INDEED JOB SCRAPING AND ANALYSIS

Colin Green

PROBLEM STATEMENT

Job hunting is tedious and time consuming

Job descriptions are ads meant to attract as many applicants as possible

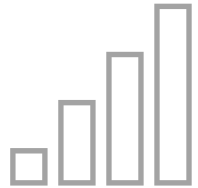
Search algorithms struggle to find key differences between similar jobs

I am looking for entry level jobs focused on data science, machine learning, and/or data analytics

Many of the job postings I see are Business Analyst positions with a focus on sales, management, or marketing

OVERALL GOAL

Streamline my job search while reducing the number of unrelated posts that I must sift through



Rank the job postings based on their similarity to jobs that I have applied to in the past

EXPECTATION

Jobs I will apply to

Python

R

SQL

Machine Learning

Data Science

Math/Physics

Entry-Level/Intern

Jobs I won't apply to

Manager

Director

Sales

Web

PhD

Machining

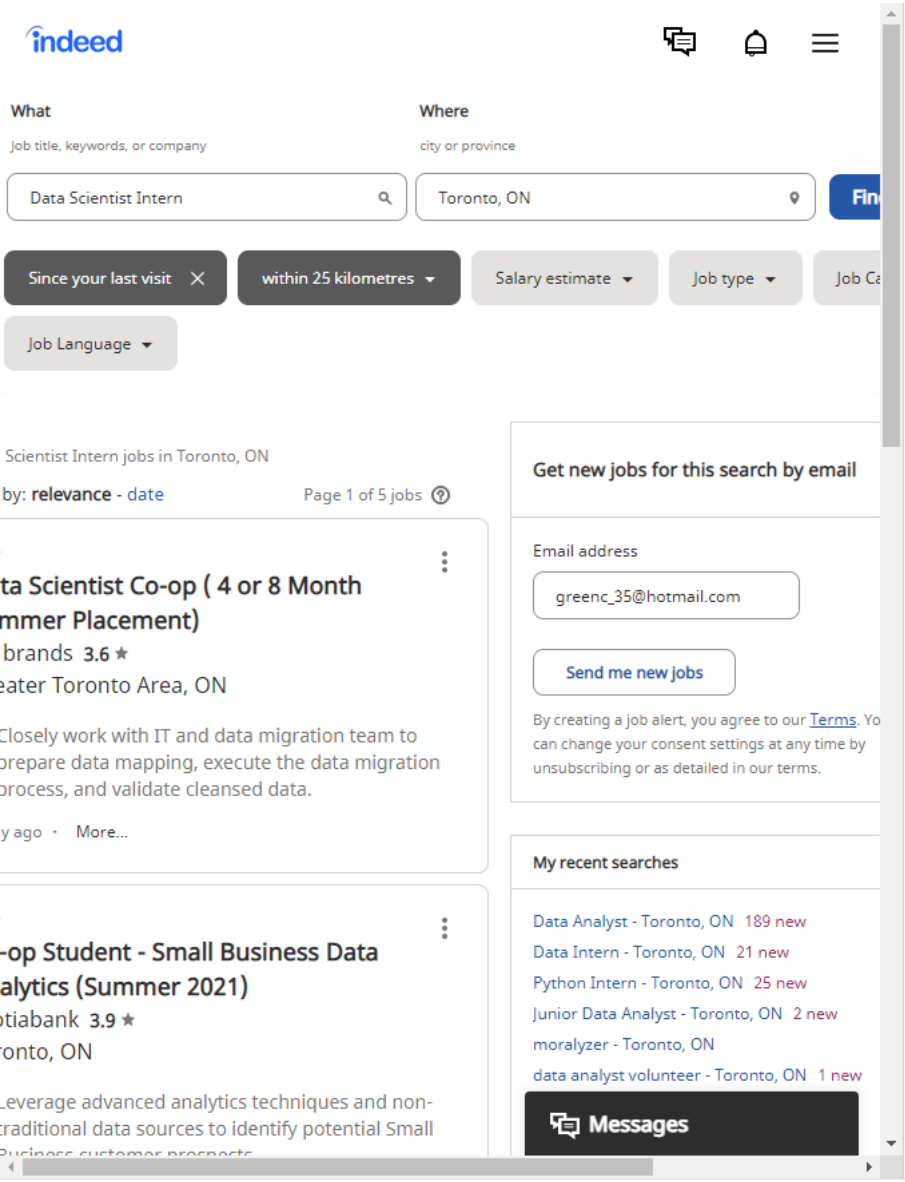
Full Stack

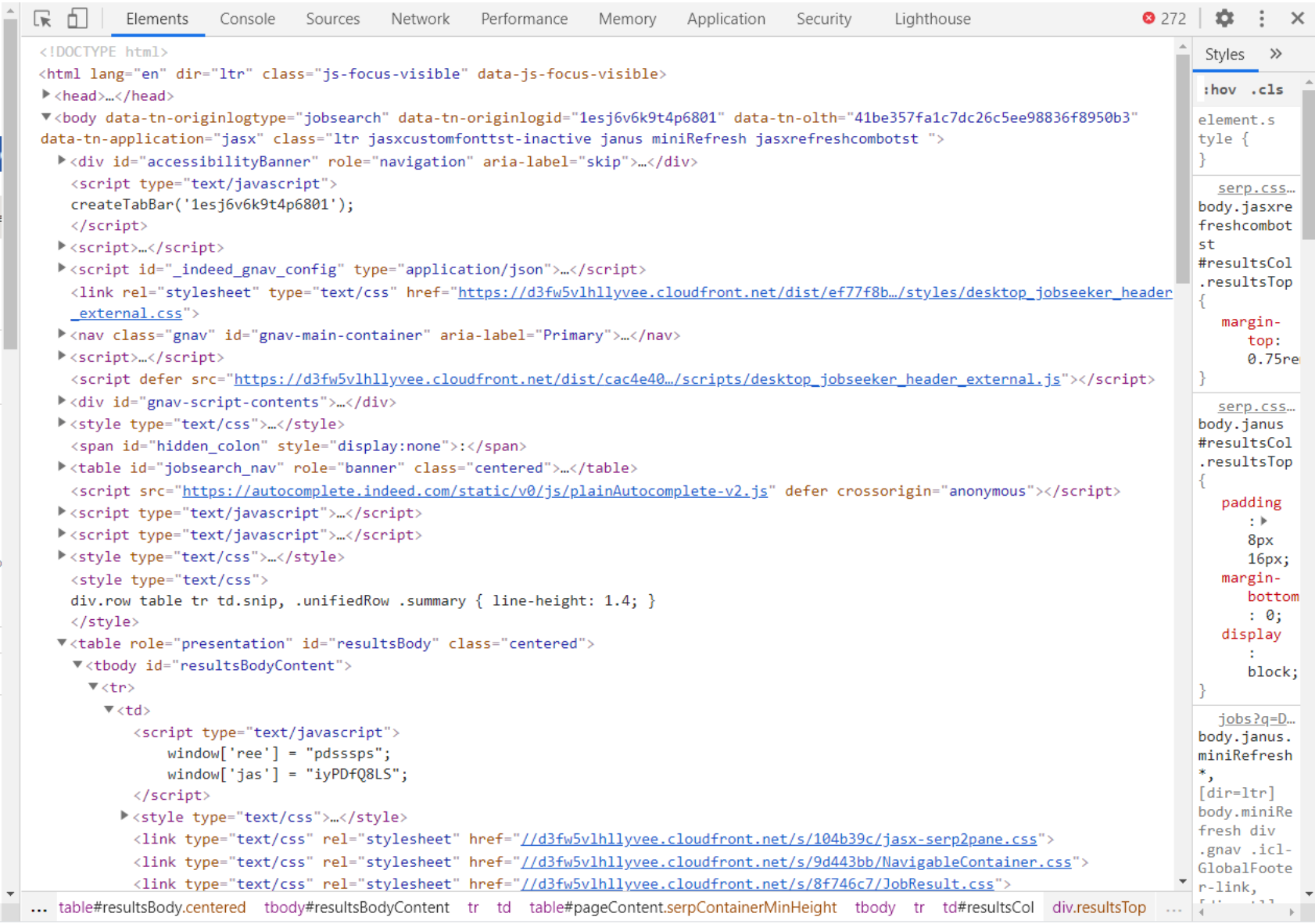
WEB SCRAPING – BEAUTIFUL SOUP

- Begin by creating URLS
  - Input includes search terms, city, province, location radius, and date range
- Outputs a list of URLs (as strings) based on the format that indeed accepts
- Next I used the requests package to retrieve the information from the web for each URL
- Using the Beautiful Soup package, I parsed each page to turn it into something readable

BASIC FUNCTION SYNTAX

```
def extract_job_title_from_result(soups):
    jobs = []
    for soup in soups:
        for div in soup.find_all(name='div', attrs={'class': 'row'}):
            for a in div.find_all(name='a', attrs={'data-tn-element': 'jobTitle'}):
                jobs.append(a['title'])
    return(jobs)
```


indeed
What
Job title, keywords, or company
Where
city or province
Data Scientist Intern
Toronto, ON
Since your last visit
within 25 kilometres
Salary estimate
Job type
Job Language
Scientist Intern jobs in Toronto, ON
by: relevance - date
Page 1 of 5 jobs
ta Scientist Co-op (4 or 8 Month
mmer Placement)
brands 3.6
eater Toronto Area, ON
Closely work with IT and data migration team to
prepare data mapping, execute the data migration
process, and validate cleansed data.
y ago · More...
-op Student - Small Business Data
alytics (Summer 2021)
otiabank 3.9
ronto, ON
Leverage advanced analytics techniques and non-
traditional data sources to identify potential Small
Get new jobs for this search by email
Email address
greenc_35@hotmail.com
Send me new jobs
By creating a job alert, you agree to our Terms. Yo
can change your consent settings at any time by
unsubscribing or as detailed in our terms.
My recent searches
Data Analyst - Toronto, ON 189 new
Data Intern - Toronto, ON 21 new
Python Intern - Toronto, ON 25 new
Junior Data Analyst - Toronto, ON 2 new
moralyzer - Toronto, ON
data analyst volunteer - Toronto, ON 1 new
Messages


Elements
Console
Sources
Network
Performance
Memory
Application
Security
Lighthouse
272
Styles
:hov .cls
<!DOCTYPE html>
<html lang="en" dir="ltr" class="js-focus-visible" data-js-focus-visible>
<head>…</head>
<body data-tn-originlogtype="jobsearch" data-tn-originlogid="1esj6v6k9t4p6801" data-tn-olth="41be357fa1c7dc26c5ee98836f8950b3" data-tn-application="jasx" class="ltr jasxcustomfonttst-inactive janus miniRefresh jasxrefreshcombotst ">
…
<script type="text/javascript">
createTabBar('1esj6v6k9t4p6801');
</script>
<script>…</script>
<script id="_indeed_gnav_config" type="application/json">…</script>
<link rel="stylesheet" type="text/css" href="https://d3fw5vlhllyvee.cloudfront.net/dist/ef77f8b…/styles/desktop_jobseeker_header_external.css">
<nav class="gnav" id="gnav-main-container" aria-label="Primary">…</nav>
<script>…</script>
<script defer src="https://d3fw5vlhllyvee.cloudfront.net/dist/cac4e40…/scripts/desktop_jobseeker_header_external.js"></script>
…
<style type="text/css">…</style>
:
<table id="jobsearch_nav" role="banner" class="centered">…</table>
<script src="https://autocomplete.indeed.com/static/v0/js/plainAutocomplete-v2.js" defer crossorigin="anonymous"></script>
<script type="text/javascript">…</script>
<script type="text/javascript">…</script>
<style type="text/css">…</style>
<style type="text/css">
div.row table tr td.snip, .unifiedRow .summary { line-height: 1.4; }
</style>
<table role="presentation" id="resultsBody" class="centered">
<tbody id="resultsBodyContent">
<tr>
<td>
<script type="text/javascript">
window['ree'] = "pdsssps";
window['jas'] = "iyPDfQ8LS";
</script>
<style type="text/css">…</style>
<link type="text/css" rel="stylesheet" href="//d3fw5vlhllyvee.cloudfront.net/s/104b39c/jasx-serp2pane.css">
<link type="text/css" rel="stylesheet" href="//d3fw5vlhllyvee.cloudfront.net/s/9d443bb/NavigableContainer.css">
<link type="text/css" rel="stylesheet" href="//d3fw5vlhllyvee.cloudfront.net/s/8f746c7/JobResult.css">
table#resultsBody.centered tbody#resultsBodyContent tr td table#pageContent.serpContainerMinHeight tbody tr td#resultsCol div.resultsTop
serp.css…
body.jasxrefreshcombotst #resultsCol .resultsTop {
margin-top: 0.75re
}
serp.css…
body.janus #resultsCol .resultsTop {
padding: 8px 16px;
margin-bottom: 0;
display: block;
}
jobs?q=D…
body.janus.miniRefresh *, [dir=ltr] body.miniRefresh div .gnav .icl-GlobalFooter-link,

EXTRACTING INFORMATION

- From this information I extracted:
  - Job Titles
  - Company Names
  - Locations
  - HREF Link
- Job descriptions (the goal) are not saved on the main search results page
- I can use the first function and the HREF links to get the full information for each job listed

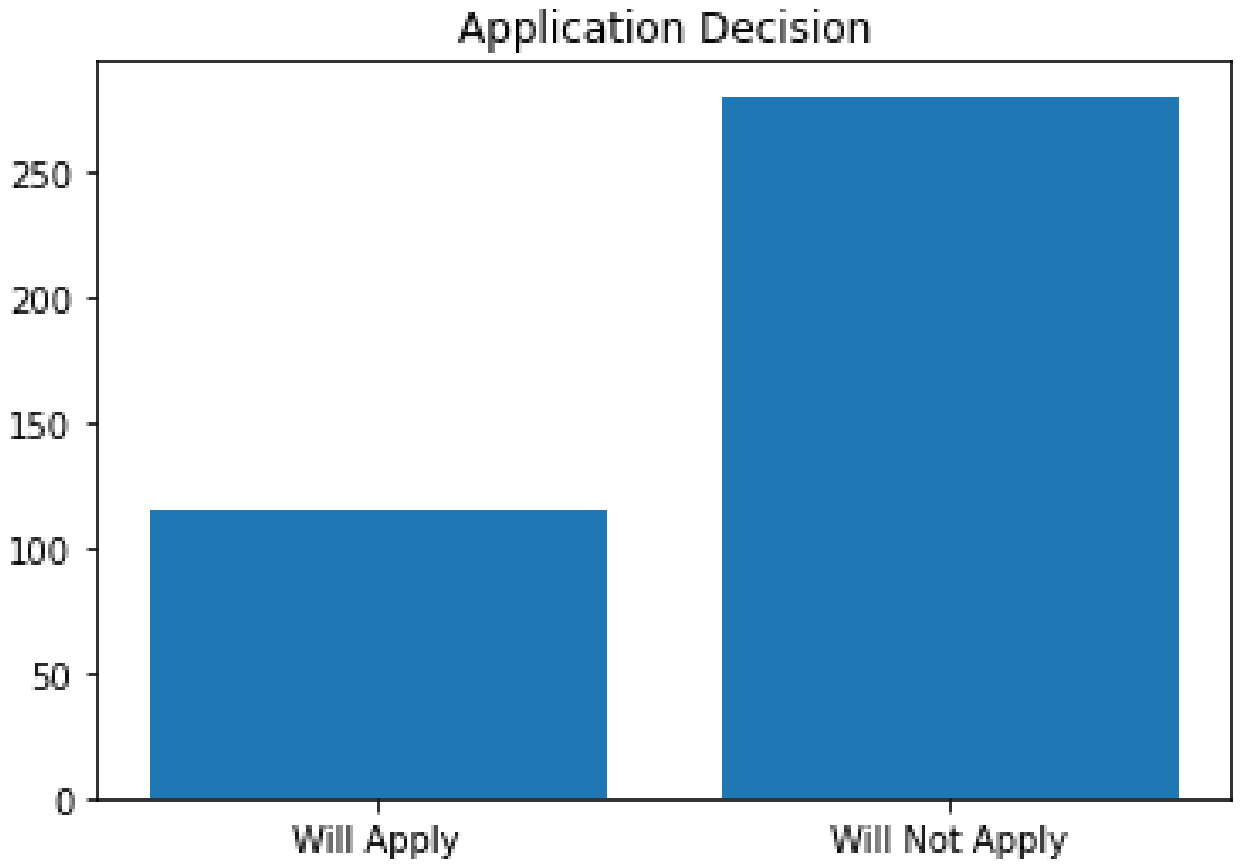
REMOVE DUPLICATES

- Many jobs will show up in more than one search
- After retrieving the job description, I pass it through a function that creates a dataframe using all of the scraped information
- I removed duplicates based on job title and company name
- Finally I saved the data as a csv file

DATA LABELING

- I went through and labeled each job posting (-400 postings)
- I labeled them as “Yes” I would apply for the job or “No” I would not apply
- This was time consuming but I did it slowly over the last month
- In the future I would consider trying out some form of clustering in order to avoid this
- I had to skim through HTML and may not have been as consistent as I should have been

DATA LABELS




```
Will Apply:  115 ( 29.11 %)
Will Not Apply:  280 ( 70.89 %)
```

PREPROCESSING

	job_title	company	location	href	description	apply
0	Data Analyst Co-op (Spring term)	Ridley College (Canada)	St. Catharines, ON	https://ca.indeed.com/rc/clk?jk=4aafa08c370b87...	[[Position Title: Data Analyst Co-o...	Yes
1	Data Analytics Associate Summer Intern (MBA)	Johnson & Johnson Family of Companies	Toronto, ON	https://ca.indeed.com/rc/clk?jk=1d1d136f3b5263...	[[Data Analytics Associate Intern –...	No
2	Data Analyst, Summer 2021 Student Opportunities	RBC	Toronto, ON	https://ca.indeed.com/rc/clk?jk=5bc75ed7e05b22...	[[What is the opportunity? ...	Yes
3	Data Scientist, Summer Student 2021 Opportunities	RBC	Toronto, ON	https://ca.indeed.com/rc/clk?jk=1bdf42b3d5b3e4...	[[What is the opportunity? ...	Yes
4	Business/Operations Analyst, Summer 2021 Stude...	RBC	Toronto, ON	https://ca.indeed.com/rc/clk?jk=76d9a17c168e02...	[[What is the opportunity? <...	Yes

PREPROCESSING

- HTML is really messy
- I created a cleanHTML function that used a regex to remove all of the tags ( )
- I also had to remove special characters (“, ”, â€™, \n, etc...)
- Next I ran it through a tokenizer that would remove punctuation, braces, excess whitespace and set everything to lowercase
- Finally I removed all of the stopwords and numbers

A NOTE ON NUMBERS

Numbers could be really useful (ie, 5 years of experience)

But there were enough numbers caused by the messiness of the HTML that the problems outweighed the benefits

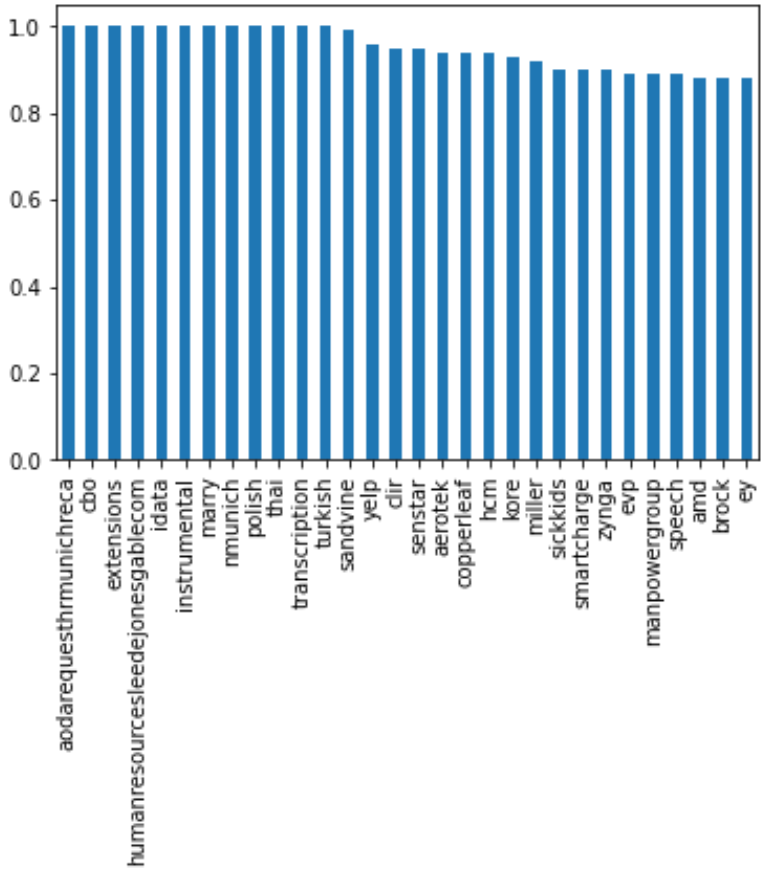
DATASETS

I compiled the data into 3 different formats

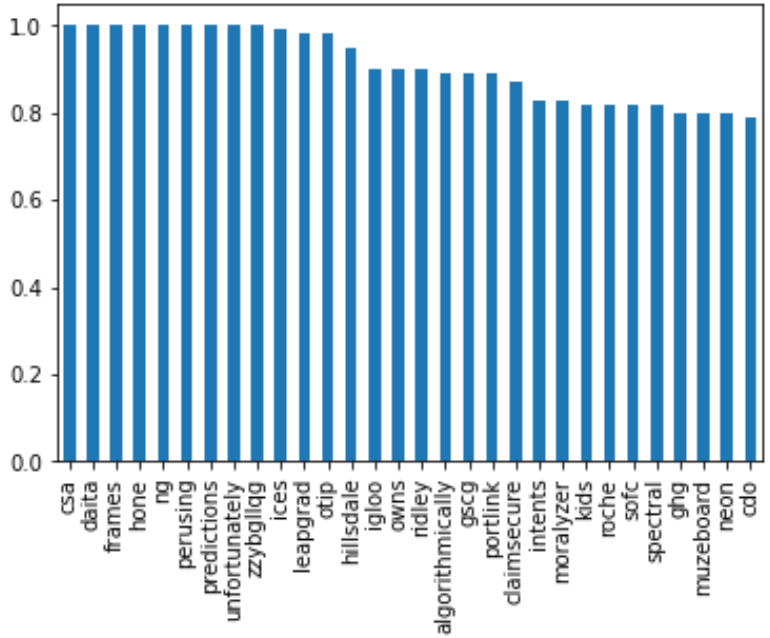
- TFIDF
  - 3870 Features
- Bag of Words
  - 9780 Features
- nGrams (n=2)
  - 87625 Features

DATA EXPLORATION – TF-IDF

Will Not Apply

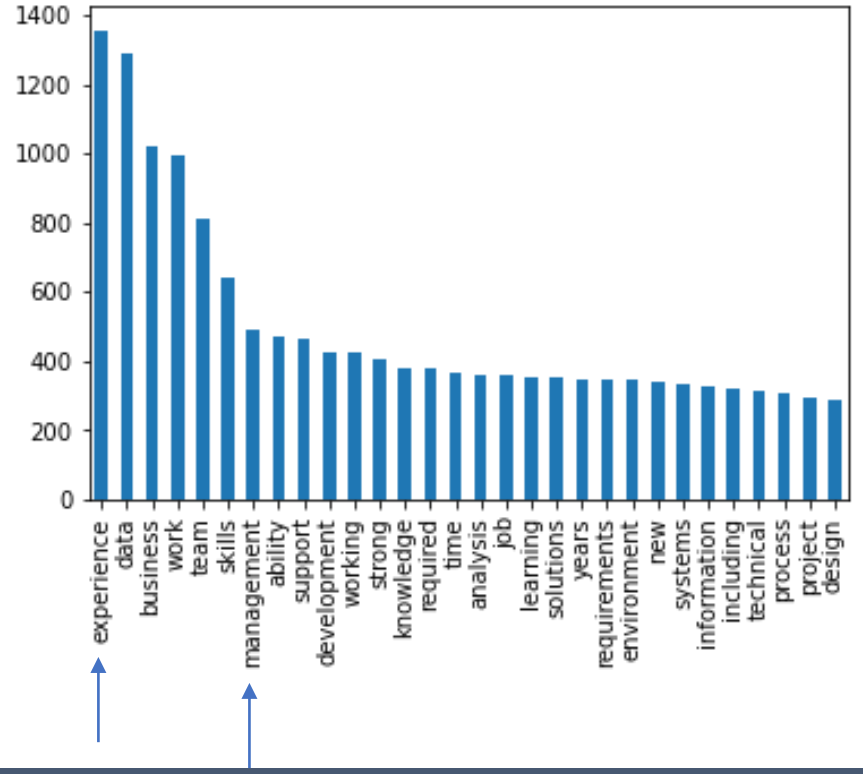


Will Apply

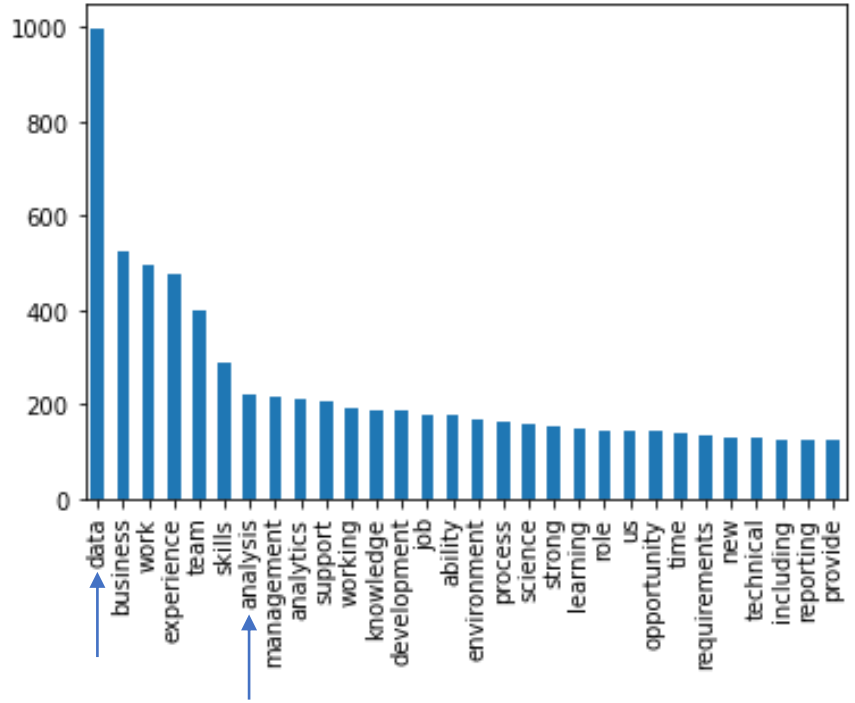


DATA EXPLORATION - BOW

Will Not Apply

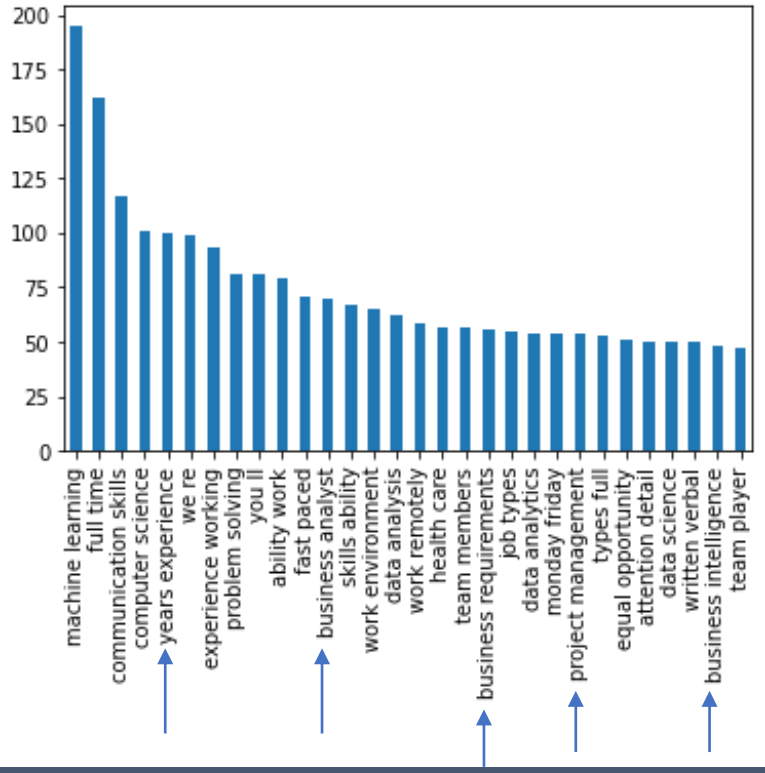


Will Apply

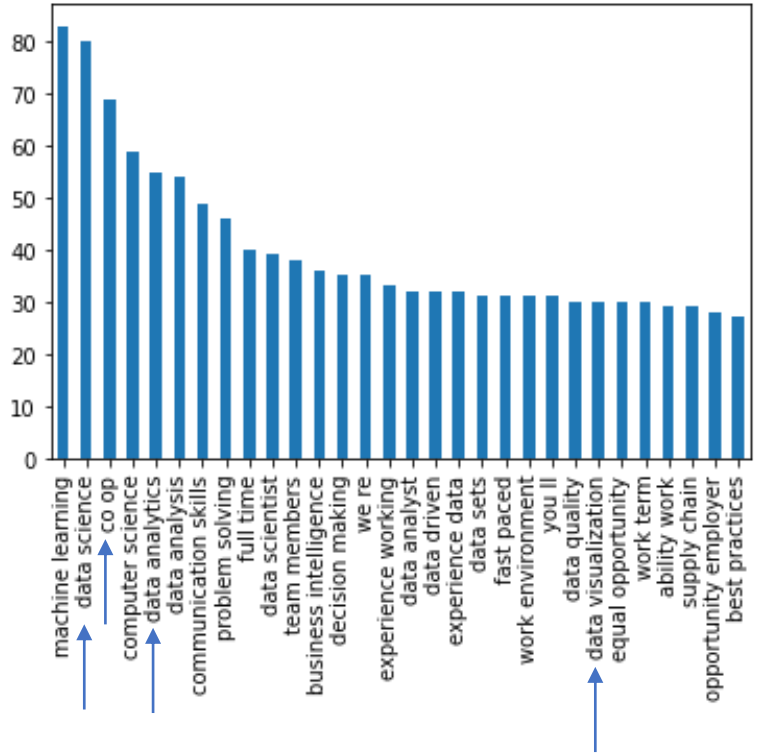


DATA EXPLORATION - NGRAMS

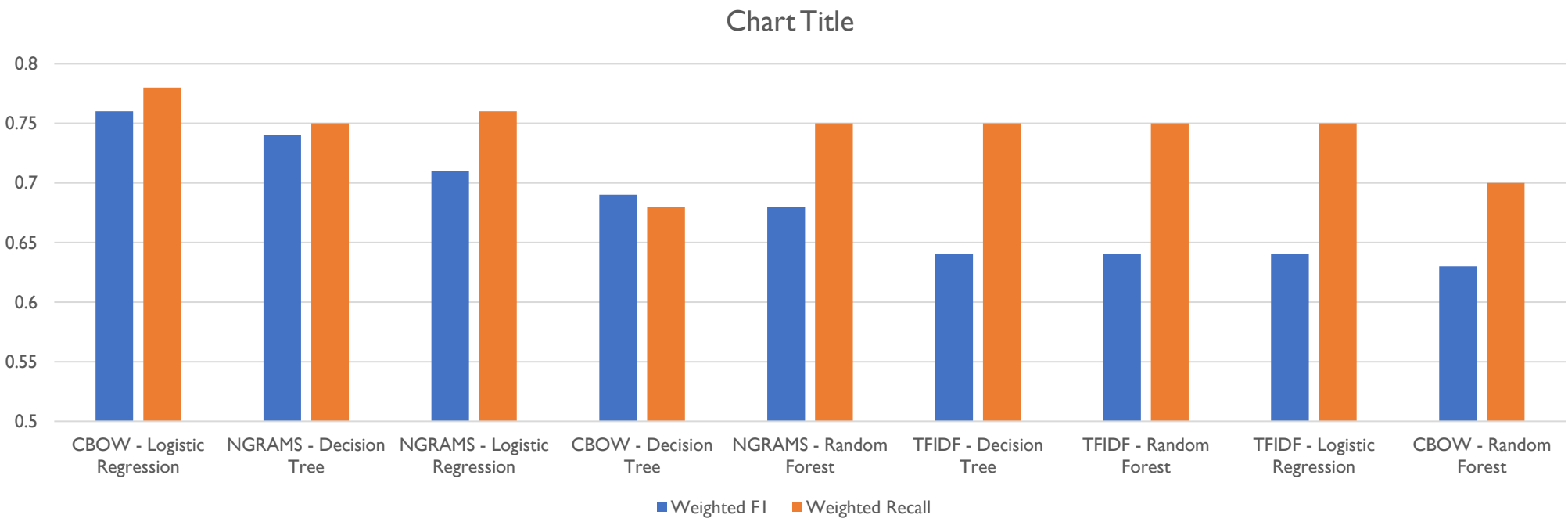
Will Not Apply



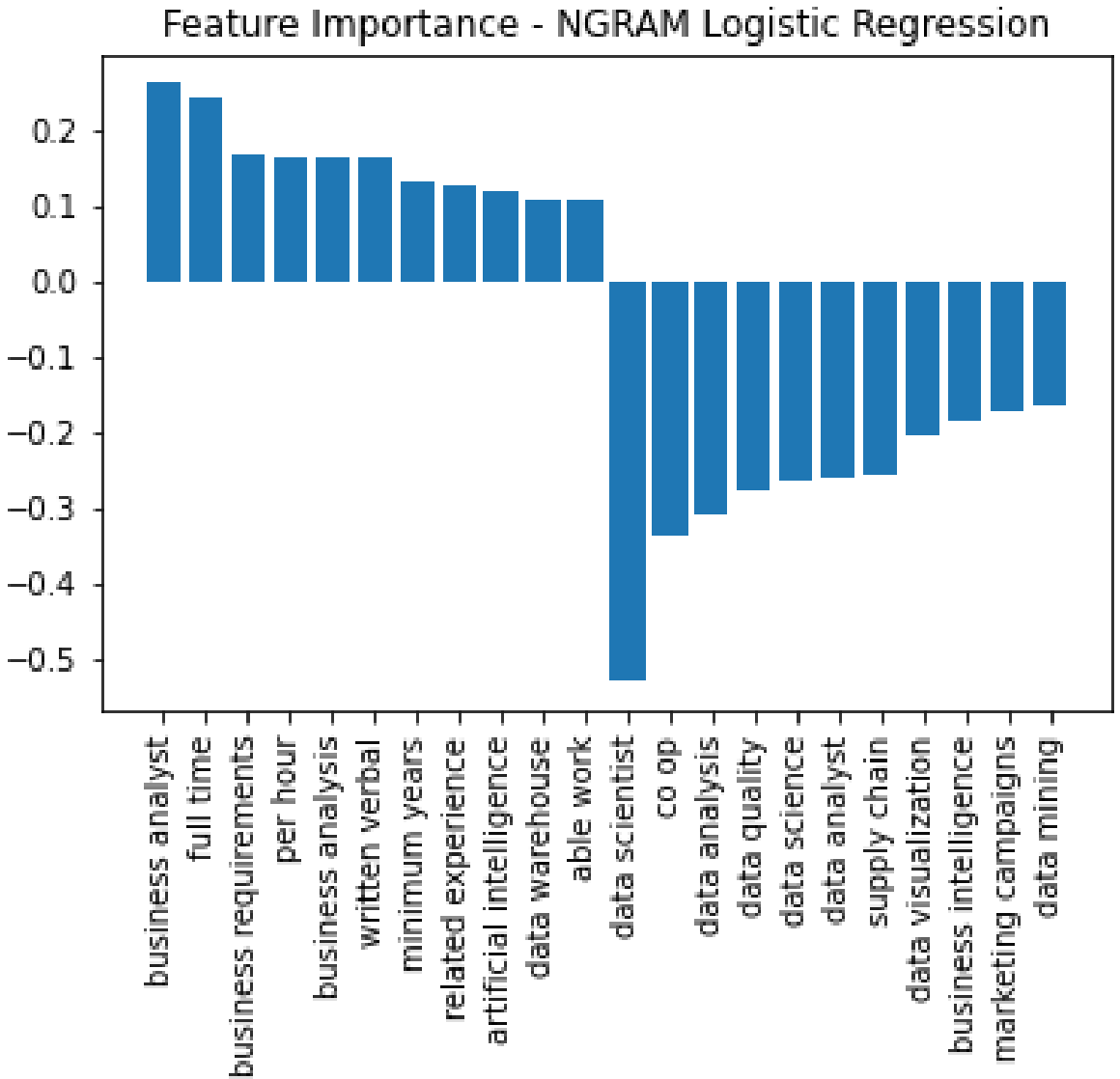
Will Apply



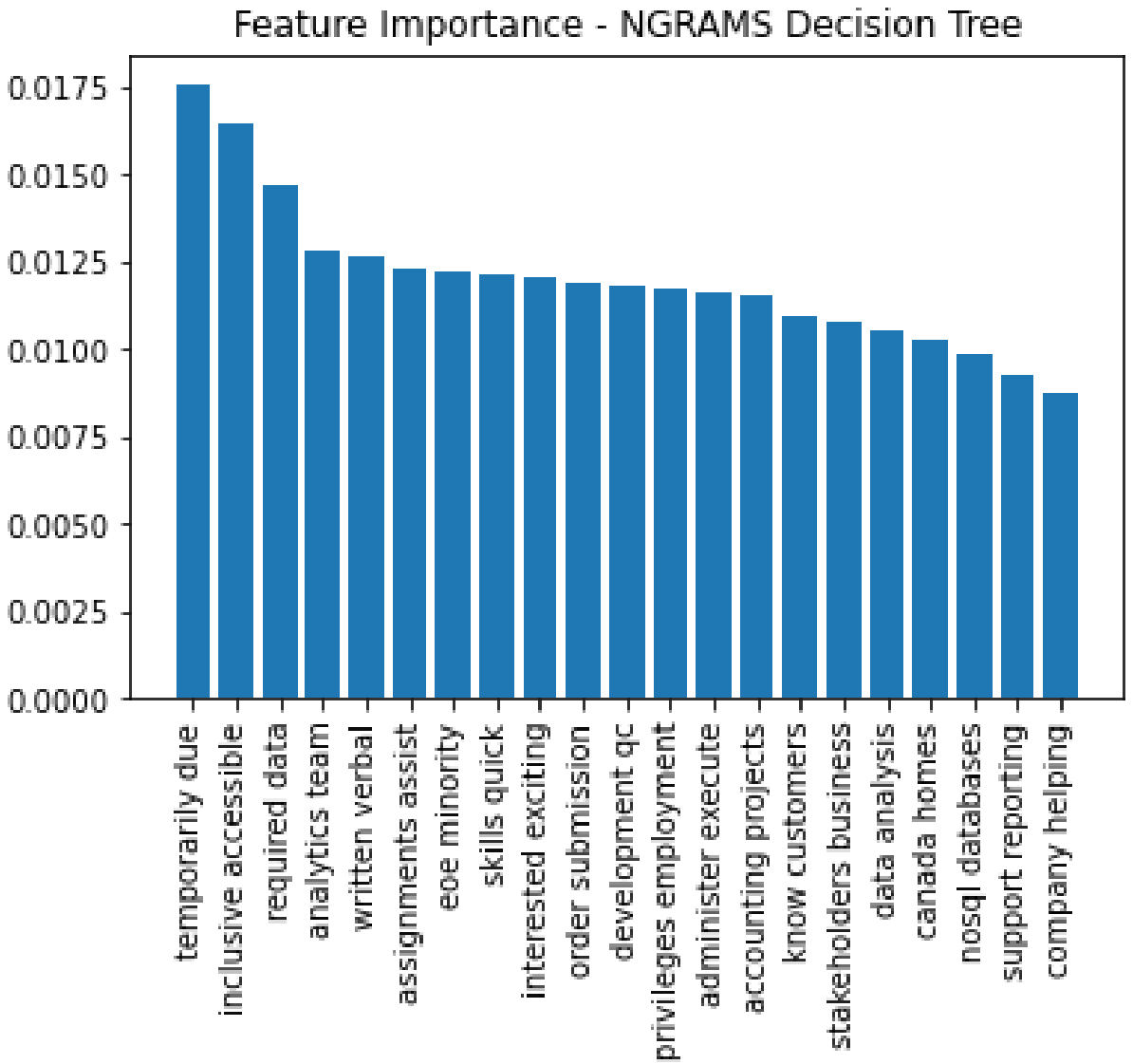
PRELIMINARY MODELING



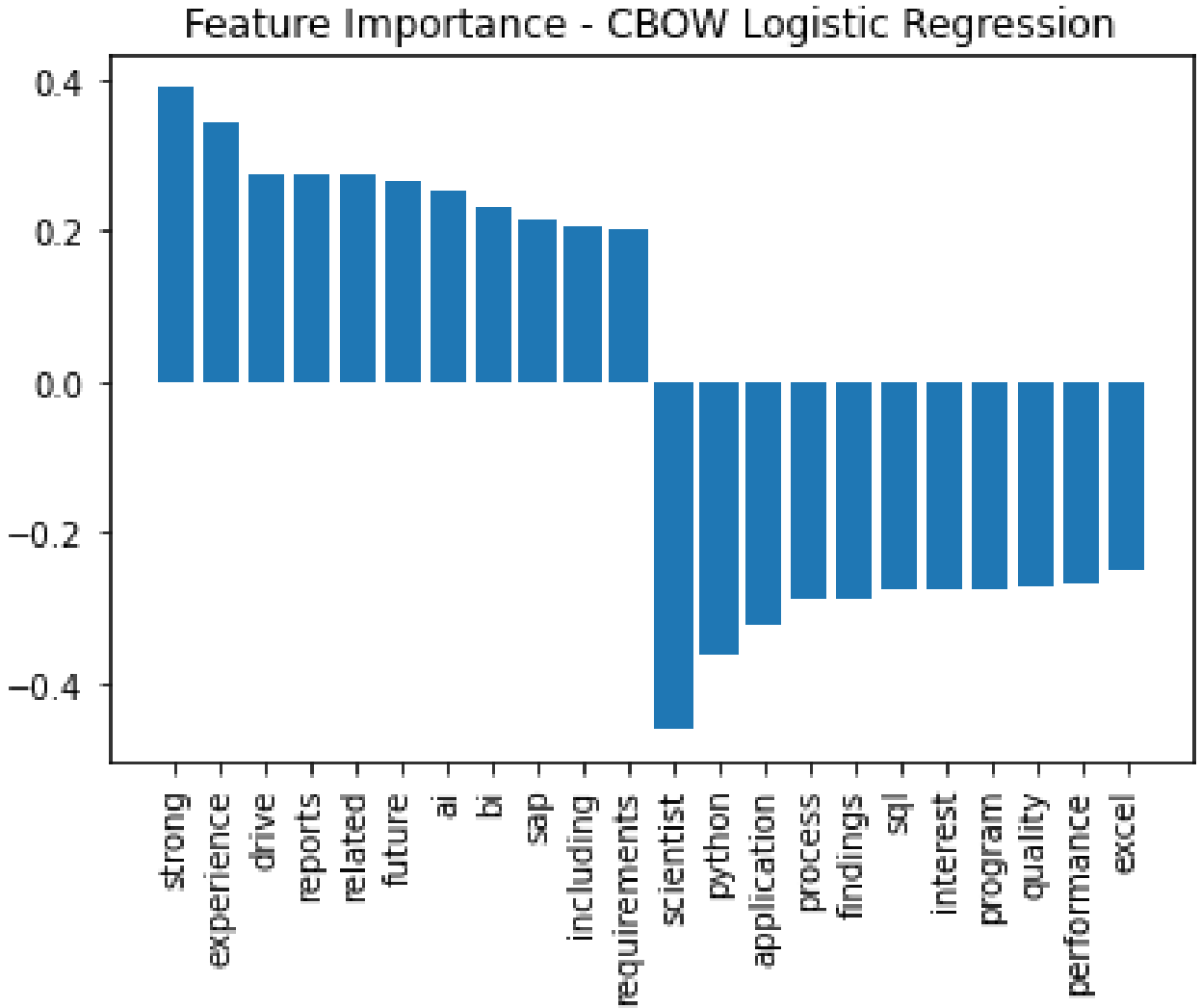
FEATURE IMPORTANCE



FEATURE IMPORTANCE



FEATURE IMPORTANCE



SUMMARY

- Successfully scraped the information needed from indeed
- Parsed through and cleaned the data
- CBOW and nGrams outperformed TF-IDF in preliminary modeling
- Successfully identified features that matched my expectations
- Limits:
  - Only able to search on Indeed
  - 400 Jobs from the last 7 weeks
  - 30%/70% label split
  - Labels have very similar data
  - Job Boards have additional data points that I don't have access to