Automating Job Search with Aggregated Data

Anna Gaudette, Priyanka Priyanka, Akansha Bansal

Table of Contents

Executive Summary

A single job search for a position"analyst" on Indeed.com (a job search website) leads to over 180k search results. There are roughly 15-20 job aggregator websites out there making it exceptionally tedious to decide where and how to search for relevant job openings. With the goal of automating and simplifying the job search process we decided to create a job details database containing job listings from Indeed and Google Search API. We have demonstrated a use case to build a custom job board for “Analyst” positions which can be extended to any job role just by replacing the search term.

We started with calling the google search API to extract fields such as company name, job title, location , recency of the post and schedule type. We supplemented the data with web scraped information on salary ranges from Indeed. We also created a mongodb collection to import this data in the form of documents to make it more accessible for further analysis. We added a text index on job title and location to speed up search.

This data will help both job seekers and providers by giving them insights about the current job market - most frequent job postings, popular job titles, salary ranges and job locations. These insights will help job seekers refine their job search and help employers by staying ahead of the market.

Background

Some of the challenges of finding accurate job postings is that most of the information is scattered. Be it across company specific job portals making it hard to individually track new postings or searching on job aggregator websites that charge job publishers by usage and prioritize promoted jobs.With this idea in mind we decided to evaluate different resources for collecting job postings and collating them to achieve a compact data store that enables us to keep track of relevant job listings. We looked at various websites such as glassdoor, usajobs, monster and indeed to understand the ones that can be scraped vs those that had APIs available. We found that there was no one perfect website and we had to use a mix of API and web scraping techniques to improve our data store.

Data Sources

We found that while there is a lot of data on job postings, extracting relevant fields is often not as simple. Glassdoor for instance is only a job review website and doesn't offer much in terms of unique fields or sufficient number of fields such as job title, description, nature of the job, location for a relevant job search. The data was riddled with redundant information and duplicate entries. Then there are federal websites such as usajobs which due to security concerns are not viable to scrape. As for the websites that provided an API, a lot of them were paid and the ones that are free offer only limited data and search fields. So we decided to go with Google Jobs API provided by a firm SerpAPI.com. (SERP stands for search engine results pages) .This API only allows 100 google searches per month for free and can become a bottleneck to scale. Hence, we decided to add more data from indeed.com to supplement our findings.

Google Jobs API

We extracted below mentioned fields for 100 top job postings in the United States for Analyst roles using the Google Jobs API (SerpAPI) which allows scraping results from Google Jobs search. The API supports JSON responses. We selected the following fields from the API response:

- Company Name
- Job Title
- Job Location
- Posted (when was the job posted/ recency of the job posting)
- Job Type (Full time/ Part time/ Contract)

We used the following API url and the embedded api key to search for "Analyst" positions in the United States:

```
url = "https://serpapi.com/search.json?engine=google_jobs&q=analyst&hl=en&start="+
str(pageNum) +"&api_key=783f3db41babc418337fa011423f23591698d6433f3f04e2acd3ec781c7fd48c"
```

Indeed

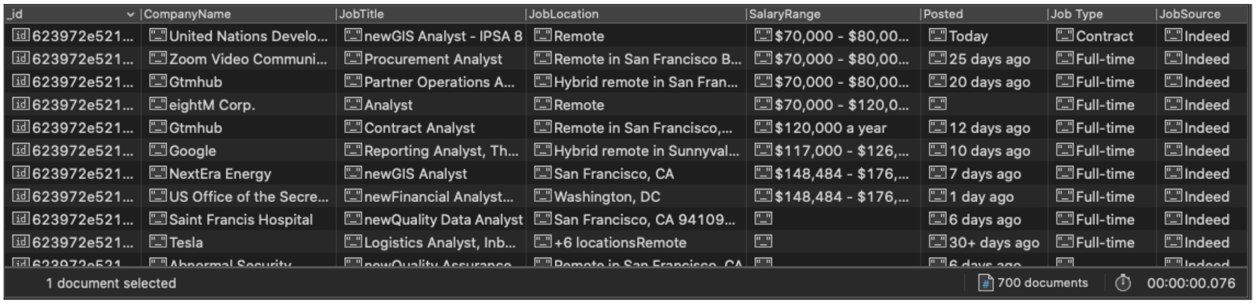
We web-scraped 600 top job postings in the United States for Analyst roles out of over 180k job postings with an additional field of **salary range**. We identified that a lot of the job postings appear multiple times across indeed's pages and due to the dynamic nature of the job postings it can be difficult to keep track of them. We used Python with BeautifulSoup to scrape the job details - Company Name, Job Title, Job Location, Salary Range, Posted (when was the job posted/ recency of the job posting) and Job Type (Full time/ Part time/ Contract).

We used the following to search for "Analyst" positions in the US:

```
URL = "https://www.indeed.com/jobs?q=analyst&start="+ str(pageNum) + "&vjk=ce8b0fe1bba304ca"
```

Database Design

We decided to combine this data into a mongodb collection called "jobs_data". The jobs_data collection consists of various documents with fields: CompanyName, JobTitle, JobLocation, SalaryRange, Posted, JobType, JobSource. Since we have data coming from two different platforms – google jobs API and indeed web scraping with different fields, mongodb allows for easy integration of the two varying data sources.Additionally, if we want to enrich the data from additional sources mongodb allows for better scalability(both vertically and horizontally) unlike relational databases.



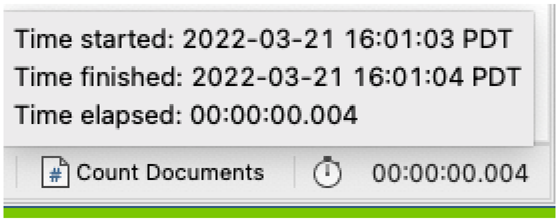
_id	CompanyName	JobTitle	JobLocation	SalaryRange	Posted	Job Type	JobSource
623972e521...	United Nations Develo...	newGIS Analyst - IPSA 8	Remote	$70,000 - $80,00...	Today	Contract	Indeed
623972e521...	Zoom Video Communi...	Procurement Analyst	Remote in San Francisco B...	$70,000 - $80,00...	25 days ago	Full-time	Indeed
623972e521...	Gtmhub	Partner Operations A...	Hybrid remote in San Fran...	$70,000 - $80,00...	20 days ago	Full-time	Indeed
623972e521...	eightM Corp.	Analyst	Remote	$70,000 - $120,0...		Full-time	Indeed
623972e521...	Gtmhub	Contract Analyst	Remote in San Francisco,...	$120,000 a year	12 days ago	Full-time	Indeed
623972e521...	Google	Reporting Analyst, Th...	Hybrid remote in Sunnyval...	$117,000 - $126,...	10 days ago	Full-time	Indeed
623972e521...	NextEra Energy	newGIS Analyst	San Francisco, CA	$148,484 - $176,...	7 days ago	Full-time	Indeed
623972e521...	US Office of the Secre...	newFinancial Analyst...	Washington, DC	$148,484 - $176,...	1 day ago	Full-time	Indeed
623972e521...	Saint Francis Hospital	newQuality Data Analyst	San Francisco, CA 94109...		6 days ago	Full-time	Indeed
623972e521...	Tesla	Logistics Analyst, Inb...	+6 locationsRemote		30+ days ago	Full-time	Indeed
623972e521...	Abnormal Security	newQuality Assurance	Remote in San Francisco, CA		6 days ago		Indeed

1 document selected 700 documents 00:00:00.076

To speed up the search for job titles we decided to put a text index on Job Title.We achieved search results in 4 ms instead of 65ms in a non-indexed regex search, a ~(10-20)x improvement in speed.

With Index

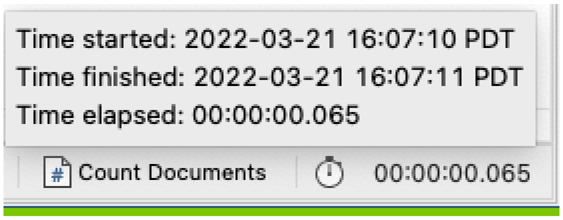
```
db.jobs_data.createIndex({JobTitle: "text"})
db.jobs_data.getIndexes()
db.jobs_data.find({$text:{$search: "Data"}}).pretty()
```



```
Time started: 2022-03-21 16:01:03 PDT
Time finished: 2022-03-21 16:01:04 PDT
Time elapsed: 00:00:00.004
```

Without Index

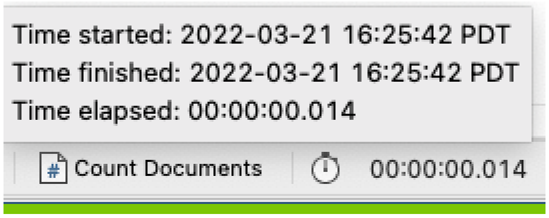
```
db.jobs_data.find({JobTitle:{$regex:"Data"}}).pretty()
```



```
Time started: 2022-03-21 16:07:10 PDT
Time finished: 2022-03-21 16:07:11 PDT
Time elapsed: 00:00:00.065
```

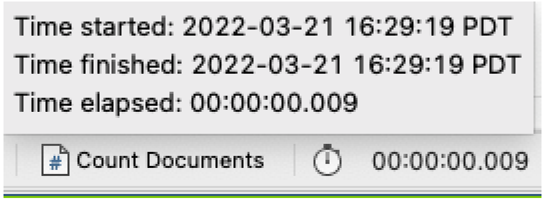
We used a text index on multiple fields such as JobTitle and JobLocation , to allow users to search for either location or job type in far less time.

```
db.jobs_data.createIndex({JobTitle: "text",JobLocation: "text"})
db.jobs_data.find({$text:{$search: "San Francisco"}}).pretty()
```



```
Time started: 2022-03-21 16:25:42 PDT
Time finished: 2022-03-21 16:25:42 PDT
Time elapsed: 00:00:00.014
```

```
db.jobs_data.find({$text:{$search: "Data Analyst" }}).pretty()
```



```
Time started: 2022-03-21 16:29:19 PDT
Time finished: 2022-03-21 16:29:19 PDT
Time elapsed: 00:00:00.009
```

Business Insights

We decided to pick MongoDB because document databases provide a lot of flexibility in terms of definition of the schema, adding new fields and storing fields with incomplete information. Our target audience includes both companies that are posting jobs as well as job seekers searching for jobs. The jobs database can be used for analysis of the job market and addresses the below the questions

- Which job titles are in demand?
- Which locations are most of the jobs concentrated in ?
- What jobs are frequently posted ?
- How many job roles are Remote vs In-person?
- Whether the listed jobs are full time or part-time ?
- Which locations explicitly specify the salary range and which jobs have higher salary offerings?
- Which companies are offering analyst based job roles?
- Which companies offer contract vs full-time positions for analyst roles?

This information helps the job seekers in gaining insights about the job roles in demand and expected salary ranges. For employers, this information will help them better understand their competitors, job titles they are using and the salaries they are offering for those positions Employers can use this data to estimate salary ranges based on market demand for roles that have more frequent postings. Job attrition can be measured by calculating how frequently a job posting appears and the number of reposts for a certain role over a period of 3 to 6 months. This can also be indicative of which positions stay vacant and/or are tougher to fill. Job seekers can accordingly prioritize their job search process by opting for frequently posted jobs. We can also use this data to build our own customized job aggregator repository (in this report we illustrated the "analyst" roles as an example). This database, in particular, can be used for instance by the UC Davis MSBA cohort or any Analytics professional for that matter, to refine their job search process.

Conclusion

Job data is easily available on the internet but searching for the right job is always time consuming. The same goes for employers when they are trying to gauge the competition and keep up with the market trends. Our project shows a prototype of how to build a customized job database incorporating data from various job listing sources. While creating this database, we also showed the challenges that we faced in scraping data and how we navigated them. This database can be further enriched by scraping and adding additional fields like job descriptions which can be used for text/ sentiment analysis to identify skills that are in demand. This webscraper can be a one-stop shop for both job seekers and job providers.