Indeed Web Scraping Report

LIS 875

Baolu Yu & Shaoxuan Wang

05.08.2022

INTRODUCTION

Our objective of this project is to give MS Information Program students job-seeking instructions, we analyzed 3 job positions based in 3 cities: Software Developer, UX Programmer, Data Analyst on the Indeed website in San Francisco, Chicago, and New York. In this project, we collected 1500 job data and discussed topics in five aspects:

- Urgent hiring company in recent months
- Talent required at different levels in the job market
    - e.g., entry-level, senior-level, staff level, principal level, etc.
- The top 10 listed skills for DA, UX, SDE
    - SQL,tableau,python…
- The top 10 popular reviewed employers
- Which position is hard to find a job

We accomplished our goal by crawling the data through the site, cleaning it up, and analyzing and visualizing it. The following techniques are applied: Pandas, URlib, RE, requests, Itertools, BS4, Time, Numpy, Plotly, NLP, and Gensim.

CRAWLING DATA

1. Make a permutation list of all positions and locations, and take out the job URLs on the first 10 pages based on these two keywords
2. Extract specific information from each job URL: company name, description, views, post day, reviews
3. Cleaning data (de-duplication, de-null, replace missing values by mod)
4. Specific analysis and mapping

GENERATE JOB LINKS:

We found that ‘q’ stands for position and ‘l’ stands for location in Indeed's URL. And every job has a ‘jobid’ stored in jobKeysWithInfo, which we pull out using regular expressions and request. During the search, we made a list of different expressions for each position, For example, software developers can be expressed as front-end Web development Engineer',' back-end Web development Engineer',' full-stack development ' '. 'Data Analyst' can also be searched by 'Data Engineer ', 'Data scientist '. 'UX Designer 'can also be 'UX programmer'. By doing this we obtained abundant information.

DATA:

	keyword	location	joblink
0	Front-end web development Engineer	Chicago	https://www.indeed.com/viewjob?jk=64e0a5fbf103...
1	Front-end web development Engineer	Chicago	https://www.indeed.com/viewjob?jk=bb475fa2a09b...
2	Front-end web development Engineer	Chicago	https://www.indeed.com/viewjob?jk=4acf51e325e2...
3	Front-end web development Engineer	Chicago	https://www.indeed.com/viewjob?jk=fbe5b541f511...

PARSE JOB LINKS:

In the process of crawling, we first utilized beautiful soup to parse the job links. And then we applied the "find()" function to find "div" and labels in the HTML pages for the information we wanted to grab. Next, we used "text()"" to obtain the text and store them in a dictionary, where the key value is the name of the information, and the values are the text we obtained of the information from the URLs.

Some of the URLs did not work, probably because indeed use different programmers to place job data, so their placement habits are different. We used exception, and when a crawling error occurs, we skip over and give an error ID.

But then we also found that there would be too many null values, especially for the job description field. This is because the number of "p" labels and the name of the "div" classes are different on each page. So instead of scrapping the text in paragraphs between "p" labels, we took out all the text information under the title of "Job Description Details". Also, we implemented the "find_next_siblings()" function to grab text that is not in any labels.

DATA:

	position	company	discription	post_day	reviews	job_link	error id
0	Senior Front End Engineer – Marketing Web	Glassdoor	Why Glassdoor? Our mission is to help people e...	Posted 24 days ago	47 reviews	https://www.indeed.com/viewjob?jk=64e0a5fbf103...	NaN
1	Front-End Engineer (FEE), Amazon Ads	Amazon.com Services LLC	Professional non-internship experience with f...	Posted 6 days ago	81,599 reviews	https://www.indeed.com/viewjob?jk=bb475fa2a09b...	NaN
2	Senior Software Engineer, Front End	Google	Google's software engineers develop the next-...	Posted 30+ days ago	4,083 reviews	https://www.indeed.com/viewjob?jk=4acf51e325e2...	NaN
3	Front End Developer	Dermacare	Dermacare LLC (DBA BlueChew) is on a mission t...	Hiring 1 candidate for this role		https://www.indeed.com/viewjob?jk=fbe5b541f511...	NaN

DATA CLEANING:

We then combine the two tables together so that the new table contains the keyword and

location we searched, then the specific job name, along with the company, description, number of views, and post day since the job was posted.

For the ‘review’ column, we removed sentences "reviews" and replaced null values with modes. For the result of Post day, we changed the result of 'Posted today' to the number 0. For the result similar to 'Posted 7 days ago', we made a judgment and only kept the number. Some information has a number that has nothing to do with post day, such as' Hiring 1 candidate ', and we made a judgment and dropped the value.

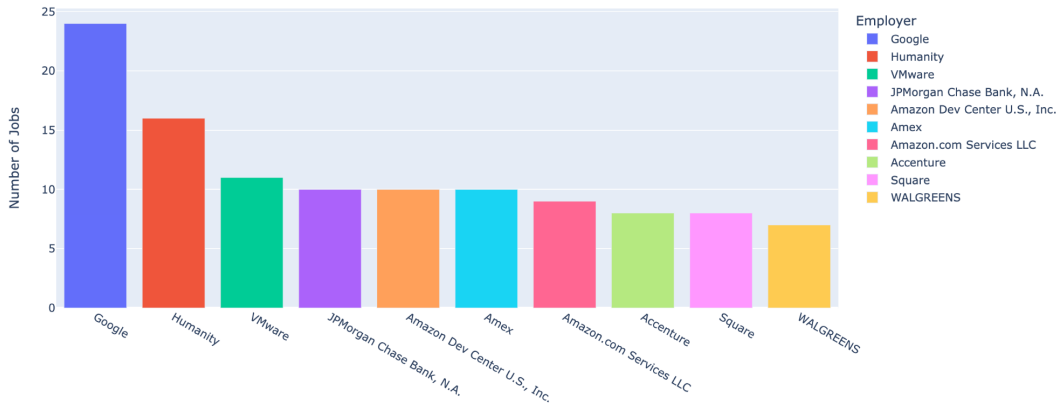
As for Description/Company, we removed ‘/n’ in the previous step of crawling, so the data was relatively clean. However, we dropped all the Company information that was not crawled out, and the error ID used to be utilized for correct sorting, now it is no longer needed after merging the tables.

DATA:

	keyword	location	position	company	discription	post_day	reviews	job_link	category
0	UX Designer	Chicago	UX Designer	Broadcom	Hi, the Broadcom Mainframe UX team is looking...	7.0	819	https://www.indeed.com/viewjob?jk=4a2a744726e9...	ux
1	UX Designer	Chicago	UX Designer	Casechek	Full–Time THE OPPORTUNITYCasechek is a f...	8.0	2	https://www.indeed.com/viewjob?jk=0ce456ef8d73...	ux
2	UX Designer	Chicago	UX Researcher/Designer	Buildout	In 2010 Buildout pioneered commercial real es...	0.0	2	https://www.indeed.com/viewjob?jk=451fd01159b1...	ux
3	UX Designer	Chicago	UX Designer	Deloitte	Are you a creative thinker who loves to be o...	30.0	10619	https://www.indeed.com/viewjob?jk=c68b8b2c26d9...	ux
4	UX Designer	Chicago	Designer, UX (Mobile Games)	WarnerMedia	Company Overview WarnerMedia showcases some of...	30.0	439	https://www.indeed.com/viewjob?jk=c2b32d5fd67c...	ux

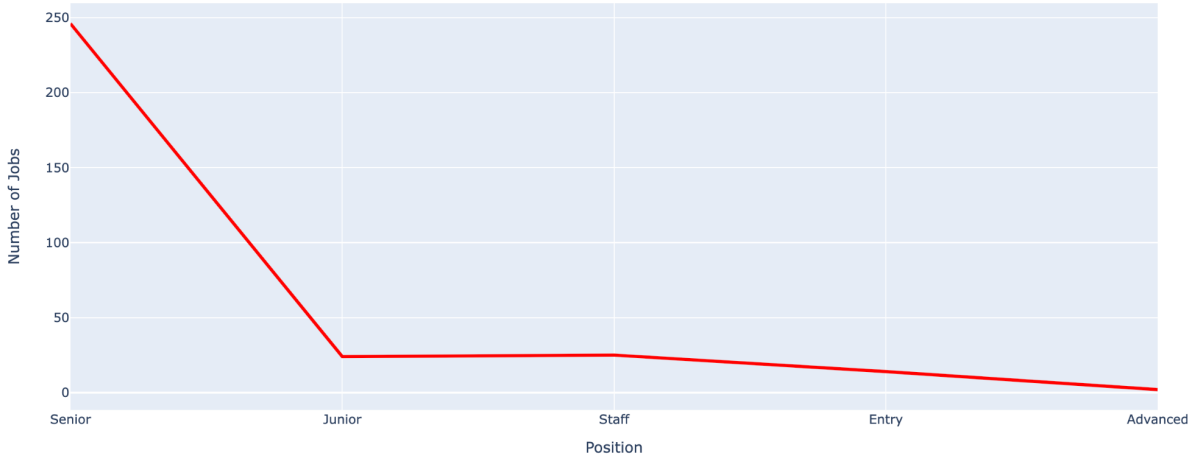
DATA ANALYSIS AND VISUALIZATION

URGENT HIRING COMPANY IN RECENT MONTHS:



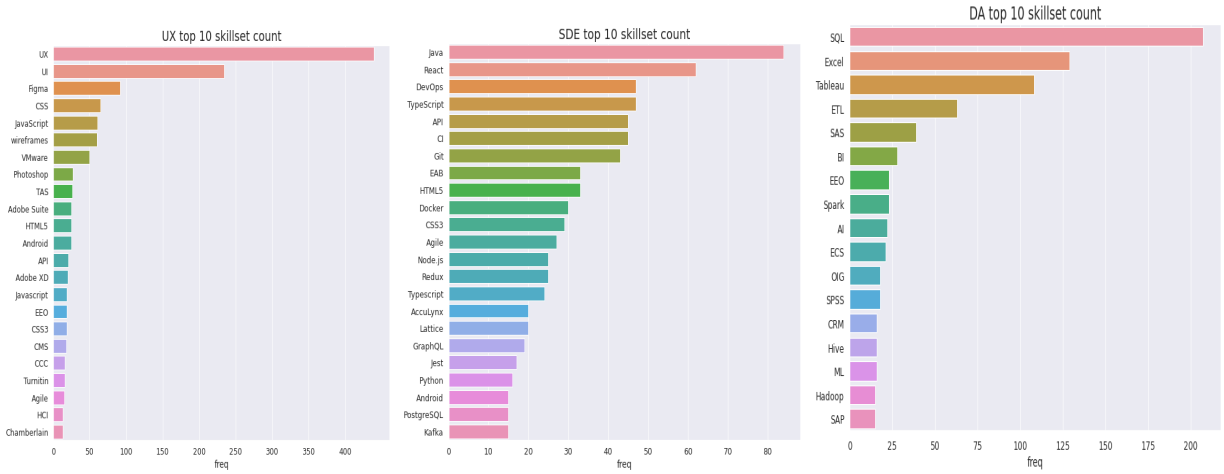
We sorted companies based on post days which were less than 30 days, this means we are looking for the top 10 diligent companies in releasing jobs last month. These urgent hiring companies are more likely to pass the interview.

TALENT REQUIRED AT DIFFERENT LEVELS IN THE JOB MARKET:



We got those keywords like entry, senior, and junior in the position column, and counts their value. We lower their case and check a lot of abbreviations. To make sure we didn't miss those data. We did not judge the level according to the number of years experience, because many people may not get promoted even after working for many years, so we only counted the data that the company emphasized this level in job post. The highest one is actually senior: Senior people are likely to be the main labor force in a tech company, kind of shortage! We use bag of words model and count tokens with key entity.

THE TOP 10 LISTED SKILLS FOR DA, UX, SDE:

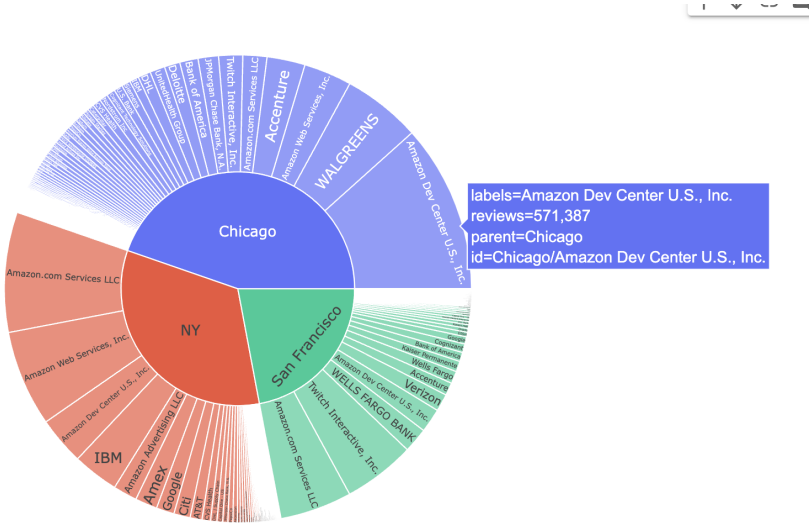


Sql, excel, tableau skills are kind of mandatory if students want to be good data analysts. Even though lots of people say that only old companies will still use SAS, what we found is SAS still hasn't been out of date.

If students want to apply for a job as a software developer, please learn more about java. It is popular and React is a good framework. some of the skillset requirement is even soft skills like DevOps means development department cooperates with the operational department.

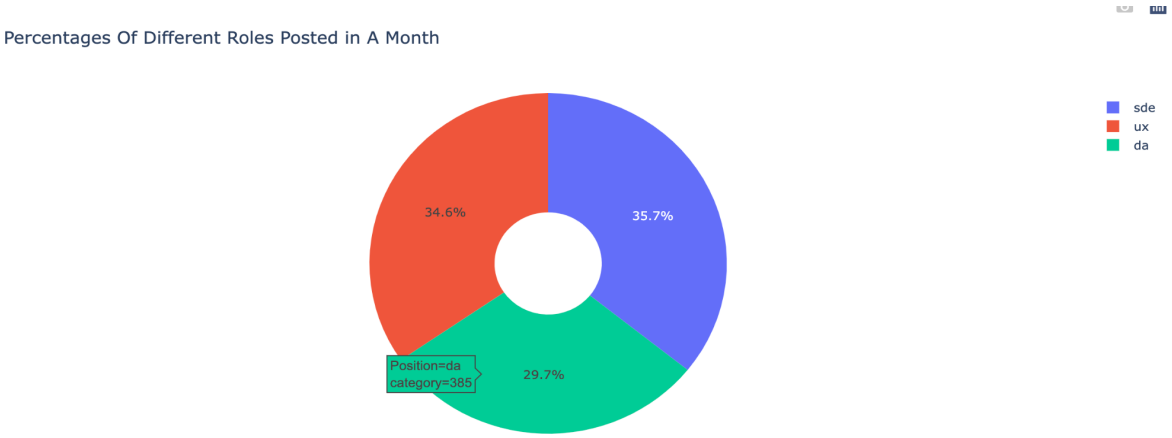
For the UX programmer, UX and UI is the basic skill, Figma, CSS is important. If students have any Ability in editing pictures or designing ability using photoshop, that would be a plus.

THE TOP 10 POPULAR REVIEWED EMPLOYERS:



The popularity of the company is judged by the number of reviews, and we ranked them by each region for easier observation. Amazon is the most popular one in any of these cities. The fun thing is when We looked into where Walgreen was founded, and it was in Chicago. This result is consistent with our statistics that Walgreens is very popular in Chicago. For the same purpose, we started searching about Twitch and IBM, their parent company is actually based in San Francisco and New York. So our graph makes sense.

WHICH POSITION IS HARD TO FIND A JOB:



After classifying all the positions, base on postday, we studied the positions in the market of DA, SDE and UX in recent 30days: DA positions are relatively few and the market competition is much more fierced when comparing to others.

CONCLUSION

1. If student want to find a job, whatever it is, try Google, Humanity, VMware to apply
2. If student do have a job, please try to become a senior title to be competitive
3. Job skillset:
    a. To work as a data analyst, learn SQL, EXCEL, and Tableau.
    b. Looking for a job as a UX designer, in addition to the essential UX and UI skills, llook at Figma, CSS, and Javascript.
    c. Learn JAVA, REACT, TYPESCRIPT for SDE jobs.
4. If the company is headquartered in a city, the company will be more popular and attract more attention in that city.
5. Data analysis is comparatively fiereced competition. If you are interested, you can apply for more SDE instead of DA position.