CS-GY 6513 Project Report:
Resume Matching using Text Processing Techniques

Elastic Computing Infrastructure sponsored and underwritten by IBM Power Systems Academic Initiative

Student name: Shuyang Cao
New York University
Brooklyn, NY
mail_id: sc7813

Student name: Jianfei Zhao
New York University
Brooklyn, NY
mail_id: jz3766

1. Introduction

In recent years, text processing techniques has become very useful to solve some practical problems. For example, text classification, text searching, sentiment analysis, recommending system etc. After transferring texts into vector space, we can use multiple techniques to find some properties of the texts or quickly searching some crucial information of the given requests. These techniques include word embedding, N-Gram, TF-IDF or base on some machine learning model such as SVM, K-Means, Random Forests etc.

In this project, we aim to match a given resume to some job information which were crawled from a job hunting website. This will help the job-hunters to find some jobs which is most related to their skills, background and experience. We also plan to do some analysis on the data that collected from the website, for example, what kind of techniques are needed most in a job? The location distribution of job demanding etc. We will visualize the results and find some interesting patterns of these data.

2. Methods

The data for the project is gathered from a job hunting website http://www.monster.com. This website provides us with tons of job information by querying the job and preferred working location. In the project, we collect the information of 20 kinds of job. By searching one job we can get tens of thousands of records. Gather detailed information and store them into .csv file and MySQL in Azure.

Crawling in python use modules of requests, lxml and multi-threading. The website of Monster is using Ajax. Thus we simulate the GET response with the headers. There will be 25 records in one page. Traversing all the pages it come out can get a list of job detail links. Access all the urls and enter second-level page. Using xpath to match the elements such as job title, company name, working location, detailed description etc. Clear the data. And store them in different tables named by the jobs in an Azure MySQL database. Write a client to query the table base on location, job, company, description.

After gathering and storing the data, we can use some text processing techniques to finish our tasks.

We analyse word and document frequency using a technique called TF-IDF. TF stands for term frequency, that is how frequently a word occurs in a document. There are words in a document, however, that occur many times but may not be important. in English, these are probably words like “the”, “is”, “of”, and so forth. We might take the approach of adding words like these to a list of stop words and removing them before analysis, but it is possible that some of words might be more important in some documents than others. To distinguish these words, we use another method called inverse document frequency (IDF), which decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents.

In practical, we define a new vector space representation. For document i, we construct a vector x_i such that the j-th coordinate is given by:

$$x_i(j) = tf_i(j) \cdot idf(j)$$

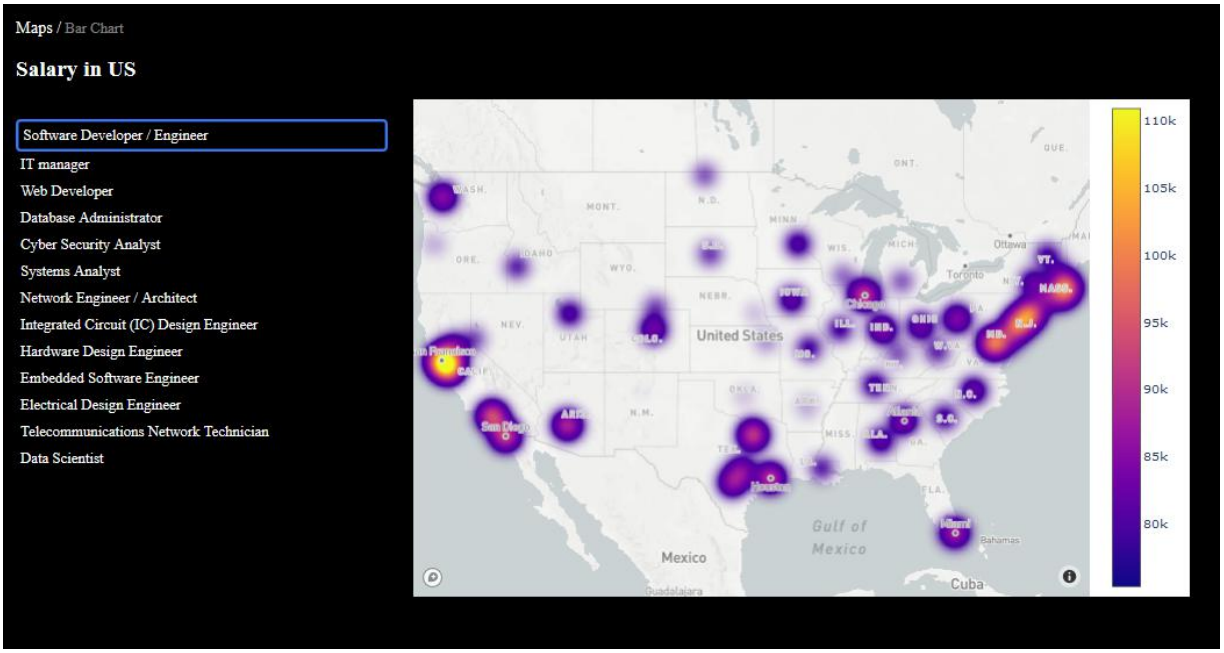
The term $tf_i(j)$ represents term-frequency, which counts the number of occurrences of word j in document i. The term idf(j) represents inverse-document frequency, and is defined as follows.

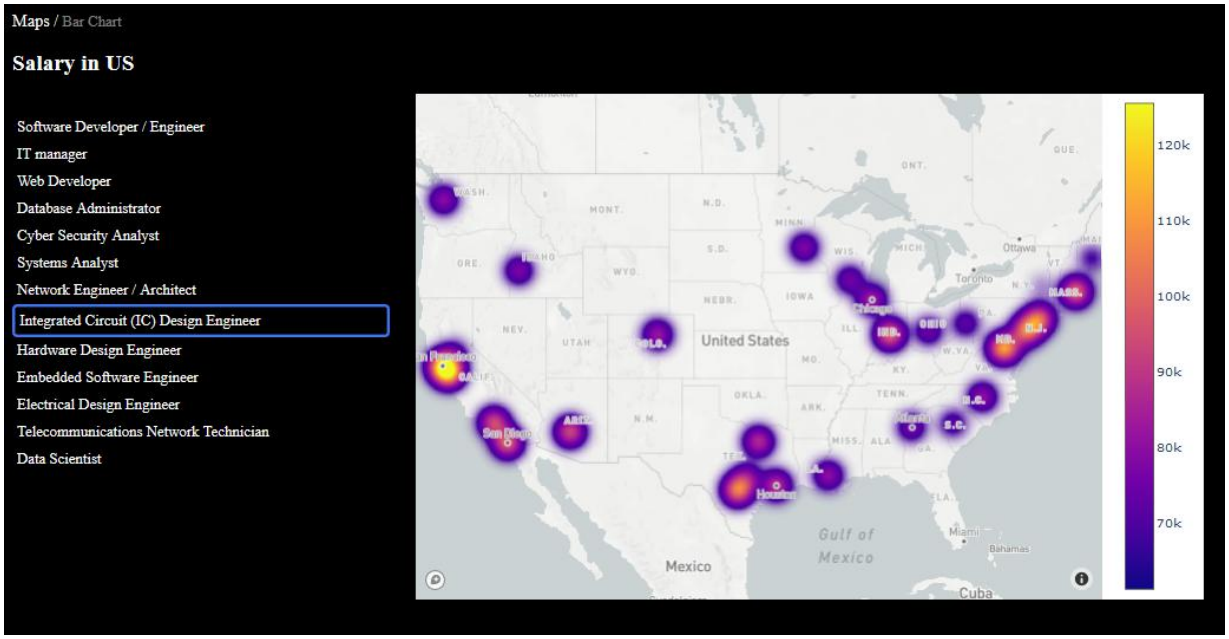
Let n_j be the number of documents in the database which contain at least one occurrence of word j. Then,

$$\text{idf(j)} = \log \frac{\text{n}}{\text{n}_\text{j} + 1}$$

Finally, we calculate the cosine-similarity between query vector and the TF-IDF matrix which is defined above.

3. Visualization





4. Results

In this project, all the jobs we get from the websites are related to field in Electronic and Computer Engineering(ECE) or Computer Science(CS). Finally we use about 20,000 jobs for matching the resume, and the resume we used are also from ECE major or CS major.

For implementing the resume matching system, we use PySpark in Python, which is a great language for performing exploratory data analysis at scale, building machine learning pipelines, and creating ETLs for a data platform. The first benefit of using PySpark is that it has an in-memory computation architecture, which helps us increase the speed of processing. Secondly, Being dynamic in nature, it helps us to develop a parallel application, as Spark provides 80 high-level operators. For these reasons, data scientists always use it to build a recommendation system or machine learning system.

These are some jobs recommended to a resume which shows a solid background in the field of Machine Learning:

Top 20 matched jobs:

job	company	location	similarity
Machine Learning ...	CyberCoders	New York, New Yor...	0.23400453719699693
computer vision e...	Brambles USA Inc	Orlando, Florida	0.23260229031089125
Machine Learning ...	Brambles USA Inc	Orlando, Florida	0.23260229031089125
data-scientist	Microsoft Corpora...	Redmond, Washingt...	0.23218043070360098
Python Software E...	CHEP	Orlando, Florida ...	0.2318661081229323
computer vision e...	CHEP	Orlando, Florida ...	0.2318661081229323
Machine Learning ...	CHEP	Orlando, Florida ...	0.23161338123623557
data-scientist	Randstad	Raleigh, North Ca...	0.2265064669531539
data-scientist	ServiceNow, Inc.	Kirkland, Washing...	0.22312532521787307
data-scientist	SPECTRUM	Golden, Colorado	0.21862107413506884
Spark Engineer	SPECTRUM	Englewood, Colorado	0.21862107413506884
Machine Learning ...	SPECTRUM	Wheat Ridge, Colo...	0.21862107413506884
FPGA Engineer	SPECTRUM	Pine, Colorado	0.21862107413506884
NLP engineer	SPECTRUM	Englewood, Colorado	0.21862107413506884
data-scientist	Paycom	Oklahoma City, Ok...	0.21478529137798136
Machine Learning ...	Paycom	Oklahoma City, Ok...	0.21478529137798136
data-scientist	Randstad Technolo...	San Francisco, Ca...	0.21284766243301098
data-scientist	Exelon	OAK BROOK, Illinois	0.21057569022732878
NLP engineer	Trexquant Investment	Stamford, Connect...	0.20802822570551724
NLP engineer	Tailored Management	South San Francis...	0.20401481468053828

(1) As is shown above, the most similar jobs are Machine Leaning, Computer Vision Engineer and Data Scientist .etc, which are all in the field of Machine Learning.

The following are all job recommendations for different kinds of resumes:

Top 20 matched jobs:

job	company	location	similarity
SQL Developer	AbleForce, Inc.	SAN DIEGO, Califo...	0.264033861383034
Database Engineer	AbleForce, Inc.	SAN DIEGO, Califo...	0.264033861383034
SQL Developer	Tekmark Global So...	Ohio	0.26021566245042727
SQL Developer	Accede Solutions Inc	Buffalo Grove, Il...	0.23786697071082044
SQL Developer	Shulman Fleming a...	Jersey City, New ...	0.23768455539319813
Database Engineer	Shulman Fleming a...	Jersey City, New ...	0.23768455539319813
Python Software E...	Shulman Fleming a...	Jersey City, New ...	0.23768455539319813
database administ...	AbleForce, Inc.	SAN DIEGO, Califo...	0.2314174228770613
IT manager	Zachary Piper LLC	Norristown, Penns...	0.22206289327132797
SQL Developer	Questa Technology...	Durham, North Car...	0.2212274635445375
Database Engineer	Randstad Technolo...	Brea, California ...	0.20937696887132082
SQL Developer	Randstad Technolo...	Brea, California ...	0.20937696887132082
database administ...	Netsource, Inc.	Bellevue, Washington	0.20896717872484252
Database Engineer	Randstad Technolo...	Phoenix, Arizona ...	0.1854604088665849
SQL Developer	Randstad Technolo...	Phoenix, Arizona ...	0.1854604088665849
database administ...	The Maxis Group	Phoenix, Arizona ...	0.18379570062459383
python	Shulman Fleming a...	Jersey City, New ...	0.18032984907850536
SQL Developer	Randstad	Brea, California ...	0.17606677224380562
Database Engineer	Randstad	Brea, California ...	0.17606677224380562
Spark Engineer	Zachary Piper LLC	Horsham, Pennsylv...	0.1749505043029096

(2) Resume for a Database Engineer employee.

Top 20 matched jobs:

job	company	location	similarity
Electrical Design...	Gulfstream Strate...	Charlotte, North ...	0.19837398245555088
Electrical Design...	Gulfstream Strate...	Raleigh, North Ca...	0.19837398245555088
Electrical Design...	Gulfstream Strate...	Jacksonville, Flo...	0.19837398245555088
Electrical Design...	Gulfstream Strate...	Jacksonville, Flo...	0.19837398245555088
Electrical Design...	Gulfstream Strate...	Raleigh, North Ca...	0.19837398245555088
Electrical Design...	Gulfstream Strate...	Sacramento, Calif...	0.19837398245555088
Electrical Design...	Gulfstream Strate...	Charlotte, North ...	0.19837398245555088
Electrical Design...	Gulfstream Strate...	Sacramento, Calif...	0.19837398245555088
FPGA Engineer	Surf Search	Cincinnati, Ohio	0.16304672360326572
Electrical Design...	Surf Search	Cincinnati, Ohio	0.16304672360326572
electrical test	Northern Wind, Inc.	NEW BEDFORD, Mass...	0.15299152687517378
Electrical Design...	Synerfac Technica...	Columbus, Ohio 43235	0.1504037889182138
PLC Technician	Hunt, Guillot and...	Erie, Pennsylvania	0.13440340264780684
electrical test	Tech USA	Huntsville, Alaba...	0.1342985896378198
electrical test	Sterling Machiner...	SOUTH EL MONTE, C...	0.13388333096894836
Electrical Design...	Synerfac Technica...	Hampstead, Maryla...	0.13380706384992894
electrical test	Lowell Obervatory	FLAGSTAFF, Arizon...	0.13007857259261205
Electrical Design...	CyberCoders	Dallas, Texas 76001	0.12707251887663865
electrical test	Spherion	Hope Hull, Alabam...	0.12694260039065755
electrical test	Randstad	Perth Amboy, New ...	0.1263053518144696

(3) Resume for a Electrical Engineering major employee.

Top 20 matched jobs:

job	company	location	similarity
Machine Learning ...	Southwest Researc...	San Antonio, Texas	0.08116147399903528
FPGA Engineer	Microchip Technology	San Jose, California	0.0806467142977962 7
FPGA Engineer	Lam Research Corp...	Fremont, Californ...	0.08037405195916836
DSP engineer	Lam Research Corp...	Fremont, Californ...	0.08037405195916836
FPGA Engineer	Randstad	Ashaway, Rhode Is...	0.07606167591026433
Embedded Systems ...	Southwest Researc...	San Antonio, Texas	0.071527836903023
DSP engineer	ON Semiconductor	San Jose, California	0.06983257516394467
FPGA Engineer	Odyssey Systems C...	Lexington, Massac...	0.06794424029819622
Embedded Systems ...	Odyssey Systems C...	Lexington, Massac...	0.06794424029819622
DSP engineer	Odyssey Systems C...	Lexington, Massac...	0.06794424029819622
ARM engineer	Odyssey Systems C...	Lexington, Massac...	0.06794424029819622
Circuit Design En...	Randstad Technolo...	Ashaway, Rhode Is...	0.06712068324775063
FPGA Engineer	Randstad Technolo...	Ashaway, Rhode Is...	0.06712068324775063
FPGA Engineer	Dynetics	Huntsville, Alabama	0.06575373417683203
Machine Learning ...	Southwest Researc...	San Antonio, Texas	0.06496953463692742
computer vision e...	Southwest Researc...	San Antonio, Texas	0.06496953463692742
computer vision e...	CyberCoders	Irvine, Californi...	0.06496923964857998
Python Software E...	CyberCoders	Irvine, Californi...	0.06496923964857998
Machine Learning ...	Southwest Researc...	San Antonio, Texas	0.0647745018675982
computer vision e...	Southwest Researc...	San Antonio, Texas	0.0647745018675982

(4) Resume for a Computer Engineering major employee.

In addition, we can also use SQL query in PySpark to find some recommended jobs in a specific city, you can choose any big city you like in USA, for example, the recommended jobs in New York City for a CE major student is shown below:

job	company	location	similarity
Telecommunication...	Clarapath Inc.	New York City, Ne...	0.04988706287672642
FPGA Engineer	Clarapath Inc.	New York City, Ne...	0.04988706287672642
python	Synechron	New York City, Ne...	0.035548372632214956
Telecommunication...	Clarapath Inc.	New York City, Ne...	0.034945310902609356
web developer	Clarapath Inc.	New York City, Ne...	0.034945310902609356
Database Engineer	Synechron	New York City, Ne...	0.03454556273853086
Python Software E...	Synechron	New York City, Ne...	0.03454556273853086
Architect	Synechron	New York City, Ne...	0.03454556273853086
Software Product ...	Technovision	New York City, Ne...	0.03115421079601298
database administ...	Business Informat...	New York City, Ne...	0.029346975702960926
Test Automation E...	Park Hudson Inter...	New York City, Ne...	0.027419386578140224
DSP engineer	CloudFlare	New York City, Ne...	0.02657937467733803
data-scientist	Grow	New York City, Ne...	0.023481358658936607
Machine Learning ...	Grow	New York City, Ne...	0.023481358658936607
Spark Engineer	Apex Systems	New York City, Ne...	0.021909819169386807
data-scientist	The NPD Group, Inc.	New York City, Ne...	0.020684520483680347
Java Software Eng...	Data Inc	New York City, Ne...	0.02004689255561453
Data Analyst	GRANT PETERS ASSO...	New York City, Ne...	0.019200753034171885
JavaScript Developer	Case Interactive	New York City, Ne...	0.01902748198912478
Python Software E...	Case Interactive	New York City, Ne...	0.01902748198912478

Another SQL query example is to find the rank of a specific job given your resume. For example, a CE major student could find some similar job information for Embedded Systems Engineer:

job	company	location	similarity
Embedded Systems ...	Southwest Researc...	San Antonio, Texas	0.11408710602803729
Embedded Systems ...	US ARMY Ground Ve...	WARREN, Michigan ...	0.10491342174520059
Embedded Systems ...	Abbott	Alameda, California	0.09937305121916463
Embedded Systems ...	Talentlab	OTTAWA, ON	0.09733485801632007
Embedded Systems ...	Oculii Corp.	Beavercreek, Ohio...	0.0955485982066785
Embedded Systems ...	Kumu Networks	Sunnyvale, Califo...	0.0954019401067412
Embedded Systems ...	D3 Engineering	Rochester, New Yo...	0.09532752286044031
Embedded Systems ...	NVIDIA Corporation	Santa Clara, Cali...	0.09069194568483023
Embedded Systems ...	CyberCoders	San Antonio, Texas	0.08739002114928836
Embedded Systems ...	Southwest Researc...	San Antonio, Texas	0.08734039335372538
Embedded Systems ...	Neteffects	Creve Coeur, Miss...	0.0839969616373131
Embedded Systems ...	Novanta	North Syracuse, N...	0.08396024753874272
Embedded Systems ...	CyberCoders	Hayward, Californ...	0.0820860831573231
Embedded Systems ...	CyberCoders	Colorado Springs,...	0.08191617468147264
Embedded Systems ...	IntelliPro Group ...	San Jose, California	0.08093664616821399
Embedded Systems ...	GCR Professional ...	Goleta, California	0.07659893476745336
Embedded Systems ...	Enphase Energy	Austin, Texas	0.07594776906368453
Embedded Systems ...	CyberCoders	Eugene, Oregon	0.07497889302584168
Embedded Systems ...	CyberCoders	Annapolis Junctio...	0.0745246835067582
Embedded Systems ...	Randstad	Brighton, Massach...	0.074523565258464

Finally, for the running time of this resume matching system, it will take about 30 seconds to finish the recommendation for you, and the query time is within 10 seconds. This shows the benefit of using Spark to process data.

5. Further Work

There are so many ways to improve our project and a lot of additional work we can do to make this project more complex and wonderful. For example, we can add more different kinds of jobs to our SQL system to match more resumes from other majors. In this case, we should modify our code to process huge amount of data in short time. Currently the volume our data is not so big and the processing time need to be speed up as well.