What you need to get a job at google

Mining Job Dataset from Google

1. Members:

r04921037 朱柏澂

r06921048 李友岐

r06921106 莊世聿

2. Moltivation:

Google is one of the best company in the world and lots of people wish to join them. As a result, we want to figure out what skill you need to get a job at Google by mining the Google Jobs Site.

Also we want to utilize the open dataset from GitHub to know what programming skill sets aren’t popular, but highly demanded in Google’s job requirements. By comparing the demand and supply, we can find out which skill is the worthiest to invest.

3. Introduction:

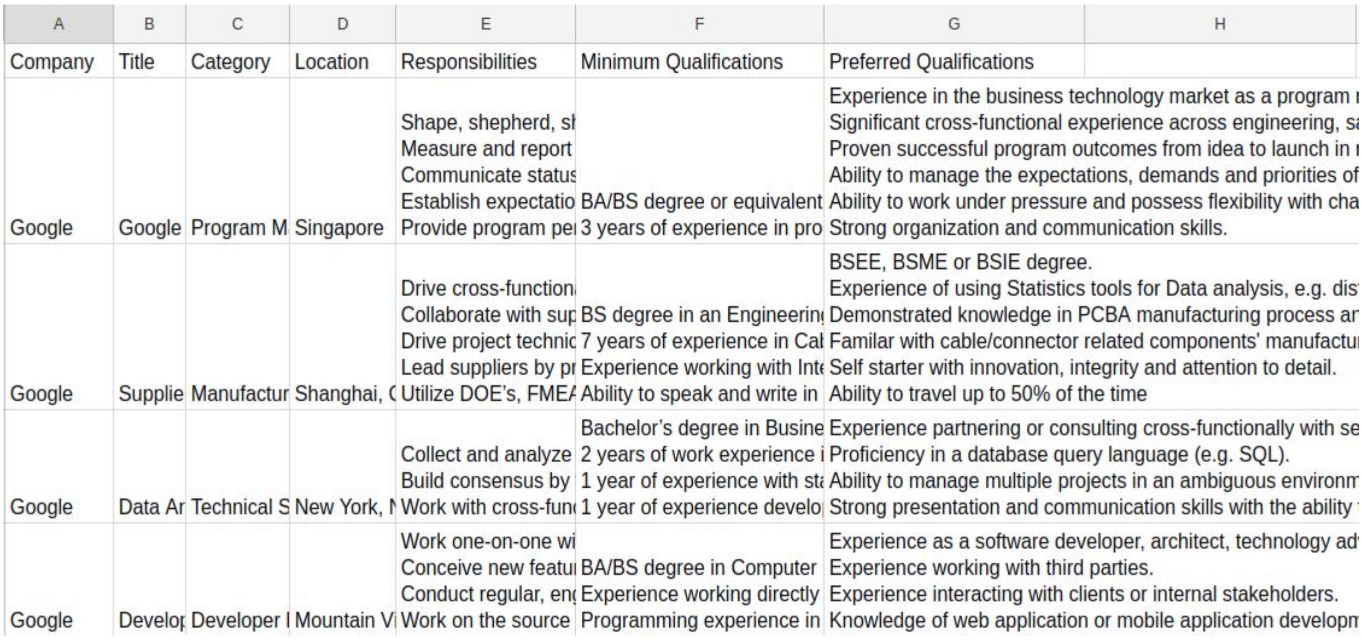
We obtain two datasets from kaggle, which are Google jobs skills and Github Repos, respectively. The first one shows all job openings of Google and the corresponding description, including job category, location, responsibility, minimum qualification and prefered qualification. The other one shows the information of each Github repository, such as the programming language of it.

First, we find out the frequent itemsets of Google jobs data by Eclat algorithm. By analyzing these frequent itemsets, we can know what ability Google highly demands (e.g. programming language, academic background and professional experience). After that, we find out the frequent itemsets of Github repositories by Eclat algorithm. These frequent itemsets can show us which programming language is popular, which means a lot of people able to write it. Finally, we compare the results of mining the data of Google jobs and Github repositories to check whether they matches.

4. Data source:

Google Job Skills Dataset (Total 1.2k records)

https://www.kaggle.com/niyamatalmass/google-job-skills



A	B	C	D	E	F	G	H
Company	Title	Category	Location	Responsibilities	Minimum Qualifications	Preferred Qualifications	
Google	Google	Program M	Singapore	Shape, shepherd, sl Measure and report Communicate status Establish expectatio Provide program pe	BA/BS degree or equivalent 3 years of experience in pro	Experience in the business technology market as a program r Significant cross-functional experience across engineering, sa Proven successful program outcomes from idea to launch in r Ability to manage the expectations, demands and priorities of Ability to work under pressure and possess flexibility with cha Strong organization and communication skills.	
Google	Supplie	Manufactur	Shanghai, (	Drive cross-function Collaborate with sup Drive project technic Lead suppliers by pr Utilize DOE's, FMEA	BS degree in an Engineerin 7 years of experience in Cal Experience working with Inte Ability to speak and write in	BSEE, BSME or BSIE degree. Experience of using Statistics tools for Data analysis, e.g. dis Demonstrated knowledge in PCBA manufacturing process ar Familar with cable/connector related components' manufactu Self starter with innovation, integrity and attention to detail. Ability to travel up to 50% of the time	
Google	Data Ar	Technical S	New York, N	Collect and analyze Build consensus by Work with cross-fun	Bachelor's degree in Busine 2 years of work experience i 1 year of experience with st 1 year of experience develo	Experience partnering or consulting cross-functionally with se Proficiency in a database query language (e.g. SQL). Ability to manage multiple projects in an ambiguous environm Strong presentation and communication skills with the ability	
Google	Develop	Developer I	Mountain Vi	Work one-on-one wi Conceive new featu Conduct regular, en Work on the source	BA/BS degree in Computer Experience working directly Programming experience in	Experience as a software developer, architect, technology ad Experience working with third parties. Experience interacting with clients or internal stakeholders. Knowledge of web application or mobile application developm	

Each row provides the information of a job opening, which includes the Title, Category, Location, Responsibility, Minimum and Preferred Qualifications. In this project, we focus on mining the data in Minimum and Preferred Qualifications, which consists of the expected academic background, professional experience and programming language.

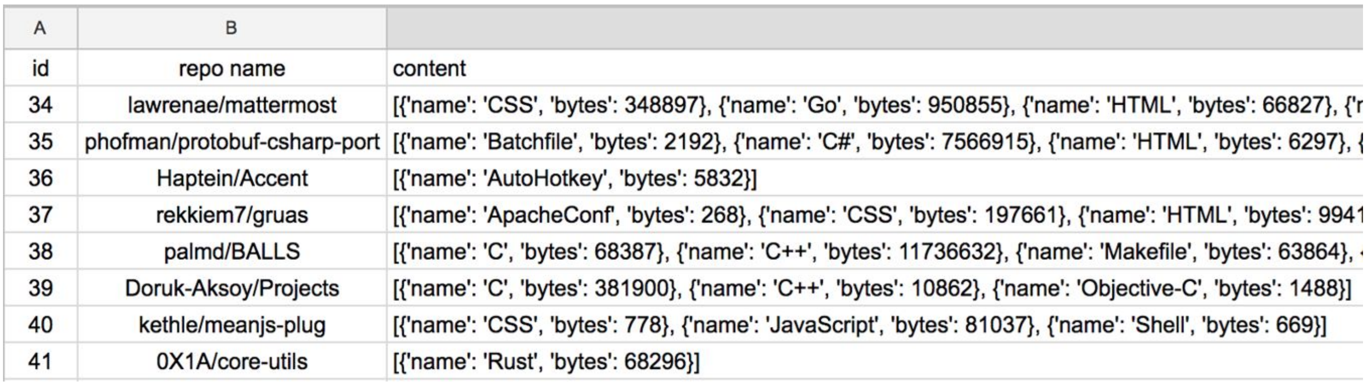
GitHub Repos Dataset (Total 3M records)

https://www.kaggle.com/github/github-repos

The GitHub Dataset is stored in a BigQuery database. Therefore, we use an open-sourced project BigQuery Helper (https://github.com/SohierDane/BigQuery_Helper) as our API Client. By the abstraction it provides, we can use SQL-like syntax to query the result we want.

The total dataset consists of all commit logs, language usage, file names, code content and licenses. Here we utilize language usage for our topic: to find the most significant skill in programming area.

After dumping the language-repo data entries, we save it as CSV file for reuse.



A	B	
id	repo name	content
34	lawrenae/mattermost	[{'name': 'CSS', 'bytes': 348897}, {'name': 'Go', 'bytes': 950855}, {'name': 'HTML', 'bytes': 66827}, {'r
35	phofman/protobuf-csharp-port	[{'name': 'Batchfile', 'bytes': 2192}, {'name': 'C#', 'bytes': 7566915}, {'name': 'HTML', 'bytes': 6297}, {
36	Haptein/Accent	[{'name': 'AutoHotkey', 'bytes': 5832}]
37	rekkiem7/gruas	[{'name': 'ApacheConf', 'bytes': 268}, {'name': 'CSS', 'bytes': 197661}, {'name': 'HTML', 'bytes': 9941
38	palmd/BALLS	[{'name': 'C', 'bytes': 68387}, {'name': 'C++', 'bytes': 11736632}, {'name': 'Makefile', 'bytes': 63864},
39	Doruk-Aksoy/Projects	[{'name': 'C', 'bytes': 381900}, {'name': 'C++', 'bytes': 10862}, {'name': 'Objective-C', 'bytes': 1488}]
40	kethle/meanjs-plug	[{'name': 'CSS', 'bytes': 778}, {'name': 'JavaScript', 'bytes': 81037}, {'name': 'Shell', 'bytes': 669}]
41	0X1A/core-utils	[{'name': 'Rust', 'bytes': 68296}]

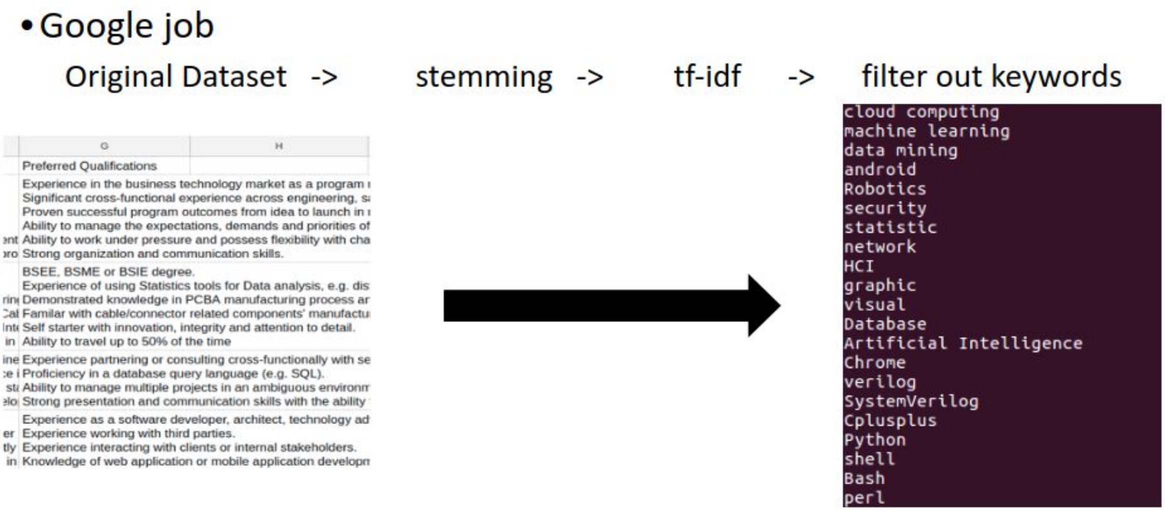
Each row represents a repository in Github and the programming languages of it are shown in content column.

5. The methodology and implementation of data processing :

In order to mine the hidden information of two data sources, we first transform them into transaction dataset just like the one provided in HW1 and HW2 by TA. Next, we find the frequent itemsets of them by Eclat algorithm and compare with each other. Therefore, we are able to check whether the results match. Our proposed method contains three steps:

1. Define the keywords
2. Transform data by matching keywords
3. Run Eclat algorithm

5-1. Define the keywords:



- Google job

Original Dataset -> stemming -> tf-idf -> filter out keywords

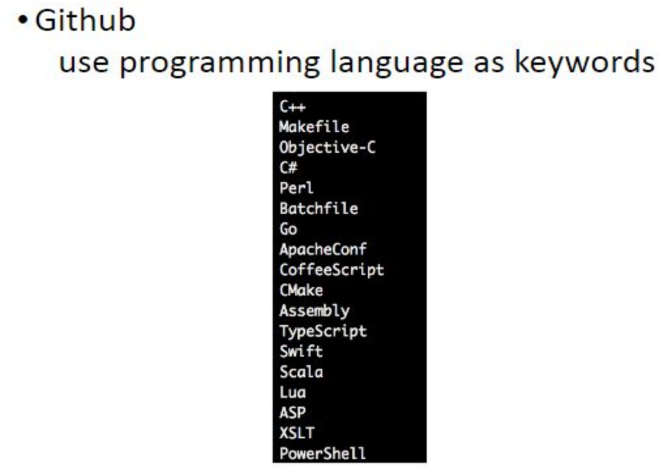
G	H
Preferred Qualifications	
Experience in the business technology market as a program Significant cross-functional experience across engineering, s Proven successful program outcomes from idea to launch in Ability to manage the expectations, demands and priorities of Ability to work under pressure and possess flexibility with cha Strong organization and communication skills.	
BSEE, BSME or BSIE degree. Experience of using Statistics tools for Data analysis, e.g. dis Demonstrated knowledge in PCBA manufacturing process ar Familar with cable/connector related components' manufactu Self starter with innovation, integrity and attention to detail. Ability to travel up to 50% of the time	
Experience partnering or consulting cross-functionally with se Proficiency in a database query language (e.g. SQL). Ability to manage multiple projects in an ambiguous environm Strong presentation and communication skills with the ability	
Experience as a software developer, architect, technology ad Experience working with third parties. Experience interacting with clients or internal stakeholders. Knowledge of web application or mobile application developn	

```
cloud computing
machine learning
data mining
android
Robotics
security
statistic
network
HCI
graphic
visual
Database
Artificial Intelligence
Chrome
verilog
SystemVerilog
Cplusplus
Python
shell
Bash
perl
```

Stemming is a common technique in text preprocessing. It will erase some common suffixes and preserve the stem of a word. For example, automate, automation, automatic will become automat. With stemming, different words with similar meaning can be transform into same token and make following analysis more convenient.

Tf-idf is another technique for text mining. In our original implementation, we try to transform dataset into keyword set by tf-idf only but we found out that there are some similar words with different part of speech. As a result of different part of speech, the tf-idf result will be inaccurate. So we combined the tf-idf with stemming and make the result of data preprocessing be more practical. After stemming and tf-idf preprocessing, the keyword set is small enough so we can figure out the real keyword we want.

For Google jobs dataset, we want to focus on popular CS topics and common programming languages. Thus, we filter out the keywords related to other topics (e.g. Accountant and Management).



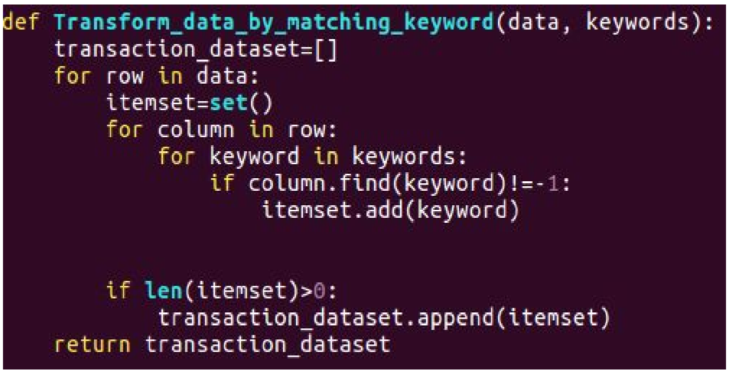
- Github
  use programming language as keywords

```
C++
Makefile
Objective-C
C#
Perl
Batchfile
Go
ApacheConf
CoffeeScript
CMake
Assembly
TypeScript
Swift
Scala
Lua
ASP
XSLT
PowerShell
```

For Github repository, we just define common programming languages as our keywords.

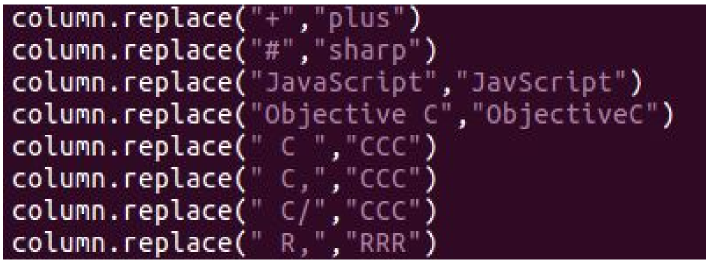
5-2. Transform data by matching keywords:

For Google Job Dataset, we only use Minimum and Preferred Qualifications column of each row. For Github Dataset, we only use Content column of each row.



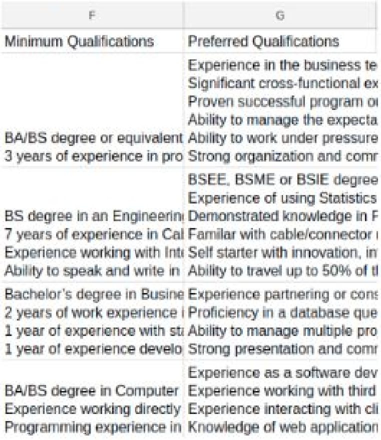
```python
def Transform_data_by_matching_keyword(data, keywords):
    transaction_dataset=[]
    for row in data:
        itemset=set()
        for column in row:
            for keyword in keywords:
                if column.find(keyword)!=-1:
                    itemset.add(keyword)


        if len(itemset)>0:
            transaction_dataset.append(itemset)
    return transaction_dataset
```

After defining keywords, we can start transforming data into transaction dataset. For each row (job), it has some columns including lots of texts. For each keyword we defined, we check whether this keyword is contained in the texts of each row. If the keyword is contained in this row, then we add the keyword into the itemset of this row. All of the work is implemented in Python. To check whether a keyword can be found in one row, we use the built-in function find() of Python. The algorithm is shown in above. However, we should do some replacements to avoid the mismatch just like the below figure. For example, C++ and C# become Cplusplus and Csharp, respectively. In addition, we replace JavaScript as JavScript since JavaScript contains Java, which may lead to mismatch. After all replacements finished, we change every letter to lowercase.

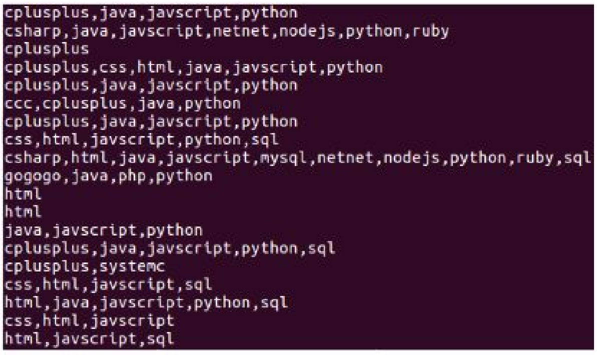


```
column.replace("+","plus")
column.replace("#","sharp")
column.replace("JavaScript","JavScript")
column.replace("Objective C","ObjectiveC")
column.replace(" C ","CCC")
column.replace(" C,","CCC")
column.replace(" C/","CCC")
column.replace(" R,","RRR")
```

- Google



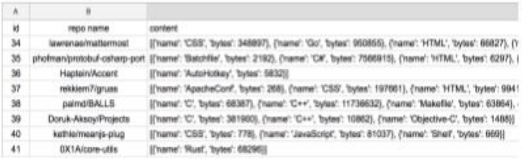
F	G
Minimum Qualifications	Preferred Qualifications
BA/BS degree or equivalent 3 years of experience in pro	Experience in the business te Significant cross-functional ex Proven successful program o Ability to manage the expecta Ability to work under pressure Strong organization and comr
BS degree in an Engineerin 7 years of experience in Cal Experience working with Int Ability to speak and write in	BSEE, BSME or BSIE degree Experience of using Statistics Demonstrated knowledge in F Familar with cable/connector Self starter with innovation, in Ability to travel up to 50% of t
Bachelor's degree in Busine 2 years of work experience i 1 year of experience with st 1 year of experience develo	Experience partnering or cons Proficiency in a database que Ability to manage multiple pro Strong presentation and comr
BA/BS degree in Computer Experience working directly Programming experience in	Experience as a software dev Experience working with third Experience interacting with cl Knowledge of web application



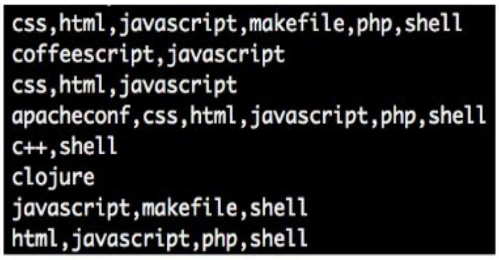
```
cplusplus,java,javscript,python
csharp,java,javscript,netnet,nodejs,python,ruby
cplusplus
cplusplus,css,html,java,javscript,python
cplusplus,java,javscript,python
ccc,cplusplus,java,python
cplusplus,java,javscript,python
css,html,javscript,python,sql
csharp,html,java,javscript,mysql,netnet,nodejs,python,ruby,sql
gogogo,java,php,python
html
html
java,javscript,python
cplusplus,java,javscript,python,sql
cplusplus,systemc
css,html,javscript,sql
html,java,javscript,python,sql
css,html,javscript
html,javscript,sql
```

Google job data is transformed into transaction dataset. In this project, we focus on popular CS topics and common programming languages. As a result, each row of the transaction dataset consists of CS topics and programming languages.

- Github



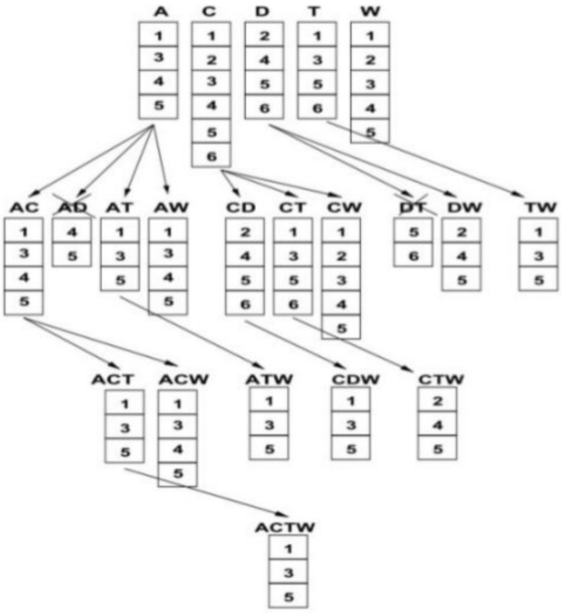
A	B	
id	repo name	content
34	lawrenas/mattermost	[{'name': 'CSS', 'bytes': 348897}, {'name': 'Go', 'bytes': 950855}, {'name': 'HTML', 'bytes': 66827}, {'
35	phofman/protobuf-csharp-port	[{'name': 'Batchfile', 'bytes': 2192}, {'name': 'C#', 'bytes': 7566915}, {'name': 'HTML', 'bytes': 6297},
36	Haptein/Accent	[{'name': 'AutoHotkey', 'bytes': 5832}]
37	rekkiem7/gruas	[{'name': 'ApacheConf', 'bytes': 268}, {'name': 'CSS', 'bytes': 197661}, {'name': 'HTML', 'bytes': 994
38	palmd/BALLS	[{'name': 'C', 'bytes': 68387}, {'name': 'C++', 'bytes': 11738632}, {'name': 'Makefile', 'bytes': 63864},
39	Doruk-Aksoy/Projects	[{'name': 'C', 'bytes': 381900}, {'name': 'C++', 'bytes': 10862}, {'name': 'Objective-C', 'bytes': 1488}]
40	kathle/meanjs-plug	[{'name': 'CSS', 'bytes': 778}, {'name': 'JavaScript', 'bytes': 81037}, {'name': 'Shell', 'bytes': 669}]
41	0X1A/core-utils	[{'name': 'Rust', 'bytes': 68296}]



```
css,html,javascript,makefile,php,shell
coffeescript,javascript
css,html,javascript
apacheconf,css,html,javascript,php,shell
c++,shell
clojure
javascript,makefile,shell
html,javascript,php,shell
```

Github repository data is transformed into transaction dataset, which consists of common programming languages.

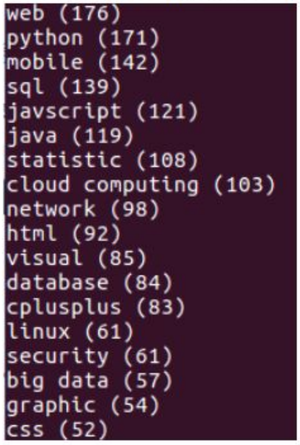
5-3. Run Eclat algorithm:



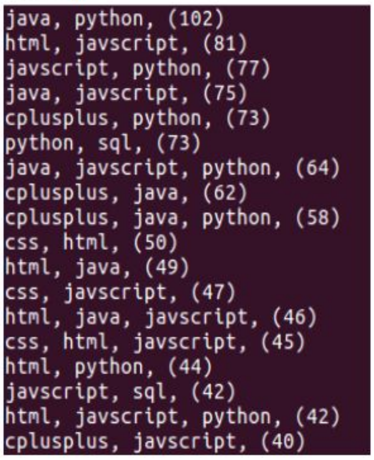
After transforming origin data into transaction dataset, we can directly apply our Eclat Python script from HW1 to find the frequent itemsets. The only difference is that the dataset provided by TA in HW1 is composed of integer while these two datasets (transformed from Google job and Github repository data source) are composed of strings.

Above is an illustration of Eclat algorithm. Python's built-in container set() is the main data structure we use. In Eclat algorithm, we first read in the transaction dataset and store the information with vertical data layout. Each keyword has a rowset to record which rows contain it. For example, if a keyword shows up in row 2, row 3 and row 5, then its rowset is equal to {2, 3, 5}. If we want to check whether the new itemset generated from these two keywords is frequent, we can just compute the intersection of the rowset of these two keywords. The support number of new itemset is the length of its rowset, which is the intersection of the rowset of these two items.

• Google (min sup : 1%)

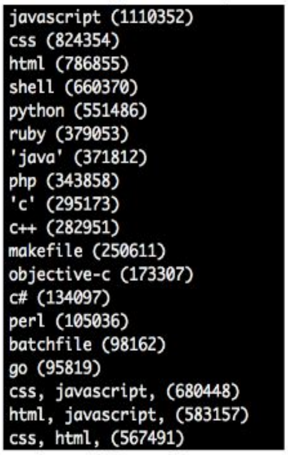


```
web (176)
python (171)
mobile (142)
sql (139)
javscript (121)
java (119)
statistic (108)
cloud computing (103)
network (98)
html (92)
visual (85)
database (84)
cplusplus (83)
linux (61)
security (61)
big data (57)
graphic (54)
css (52)
```



```
java, python, (102)
html, javscript, (81)
javscript, python, (77)
java, javscript, (75)
cplusplus, python, (73)
python, sql, (73)
java, javscript, python, (64)
cplusplus, java, (62)
cplusplus, java, python, (58)
css, html, (50)
html, java, (49)
css, javscript, (47)
html, java, javscript, (46)
css, html, javscript, (45)
html, python, (44)
javscript, sql, (42)
html, javscript, python, (42)
cplusplus, javscript, (40)
```

• Github (min sup : 2.5%)



```
javascript (1110352)
css (824354)
html (786855)
shell (660370)
python (551486)
ruby (379053)
'java' (371812)
php (343858)
'c' (295173)
c++ (282951)
makefile (250611)
objective-c (173307)
c# (134097)
perl (105036)
batchfile (98162)
go (95819)
css, javascript, (680448)
html, javascript, (583157)
css, html, (567491)
```

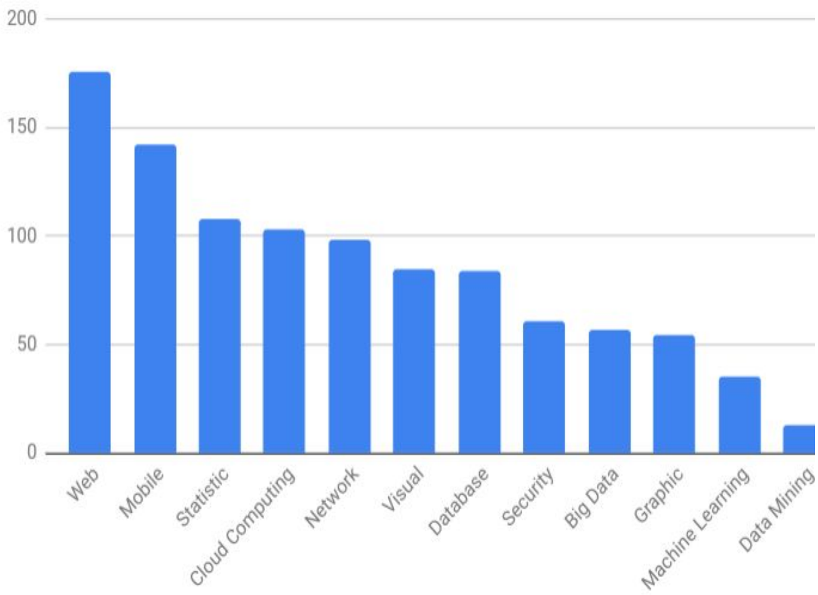
The results of mining frequent itemsets are shown in above. Each row shows a frequent itemset, which consists of the keywords we defined and the support number. Note that the number of total rows of Google jobs data is 1200 while that of Github repository is around 3 million. Therefore, we set the minimum support of the Google jobs data to 0.1%, which means the support number of each frequent itemset should be at least 2. On the other hand, the minimum support of Github repository data is set to 2.5%.

6. Our finding:

Finally, we compare the results we mined to check whether there exists difference.

6-1. Popular CS keyword:

CS keyword	Support
Web	176
Mobile	142
Statistic	108
Cloud Computing	103
Network	98
Visual	85
Database	84
Security	61
Big Data	57
Graphic	54
Machine Learning	35
Data Mining	13



Although Machine Learning and Data Mining are two popular topics in computer science, they are not highly demanded by Google as we thought. Instead, the top three popular keywords are Web, Mobile and Statistic.

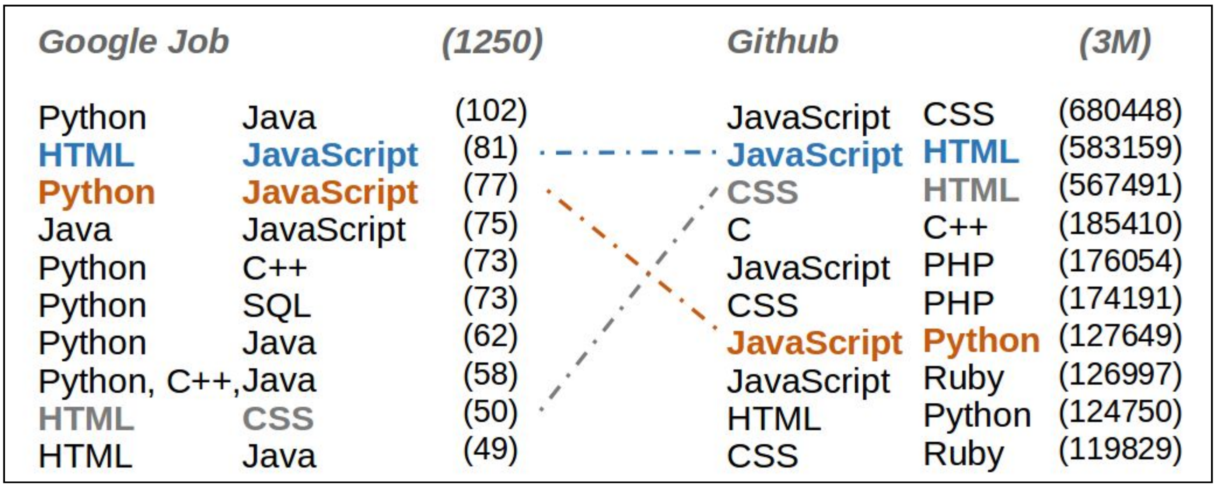
6-2. Popular programming language:

Rank	*Google Job (1250)*		*Github Repository (3M)*	
	Language	Support	Language	Support
1	Python	171	JavaScript	1110352
2	**SQL**	139	CSS	824161
3	JavaScript	121	HTML	786855
4	Java	119	Shell	641478
5	HTML	92	Python	551486
6	C++	83	Ruby	378531
7	CSS	52	Java	371812
8	**R**	41	PHP	343858
9	PHP	28	C	295173
10	**Matlab**	27	C++	281009
11	C	26	Makefile	250611
12	Ruby	20	Objective-C	190618
13	C#	19	C#	134097
14	Perl	17	Perl	103327
15	Shell	13	Go	90948

By comparing L_1 itemsets, we find that SQL, R and Matlab are highly demand in Google job dataset but not popular in Github. From Google job dataset, we further observe that 83% of SQL appear with either DataBase, Web or Statistic, 90% of R appear with Statistic and 93% of Matlab appear with Statistic. Thus, we think these three programming languages are the worthiest to invest. In addition, Statistic is also worthwhile to learn since it is highly related to SQL, R and Matlab.

On the other hand, we find that Makefile, Go and Objective-C are popular in Github but not highly demanded by Google. As a result, student shouldn't spend too much time researching these three programming languages if obtaining a job at Google is the first option of him or her.

6-3. Popular programming language combination:



Google Job		(1250)	Github		(3M)
Python	Java	(102)	JavaScript	CSS	(680448)
HTML	JavaScript	(81)	JavaScript	HTML	(583159)
Python	JavaScript	(77)	CSS	HTML	(567491)
Java	JavaScript	(75)	C	C++	(185410)
Python	C++	(73)	JavaScript	PHP	(176054)
Python	SQL	(73)	CSS	PHP	(174191)
Python	Java	(62)	JavaScript	Python	(127649)
Python, C++,	Java	(58)	JavaScript	Ruby	(126997)
HTML	CSS	(50)	HTML	Python	(124750)
HTML	Java	(49)	CSS	Ruby	(119829)

Next, we compare L_2 and L_3 itemsets to check whether popular programming language combinations match. The pair {HTML, JavaScript} is not only popular in Google job and Github but also ranks 2nd in both. Therefore, we should invest some time in HTML when learning the JavaScript.

6-4. Strong relation between popular programming languages:

We further compute the probability of appearance between popular programming languages in Google jobs dataset, the relations with large confidence are shown in below.

Relation	Confidence
C→ C++	100%
C# →Python	100%
Perl →Python	100%
Ruby → Python	100%
Shell →Python	100%
CSS →HTML	96%
C# → (Python ∧ JavaScript)	95%
PHP →Java	93%
R →SQL	93%

Obviously, we can observe that C always appears with C++, which may be due to the fact that the job requirement is C/C++ most time. In addition, Python is very often to appear with other programming languages. When one of the following four C#, Perl, Ruby and Shell appear, Python always appears.

7. Conclusion and future work:

In this project, we obtain two datasets from kaggle, which are Google jobs skills and Github Repos. Since Github Repos contains 3M records, we use programming skill to process this large dataset successfully. In addition, we use data mining technique to find some programming languages which are highly demand by Google but not popular on Github, which are SQL, R and Matlab. Therefore, it is worthwhile to learn these programming languages. Besides, We also find that these three programming languages are all related with statistic. On the other hand, there exists an interesting fact that Go is developed by Google but not highly demand by Google itself. In order to explore the actual trend of the world, we should analyze more company's job opening (e.g. Facebook, Amazon and LinkedIn). Thus, our future work is to apply the model on other job datasets.

8. Reference:

1. Google Job Skills Dataset
   https://www.kaggle.com/niyamatalmass/google-job-skills
2. GitHub Repos Dataset
   https://www.kaggle.com/github/github-repos
3. Eclat Algotirhm
   http://mlwiki.org/index.php/Eclat
4. TF-IDF
   http://mlwiki.org/index.php/TF-IDF
5. Python Numpy
   http://cs231n.github.io/python-numpy-tutorial
6. Python Scikit-Learn
   https://scikit-learn.org/stable
7. Google Jobs on Glassdoor
   https://www.glassdoor.com/Job/google-jobs-SRCH_KE0,6_IP6.htm