Google Job Listing Data Mining

DTSA5504 Final Report

Zehu Cai
Master of Science, in Data Science
University of Colorado, Boulder

Abstract

Google is one of my dream companies to get into, many people are getting into Google every year and there are many blogs posted that teach people how to get into Google too. But most of the posts didn't post about how they come to conclusion and what evidence supports their opinion. In the project, Web scraping will be used to get data from Google job listings, and perform data mining and analysis on it, at the end, data will be deployed as an online web tool to build a data visualization application.

Introduction

Google is one of the largest tech companies in the world, it's also one of the most reputable employers [1]. In 2019, Google received over 3.3 million applications and hire 20,000 employees [2, 3]. So, each applicant only has a 0.61% chance of getting hired. The maximize the probability of being hired by Google, there are many things an applicant can do. There's a list provided by google careers [4]. Self-reflection, think about previous experience and summarize the most reward elements to offer for the next opportunity. Job search, do research on the company and the role you're interested in. Resume, create and update your resume. Interviews include online assessments, short virtual chats, project work, and in-depth interviews. The list provided by Google gives a rough outline of the hiring process, but it's hard to decide what skills and qualities they are looking for.

There are many logs posted giving suggestions on how to get a job in Google. Most of them give general guidelines [5] but didn't specify the skills or didn't provide the evidence behind the suggestion.

In this project, Google listing data will be used to drive insight for Google job seekers, to help them figure out what was required for their ideal role. It will include tech stack for software engineers or data scientists, skill set for other hires, qualities, requirements, etc. Such that, job seekers that look for jobs opportunity at Google will know exactly what they need to be preparing and learning to maximize the chance of getting their dream job.

Related Work

What has been done?

Many studies have been done on job skills classification and prediction. One common tool to use is machine learning, the research from Andrea M., etc. [6], uses machine learning to help characterize each job family with the appropriate level of competence required within each Big Data skill set. A neural network is also one of the most common tools for conducting research, research from Sun, Y., etc. [7] uses a neural network to separate job skills and measure their value based on massive job postings. Some research use reinforcement learning, and research also from Sun, Y., etc. [8] use deep reinforcement learning to estimate the long-term skill learning utilities.

This project will build on prior work using data collecting, processing, and modeling methods. Many methods will be used to drive better data insights compared to prior work.

Proposed Work

The pipeline of the project will be data collection, data processing, evaluation data analysis, data modeling, and data deployment.

For data collection, Selenium will be used to scrap data from the google job listing. There are many tools for web scraping, such as Beautiful Soup and Scrapy. Selenium is a powerful web automation tool, it works well for web pages that were rendered in real-time in the browser, like google job listings.

For data processing, there were many steps involved. The first is to data clean. In this step, the data collected from the web page will be cleaned and parsed. Then Pandas will be used to prepare data analysis and modeling. This step aims to clean the data further and deal with null values and outliers.

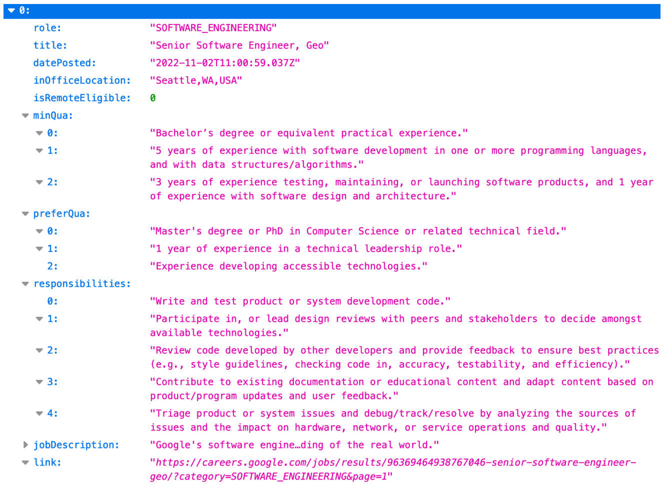
For data modeling, Sklearn and TensorFlow will be used. The goal of this step is to identify the necessary skill sets and determine the rank of each of them, using topic modeling to determine what are the different qualities and requirements of the job description.

For data deployment, React.js will be mainly used to create an interactive data visualization webpage.

Project Pipeline

1. Data collection:

Successfully used Selenium to get Google job listings by different roles. Check out the GitHub link of the web scraper source code[9].



```
▼ 0:
    role:                 "SOFTWARE_ENGINEERING"
    title:                "Senior Software Engineer, Geo"
    datePosted:           "2022-11-02T11:00:59.037Z"
    inOfficeLocation:     "Seattle,WA,USA"
    isRemoteEligible:     0
  ▼ minQua:
    ▼ 0:                  "Bachelor’s degree or equivalent practical experience."
    ▼ 1:                  "5 years of experience with software development in one or more programming languages,
                          and with data structures/algorithms."
    ▼ 2:                  "3 years of experience testing, maintaining, or launching software products, and 1 year
                          of experience with software design and architecture."
  ▼ preferQua:
    ▼ 0:                  "Master's degree or PhD in Computer Science or related technical field."
    ▼ 1:                  "1 year of experience in a technical leadership role."
      2:                  "Experience developing accessible technologies."
  ▼ responsibilities:
      0:                  "Write and test product or system development code."
    ▼ 1:                  "Participate in, or lead design reviews with peers and stakeholders to decide amongst
                          available technologies."
    ▼ 2:                  "Review code developed by other developers and provide feedback to ensure best practices
                          (e.g., style guidelines, checking code in, accuracy, testability, and efficiency)."
    ▼ 3:                  "Contribute to existing documentation or educational content and adapt content based on
                          product/program updates and user feedback."
    ▼ 4:                  "Triage product or system issues and debug/track/resolve by analyzing the sources of
                          issues and the impact on hardware, network, or service operations and quality."
  ▶ jobDescription:       "Google's software engine…ding of the real world."
  ▼ link:                 "https://careers.google.com/jobs/results/96369464938767046-senior-software-engineer-
                          geo/?category=SOFTWARE_ENGINEERING&page=1"
```

The image shows a single entry of the raw data collected. And it’s stored in a JSON file.

2. Data Understanding:

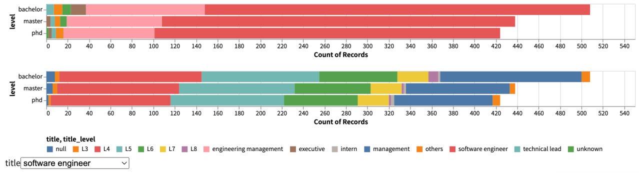
Each data entry is stored with a dictionary data structure and each JSON file is a list of dictionaries. For the role of software engineering, there are 543 entries. Each entry contains many different attributes from the listing. Including location, title, date posted, etc.

3. Data Processing:

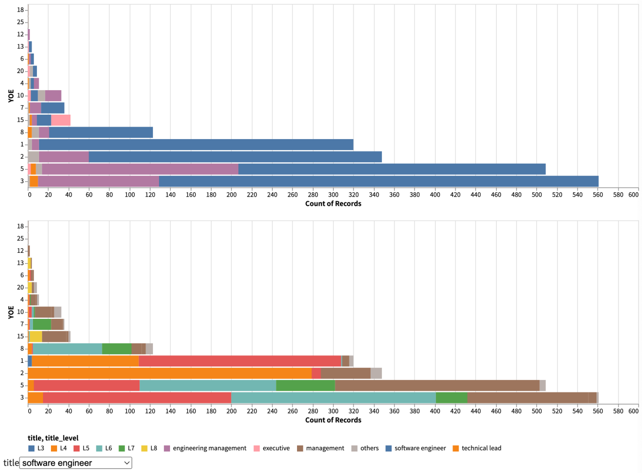
Use the data for further mining, there are many steps involved in cleaning the data. The code for cleaning the data is also in the GitHub link. Each attribute required a different approach to cleaning. For the scalability of the Data Cleaner, it will only perform basic cleanings, such as parsing the title into the main title and the field of the title and parsing the location into the country, state, and city. Extract degree requirements and years of experience from minimum qualification and preferred qualification. After the processing, merged all the data into a single JSON file for EDA and data mining.

4. EDA

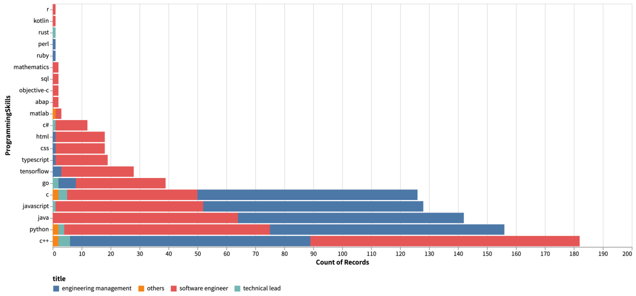
After collecting and processing all the raw data, using a data visualization tool is a good way to drive insight from data. Such as degree requirements, job location, programming languages, and years of experience. The dataset contains many different roles that belong to software engineering, such as software engineer, engineer manager, executive, technical lead, etc. It will be more informative if when a different role was selected, the detail of that role shows up. Altair is a powerful python package, it’s used to create interactive data visualization and pair plot. Using Altair, customize utilities can be added to the data visualization, such as a tooltip, and data filtering.



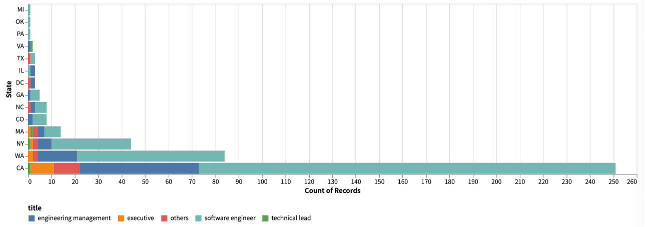
The plot above shows the count of different levels of degrees required at the job listing and the role and role level response to that. Bachelor's degree is most common overall, but the ratio of Ph.D. and master's is also talking over a large ratio of listing requirements.



The plot above is the year of experience required for different roles and levels. As a software engineer, the most required YOE is 1-5 years, for the L3 level, it required 1 year of YOE, and L4 required 1-5 years.



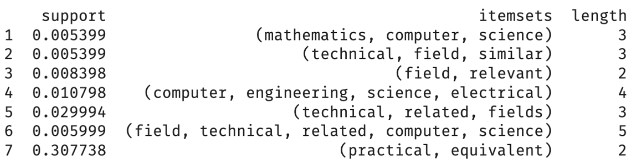
The plot above shows the count of each programming skill mentioned on the listing. C/C++ is most frequent, followed by Python, Java, JavaScript, and Go.



The plot above shows the count of job opening locations in the US, in the state of California, Washington, and New York has most of the job openings.

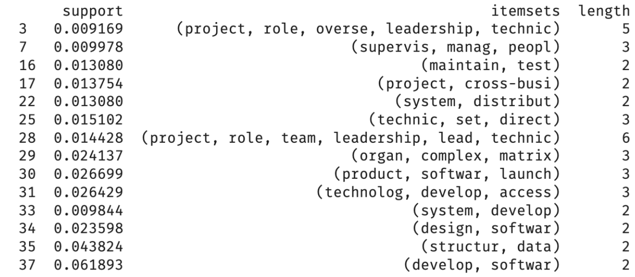
5. Frequent Pattern Analysis

FP max is a variation of FP growth, it's more efficient than FP growth. It will be used to mine the frequent degree field and experience field.



```
     support                                   itemsets  length
1  0.005399                (mathematics, computer, science)       3
2  0.005399                   (technical, field, similar)       3
3  0.008398                             (field, relevant)       2
4  0.010798    (computer, engineering, science, electrical)       4
5  0.029994                   (technical, related, fields)       3
6  0.005999  (field, technical, related, computer, science)       5
7  0.307738                        (practical, equivalent)       2
```

The form above shows the frequent items set in the degree field on the job listing. The most mentioned degrees field are computer science, engineering, and mathematics. Practical equivalent, the related technical field also mentions a lot.



```
      support                                       itemsets  length
3   0.009169      (project, role, overse, leadership, technic)       5
7   0.009978                          (supervis, manag, peopl)       3
16  0.013080                                  (maintain, test)       2
17  0.013754                             (project, cross-busi)       2
22  0.013080                               (system, distribut)       2
25  0.015102                            (technic, set, direct)       3
28  0.014428  (project, role, team, leadership, lead, technic)       6
29  0.024137                          (organ, complex, matrix)       3
30  0.026699                        (product, softwar, launch)       3
31  0.026429                      (technolog, develop, access)       3
33  0.009844                                 (system, develop)       2
34  0.023598                                 (design, softwar)       2
35  0.043824                                  (structur, data)       2
37  0.061893                                (develop, softwar)       2
```

The form above shows the frequent items set of experience field. The words are stemmed before the frequent pattern mining.

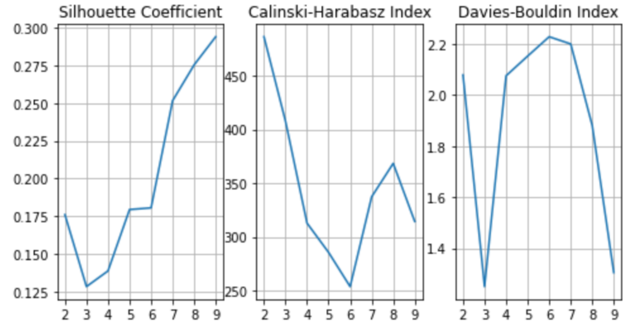
6. Topic Modeling

Understanding job responsibility is always important when applying for a job. To maximize the understanding of general job responsibilities, topic modeling is a good tool to categorize the different aspects of all of the responsibilities.
Mini Batch and hierarchical clustering were chosen for the clustering task.

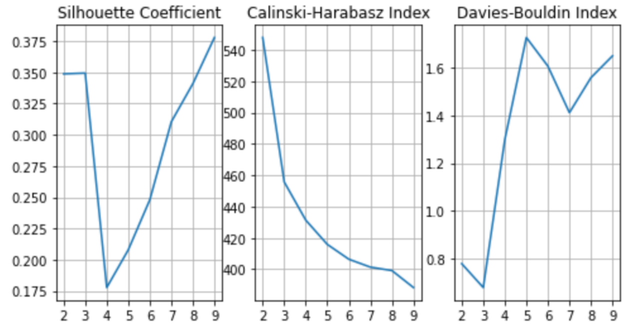
7. Evaluation

For unlabeled data, the evaluation must be performed using the model itself.
(1) The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. A Higher Silhouette Coefficient score relates to a model with better-defined clusters.
(2) The Calinski-Harabasz index is the ratio of the sum of between-clusters dispersion and within-cluster dispersion for all clusters, The score is higher when clusters are dense and well separated.
(3) The Davies-Bouldin score is defined as the average similarity measure of each cluster with its most similar cluster, the minimum score is zero, with lower values indicating better clustering.

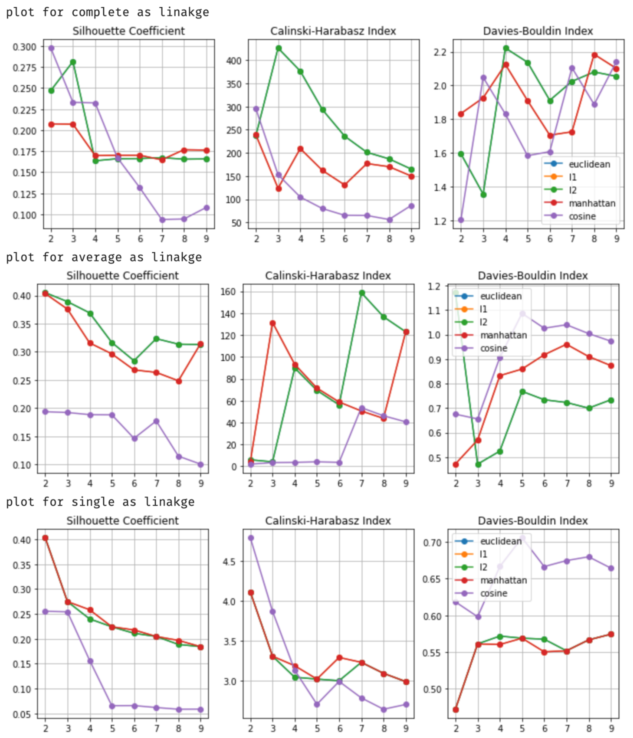
For getting the best model, these 3 evaluation metrics will be used to determine which model and how many clusters to use.

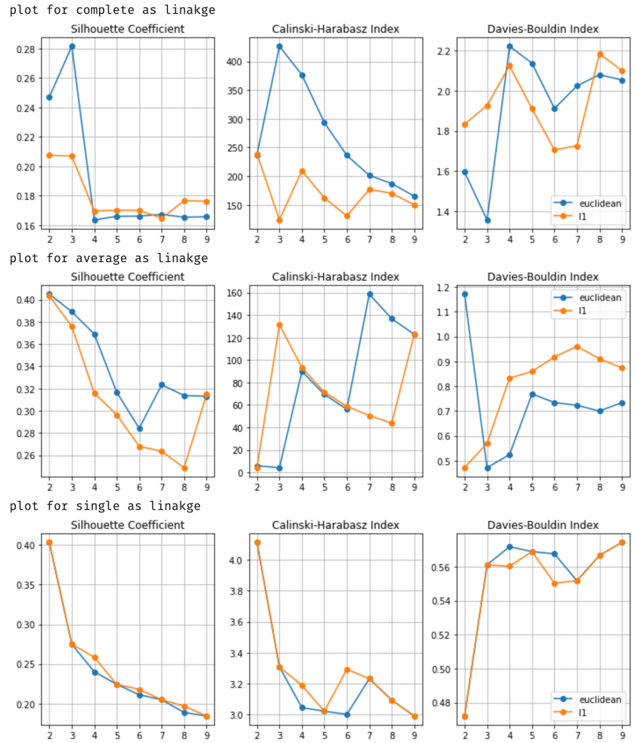


The plot above shows the result from minibatch kmeans.



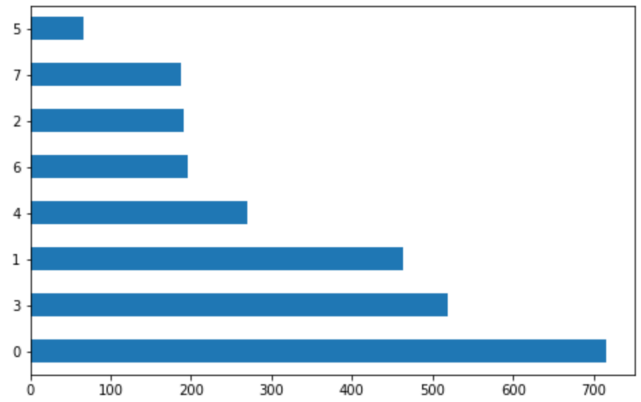
The plot above shows the result of hierarchical clustering.



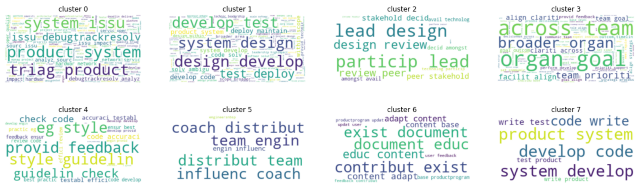


The above two plot shows the hyperparameter tuning result from hierarchical clustering. For some reason, matplotlib are unable to put all the affinity line in the same plot, so the Euclidean and L1 have to show separately.

Based on the evaluation plot, 8 clusters, ward linkage, and Euclidean affinity are good fits for the data. So it will be selected as the best model.



Using the model to perform clustering and count the values of each cluster. The plot shows the result.
It is hard to determine what each topic is about, but creating a word cloud will be helpful for understanding keyword of each cluster.



Discussion

The challenges I encounter includes many aspects. First, when collecting data using selenium is tricky. When the project started, it was planned to scrap all the data automatically without too much manual work once the web scraper is properly programmed, but it turns out to be lot more work involved.

Conclusion

For conclusion, there are many interesting information recovered by this project. It's competitive to get a job at google but the requirement is relatively general and does; t seem unreasonable. Google provides many opportunities for various degree levels with computer science or relevant field candidates. 1 year of experience is common for applying for an entry-level position, but higher years of experience also apply. Experiences at design software, system development, launching softest product, disturbing systems, and test maintenance, are the most common technical experience requirement, working with complex matrix organizations, leadership, and cross-business project, are the most common management skills requirements. For programming languages, C/C++ is surprisingly the most asked language. And there are not surprised that California is where most of the openings are.

References

[1] Despite Worker Complaints, Google Is Still World's Most Reputable Employer. https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/google-worlds-most-reputable-employer.aspx

[2] Google received 3.3 million job applications in 2019. https://www.axios.com/2020/01/09/google-2019-applications-backlash

[3] Google is slowing down hiring through 2020 amid the COVID-19 pandemic. https://www.theverge.com/2020/4/15/21222942/google-slowing-down-hiring-through-2020-covid-19-pandemic

[4] How we hire – Google Careers. https://careers.google.com/how-we-hire/

[5] How to Get a Job at Google – wikiHow. https://www.wikihow.com/Get-a-Job-at-Google

[6] De Mauro, Andrea & Greco, Marco & Grimaldi, Michele & Ritala, Paavo. (2017). Human resources for Big Data professions: A systematic classification of job roles and required skill sets. Information Processing & Management. 54. 10.1016/j.ipm.2017.05.004.

[7] Sun, Y., Zhuang, F., Zhu, H. et al. Market-oriented job skill valuation with cooperative composition neural network. Nat Commun 12, 1992 (2021). https://doi.org/10.1038/s41467-021-22215-y

[8] Ying Sun, Fuzhen Zhuang, Hengshu Zhu, Qing He, and Hui Xiong. 2021. Cost-Effective and Interpretable Job Skill Recommendation with Deep Reinforcement Learning. In Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 3827–3838. https://doi.org/10.1145/3442381.3449985

[9] GitHub link of project source code: https://github.com/lakzeee/GoogleJobsListingMining