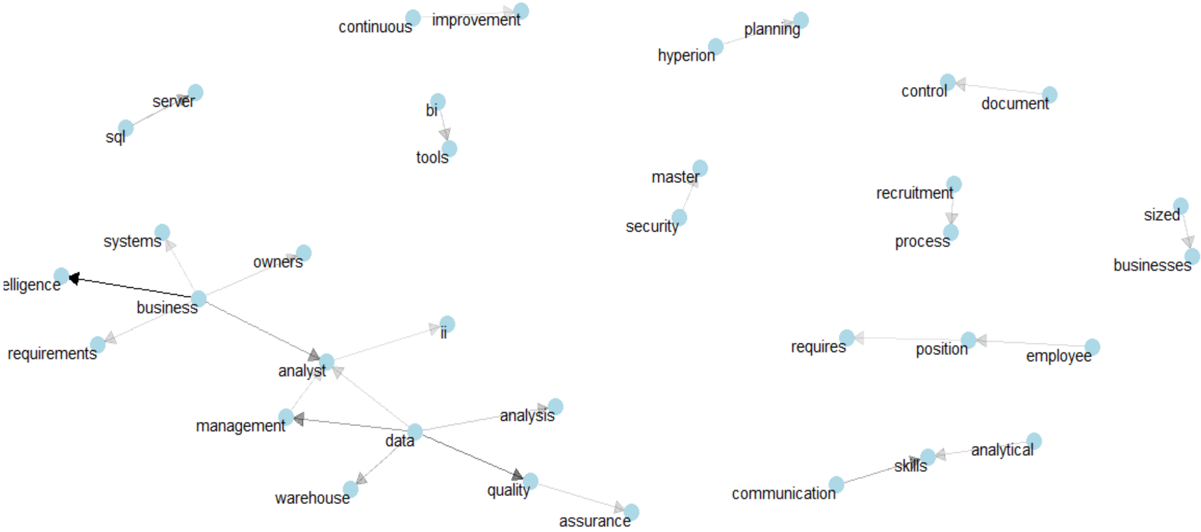
Word Cloud

Another way to look at the bigram frequency chart is the word cloud which shows some additional skills needed.



Bigram Relationship Diagram

We may be interested in visualizing all of the relationships among words simultaneously, rather than just the top few at a time. We can arrange the words into a network, or "graph." Here we'll be referring to a "graph" not in the sense of visualization, but as a combination of connected nodes.



Approach for Skills Extraction

The algorithm used for extracting key skills is the bigram analysis. N-gram models have been known to be used for text summarization. Another use for N-grams is for developing features for supervised Machine Learning models such as SVMs, Naive Bayes, etc. The idea is to use tokens such as bigrams in the feature space instead of just unigrams.

Also, for topic modeling LDA (Latent Dirichlet Allocation) has been used which is the most commonly used model for topic modeling. For the purpose of text classification, there are models available such as KNN, Naïve-Bayes etc.

Conclusion

Throughout this study we have seen various methods of analyzing unstructured text and how to derive meaning out of it. From this study it is pretty evident that there are certain sets of skills that are being asked for in the majority of the jobs. Some of them to be mentioned are “Business Intelligence”, “Data Management”, “Business Analysis” and “Communication Skills”. Having an idea of the in-demand skills for a particular job would help job aspirants to understand where they might need to focus more efforts on.