BAX422-002 PROJECT REPORT

WEB SCRAPING JOB OPPORTUNITIES FOR ANALYSTS

Sahithi Sukhavasi

Namuun Boldbaatar

Jessica Chen

Executive Summary

The job market for business analysts is highly competitive, with many candidates seeking skilled analytics professionals to drive businesses with data-backed insights. Traditionally, people have relied on methods such as job boards and networking to explore the job market and identify job opportunities. However, web scraping data from job posting sites — in our case, LinkedIn and SimplyHired — is a useful complement to these traditional methods, as it allows for timely gathering of information on the latest job postings. This can help candidates automate the job searching process, save time and resources, and stay competitive in a crowded job market. The dataset collected allows candidates to obtain a better understanding of the demand for different jobs, critical skills that are most in-demand , and salary range. For this study, we chose to use Python for web scraping and data processing, and MongoDB for data storage and retrieval, resulting in efficient data collection and useful insights into the job market.

The study aims to gather important information on job opportunities in the data analytics field by leveraging multiple platforms and storing the data in one accessible place. The web scraping routine involved using search terms on LinkedIn and SimplyHired, collecting job URLs, downloading each job posting as an HTM file, and extracting relevant attributes such as job title, company name, salary, location, and job description. The resulting dataset can be used to answer business-relevant questions related to the job market, such as demand for various job roles, most in-demand skills, and median salaries. The dataset was stored in MongoDB due to its ability to handle structured and semi-structured data and its flexibility in scaling and replicating data. The analysis of the data revealed that business analysts had the highest number of job postings, followed by data scientists, and that data analysis, communication, and problem solving were the top three skills in demand. This information can be used by job seekers to identify promising job roles and skills, and by employers to tailor their hiring strategies.

Business Idea and Context

The job market for business analysts is highly competitive, with many companies looking for skilled professionals to help them analyze and make sense of complex data. Our project revolves around the idea of collecting and creating a database that contains useful job related information for analytical roles. To achieve this we have scraped and extracted job postings from LinkedIn and SimplyHired which would help stay up to date on latest job postings and identify potential employment opportunities.

Why Analytical Roles?

Analytical roles cover a range of roles like business analyst, data scientist, data analyst, finance analyst, and the demand for data driven analysis and decision making is becoming more prominent in recent years. The need for leveraging data as a growth tool in companies has created many opportunities for individuals with analytical skills and experience to build rewarding careers across a range of industries. Analyst roles are in high demand across a variety of industries, from finance to healthcare to technology. The main purpose served by these roles is to analyze data, identify trends and patterns, and provide insights to help companies make more informed decisions. There are many companies that offer analyst positions, including large corporations like Microsoft and Amazon, as well as startups and consulting firms. The competition for these positions can be intense, as many qualified candidates are vying for these positions. It is a critical role to take up in a company, so the company's are also picky about choosing their candidates.

What services are provided?

The project at hand tries to make the job search process easier by collecting information about job opportunities that are available under the titles business analyst, data scientist, data analyst, business intelligence engineer, and market analyst. These roles and the given information on the role like job description, company, location, salary, skills, and few more attributes are extracted and saved in a database which is readily usable. For now the scope of the project covers information from the top 300 jobs that showed up for each of the five mentioned analytical job titles from LinkedIn and

SimplyHired. Just like how companies have requirements to hire an individual, the same individual might be looking for something specific when looking for jobs. The individual's requirements may be based on location of the company, employment type - full time or part time, remote or in person, salary expectations, seniority level and other factors. The user can query for roles with the specific needs that they might be looking for and immediately retrieve some available options. The search query will return job related information along with the job url, the user can quickly jump to the specific url that interests him/her and start applying without delay. Readily providing the application url eases the process of job applications.

Project Journey

Introduction of the Data Sources

Traditionally, individuals relied on a variety of methods to identify potential job opportunities. This may include checking job boards, networking with other professionals in the industry, and approaching recruitment agencies or headhunters. Job searching and making connections in a professional network are crucial events that help land a job. Companies like LinkedIn, SimplyHired, Glassdoor, Indeed and many more companies are ruling this online space. The aim of the project is to leverage these multiple platforms to gather important information on jobs and store it in one accessible place. The main focus here is around analytical roles and we have chosen a set of five roles to start with - business analyst, data scientist, data analyst, business intelligence engineer, and market analyst. We are using these roles as search terms on LinkedIn and SimplyHired where location is set to the United States. We then move onto scraping both the websites for the job urls. Other options like Glassdoor and Indeed were explored but we don’t have access to scrape those websites.

Web-Scraping Routine

Web scraping data from sites like LinkedIn and SimplyHired can be a useful complement to these traditional methods, as it allows individuals to quickly and easily gather information on the latest job availability. By using this web scraping tool, the process of job searching can be automated, freeing up time and resources for other important tasks.

The specific strategy for web scraping job postings may vary depending on what the person is looking for. The search terms or the location are variables that can take input from the user. The results will vary depending on the location or role requested. There are multiple jobs available for each title that is searched and the results are paginated. The first few page urls are collected from the initial search using BeautifulSoup library in python and we make sure we have a minimum of 300 postings. We create soup objects for these page urls and extract the first 300 job posting urls. Then we make the python code run through these urls and download each one as a htm file. This process is done to make sure that the html source is easily available to extract necessary information and avoid sending requests to the website which may get you blocked off. We then read every htm file and extract the attributes like job title, job url, company name, salary, location, job description, date of posting, benefits, seniority level, industries, job functions, etc.

The business model for using web scraping to gather data on job postings aligns well with the goals of most business analyst firms. By automating the process of job searching, companies can save time and resources while still identifying potential job opportunities. This can help them stay competitive in a crowded job market and ensure that they have access to the latest job postings in their field.

Explanation of the Dataset

The dataset we collected through web scraping can be used to answer several business-relevant questions related to the job market for data scientists and business analysts, including the demand for various job roles, the most in-demand skills, and the median salaries for different job roles. This information can be used by anyone who is seeking a job to identify the most promising job roles, and even use it to negotiate better salaries based on market trends. Employers can also use this information to identify the skills and job roles that are most in-demand, allowing them to tailor their hiring strategies. The dataset scope could be expanded depending on the requirements of the user like adding more data on other job designations, or searching for jobs in a more specific location in the USA or looking for jobs elsewhere.

Database Design Choices

We chose MongoDB as the database to store our data. It offers several advantages over alternative databases, including its ability to handle structured and semi-structured data, its flexibility in scaling and replicating data across multiple servers. These features make MongoDB ideal for handling the diverse and complex data, such as job postings which contain varying amounts of data. The design choices for this project were driven by the need for a flexible and scalable solution that could handle large volumes of data. Python and MongoDB both have a well-supported interface, which makes it easy to integrate between the two. These choices have allowed us to efficiently collect and store job postings related to the data analytics field and provide useful insights into the job market.

Analysis

To start our analysis, we first identified five job roles related to the data science field namely business analyst, data analyst, data scientist, business intelligence, and market analyst. We then conducted web scraping of two popular job search websites Linkedin and SimplyHired, to extract 3000 job URLs. Next, we analyzed the data to gain insights of the job market for analysts, the distribution of job postings across the five job roles, and found out that business analysts had the highest number of job postings, followed by data scientists. We were also able to see that the majority of the job postings were in California, New York, and Texas having more than 100 job postings.

To dig further, we looked at job descriptions and identified that data analysis, communication, and problem solving were the top three skills mentioned across all job roles. Job seekers can use this information to identify the key skills that are in demand and focus on developing those skills.

Summary and Conclusions

Web scraping is a valuable tool for gathering and analyzing data from various sources quickly and accurately. The data we collected through web scraping of Linkedin and SimplyHired provides insights into the job market for data scientists and business analysts, including the most in demand skills, location, median salaries for different roles etc. This information can be used by both recruiters

and job seekers to make informed decisions and stay competitive in the rapidly changing job market. With access to real time data, businesses can adjust their hiring strategies and make more informed decisions about the talent pool available in their target locations and job applicants can use this information to identify the most suitable career paths and the companies that are currently hiring.