INFO 6210: DB Project: Jobs Database: COVID/ Corona/ Temporary Jobs Database

Title : Jobs Database Author: Karan Soni NUID: 001448528

In this project we will contribute to building a jobs database using mysql and python. We will each focus on job domains like COVID Jobs, Corona Virus Jobs and Temporary Jobs. The database is focused around specific roles, We have used an API interface to generate RAW data files containing web scrapped listings from Glassdoor and Indeed.

Project Abstract

The growing spread of coronavirus has left many Americans in fear of losing their jobs, as businesses continue to shutter and authorities tell people to stay at home. Unemployment claims multiplied, reaching an all-time high of over 3.3 million in mid-March, with analysts expecting that number to climb. Conversely, a surge of companies – not just in the health care industry – are looking to hire additional employees to meet increased demands onset by the virus.

Data Source: Our objective was to generate data from Job dashboards like Glassdoor and Indeed, as these webistes freely do not allow web scrapping, we used an automated API to fetch data from them.

APIFY interfaces:

1. Glassdoor - alexey/glassdoor-jobs-scraper
2. Indeed - hynekhruska/indeed-scraper

In [ ]:
```
# JSON parameters for Indeed Jobs with keyword 'temporary'
# POST - https://api.apify.com/v2/actor-tasks/Kp8SPPqNvsDjhZGbL/runs?token=2y8rTxfoTByNFTWZS3S7D73XD&ui=1
{
  "query": "temporary",
  "category": "Jobs",
  "location": "boston",
  "locationstate": "MA",
  "maxResults": 100000000000,
  "proxy": {
    "useApifyProxy": true
  }
}
```

In [ ]:
```
# JSON parameters for Indeed Jobs with keyword 'corona'
{
  "query": "corona",
  "category": "Jobs",
  "location": "boston",
  "locationstate": "MA",
  "maxResults": 100000000000,
  "proxy": {
    "useApifyProxy": true
  }
}
```

In [ ]:
```
# JSON parameters for Indeed Jobs with keyword 'COVID'
{
  "query": "COVID",
  "category": "Jobs",
  "location": "boston",
  "locationstate": "MA",
  "maxResults": 100000000000,
  "proxy": {
    "useApifyProxy": true
  }
}
```

In [ ]:
```
# JSON parameters for Glassdoor Jobs with keyword 'temporary'
{
  "query": "temporary",
  "category": "Jobs",
  "location": "boston",
  "locationstate": "MA",
  "maxResults": 100000000000,
  "proxy": {
    "useApifyProxy": true
  }
}
```

In [ ]:
```
# JSON parameters for Glassdoor Jobs with keyword 'corona'
{
  "query": "corona",
  "category": "Jobs",
  "location": "boston",
  "locationstate": "MA",
  "maxResults": 100000000000,
  "proxy": {
    "useApifyProxy": true
  }
}
```

In [ ]:
```
# JSON parameters for Glassdoor Jobs with keyword 'COVID'
{
  "query": "COVID",
  "category": "Jobs",
  "location": "boston",
  "locationstate": "MA",
  "maxResults": 100000000000,
  "proxy": {
    "useApifyProxy": true
  }
}
```

Export Logs: my-task alexey/glassdoor-jobs-scraper 10 check_circleSUCCEEDED 4 minutes 2020-04-23 19:17:19
search check_circleSUCCEEDED 0.1.19 2020-04-23 19:17:19 4 minutes 4096 MB 0.2799 WEB cancelFAILED 0.1.19 2020-04-23 19:15:54 a minute 4096 MB 0.0702 WEB cancelFAILED 0.1.19 2020-04-23 19:13:13 a minute 4096 MB 0.0739 WEB check_circleSUCCEEDED 0.1.19 2020-04-23 19:11:45 a few seconds 4096 MB 0.0465 WEB check_circleSUCCEEDED 0.1.19 2020-04-23 19:10:33 a few seconds 4096 MB 0.0327 WEB check_circleSUCCEEDED 0.1.19 2020-04-23 19:02:39 3 minutes 4096 MB 0.1697 WEB cancelFAILED 0.1.19 2020-04-23 19:00:59 a minute 4096 MB 0.0765 WEB check_circleSUCCEEDED 0.1.19 2020-04-23 18:44:09 a minute 4096 MB 0.0566 WEB cancelFAILED 0.1.19 2020-04-23 18:42:59 a few seconds 4096 MB 0.0415 WEB cancelFAILED 0.1.19 2020-04-23 18:42:40 2 minutes 4096 MB 0.1465 WEB my-task-1 hynekhruska/indeed-scraper 4 check_circleSUCCEEDED 16 minutes 2020-04-23 20:11:26
se check_circleSUCCEEDED 0.1.8 2020-04-23 20:11:26 16 minutes 1024 MB 0.2626 WEB check_circleSUCCEEDED 0.1.8 2020-04-23 20:00:07 10 minutes 1024 MB 0.1493 WEB access_timeTIMED-OUT 0.1.8 2020-04-23 19:38:16 5 minutes 1024 MB 0.0835 WEB access_timeTIMED-OUT 0.1.8 2020-04-23 19:24:01 5 minutes 1024 MB 0.0835 WEB

In [3]:
```
import requests
import json
import pandas as pd
import numpy as np
from urllib.request import urlopen
```

Glassdoor data

Description of Data: glassdoor1 = /Users/karansoni/Desktop/semester 2/DMDD/jobs-projects/data/glassdoor-covid.csv glassdoor2 = /Users/karansoni/Desktop/semester 2/DMDD/jobs-projects/data/glassdoor-corona.csv glassdoor3 = /Users/karansoni/Desktop/semester 2/DMDD/jobs-projects/data/glassdoor-temporary.csv

In [4]:
```
glassdoor1= pd.read_csv(r'/Users/karansoni/Desktop/semester 2/DMDD/jobs-projects/data/glassdoor-covid.csv')
glassdoor2= pd.read_csv(r'/Users/karansoni/Desktop/semester 2/DMDD/jobs-projects/data/glassdoor-corona.csv')
glassdoor3= pd.read_csv(r'/Users/karansoni/Desktop/semester 2/DMDD/jobs-projects/data/glassdoor-temporary.csv')
glassdoor1.head()
glassdoor2.head()
glassdoor3.head()
```

Out[4]:

	companyDetails/@type	companyDetails/logo	companyDe
0	Organization	https://media.glassdoor.com/sqll/371652/johnle...	JOHNLEON
1	Organization	https://media.glassdoor.com/sqll/371652/johnle...	JOHNLEON
2	Organization	https://media.glassdoor.com/sqll/118246/simon-...	Simon-Kuch &
3	Organization	https://media.glassdoor.com/sqll/5631/trader-j...	Trader Joe&
4	Organization	https://media.glassdoor.com/sqll/371652/johnle...	JOHNLEON

5 rows × 27 columns

Data Validation

In [5]:
```
glassdoor1.isnull().any()
```

Out[5]:
```
companyDetails/@type                          False
companyDetails/logo                            True
companyDetails/name                           False
companyDetails/sameAs                          True
datePosted                                    False
employerName                                  False
employerRating                                 True
id                                            False
jobDetails                                    False
jobLocation/@type                             False
jobLocation/address/@type                     False
jobLocation/address/addressCountry/@type      False
jobLocation/address/addressCountry/name       False
jobLocation/address/addressLocality           False
jobLocation/address/addressRegion             False
jobLocation/address/postalCode                 True
jobLocation/geo/@type                         False
jobLocation/geo/latitude                      False
jobLocation/geo/longitude                     False
jobTitle                                      False
salary/@type                                   True
salary/currency                                True
salary/value/@type                             True
salary/value/maxValue                          True
salary/value/minValue                          True
salary/value/unitText                          True
url                                           False
dtype: bool
```

In [6]:
```
glassdoor2.isnull().any()
```

Out[6]:
```
companyDetails/@type                          False
companyDetails/logo                            True
companyDetails/name                           False
companyDetails/sameAs                          True
datePosted                                    False
employerName                                  False
employerRating                                 True
id                                            False
jobDetails                                    False
jobLocation/@type                             False
jobLocation/address/@type                     False
jobLocation/address/addressCountry/@type      False
jobLocation/address/addressCountry/name       False
jobLocation/address/addressLocality           False
jobLocation/address/addressRegion             False
jobLocation/address/postalCode                 True
jobLocation/geo/@type                         False
jobLocation/geo/latitude                      False
jobLocation/geo/longitude                     False
jobTitle                                      False
salary/@type                                   True
salary/currency                                True
salary/value/@type                             True
salary/value/maxValue                          True
salary/value/minValue                          True
salary/value/unitText                          True
url                                           False
dtype: bool
```

In [7]:
```
glassdoor3.isnull().any()
```

Out[7]:
```
companyDetails/@type                          False
companyDetails/logo                            True
companyDetails/name                           False
companyDetails/sameAs                          True
datePosted                                    False
employerName                                  False
employerRating                                 True
id                                            False
jobDetails                                    False
jobLocation/@type                             False
jobLocation/address/@type                     False
jobLocation/address/addressCountry/@type      False
jobLocation/address/addressCountry/name       False
jobLocation/address/addressLocality           False
jobLocation/address/addressRegion             False
jobLocation/address/postalCode                 True
jobLocation/geo/@type                         False
jobLocation/geo/latitude                      False
jobLocation/geo/longitude                     False
jobTitle                                      False
salary/@type                                   True
salary/currency                                True
salary/value/@type                             True
salary/value/maxValue                          True
salary/value/minValue                          True
salary/value/unitText                          True
url                                           False
dtype: bool
```

Data Manipulation

In [47]:
```
#All the data from the 3 tables are then stored to the file glassdoor.
csv
glassdoormain= pd.read_csv(r'/Users/karansoni/Desktop/semester 2/DMDD/
jobs-projects/data/glassdoor.csv')
```

In [48]:
```
glassdoormain.head()
```

Out[48]:

	ID	jobDetails	jobTitle	url	employerRating
0	1001	FamilyAid Boston, the citys largest human serv...	Bilingual (Spanish) Licensed/MSW Social Worker	https://www.glassdoor.com/job-listing/bilingua...	NaN
1	1002	Summary\n\n This position is in the Region 1, ...	Program Analyst	https://www.glassdoor.com/job-listing/program-...	3.3
2	1003	COVID -19 Social Compliance Officer - Military...	COVID-19 Social Compliance Officer - Military ...	https://www.glassdoor.com/job-listing/covid-19...	2.1
3	1004	If you like working early mornings, being part...	GeoCoordinates	https://www.glassdoor.com/job-listing/part-tim...	3.4
4	1005	Do you have a clean driving record and enjoy n...	GeoCoordinates	https://www.glassdoor.com/job-listing/delivery...	3.4

In [49]:
```
glassdoormain.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 511 entries, 0 to 510
Data columns (total 5 columns):
ID                511 non-null int64
jobDetails        511 non-null object
jobTitle          511 non-null object
url               511 non-null object
employerRating    478 non-null float64
dtypes: float64(1), int64(1), object(3)
memory usage: 20.0+ KB
```

In [50]:
```
# Drop attributes that we don't need and check for null values
glassdoormain.drop(['employerRating'], inplace=True, axis = 1)
glassdoormain.isnull().any()
```

Out[50]:
```
ID            False
jobDetails    False
jobTitle      False
url           False
dtype: bool
```

In [51]:
```
glassdoormain.head()
```

Out[51]:

	ID	jobDetails	jobTitle	url
0	1001	FamilyAid Boston, the citys largest human serv...	Bilingual (Spanish) Licensed/MSW Social Worker	https://www.glassdoor.com/job-listing/bilingua...
1	1002	Summary\n\n This position is in the Region 1, ...	Program Analyst	https://www.glassdoor.com/job-listing/program-...
2	1003	COVID -19 Social Compliance Officer - Military...	COVID-19 Social Compliance Officer - Military ...	https://www.glassdoor.com/job-listing/covid-19...
3	1004	If you like working early mornings, being part...	GeoCoordinates	https://www.glassdoor.com/job-listing/part-tim...
4	1005	Do you have a clean driving record and enjoy n...	GeoCoordinates	https://www.glassdoor.com/job-listing/delivery...

In [46]:
```
glassdoormain['ID'].is_unique
```

Out[46]:
```
True
```

In [54]:
```
# Writing the cleaned data into a new csv file named data1.csv
glassdoormain.to_csv(r'/Users/karansoni/Desktop/semester 2/DMDD/jobs-projects/data/data1.csv')
```

Indeed data

In [52]:
```
#All the data from the 3 tables are then stored to the file indeed.csv
indeed= pd.read_csv(r'/Users/karansoni/Desktop/semester 2/DMDD/jobs-projects/data/indeed.csv')
```

In [53]:
```
indeed.head()
```

Out[53]:

	ID	jobDetails	jobTitle	
0	2001	Are you a skilled craftsman or motivated train...	SHEETMETAL INSTALLER-ESSENTIAL COVID SAFE BUS...	https://www.indeed.com/pagead/clk?mo=r&ad=-6NY...
1	2002	Description:\nThe Lowell General Hospital Alte...	Patient Care Technician, Nursing Assistant, CN...	https://www.indeed.com/pagead/clk?mo=r&ad=-6NY...
2	2003	You will receive a $300 SIGN-ON BONUS when you...	Caregiver for COVID Clients	https://www.indeed.com/rc/clk?jk=2b9e199638761...
3	2004	Type: Temporary approximately for 1 monthShift...	Health Screener	https://www.indeed.com/company/Biokinetix/job
4	2005	CNAs and HHAs and Nurses Needed to Care for Se...	CNAs pos COVID clients PPE+ Bonuses Plymouth, ...	https://www.indeed.com/pagead/clk?mo=r&ad=-6NY...

In [55]:
```
indeed.isnull().any()
```

Out[55]:
```
ID            False
jobDetails    False
jobTitle      False
url           False
dtype: bool
```

In [56]:
```
indeed.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1996 entries, 0 to 1995
Data columns (total 4 columns):
ID            1996 non-null int64
jobDetails    1996 non-null object
jobTitle      1996 non-null object
url           1996 non-null object
dtypes: int64(1), object(3)
memory usage: 62.5+ KB
```

In [58]:
```
indeed['ID'].is_unique
```

Out[58]:
```
True
```

In [59]:
```
# Writing the cleaned data into a new csv file named data1.csv
indeed.to_csv(r'/Users/karansoni/Desktop/semester 2/DMDD/jobs-projects
/data/data2.csv')
```

Combining the data files

In [79]:
```
data1= pd.read_csv(r'/Users/karansoni/Desktop/semester 2/DMDD/jobs-pro
jects/data/data1.csv')
data2= pd.read_csv(r'/Users/karansoni/Desktop/semester 2/DMDD/jobs-pro
jects/data/data2.csv')
```

In [80]:
```
data1.head()
```

Out[80]:

	ID	jobDetails	jobTitle	url
0	1001	FamilyAid Boston, the citys largest human serv...	Bilingual (Spanish) Licensed/MSW Social Worker	https://www.glassdoor.com/job-listing/bilingua...
1	1002	Summary\n\n This position is in the Region 1, ...	Program Analyst	https://www.glassdoor.com/job-listing/program-...
2	1003	COVID -19 Social Compliance Officer - Military...	COVID-19 Social Compliance Officer - Military ...	https://www.glassdoor.com/job-listing/covid-19...
3	1004	If you like working early mornings, being part...	GeoCoordinates	https://www.glassdoor.com/job-listing/part-tim...
4	1005	Do you have a clean driving record and enjoy n...	GeoCoordinates	https://www.glassdoor.com/job-listing/delivery...

In [81]:
```
data2.head()
```

Out[81]:

	ID	jobDetails	jobTitle	
0	2001	Are you a skilled craftsman or motivated train...	SHEETMETAL INSTALLER-ESSENTIAL COVID SAFE BUS...	https://www.indeed.com/pagead/clk?mo=r&ad=-6NY...
1	2002	Description:\nThe Lowell General Hospital Alte...	Patient Care Technician, Nursing Assistant, CN...	https://www.indeed.com/pagead/clk?mo=r&ad=-6NY...
2	2003	You will receive a $300 SIGN-ON BONUS when you...	Caregiver for COVID Clients	https://www.indeed.com/rc/clk?jk=2b9e199638761...
3	2004	Type: Temporary approximately for 1 monthShift...	Health Screener	https://www.indeed.com/company/Biokinetix/job
4	2005	CNAs and HHAs and Nurses Needed to Care for Se...	CNAs pos COVID clients PPE+ Bonuses Plymouth, ...	https://www.indeed.com/pagead/clk?mo=r&ad=-6NY...

In [82]:
```
alljobs = pd.concat([data1, data2], ignore_index=True)
```

In [83]:
```
alljobs.head()
```

Out[83]:

	ID	jobDetails	jobTitle	url
0	1001	FamilyAid Boston, the citys largest human serv...	Bilingual (Spanish) Licensed/MSW Social Worker	https://www.glassdoor.com/job-listing/bilingua...
1	1002	Summary\n\n This position is in the Region 1, ...	Program Analyst	https://www.glassdoor.com/job-listing/program-...
2	1003	COVID -19 Social Compliance Officer - Military...	COVID-19 Social Compliance Officer - Military ...	https://www.glassdoor.com/job-listing/covid-19...
3	1004	If you like working early mornings, being part...	GeoCoordinates	https://www.glassdoor.com/job-listing/part-tim...
4	1005	Do you have a clean driving record and enjoy n...	GeoCoordinates	https://www.glassdoor.com/job-listing/delivery...

In [84]:
```
alljobs.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2507 entries, 0 to 2506
Data columns (total 4 columns):
ID            2507 non-null int64
jobDetails    2507 non-null object
jobTitle      2507 non-null object
url           2507 non-null object
dtypes: int64(1), object(3)
memory usage: 78.4+ KB
```

In [85]:
```
alljobs = alljobs.to_csv(r'/Users/karansoni/Desktop/semester 2/DMDD/jobs-projects/data/jobs.csv')
```

We have now generated a collection of 2507 job listings extracted from Glassdoor and Indeed.