Project Report for Info 6210- Data Management and Data Design

AI Skunkworks Project

JOB DATABASE



Harshita Gupta

Ritvik Chauhan

**Summary:**

This project is aimed to create a list of companies with their link and websites including the job postings. The project will be dealing with two concerns First, matching the job seekers with the right employers and second, provide guidance to aspiring job seekers on the skills that are in demand so that they can build them to stay relevant in the job market. For this study, the data is scraped from www.indeed.com a job listing site for different companies. The relevant variables that will be scrapped due to the nature of the site include: Job Location, Job description, Salary. After collecting the data, data cleaning will be conducted and then to predict factors such as salary according to region, skill required for various jobs and future trend of a job, linear regression, text classifier and logistic regression is used.

The main goal of this project is to have a list of at least 1500 companies with their website and media pages links for domains such as Finance, Health Care, Digital Marketing, Engineering – Computer vision and E-commerce. Apart from that, we will also have the list of relevant skills required for a domain.

**Methods:**

The very first step we followed is scraping data from Indeed.com by using the python module Beautiful Soup. Since the site is relatively well structured and a simple text page hence easy to find the relevant data. Below is the sample view of scrapped data:



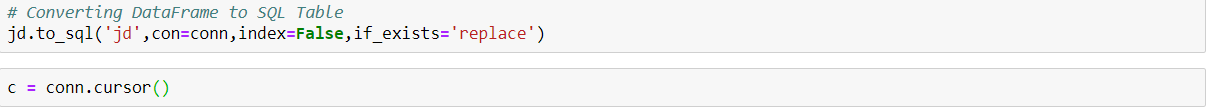
Later we cleaned the data and stored the data in the form of CSV file:



Once the scrapping and cleaning of data is done, we came up with all the below columns:

1. City: Where the job vacancy is available
2. Job: The title of job such as Business Intelligence engineer, Quantitive Analyst manymore
3. Salary: The annual amount of salary
4. Company: The name of company who is hiring
5. Expertise: The field/ skill or area of expertise for a Job posting
6. Recruiter: Name of the hiring manager for a Job Posting

In the next step, we converted the data frame into SQL table.

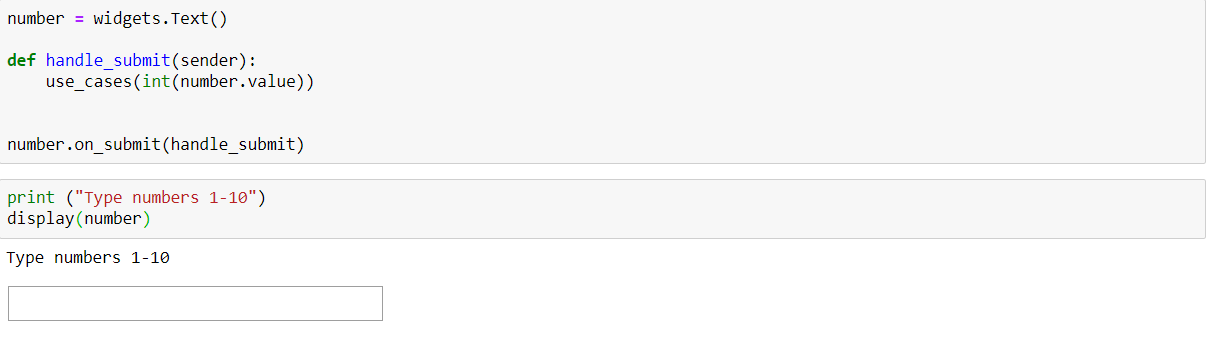


After converting the data frame into SQL table, we performed ten use cases on the Data frame, with the help of creating an interface by python switching. One of the use case we performed are below:

If someone wants to figure out the number of Jobs available according the city, so the first use case is based on the approach of a city like, if you put New York, you will get all the jobs where city is New York and we assigned it as 1. So, when you put 1 in the box below, it will display the result showing all the jobs from New York. Likewise, we did for all the 10 use cases

Use-Cases: 10 use cases have been covered.

1. All jobs in New York
2. All jobs with six figure or more salary
3. All jobs in Seattle and Los Angeles .
4. All jobs with the keyword data
5. All jobs by the recruiter Sophia Corde
6. All jobs related to Data
7. All jobs in New York with at least 100,000$ salary
8. All Analyst jobs in New York
9. All Engineering jobs with at least 100,000$ salary
10. All jobs by any recruiter named John



So, we created the interface such that, if you put any number between 1- 10, you will get the result corresponding to that use case.

**Word Cloud**: The next step we did is for the most relevant skills we created a word cloud that looks like:



**Results and Findings:**

The results and findings, we got from this project is that we got the list of companies with their website and media pages links for domains such as Finance, Health Care, Digital Marketing, Engineering – Computer vision and E-commerce. Apart from that, we will also have the list of relevant skills required for a domain. Also, we had an interface with the help of Python Switching to get the output from the use cases, which made it easier to look for some particular information, For example jobs according to cities.

**Statement of Contribution:**

Our contribution: 75% (converting basic code to required form, string to binary, text classifier, regression methods, web scrapping etc.)

External contribution: 25%

**References**

1. <https://stackoverflow.com/questions/41792471/installing-wordcloud-using-jupyter-notebook>
2. <https://www.kaggle.com/adiljadoon/word-cloud-with-python>
3. <https://github.com/jupyter/notebook/issues/1892>
4. <https://www.fernandomc.com/posts/using-requests-to-get-and-post/>
5. <https://stackoverflow.com/questions/46253288/scrape-with-correct-character-encoding-python-requests-beautifulsoup>
6. <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html>
7. <https://towardsdatascience.com/the-dummys-guide-to-creating-dummy-variables-f21faddb1d40>
8. <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html>
9. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
10. <https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/>
11. <https://www.vertabelo.com/blog/technical-articles/designing-a-database-for-an-online-job-portal>