Individual Project:**Career options/ Skill set recommendation for a User candidate**

Name: Sagar Bonde

SJSU Id: 013761566

**Introduction:**

Choosing right career choices and developing profession skills can be confusing at times, especially to the fresh graduates. Myraid number of Job listing portals available across internet, adds to the confusion of which careers choices to choose and which skills are required by which companies.

This project is about recommending best possible careers paths and skillsets to that one dream job or best possible jobs as per his own interest.

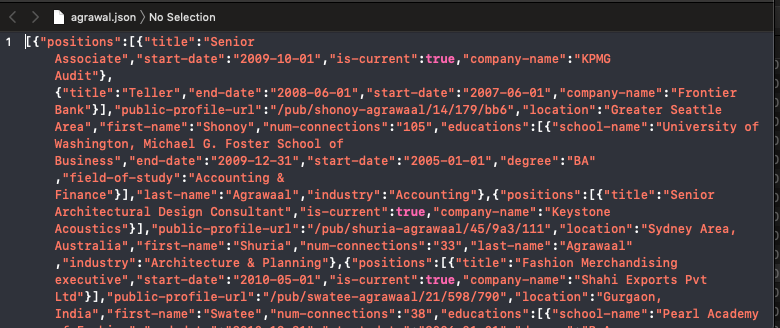
Usually the candidate have to do some research/survey to find out similar job postions that of the Job position that he / she desires, or the filed of choice that the candidate wants to work into.

This research/survey is ofcourse is a manual task and thus may consume a considerable amount of time. This the exact problem tha is being addressed by this project. With the help of gathering such data from LinkedIn a Text analysis method is being implemented to find out similar Job postings. Same thing is being repeated for to find out similar skills and education backgrounds of candidates. Further the similar the past job positions and the qualifications skills and expertise linked to those postions are compared to find out similar skillsets requied. Through text analysis algorithm a score has been given to the each and every candidate record. In the end the skills sets and educational qualifications that best match with the particular queried job position or Company description can be reccommended based on the highest scores.

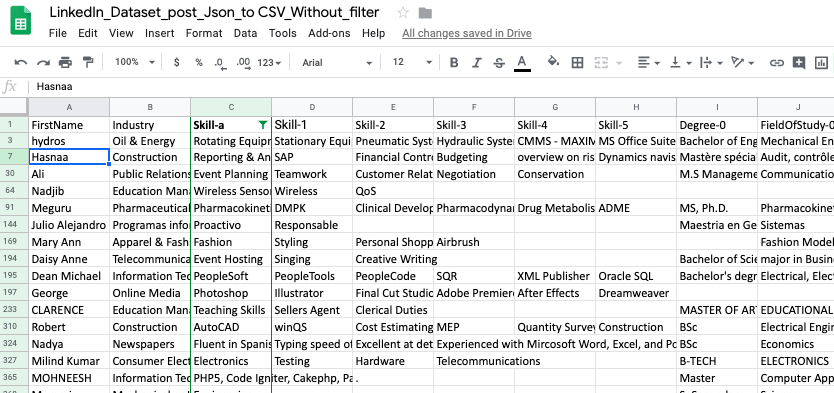
**Data Collection:**

The most popular personal portfolio site today is Linkedin so the linkedin profiles data is being used for this project. With the help of APIs various data attributes have already been collected. For example user profile data for a particular name. The collected data was initially in JSON format. Which has been converted to CSV format.

A sample dataset in JSON format looks like as following:



And this data after converting to CSV looks like as follows:



Further CSV data was cleaned to get proper/ desirable values of atttributes such as Skills, Education details, Job positions etc. After being cleaned this data can be used as explicit inputs to analyse patterns between academic credentials, field of work experience, expertise etc.

Similarly creating a dataset for Job positions descriptions, Job titles and Company details to find job positions and anticipated candidate qualifications.

The Final Data has following attributes:

**Skills Job positions held**

**Education Degree title Company details at which Job positions held**

**Education Field of study Industry / field of work experience**

**Data Analysis & Implementation:**

Detecting the similarity between job positions and qualifications required is important for skills recommendation systems as it allows, for example, the application of item-to-item based recommendations.

For best results the job titles can be combined with full-text job descriptions to detect important words multiple models and weighting functions, can be tested for the best performance picking out the best skill sets and career options

For building the dataset, we assume that two jobs are similar, and the skills sets required are also similar, if two or more users have the same job title in their past work experience and have the same skill sets. However, this source of information can be noisy, due to random profiles, and multiple data formats or when job postings are saved for a friend and not for the profile owner. Hence, by selecting only jobs where several users co-interacted with, we can increase the probability that such jobs are similar.

To accomplish the above mentioned goals a content based reccomendation system using TF-IDF feature extraction method is used here to analyse the record text.

The TF-IDF for a particular Skill or education description **t**, with respect to skills or educational qualifications from some other record **d** can be given as:

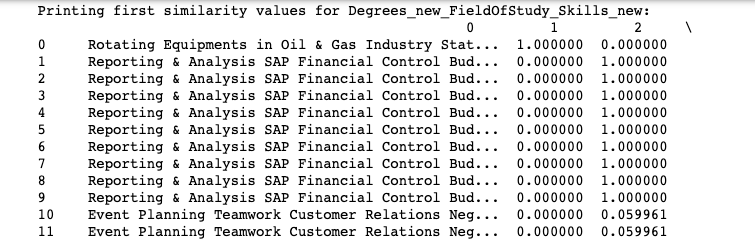


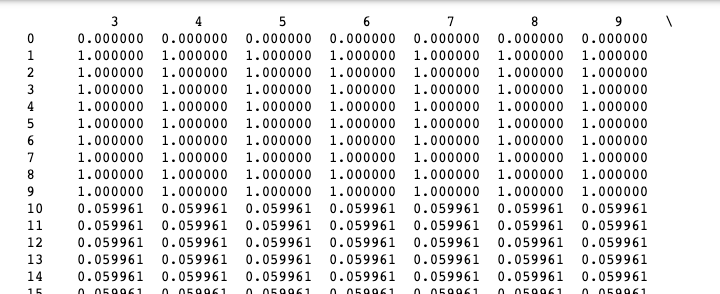
a vector of raw tf-idfs:

Educational A vector of raw Tf-IDF values is being calculated with the help of TfidfVectorizer() from sklearn library.

Further a cosine similarity function gives us the final score of similarity between two skills records.

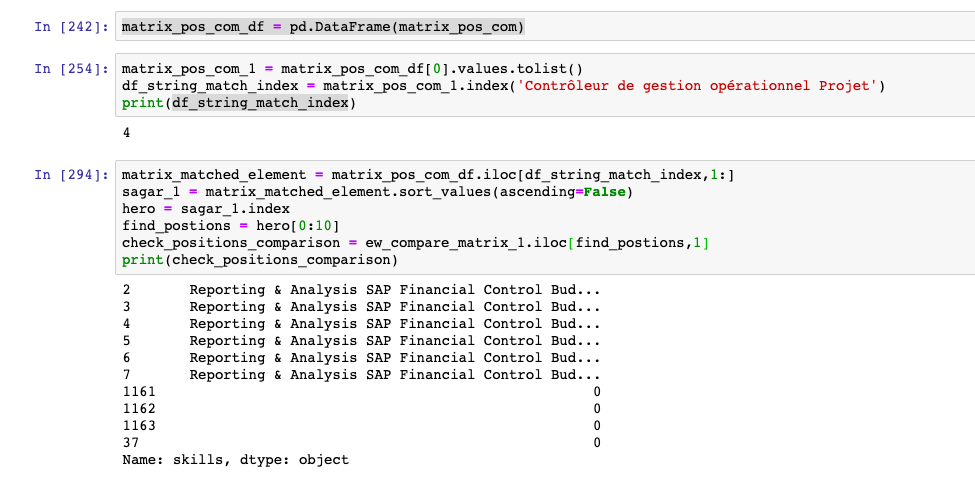
The matrix which depicts the similarity scores for each record with every other record is shows as follows:





Furthr sorting for each of the skill+ educational+filed of study record we get a rated scores in decending , which can be used for recommendations.

The Indexs of top n records can be exracted which looks as follows:



**These indexes are the indexes for which the job titles can be matched to the skillset and educational qualifications.**