

Make My Career

Description

Searching for a new job is a huge task. Making a career change is monumental. When you want a fresh start or a change in your career, you will need help in the form of advice and an outside perspective. When you want to go beyond your current role, and explore new roles, you can achieve it by simply adding a few more skills to the ones you already have under your belt. We have a system that will help you achieve this change.

MakeMyCareer is a system that will provide suggestions to a user in the form of prospective careers, positions, skills, and resources to add to their profile based on their current skills, expectations, and the job market. Also, the user will be able to make their profile impressive and enticing to prospective recruiters in terms of their resume, and their online presence for an easy hiring process.

The System will provide -

- Prospective Positions or Profiles based on the current skills a user already has, their work experience, expectations, and the demand in the job market. These expectations can include requirements like salary, work-life balance, and ideologies.
- Based on a Target Position, the user will want to know any additional skills missing from their profile. These will be looked up and filtered using the latter parameters, and the user's target companies.
- Once the user knows the skills required, they would need trending and quality resources to learn these skills.

Team:

Shabina Singh singh.shab@northeastern.edu 002652525

Steps

Sources :

For this project, one of the top websites which could provide a substantial amount and clean information on job openings and additional information related to job searches was LinkedIn and was scraped to obtain a snapshot of data on current job openings. Also Coursera was used to scrape data on skills and resources.

Scraping:

Two python scripts were used along with BeautifulSoup library to download data from LinkedIn and Coursera directories, hit the subsequent page for each job/course and extract information from each HTML based on the right tags. This data was stored in their respective CSV files to be cleaned by a different python script.

Files:

<https://github.com/anibahs/MakeMyCareer/blob/main/FinalProject/Scripts/CourseraScraper.ipynb>

<https://github.com/anibahs/MakeMyCareer/blob/main/FinalProject/Scripts/LinkedInScraper.ipynb>

Data Cleaning:

Another python script along with pandas, numpy and matplotlib and other libraries was used to clean the data.

It ensured the data does not contain null values - empty values were converted to Nan and completed empty rows were deleted.

Any data in different languages was removed.

Multi-valued entries were removed by creating multiple entries for each attribute. In jobs data, industries, companies and skills were found to be multivalued. In courses, skills were multivalued.

Data Quality Analysis:

Data visualization in the form of countplot were created to check if the number of jobs were accurate and consistent with the industry and skills in the job market. Data completeness was ensured by removing null values and multivalues.

File -

<https://github.com/anibahs/MakeMyCareer/blob/main/FinalProject/Scripts/DataExtraction.ipynb>

Data Extraction:

Using a sample set of technical skills the model MultinomialNB was trained on this data to predict additional technical skills and some of the top skills were used to relate to the job positions.

Database Creation:

All pandas dataframes on jobs, courses, skills, and locations were formed, cleaned and also linked with relevant attributes.

Tables were created using SQL queries with all the necessary constraints for primary and foreign key relationships.

This data was then inserted into a database on MySQL Workbench through a connection with python using mysql libraries..

Normalization:

Once the database was set up, the data was found to be already in First Normal Form. It was then converted into Second Normal Form by moving data with partial dependencies (company, industry) into new tables . Location and Skills were already removed and linked with foreign keys.

It was then converted into 3rd normal form by removing redundant data from the tables and eliminating transitive dependencies.

Additional Steps:

As part of assignment 3 as well as initial steps, twitter was scraped again using beautiful soup and a database was created,, however the data was not found enough nor contained all the details pertaining to job openings or any related topics. Details such as the job poster, skill sets, salary requirements etc were missing. It was therefore not

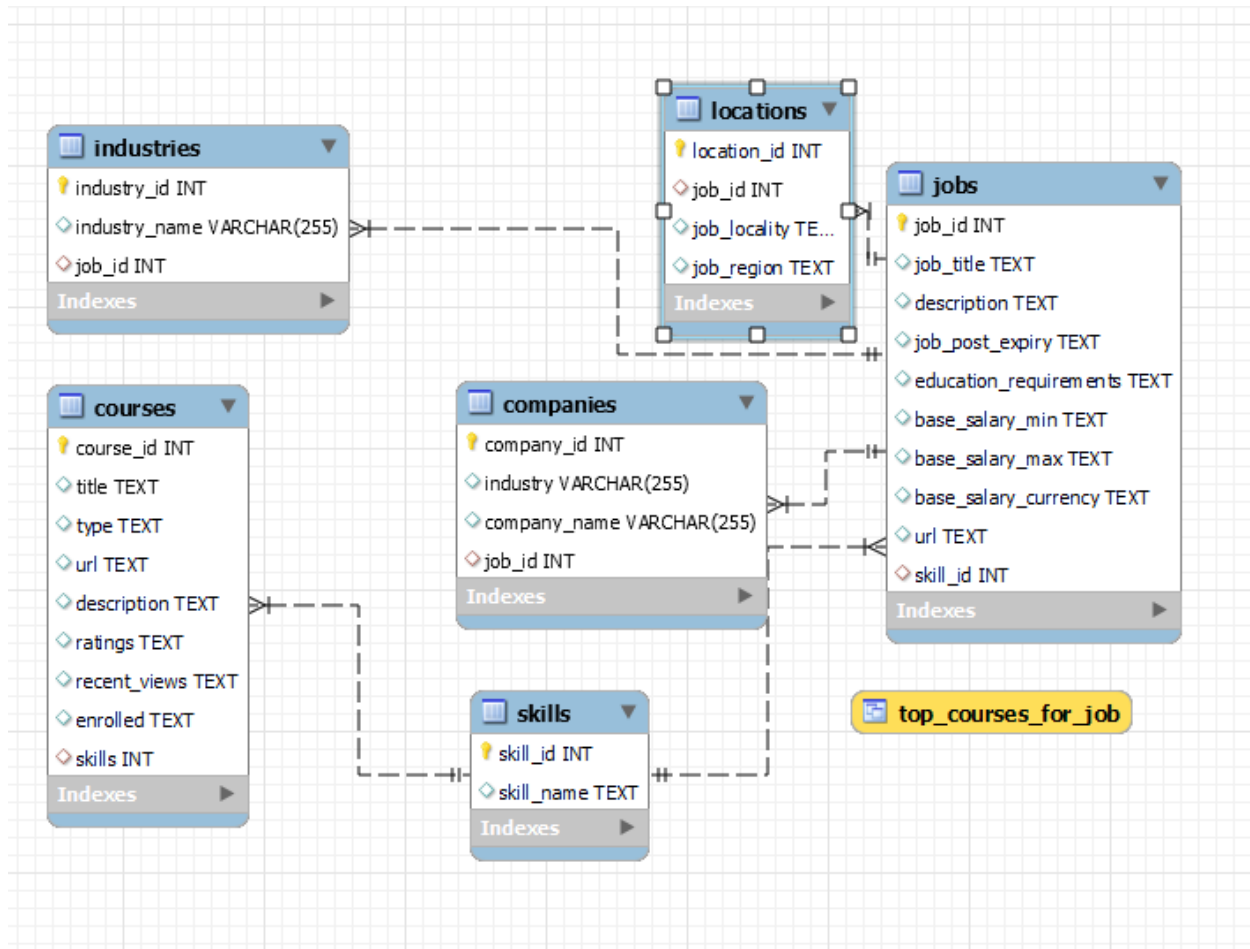
SQL scripts used -

<https://github.com/anibahs/MakeMyCareer/blob/main/FinalProject/Scripts/1NF.sql>

<https://github.com/anibahs/MakeMyCareer/blob/main/FinalProject/Scripts/2NF.sql>

<https://github.com/anibahs/MakeMyCareer/blob/main/FinalProject/Scripts/3NF.sql>

Entity Relationship Diagram



Sample Data

Jobs :

job_id	job_title	description	job_post	education_re	base_si	base_	base_ url	skill_id
0	Data Scientist Internship	iheartmediacurrent employee contingent ...	2022-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	0
1	Data Scientist Internship	iheartmediacurrent employee contingent ...	2022-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	16
2	Data Scientist Internship	iheartmediacurrent employee contingent ...	2022-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	18
3	Data Scientist Intern	task internship title interndept name risk ...	2022-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	0
4	Data Scientist Internship...	2022 75202asurion science internship or...	2023-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	0
5	Data Scientist Internship...	2022 75202asurion science internship or...	2023-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	4
6	Data Scientist Internship...	2022 75202asurion science internship or...	2023-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	7
7	Data Scientist Internship...	2022 75202asurion science internship or...	2023-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	15
8	Data Scientist Internship...	2022 75202asurion science internship or...	2023-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	16
9	Data Scientist Internship...	2022 75205asurion science internship ml...	2023-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	0
10	Data Scientist Internship...	2022 75205asurion science internship ml...	2023-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	4
11	Data Scientist Internship...	2022 75205asurion science internship ml...	2023-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	7
12	Data Scientist Internship...	2022 75205asurion science internship ml...	2023-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	15
13	Data Scientist Internship...	2022 75205asurion science internship ml...	2023-...	bachelor ...	NULL	NULL	https://www.linkedin.com/jobs/view/data-sci...	16

Courses :

course_id	title	type	url	description	ratings	recent_views	enrolled	skills
0	Accounting for Merg...	Course	https://www.coursera.org/learn/accounting-fo...	course aim assi...	NULL	9,141 recent views	NULL	4
1	Assisting Public Sect...	Course	https://www.coursera.org/learn/assist-public-s...	develop data a...	NULL	7,607 recent views	NULL	4
2	Assisting Public Sect...	Course	https://www.coursera.org/learn/assist-public-s...	develop data a...	NULL	7,607 recent views	NULL	5
3	Atenci�n prehospital...	Course	https://www.coursera.org/learn/ictus-agudo-e...	el ictus e una e...	231 ratings	3,625 recent views	4,251 already enrolled	5
4	Atenci�n prehospital...	Course	https://www.coursera.org/learn/ictus-agudo-e...	el ictus e una e...	231 ratings	3,625 recent views	4,251 already enrolled	17
5	Build, Train, and De...	Course	https://www.coursera.org/learn/ml-pipelines-bert	second course ...	115 ratings	17,709 recent views	9,797 already enrolled	4
6	Build, Train, and De...	Course	https://www.coursera.org/learn/ml-pipelines-bert	second course ...	115 ratings	17,709 recent views	9,797 already enrolled	5
7	Build, Train, and De...	Course	https://www.coursera.org/learn/ml-pipelines-bert	second course ...	115 ratings	17,709 recent views	9,797 already enrolled	15
8	Build, Train, and De...	Course	https://www.coursera.org/learn/ml-pipelines-bert	second course ...	115 ratings	17,709 recent views	9,797 already enrolled	17
9	Build, Train, and De...	Course	https://www.coursera.org/learn/ml-pipelines-bert	second course ...	115 ratings	17,709 recent views	9,797 already enrolled	18
10	Comercio, Inmigraci...	Course	https://www.coursera.org/learn/comercio-inmi...	este e el segu...	77 ratings	3,112 recent views	2,901 already enrolled	5
11	FPGA computing sys...	Course	https://www.coursera.org/learn/fpga-intro	course anyone...	224 ratings	7,362 recent views	13,249 already enrolled	4

Skills :

skill_id	skill_name
0	django
1	php
2	java
3	numpy
4	ai
5	ui
6	mysql
7	flask
8	pytorch
9	spark
10	tensorflow
11	tableau

Locations :

	location_id	job_id	job_locality	job_region
▶	0	0	Texas	United States
	1	1	Farmington Hills	MI
	2	2	Nashville	TN
	3	3	Nashville	TN
	4	4	Home	MN
	5	5	Wayne	NJ
	6	6	Parsippany	NJ
	7	7	Raleigh	NC
	8	8	Phoenix	AZ
	9	9	Los Angeles	CA
	10	10	Home	PA
	11	11	Los Angeles	CA

Companies :

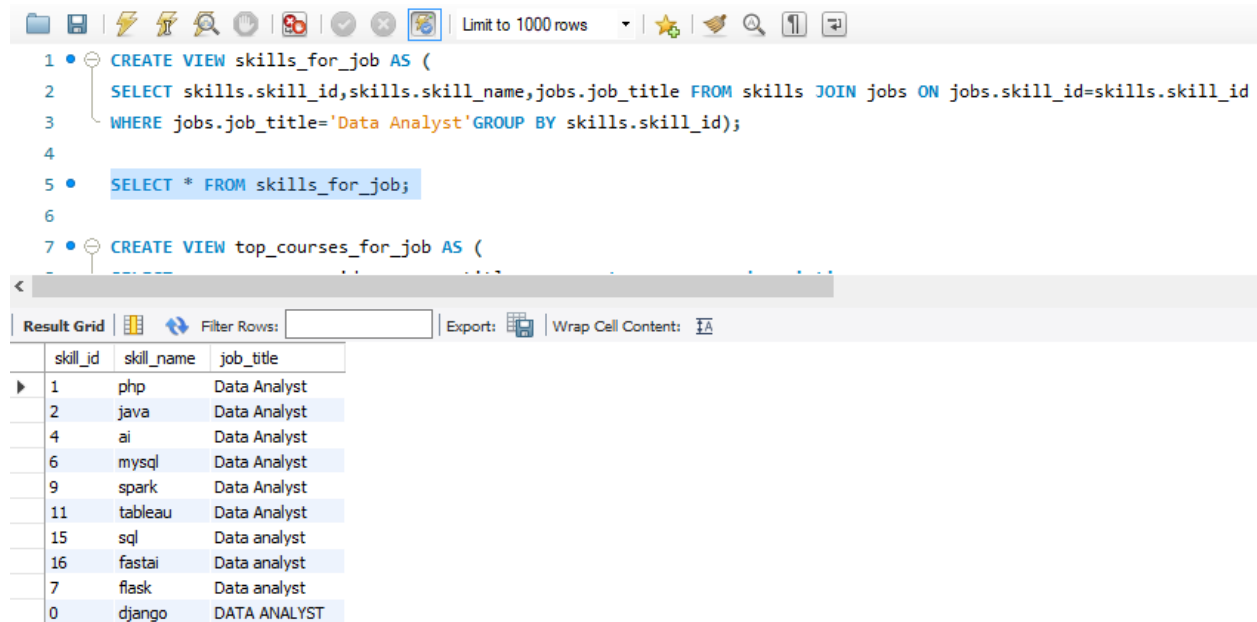
	company_name
▶	iHeartMedia
	Mercedes-Benz Financial Services USA LLC
	Asurion
	Digi-Key Electronics
	Valley Bank
	Langan Engineering & Environmental Services
	IBM
	Amgen
	Highmark Health
	BioSpace
	Dana-Farber Cancer Institute
	US 10000

Industries

	industry_name
▶	Radio and Television Broadcasting
	Financial Services
	Insurance
	Appliances and Electrical and Electronics Manuf...
	Banking
	Civil Engineering
	Computer Hardware Manufacturing
	Biotechnology Research
	Pharmaceutical Manufacturing
	Hospitals and Health Care
	Online Audio and Video Media
	Transportation and Logistics

Views based on Use Cases

Find skills required for a particular Job Position -



```
1 CREATE VIEW skills_for_job AS (  
2   SELECT skills.skill_id,skills.skill_name,jobs.job_title FROM skills JOIN jobs ON jobs.skill_id=skills.skill_id  
3   WHERE jobs.job_title='Data Analyst'GROUP BY skills.skill_id);  
4  
5 SELECT * FROM skills_for_job;  
6  
7 CREATE VIEW top_courses_for_job AS (  
8   SELECT courses.course_id, courses.title,courses.type,courses.description,  
9   courses.ratings,courses.recent_views,courses.enrolled,skills.skill_name,jobs.job_title  
10  FROM courses JOIN (skills JOIN jobs ON jobs.skill_id=skills.skill_id ) ON courses.skills=skills.skill_id  
11  WHERE jobs.job_title in ('Data Analyst') ORDER BY courses.ratings);  
12 SELECT * FROM top_courses_for_job;
```

Result Grid

	skill_id	skill_name	job_title
1	php	Data Analyst	
2	java	Data Analyst	
4	ai	Data Analyst	
6	mysql	Data Analyst	
9	spark	Data Analyst	
11	tableau	Data Analyst	
15	sql	Data analyst	
16	fastai	Data analyst	
7	flask	Data analyst	
0	django	DATA ANALYST	

Find top courses for my job position -

```
6 CREATE VIEW top_courses_for_job AS (  
7   SELECT courses.course_id, courses.title,courses.type,courses.description,  
8   courses.ratings,courses.recent_views,courses.enrolled,skills.skill_name,jobs.job_title  
9   FROM courses JOIN (skills JOIN jobs ON jobs.skill_id=skills.skill_id ) ON courses.skills=skills.skill_id  
10  WHERE jobs.job_title in ('Data Analyst') ORDER BY courses.ratings);  
11  
12 SELECT * FROM top_courses_for_job;
```

Result Grid

	course_id	title	type	description	ratings	recent_views	enrolled	skill_name	job_title
1	1316	Architecting Solution...	Course	looking get technical looking ...	NULL	51,491 recent views	2,230 already enrolled	ai	Data Analyst
	320	Advanced Data Mod...	Course	develop working knowledge f...	NULL	4,923 recent views	NULL	mysql	Data Analyst
	1295	Aprendizado de mÃi...	Course	este curso mergulha no fund...	NULL	2,221 recent views	NULL	ai	Data Analyst
	1341	Arquitecturas de Big ...	Course	el curso de arquiteturas de ...	NULL	2,002 recent views	NULL	spark	Data Analyst
	924	Advanced Data Mod...	Course	develop working knowledge f...	NULL	4,923 recent views	NULL	mysql	Data Analyst
	1310	Architecting Google ...	Course	questo curso architecting go...	NULL	NULL	NULL	ai	Data Analyst
	1311	Architecting Google ...	Course	questo curso architecting go...	NULL	NULL	NULL	ai	Data Analyst
	1313	Architecting Google ...	Course	questo curso architecting go...	NULL	NULL	NULL	ai	Data Analyst
	1437	Attracting and Sourc...	Course	shortlisting interviewing proc...	NULL	1,750 recent views	NULL	ai	Data Analyst
	1388	Arilha Networkinn Ba...	Course	arilha networkinn essential v...	NULL	1.889 recent views	NULL	ai	Data Analyst

Find jobs for set of skills in user's profile-

```

13 CREATE VIEW jobs_for_skill AS (
14 SELECT courses.course_id, courses.title,courses.type,courses.description,
15 courses.ratings,courses.recent_views,courses.enrolled,skills.skill_name,jobs.job_title
16 FROM courses JOIN (skills JOIN jobs ON jobs.skill_id=skills.skill_id )
17 ON courses.skills=skills.skill_id WHERE skills.skill_name in ('ai', 'python', 'sql')
18 GROUP BY skills.skill_name);
19
20 SELECT * FROM top_courses_for_job;

```

course_id	title	type	description	ratings	recent_views	enrolled	skill_name	job_title
1316	Architecting Solution...	Course	looking get technical looking ...	NULL	51,491 recent views	2,230 already enrolled	ai	Data Analyst
320	Advanced Data Mod...	Course	develop working knowledge f...	NULL	4,923 recent views	NULL	mysql	Data Analyst
1295	Aprendizado de mÃi...	Course	este curso mergulha no fund...	NULL	2,221 recent views	NULL	ai	Data Analyst
1341	Arquitecturas de Big ...	Course	el curso de arquiteturas de ...	NULL	2,002 recent views	NULL	spark	Data Analyst
924	Advanced Data Mod...	Course	develop working knowledge f...	NULL	4,923 recent views	NULL	mysql	Data Analyst
1310	Architecting Google ...	Course	questo curso architecting go...	NULL	NULL	NULL	ai	Data Analyst
1311	Architecting Google ...	Course	questo curso architecting go...	NULL	NULL	NULL	ai	Data Analyst
1313	Architecting Google ...	Course	questo curso architecting go...	NULL	NULL	NULL	ai	Data Analyst
1437	Attracting and Sourc...	Course	shortlisting interviewing proc...	NULL	1,750 recent views	NULL	ai	Data Analyst
1388	Aruba Networking Ba...	Course	aruba networking essential v...	NULL	1,889 recent views	NULL	ai	Data Analyst

Which courses are available on AI which provide a certificate?

```

22
23 CREATE VIEW cert_courses_for_skill AS (
24 SELECT * FROM courses JOIN skills ON courses.skills=skills.skill_id
25 WHERE skills.skill_name='ai' and type = 'Course');
26
27 SELECT * FROM cert_courses_for_skill;
28

```

course_id	title	type	url	description	ratings	recent_views	enrolled	skills	skill_id	skill_nar
0	Accounting for Mergers and Acquisitions: Foun...	Course	https://www.coursera...	course aim assisting in...	NULL	9,141 recent views	NULL	4	4	ai
1	Assisting Public Sector Decision Makers With Pol...	Course	https://www.coursera...	develop data analysis ...	NULL	7,607 recent views	NULL	4	4	ai
5	Build, Train, and Deploy ML Pipelines using BER...	Course	https://www.coursera...	second course practic...	115 ratings	17,709 recent views	9,797 alre...	4	4	ai
11	FPGA computing systems: Background knowled...	Course	https://www.coursera...	course anyone passio...	224 ratings	7,362 recent views	13,249 alr...	4	4	ai
13	Future Healthcare Payment Models Coursera	Course	https://www.coursera...	course review driver h...	34 ratings	3,911 recent views	3,719 alre...	4	4	ai
14	Improving Immunity Based on Traditional Easte...	Course	https://www.coursera...	exercise medicine regu...	61 ratings	6,906 recent views	3,718 alre...	4	4	ai
17	Incrementar - Parte 2 y Controlar Coursera	Course	https://www.coursera...	el curso explora herra...	NULL	5,330 recent views	NULL	4	4	ai
19	IntelÃ© Network Academy - Network Transfor...	Course	https://www.coursera...	welcome intelÃ© networ...	473 ratings	20,587 recent views	29,575 alr...	4	4	ai
20	Introduction to Academic Writing Coursera	Course	https://www.coursera...	welcome introduction ...	NULL	19,604 recent views	NULL	4	4	ai
21	Know Thyself - The Value and Limits of Self-Kno...	Course	https://www.coursera...	according legend inscri...	507 ratings	32,696 recent views	51,281 alr...	4	4	ai
22	Les bits et les octets des rÃ©seaux informatiq...	Course	https://www.coursera...	ce cours est conÃ©u po...	NULL	4,896 recent views	NULL	4	4	ai
24	Mandarin Chinese for Intermediate Learners: P...	Course	https://www.coursera...	mandarin chinese 1 chi...	102 ratings	4,579 recent views	9,988 alre...	4	4	ai

Which is the most trending course on mysql?


```

29 • CREATE VIEW courses_for_skill AS (
30   SELECT * FROM courses JOIN skills ON courses.skills=skills.skill_id
31   WHERE skills.skill_name='ai' ORDER BY recent_views DESC);
32
33 • SELECT * FROM courses_for_skill LIMIT 1;

```

Result Grid										
Filter Rows: <input type="text"/>										
Export: Wrap Cell Content: Fetch rows:										
course_id	title	type	url	description	ratings	recent_views	enrolled	skills	skill_id	skill_name
▶ 1003	Advanced Styling wit...	Course	https://www.cours...	used case eve...	4,297 ratings	99,922 recent views	102,646 already enrolled	4	4	ai

Scripts used :

https://github.com/anibahs/MakeMyCareer/blob/main/FinalProject/mmc_views.sql

Use Cases

1. Find skills required for a Data Analyst Job Position
2. Find courses for my job position
3. Find jobs for set of skills in user's profile
4. How many paid courses are available on AI which provide a certificate?
5. Which is the most trending course on Natural Language Processing?

Queries -

```
query1 = "SELECT * \
FROM skills JOIN jobs ON jobs.skills=skills.skill_id \
WHERE jobs.job_title='Data Analyst' GROUP BY skills.skill_id;"
query2 = "SELECT * \
FROM courses JOIN (skills JOIN jobs ON jobs.skills=skills.skill_id) \
ON courses.skills=skills.skill_id WHERE jobs.job_title in ('Data Analyst') \
GROUP BY courses.skills ORDER BY courses.ratings"
query3 = "SELECT * \
FROM jobs JOIN skills ON jobs.skills=skills.skill_id \
WHERE skills.skill_name in ('ai','python', 'sql') GROUP BY skills.skill_name; "
query4 = "SELECT * \
FROM courses JOIN skills ON courses.skills=skills.skill_id \
WHERE skills.skill_name='ai' and type = 'Professional Certificate';"
query5 = "SELECT * FROM courses JOIN skills ON courses.skills=skills.skill_id \
WHERE skills.skill_name='ai' ORDER BY recent_views DESC LIMIT 1;"
```

Twitter bot use cases

Use Case: User can look for opening for their target job position Description: User can look for opening for a position named "Engineer"

Actor: User

Precondition: User should have a valid target position name

Steps:

Actor action: User request for list of job openings for his target position.

System Responses: If the position exists, the system will return a list of job openings posted.

Post Condition: List of job openings suggested

Alternate Path: The user request is not correct and system throws an error

Error: User information is incorrect

Use Case: User can look for openings posted by their dream company handle
Description: Search for job posts posted by a particular user

Actor: User

Precondition: User should have a company name user is target

Steps:

Actor action: User request for list of job openings for his target position.

System Responses: If the company has posted job openings, the system will return the list.

Post Condition: List of job openings suggested

Alternate Path: The user request is not correct and system throws an error

Error: User information is incorrect

Use Case: User can look for openings posted within last 5 days and for a particular position
Description: Search for job posts posted within last 5 days

Actor: User

Precondition: User should have a valid target position name

Steps:

Actor action: User request for list of job openings for his target position.

System Responses: The system will return a list of job posts.

Post Condition: List of job openings suggested

Alternate Path: The user request is not correct and system throws an error

Error: User information is incorrect

Use Case: User can assess which job positions are more in demand
Description: Search for job posts for different job positions

Actor: User

Precondition: User should have a valid target position name

Steps:

Actor action: User request for list of job openings for his target position.

System Responses: The system will return a count of job posts for a position.

Post Condition: List of job openings suggested

Alternate Path: The user request is not correct and system throws an error

Error: User information is incorrect

Use Case: User can assess which companies are posting more jobs

Description: Search for job posts for job positions by different companies

Actor: User

Precondition: User should have a valid target position name and target company

Steps:

Actor action: User request for list of job openings for his target position posted by company.

System Responses: The system will return a count of job posts posted by company handle.

Post Condition: Count of job openings suggested

Alternate Path: The user request is not correct and system throws an error

Error: User information is incorrect

File: <https://github.com/anibahs/MakeMyCareer/blob/main/FinalProject/Scripts/TwitterBot.ipynb>

```

!pip install langdetect
#!pip install google_trans_new
#!pip install virtualenv
#!pip install google-cloud-translate
#!pip install deep_translator

!pip install urllib3==1.25.0
!pip install wget
!pip install pip

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: langdetect in /usr/local/lib/python3.8/dist-packages (1.0.9)
Requirement already satisfied: six in /usr/local/lib/python3.8/dist-packages (from langdetect) (1.15.0)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: urllib3==1.25.0 in /usr/local/lib/python3.8/dist-packages (1.25)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: wget in /usr/local/lib/python3.8/dist-packages (3.2)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pip in /usr/local/lib/python3.8/dist-packages (21.1.3)

from google.colab import files
from google.colab import auth
from google.colab import drive
auth.authenticate_user()
drive.mount('/content/drive')

↳ Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

import pip

for each in ["selenium","requests","beautifulsoup4"]:
    pip.main(['install', each])

WARNING: pip is being invoked by an old script wrapper. This will fail in a future version of pip.
Please see https://github.com/pypa/pip/issues/5599 for advice on fixing the underlying issue.
To avoid this problem you can invoke Python with '-m pip' instead of running pip directly.
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting selenium
  Downloading selenium-4.7.2-py3-none-any.whl (6.3 MB)
Requirement already satisfied: certifi>=2021.10.8 in /usr/local/lib/python3.8/dist-packages (from selenium) (2022.9.24)
Collecting urllib3[socks]>=1.26
  Downloading urllib3-1.26.13-py2.py3-none-any.whl (140 kB)
Collecting trio>=0.17
  Downloading trio-0.22.0-py3-none-any.whl (384 kB)
Collecting trio-websocket>=0.9
  Downloading trio_websocket-0.9.2-py3-none-any.whl (16 kB)
Collecting exceptiongroup>=1.0.0rc9
  Downloading exceptiongroup-1.0.4-py3-none-any.whl (14 kB)
Collecting async-generator>=1.9
  Downloading async_generator-1.10-py3-none-any.whl (18 kB)
Requirement already satisfied: attrs>=19.2.0 in /usr/local/lib/python3.8/dist-packages (from trio>=0.17->selenium) (22.1.0)
Requirement already satisfied: idna in /usr/local/lib/python3.8/dist-packages (from trio>=0.17->selenium) (2.10)
Requirement already satisfied: sortedcontainers in /usr/local/lib/python3.8/dist-packages (from trio>=0.17->selenium) (2.4.0)
Collecting sniffio
  Downloading sniffio-1.3.0-py3-none-any.whl (10 kB)
Collecting outcome
  Downloading outcome-1.2.0-py2.py3-none-any.whl (9.7 kB)
Collecting wsproto>=0.14
  Downloading wsproto-1.2.0-py3-none-any.whl (24 kB)
Requirement already satisfied: PySocks!=1.5.7,<2.0,>=1.5.6 in /usr/local/lib/python3.8/dist-packages (from urllib3[socks]>=1.26->
Collecting h11<1,>=0.9.0
  Downloading h11-0.14.0-py3-none-any.whl (58 kB)
Installing collected packages: sniffio, outcome, h11, exceptiongroup, async-generator, wsproto, urllib3, trio, trio-websocket, se
  Attempting uninstall: urllib3
    Found existing installation: urllib3 1.25
    Uninstalling urllib3-1.25:
      Successfully uninstalled urllib3-1.25
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the
requests 2.23.0 requires urllib3!=1.25.0,!1.25.1,<1.26,>=1.21.1, but you have urllib3 1.26.13 which is incompatible.
Successfully installed async-generator-1.10 exceptiongroup-1.0.4 h11-0.14.0 outcome-1.2.0 selenium-4.7.2 sniffio-1.3.0 trio-0.22.
WARNING: pip is being invoked by an old script wrapper. This will fail in a future version of pip.
Please see https://github.com/pypa/pip/issues/5599 for advice on fixing the underlying issue.
To avoid this problem you can invoke Python with '-m pip' instead of running pip directly.
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: requests in /usr/local/lib/python3.8/dist-packages (2.23.0)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.8/dist-packages (from requests) (3.0.4)
Collecting urllib3!=1.25.0,!1.25.1,<1.26,>=1.21.1
  Downloading urllib3-1.25.11-py2.py3-none-any.whl (127 kB)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.8/dist-packages (from requests) (2022.9.24)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.8/dist-packages (from requests) (2.10)
Installing collected packages: urllib3
  Attempting uninstall: urllib3
    Found existing installation: urllib3 1.25
    Can't uninstall 'urllib3'. No files were found to uninstall.
Successfully installed urllib3-1.26.13

```

WARNING: pip is being invoked by an old script wrapper. This will fail in a future version of pip.
 Please see <https://github.com/pypa/pip/issues/5599> for advice on fixing the underlying issue.
 To avoid this problem you can invoke Python with '-m pip' instead of running pip directly.
 Looking in indexes: <https://pypi.org/simple>, <https://us-python.pkg.dev/colab-wheels/public/simple/>

```
#scraping indeed for positions according to candidate profile
```

```
import requests
from bs4 import BeautifulSoup
```

```
import datetime, time
```

```
import wget
```

```
import csv, re
import logging
```

```
import pandas as pd
import sqlite3
import json
from pytz import timezone
```

```
from requests.auth import HTTPBasicAuth
import requests
from requests.exceptions import HTTPError
```

```
NumExpr defaulting to 2 threads.
```

```
headers = {'Accept': 'application/json'}#, "Accept-Language":"en-US,en;q=0.9"}
courses_df = pd.DataFrame(columns=["title", "type", "url", "description", "ratings", "recent_views", "enrolled"])
```

```
list_of_links = []
```

```
try:
```

```
    s = requests.Session()
```

```
    for i in range(0,10):
```

```
        coursera_url = "https://www.coursera.org/directory/courses?page="+str(i)
```

```
        time.sleep(1)
```

```
        html_text = s.get(coursera_url, headers=headers).text
```

```
        soup = BeautifulSoup(html_text, "html.parser")
```

```
        for meat in soup.find_all('a'):
```

```
            if (meat.get('href') != None):
```

```
                t = meat.get('href')
```

```
                if "learn/" in t:
```

```
                    list_of_links.append("https://www.coursera.org" + t)
```

```
    for i in range(0,10):
```

```
        coursera_url = "https://www.coursera.org/directory/specializations?page="+str(i)
```

```
        time.sleep(1)
```

```
        html_text = s.get(coursera_url, headers=headers).text
```

```
        soup = BeautifulSoup(html_text, "html.parser")
```

```
        for meat in soup.find_all('a', href=True):
```

```
            if (meat.get('href') != None):
```

```
                t = meat.get('href')
```

```
                print(t)
```

```
                if "specializations/" in t:
```

```
                    list_of_links.append("https://www.coursera.org" + t)
```

```
    for i in range(0,10):
```

```
        coursera_url = "https://www.coursera.org/directory/certificates?page="+str(i)
```

```
        time.sleep(1)
```

```
        html_text = s.get(coursera_url, headers=headers).text
```

```
        soup = BeautifulSoup(html_text, "html.parser")
```

```
        for meat in soup.find_all('a'):
```

```
            if (meat.get('href') != None):
```

```
                t = meat.get('href')
```

```
                if "professional-certificates/" in t:
```

```
                    print(t)
```

```
                    list_of_links.append(t)
```

```
    print(list_of_links)
```

```
except HTTPError as http_err:
```

```
    print(f'HTTP error occurred: {http_err}')
```

```
except Exception as err:
```

```
    print(f'Other error occurred: {err}')
```

```
/
```

```
/browse
```

```
/courses
```

```
/degrees
```

```
/enterprise
```

```
https://blog.coursera.org/
```

```
/
```

```

/degrees
/mastertrack
/certificates/learn
/career-academy/?trk_ref=globalnav
/business/?utm_campaign=website&utm_content=corp-to-home-for-enterprise&utm_medium=coursera&utm_source=header&utm_term=a-out
https://www.coursera.org/campus?utm\_campaign=website&utm\_content=corp-to-landing-for-campus&utm\_medium=coursera&utm\_source=header
/browse
/courses
/directory/specializations?authMode=login&page=1
/directory/specializations?authMode=signup&page=1
/
/directory/degrees
/directory/mastertracks
/directory/certificates
/directory/courses
/directory/partners
/directory/instructors
/directory/languages
/directory/topics
/directory/videos
/directory/queries
/directory/collections
/directory/course-reviews
/specializations/3d-printing-additive-manufacturing
/specializations/ai-for-business-wharton
/specializations/ai-foundations-for-everyone
/specializations/ai-product-management-duke
/specializations/ai-for-medicine
/specializations/ai-healthcare
/specializations/abnormal-psychology
/specializations/academic-english
/specializations/academic-skills
/specializations/accounting-data-analytics
/specializations/active-optical-devices
/specializations/adapting-career-development
/specializations/addressing-racial-health-inequity-in-healthcare
/specializations/administracion-de-empresas
/specializations/administracion-proyectos
/specializations/advanced-app-android
/specializations/advanced-data-science-ibm
/specializations/advanced-machine-learning-tensorflow-gcp
/specializations/advanced-machine-learning-tensorflow-gcp
/specializations/advanced-python-scripting-for-cybersecurity
/specializations/advanced-spacecraft-dynamics-control
/specializations/advanced-statistics-data-science
/specializations/advanced-system-security-design
/specializations/agile-development
/specializations/agile-development
/specializations/agile-leadership-change-management

```

```

course_data=[]
print(len(list_of_links))

for i in range(len(list_of_links)):
    course_details = []
    course_req = requests.Session()
    course_url=list_of_links[i]
    print(course_url)
    course_text = course_req.get(course_url, headers=headers).text
    course_soup = BeautifulSoup(course_text, "html.parser")

    course_title = course_soup.find('title', {'data-react-helmet':"true"})
    if(course_title):
        if(course_title.text):
            course_details.append(course_title.text)
            if("learn/" in course_url):
                course_details.append("Course")
            elif("professional-certificates/" in course_url):
                course_details.append("Professional Certificate")
            elif("specializations/" in course_url):
                course_details.append("Specialization")
            else:
                course_details.append(None)

        if(course_url):
            course_details.append(course_url)
        else:
            course_details.append(None)
    else:
        course_details.append(None)
        course_details.append(None)

    recent_views = course_soup.find('div',{'class':'content-inner'})
    if(recent_views):
        course_details.append(recent_views.text)

```

```

else:
    course_details.append(None)

course_rating = course_soup.find('span', {'data-test': 'ratings-count-without-asterisks'})
if(course_rating):
    course_details.append(course_rating.text)
else:
    course_details.append(None)

recent_views = course_soup.find('div',{'class': '_bd90rg'})
if(recent_views):
    course_details.append(recent_views.text)
else:
    course_details.append(None)

enrolled_for_course = course_soup.find('div',{'class': '_1fpia2'})
if(enrolled_for_course):
    course_details.append(enrolled_for_course.text)
else:
    course_details.append(None)

course_data.append(course_details)
time.sleep(1)

print(course_data)
1630
https://www.coursera.org/learn/accounting-for-ma-1
https://www.coursera.org/learn/add-gore-game-unity
https://www.coursera.org/learn/ragdoll-effect-unity
https://www.coursera.org/learn/financial-engineering-advancedtopics
https://www.coursera.org/learn/approve-social-media-zapier-trello
https://www.coursera.org/learn/aprendizaje-automatico-sin-codigo-azureml-designer
https://www.coursera.org/learn/assist-public-sector-decision-makers-through-policy-analysis
https://www.coursera.org/learn/ictus-agudo-escala-race
https://www.coursera.org/learn/automate-blog-advertisements-zapier
https://www.coursera.org/learn/bases-datos-nosql-azure
https://www.coursera.org/learn/basic-desc-r-cmdr
https://www.coursera.org/learn/build-a-social-media-presence-for-your-business-using-canva
https://www.coursera.org/learn/build-social-awareness-content-for-twitter-with-canva
https://www.coursera.org/learn/ml-pipelines-bert
https://www.coursera.org/learn/cluster-analysis-rcmdr
https://www.coursera.org/learn/comercio-inmigracion-tipos-de-cambio
https://www.coursera.org/learn/crea-contenido-e-learning-canva-estudiantes
https://www.coursera.org/learn/create-a-budget-with-google-sheets
https://www.coursera.org/learn/create-attractive-infographics-with-piktochart
https://www.coursera.org/learn/creating-a-quiz-game-using-vanilla-javascript
https://www.coursera.org/learn/creation-identite-visuelle-impressionnant-utilisant-canva
https://www.coursera.org/learn/creer-des-publicites-sur-instagram
https://www.coursera.org/learn/creer-contenu-e-learning-etudiants-canva
https://www.coursera.org/learn/curso-completo-spark-databricks
https://www.coursera.org/learn/debugging-nodejs-vscode
https://www.coursera.org/learn/descubre-diferentes-formatos-anuncios-facebook
https://www.coursera.org/learn/edita-fotos-redes-sociales-canva
https://www.coursera.org/learn/enhance-organizational-communications-with-slack
https://www.coursera.org/learn/fpga-intro
https://www.coursera.org/learn/healthcare-payment-models
https://www.coursera.org/learn/gestalte-shop-facebook-shops-canva
https://www.coursera.org/learn/tanseq-el-mokhasas-w-el-sharty-fi-microsoft-excel
https://www.coursera.org/learn/aljamea-bayn-ajzaa-motaadida-min-albayanat-fi-sql
https://www.coursera.org/learn/improve-business-performance-with-google-forms-ar
https://www.coursera.org/learn/kitabab-mujaz-ibdaie-li-tahdid-almasharih-alibdaiya-ala-mail-chimp
https://www.coursera.org/learn/get-started-facebook-audience-insights
https://www.coursera.org/learn/get-shape-combine-and-merge-the-datasets-using-power-bi
https://www.coursera.org/learn/how-combine-shapes-adobe-illustrator
https://www.coursera.org/learn/how-to-create-social-media-graphics-in-canva
https://www.coursera.org/learn/rapid-api-postman
https://www.coursera.org/learn/how-edit-improve-facebook-ad
https://www.coursera.org/learn/it-security
https://www.coursera.org/learn/sjtuyoga
https://www.coursera.org/learn/impulsando-la-transformacin-del-marketing-digital
https://www.coursera.org/learn/incrementar---parte-2-y-controlar
https://www.coursera.org/learn/network-transformation-101
https://www.coursera.org/learn/introduccin-facebook-audience-insights
https://www.coursera.org/learn/introduction-to-academic-writing
https://www.coursera.org/learn/javascript-while-loop
https://www.coursera.org/learn/know-thyself-the-examined-life
https://www.coursera.org/learn/utilisation-optimal-linkedin
https://www.coursera.org/learn/les-bits-et-les-octets-des-reseaux-informatiques
https://www.coursera.org/learn/linux-tiknuluja-almaelumat-mae
https://www.coursera.org/learn/database-creation-and-modeling-using-mysqlworkbench
https://www.coursera.org/learn/machine-learning-interpretable-lime
https://www.coursera.org/learn/mandarin-chinese-intermediate-learners-1

```

```

courses_df = pd.DataFrame(course_data,
                           columns=["title", "type", "url", "description", "ratings", "recent_views", "enrolled"])

```



```
courses_df.dropna(how='all')
```

	title	type	url	description	ratings	recent_views	enrolle
0	Accounting for Mergers and Acquisitions: Foun...	Course	https://www.coursera.org/learn/accounting-for-...	This course aims at assisting you in interpret...	None	8,851 recent views	Non
1	Add Gore to Your Game in Unity	Course	https://www.coursera.org/learn/add-gore-game-u...	None	None	None	Non
2	Add Ragdoll Effect to a Character in Unity	Course	https://www.coursera.org/learn/ragdoll-effect-...	None	None	None	Non
3	Advanced Topics in Derivative Pricing Coursera	Course	https://www.coursera.org/learn/financial-engin...	This course discusses topics in derivative pri...	None	12,933 recent views	3,61 alreac enrolle
Approve							

```
courses_df
```

	title	type	url	description	ratings	recent_views	enrolle
0	Accounting for Mergers and Acquisitions: Foun...	Course	https://www.coursera.org/learn/accounting-for-...	This course aims at assisting you in interpret...	None	8,851 recent views	Non
1	Add Gore to Your Game in Unity	Course	https://www.coursera.org/learn/add-gore-game-u...	None	None	None	Non
2	Add Ragdoll Effect to a Character in Unity	Course	https://www.coursera.org/learn/ragdoll-effect-...	None	None	None	Non
3	Advanced Topics in Derivative Pricing Coursera	Course	https://www.coursera.org/learn/financial-engin...	This course discusses topics in derivative pri...	None	12,933 recent views	3,61 alreac enrolle
Approve							

```
from langdetect import detect
#from deep_translator import GoogleTranslator

for i in courses_df.index:
    if(courses_df['title'][i]):
        if(detect(courses_df['title'][i])!='en'):
            courses_df.drop(index=[i],axis=0)
    elif(courses_df['description'][i]):
```

12/16/22, 11:28 PM

CourseraScraper.ipynb - Colaboratory

```
if(detect(courses_df['description'][i])!='en'):
    courses_df.drop(index=[i],axis=0)
```

courses_df

	title	type	url	description	ratings	recent_views	enrolle
0	Accounting for Mergers and Acquisitions: Foun...	Course	https://www.coursera.org/learn/accounting-for-...	This course aims at assisting you in interpret...	None	8,851 recent views	Non
1	Add Gore to Your Game in Unity	Course	https://www.coursera.org/learn/add-gore-game-u...	None	None	None	Non
2	Add Ragdoll Effect to a Character in Unity	Course	https://www.coursera.org/learn/ragdoll-effect-...	None	None	None	Non
3	Advanced Topics in Derivative Pricing Coursera	Course	https://www.coursera.org/learn/financial-engin...	This course discusses topics in derivative pri...	None	12,933 recent views	3,61 alreac enrolle
	Approve						

```
outfile="courses_data"
outfile=outfile+'.csv'
courses_df.to_csv(outfile, sep=',', encoding='utf-8')

!cp "/content/courses_data.csv" "/content/drive/MyDrive/courses_data.csv"
```

Imports

In [610]:

```
import sys
!{sys.executable} -m pip install --ignore-installed --upgrade pip==21.3.1
import pip

!pip install pymysql
import pymysql

import pandas as pd
!pip install lxml
!pip install Corpora
!pip install requests
!pip install beautifulsoup4
!pip install TextBlob
!pip install wordcloud
import requests
from bs4 import BeautifulSoup

import nltk
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('punkt')
nltk.download('omw-1.4')
```

```
Collecting pip==21.3.1
  Using cached pip-21.3.1-py3-none-any.whl (1.7 MB)
Installing collected packages: pip
Successfully installed pip-21.3.1
Requirement already satisfied: pymysql in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (1.0.2)
Requirement already satisfied: lxml in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (4.9.1)
Requirement already satisfied: Corpora in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (1.0)
Requirement already satisfied: requests in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (2.27.1)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from requests) (1.26.13)
Requirement already satisfied: idna<4,>=2.5 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from requests) (3.4)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from requests) (2020.12.5)
Requirement already satisfied: charset-normalizer~=2.0.0 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from requests) (2.0.12)
Requirement already satisfied: beautifulsoup4 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (4.11.1)
Requirement already satisfied: soupsieve>1.2 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from beautifulsoup4) (2.3.2.post1)
Requirement already satisfied: TextBlob in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (0.17.1)
Requirement already satisfied: nltk>=3.1 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from TextBlob) (3.6.7)
Requirement already satisfied: click in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from nltk>=3.1->TextBlob) (8.0.4)
Requirement already satisfied: joblib in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from nltk>=3.1->TextBlob) (1.1.1)
Requirement already satisfied: regex>=2021.8.3 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from nltk>=3.1->TextBlob) (2022.10.31)
Requirement already satisfied: tqdm in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from nltk>=3.1->TextBlob) (4.64.1)
Requirement already satisfied: importlib-metadata in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from click->nltk>=3.1->TextBlob) (3.10.0)
Requirement already satisfied: colorama in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from click->nltk>=3.1->TextBlob) (0.4.4)
Requirement already satisfied: importlib-resources in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from tqdm->nltk>=3.1->TextBlob) (5.4.0)
Requirement already satisfied: zipp>=0.5 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from importlib-metadata->click->nltk>=3.1->TextBlob) (3.4.1)
Requirement already satisfied: typing-extensions>=3.6.4 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from importlib-metadata->click->nltk>=3.1->TextBlob) (3.7.4.3)
Requirement already satisfied: wordcloud in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (1.8.2.2)
Requirement already satisfied: matplotlib in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from wordcloud) (2.2.3)
Requirement already satisfied: pillow in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from wordcloud) (5.2.0)
Requirement already satisfied: numpy>=1.6.1 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from wordcloud) (1.19.5)
Requirement already satisfied: cycler>=0.10 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from matplotlib->wordcloud) (0.10.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from matplotlib->wordcloud) (2.4.7)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from matplotlib->wordcloud) (2.8.1)
Requirement already satisfied: pytz in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from matplotlib->wordcloud) (2021.1)
Requirement already satisfied: six>=1.10 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from matplotlib->wordcloud) (1.11.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from matplotlib->wordcloud) (1.3.1)

[nltk_data] Downloading package stopwords to
[nltk_data]   C:\Users\Shabina\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]   C:\Users\Shabina\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]   C:\Users\Shabina\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package omw-1.4 to
[nltk_data]   C:\Users\Shabina\AppData\Roaming\nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
```

True

In [611]:

```
!pip install mysql
!pip3 install mysql-connector-python
!pip install mysql-connector-python
!pip install textblob
!pip install seaborn
!pip install missingno

Collecting mysql
  Using cached mysql-0.0.3-py3-none-any.whl (1.2 kB)
Collecting mysqlclient
  Using cached mysqlclient-2.1.1.tar.gz (88 kB)
  Preparing metadata (setup.py): started
  Preparing metadata (setup.py): finished with status 'done'
Building wheels for collected packages: mysqlclient
  Building wheel for mysqlclient (setup.py): started
  Building wheel for mysqlclient (setup.py): finished with status 'error'
  Running setup.py clean for mysqlclient
Failed to build mysqlclient
Installing collected packages: mysqlclient, mysql
  Running setup.py install for mysqlclient: started
  Running setup.py install for mysqlclient: finished with status 'error'
```

```
ERROR: Command errored out with exit status 1:
  command: 'C:\Users\Shabina\.conda\envs\python-cvcourse\python.exe' -u -c 'import io, os, sys, setuptools, tokenize; sys.argv[0] =
  '''C:\Users\Shabina\AppData\Local\Temp\pip-install-5mgtccqp\mysqlclient_32cb1873604644c5956d0f99fd5b676b\setup.py'''; __file_
  _='''C:\Users\Shabina\AppData\Local\Temp\pip-install-5mgtccqp\mysqlclient_32cb1873604644c5956d0f99fd5b676b\setup.py''';f = ge
  tattr(tokenize, '''open''', open)(__file__) if os.path.exists(__file__) else io.StringIO(''''from setuptools import setup; setup
  ()''');code = f.read().replace(''''r\n''', '''\n''');f.close();exec(compile(code, __file__, '''exec'''))' bdist_wheel -d
  'C:\Users\Shabina\AppData\Local\Temp\pip-wheel-_6r92weu'
  cwd: C:\Users\Shabina\AppData\Local\Temp\pip-install-5mgtccqp\mysqlclient_32cb1873604644c5956d0f99fd5b676b\
```

In [612]:

```
import mysql.connector
```

```

In [613]:
import numpy as np
import pandas as pd
import string
from textblob import TextBlob
from nltk.corpus import stopwords
from nltk.corpus import wordnet
import os
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from wordcloud import WordCloud
import glob
import missingno as msno
import matplotlib.pyplot as plt
import seaborn as sns; sns.set()

import matplotlib.pyplot as plt
%matplotlib inline
from textblob import Word
from textblob import TextBlob, Word, Blobber
from textblob.classifiers import NaiveBayesClassifier
from textblob.taggers import NLTKTagger

import sqlite3
import re
import mysql
from textblob import TextBlob

```

```

In [614]:
def textCleaning(df):
    ## Lower case
    df['description'] = df['description'].apply(lambda x: " ".join(x.lower() for x in x.split()))

    ## remove tabulation and punctuation
    df['description'] = df['description'].str.replace('[^\w\s]', ' ')

    #remove stop words
    stop = stopwords.words('english')
    df['description'] = df['description'].apply(lambda x: " ".join(x for x in x.split() if x not in stop))

    ## Lemmatization
    df['description'] = df['description'].apply(lambda x: " ".join([Word(word).lemmatize() for word in x.split()]))

```

Read data from csv file

```
In [615]:  
  
job_df = pd.read_csv('C:\MIS_NEU\Assignments\DMDD\MakeMyCareer\Assignment4\jobs_data.csv' , keep_default_na=False)  
job_df.drop(columns = job_df.columns[0], axis = 1, inplace= True)  
job_df = job_df.dropna(how='all')  
  
textCleaning(job_df)  
aggregate = job_df.groupby(['job_title']).sum().reset_index()  
  
other_stop_words = ['junior', 'senior', 'experience', 'etc', 'job', 'work', 'company', 'technique',  
                    'candidate', 'skill', 'skills', 'language', 'menu', 'inc', 'new', 'plus', 'years',  
                    'technology', 'organization', 'ceo', 'cto', 'account', 'manager', 'data', 'scientist', 'mobile',  
                    'developer', 'product', 'revenue', 'strong']  
  
job_df['description'] = job_df['description'].apply(lambda x: " ".join(x for x in x.split() if x not in other_stop_w  
job_df.head()
```

	job_title	industry	company	description	job_locality	job_region	skills	job_post_expiry	education_requirements	ba
0	Data Scientist Internship - OR (US)	Insurance	Asurion	2022 75202asurion science internship or nashvil...	Nashville	TN		2023-01-04T03:56:32.000Z	bachelor degree	
1	Intern - Data Scientist	Medical Equipment Manufacturing	Dexcom	dexcomdexcom empowers people take control diab...	United States			2023-01-08T04:50:37.000Z	bachelor degree	
2	Data Science, Graduate Internship	Retail	Lowe's Companies, Inc.	lowe summer internship program overviewlowe 20...	Charlotte	NC		2023-01-09T14:48:39.000Z		
3	Data Scientist Internship - ML (US)	Insurance	Asurion	2022 75205asurion science internship mlnashvil...	Nashville	TN		2023-01-04T03:56:32.000Z	bachelor degree	
4	Data Science Intern	Manufacturing	Nestlé Purina North America	leader pet care industry mean ahead volume pro...	St Louis	MO		2023-01-05T20:57:11.000Z		

Train MultinomialNB for skills based on job title

In [616]:

```
## Converting text to features
vectorizer = TfidfVectorizer()
#Tokenize and build vocabulary
X = vectorizer.fit_transform(job_df.description)
y = job_df.job_title

# split data into 80% training and 20% job_df
X_train, X_job_df, y_train, y_job_df = train_test_split(X, y, test_size=0.2, random_state=109)
print("train data shape: ", X_train.shape)
print("job_df data shape: ", X_job_df.shape)

# Fit model
clf = MultinomialNB()
clf.fit(X_train, y_train)
## Predict
y_predicted = clf.predict(X_job_df)

train data shape: (2348, 33133)
job_df data shape: (587, 33133)
```


Extract Skills from job description

In [617]:

```

nltk.download('averaged_perceptron_tagger')
technical_skills = ['python', 'c', 'r', 'c++', 'java', 'hadoop', 'scala', 'flask', 'pandas', 'spark', 'scikit-learn',
                    'numpy', 'php', 'sql', 'mysql', 'css', 'mongodb', 'nltk', 'fastai', 'keras', 'pytorch', 'tensorflow',
                    'linux', 'Ruby', 'JavaScript', 'django', 'react', 'reactjs', 'ai', 'ui', 'tableau']
feature_array = vectorizer.get_feature_names()
print(feature_array)

# number of overall model features
features_numbers = len(feature_array)
## max sorted features number
n_max = int(features_numbers * 0.1)

##initialize output dataframe
output = pd.DataFrame()
for i in range(0, len(clf.classes_)):
    print("\n*****", clf.classes_[i], "*****\n")
    class_prob_indices_sorted = clf.feature_log_prob_[i, :].argsort()[::-1]
    skills = np.take(feature_array, class_prob_indices_sorted[:n_max])
    ## Extract technical skills
    top_technical_skills = list(set(technical_skills).intersection(skills))[:6]
    # transform list to string
    txt = " ".join(skills)
    blob = TextBlob(txt)
    #top 6 adjective
    top_adjectives = [w for (w, pos) in TextBlob(txt).pos_tags if pos.startswith("JJ")][:6]

    output = output.append({'job_title': clf.classes_[i],
                           'technical_skills': top_technical_skills,
                           'soft_skills': top_adjectives },
                           ignore_index=True)

```

```

[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data] C:\Users\Shabina\AppData\Roaming\nltk_data...
[nltk_data] Package averaged_perceptron_tagger is already up-to-
[nltk_data] date!

```

```

['00', '000', '00010843', '000about', '000actual', '000bdsa', '000benefits', '000check', '000city', '000comprehensive', '000do', '000duti
es', '000eligible', '000k', '000life', '000location', '000maximum', '000new', '000our', '000please', '000powered', '000qualificationsqual
ifications', '000relocation', '000reports', '000s', '000salary', '000summarythe', '000the', '000we', '001', '002', '004', '00403closing
2', '00435', '004working', '005', '00am', '00base', '00if', '00never', '00note', '00pay', '00pm', '00pmwill', '00posting', '00relocatio
n', '00required', '00requirements', '00salary', '00target', '00the', '00what', '01', '01434duration', '01752', '019', '01asalary', '02',
'02129primary', '021fax', '02210', '03', '030', '032', '04', '0433', '0455', '04clearance', '04location', '05', '0530direct', '0537', '05
778indoj', '057union00', '06', '0644', '06511203', '06511work', '0657', '06945immediate', '07', '0727', '0740employment', '079', '07936de
scription', '07completion', '07location', '07manager', '08', '0849equity', '08540', '08540tel', '08648', '08837or', '08854', '08vmware',
'09', '09completion', '09research', '0a2f', '0ability', '0at', '0candidate', '0experience', '0for', '0gxeonjv0sw', '0ideal', '0preferrede
xperience', '0pursuant', '0this', '10', '100', '1000', '100000', '1000000705hiring', '100000preferred', '1000relocation', '1001415', '100
6', '100k', '100lbs', '100m', '100msown', '100x', '101', '1012', '101319', '102', '1020', '10281', '103', '104', '105', '105579', '106',
'10601project', '107', '1083', '109', '1099', '1099work', '10development', '10experience', '10g', '10k', '10kemployee', '10m', '10mm', '1
0summary', '10th', '10trn', '10x', '11', '110', '1100', '110000', '1101429_rr00066673', '1101429_rr00066673nyu', '11030468', '1108', '110
8cal', '110k', '111', '1111', '1111systems', '112', '11246', '113', '114', '115', '115850', '115871', '115k', '115kinterview', '1172', '1
1720218', '118', '119', '11935', '11academic', '11automotive', '11th', '11a', '12', '120', '1200', '120000', '120k', '120kfull', '121',
'122', '122577', '122657', '123', '124', '1242', '125', '125562', '12564', '125m', '126', '127', '128', '128334', '128533', '129', '12911
7', '129433', '129547', '12months', '12swhy', '12th', '12yrsrequired', '13', '130', '130559', '130k', '130k10', '131', '13105activityrece
ives', '132', '1324b', '132522', '1328', '133', '13485', '135', '136', '137', '138', '139', '13educationin', '13th', '14', '140', '1401

```

In [618]:

output.head()

	job_title	soft_skills	technical_skills
0	DevOps Intern	[industrial, intern, evident, strive, enormous...	[fastai]
1	(Level Up) Software Engineer I	[android, digital, postman, swift, native, xcu...	[java, mysql, react, fastai, php]
2	(USA) 2023 Full Time: Sam's Club Data Scientist	[ar, statistic, appropriate, simple, related, ...	[spark, tensorflow, fastai, scala, python]
3	1161 - Test Engineer Intern	[descriptionsummary, skillsreasonable, problem...	[fastai]
4	2023 Intern - Software Engineer	[intern, exceptional, digital, apps, realize, ...	[fastai, java, python]

Extract Locality information from job posts

In [619]:

```

locality_df = job_df[["job_locality", "job_region"]].copy().reset_index()
#locality_df = locality_df.groupby(["job_locality", "job_region"]).agg({'index': tuple}).reset_index()
locality_df['index'] = locality_df['index'].astype('str')
locality_df = locality_df.rename(columns = {'index': 'job_id'})
locality_df.head()

```

	job_id	job_locality	job_region
0	0	Nashville	TN
1	1	United States	
2	2	Charlotte	NC
3	3	Nashville	TN
4	4	St Louis	MO

Create skills dataframe

In [622]:

```

list_of_skills = [item for sublist in list(output['technical_skills']) for item in sublist]

unique_skills = set(list_of_skills)
skills_df = pd.DataFrame(list(unique_skills),
                          columns=["skill_name"])
skills_df.reset_index()
skills_df.head()

```

	skill_name
0	java
1	spark
2	pytorch
3	sql
4	php

Generate final Jobs Dataframe

Modify Jobs dataframe to add entries per multivalues in skills

```
In [623]:  
  
import numpy as np  
from itertools import chain  
fnew_df_backup=job_df.copy()  
bad_delimiter_list=['Appliances, Electrical, and Electronics Manufacturing',  
                    'Technology, Information and Internet','Technology, Information and Media',  
                    'Movies, Videos, and Sound']  
  
fnew_df_backup.industry[fnew_df_backup.industry == 'Appliances, Electrical, and Electronics Manufacturing'] = 'Appli  
fnew_df_backup.industry[fnew_df_backup.industry == 'Technology, Information and Internet'] = 'Technology and Informa  
fnew_df_backup.industry[fnew_df_backup.industry == 'Technology, Information and Media'] = 'Technology and Informatio  
fnew_df_backup.industry[fnew_df_backup.industry == 'Movies, Videos, and Sound'] = 'Movies, Videos, and Sound'  
rm_mv_industry_df=fnew_df_backup  
  
# return list from series of comma-separated strings  
def chainer(s):  
    return list(chain.from_iterable(s.str.split(',')))  
rm_mv_industry_df.head()
```

	job_title	industry	company	description	job_locality	job_region	skills	job_post_expiry	education_requirements	ba
0	Data Scientist Internship - OR (US)	Insurance	Asurion	2022 75202asurion science internship ornashvil...	Nashville	TN		2023-01-04T03:56:32.000Z	bachelor degree	
1	Intern - Data Scientist	Medical Equipment Manufacturing	Dexcom	dexcomdexcom empowers people take control diab...	United States			2023-01-08T04:50:37.000Z	bachelor degree	
2	Data Science, Graduate Internship	Retail	Lowe's Companies, Inc.	lowe summer internship program overviewlowe 20...	Charlotte	NC		2023-01-09T14:48:39.000Z		
3	Data Scientist Internship - ML (US)	Insurance	Asurion	2022 75205asurion science internship mlnashvil...	Nashville	TN		2023-01-04T03:56:32.000Z	bachelor degree	
4	Data Science Intern	Manufacturing	Nestlé Purina North America	leader pet care industry mean ahead volume pro...	St Louis	MO		2023-01-05T20:57:11.000Z		

```
In [624]:
type(rm_mv_industry_df)
####remove multivalues for industry column

# calculate lengths of splits
lens = rm_mv_industry_df['industry'].str.split(',').map(len)
# create new dataframe, repeating or chaining as appropriate
res_df = pd.DataFrame({'job_title': np.repeat(rm_mv_industry_df['job_title'], lens),
                      'industry': chainer(rm_mv_industry_df['industry']),
                      'company': np.repeat(rm_mv_industry_df['company'], lens),
                      'description': np.repeat(rm_mv_industry_df['description'], lens),
                      'skills': np.repeat(rm_mv_industry_df['skills'], lens),
                      'job_locality': np.repeat(rm_mv_industry_df['job_locality'], lens),
                      'job_region': np.repeat(rm_mv_industry_df['job_region'], lens),
                      'job_post_expiry': np.repeat(rm_mv_industry_df['job_post_expiry'], lens),
                      'education_requirements': np.repeat(rm_mv_industry_df['education_requirements'], lens),
                      'base_salary_min': np.repeat(rm_mv_industry_df['base_salary_min'], lens),
                      'base_salary_max': np.repeat(rm_mv_industry_df['base_salary_max'], lens),
                      'base_salary_currency': np.repeat(rm_mv_industry_df['base_salary_currency'], lens),
                      'url': np.repeat(rm_mv_industry_df['url'], lens)})

res_df.reset_index()
res_df.head()
type(res_df)

pandas.core.frame.DataFrame
```

```
In [625]:
####match skills based on job title

neeeeeee = res_df
haaooooo = output

legend = pd.merge(neeeeeee, haaooooo[["job_title","technical_skills"]], on="job_title", how="left")
legend["skills"]=legend["technical_skills"]
legend=legend.drop("technical_skills", axis=1)
legend.head()
```

	job_title	industry	company	description	skills	job_locality	job_region	job_post_expiry	education_requirements	l
0	Data Scientist Internship - OR (US)	Insurance	Asurion	2022 75202asurion science internship ornashvil...	[sql, ai, fastai]	Nashville	TN	2023-01-04T03:56:32.000Z	bachelor degree	
1	Intern - Data Scientist	Medical Equipment Manufacturing	Dexcom	dexcomdexcom empowers people take control diab...	[spark, hadoop, sql, fastai, scala, python]	United States		2023-01-08T04:50:37.000Z	bachelor degree	
2	Data Science, Graduate Internship	Retail	Lowe's Companies, Inc.	lowe summer internship program overviewlowe 20...	NaN	Charlotte	NC	2023-01-09T14:48:39.000Z		
3	Data Scientist Internship - ML (US)	Insurance	Asurion	2022 75205asurion science internship mlnashvil...	[sql, ai, fastai]	Nashville	TN	2023-01-04T03:56:32.000Z	bachelor degree	
4	Data Science Intern	Manufacturing	Nestlé Purina North America	leader pet care industry mean ahead volume pro...	[sql, ai, python, fastai]	St Louis	MO	2023-01-05T20:57:11.000Z		

```
In [626]:  
  
#transpose multivalues to form multiple entries  
fnew_df = pd.DataFrame(columns=["job_title","industry","company","description","job_locality","job_region","skills",  
legend = legend.where(pd.notnull(legend), None)  
  
c=0  
for i in legend.index:  
    temp =legend.loc[i,"skills"]  
    if(temp):  
        for s in range(0,len(temp)):  
            fnew_df.loc[c]=legend.loc[i]  
            fnew_df.loc[c, "skills"]=temp[s]  
            c+=1  
  
visulize_df = fnew_df.copy()  
fnew_df.head()
```

	job_title	industry	company	description	job_locality	job_region	skills	job_post_expiry	education_requirements	ba
0	Data Scientist Internship - OR (US)	Insurance	Asurion	2022 75202asurion science internship ornashvil...	Nashville	TN	sql	2023-01-04T03:56:32.000Z	bachelor degree	
1	Data Scientist Internship - OR (US)	Insurance	Asurion	2022 75202asurion science internship ornashvil...	Nashville	TN	ai	2023-01-04T03:56:32.000Z	bachelor degree	
2	Data Scientist Internship - OR (US)	Insurance	Asurion	2022 75202asurion science internship ornashvil...	Nashville	TN	fastai	2023-01-04T03:56:32.000Z	bachelor degree	
3	Intern - Data Scientist	Medical Equipment Manufacturing	Dexcom	dexcomdexcom empowers people take control diab...	United States		spark	2023-01-08T04:50:37.000Z	bachelor degree	
4	Intern - Data Scientist	Medical Equipment Manufacturing	Dexcom	dexcomdexcom empowers people take control diab...	United States		hadoop	2023-01-08T04:50:37.000Z	bachelor degree	

```
In [627]:
job_new_df=fnew_df.copy()
job_new_df = job_new_df.drop("job_id", axis=1, errors='ignore')
job_new_df=job_new_df.rename(columns={'skills': 'skill_name'})
job_new_df=job_new_df.reset_index()
job_new_df = job_new_df.rename(columns = {'index':'job_id'})
job_new_df.head()
```

	job_id	job_title	industry	company	description	job_locality	job_region	skill_name	job_post_expiry	education_requ
0	0	Data Scientist Internship - OR (US)	Insurance	Asurion	2022 75202asurion science internship ornashvil...	Nashville	TN	sql	2023-01-04T03:56:32.000Z	bachelor degree
1	1	Data Scientist Internship - OR (US)	Insurance	Asurion	2022 75202asurion science internship ornashvil...	Nashville	TN	ai	2023-01-04T03:56:32.000Z	bachelor degree
2	2	Data Scientist Internship - OR (US)	Insurance	Asurion	2022 75202asurion science internship ornashvil...	Nashville	TN	fastai	2023-01-04T03:56:32.000Z	bachelor degree
3	3	Intern - Data Scientist	Medical Equipment Manufacturing	Dexcom	dexcomdexcom empowers people take control diab...	United States		spark	2023-01-08T04:50:37.000Z	bachelor degree
4	4	Intern - Data Scientist	Medical Equipment Manufacturing	Dexcom	dexcomdexcom empowers people take control diab...	United States		hadoop	2023-01-08T04:50:37.000Z	bachelor degree

```
In [628]:
skills_df = skills_df.drop("skill_id", axis=1, errors='ignore')
skills_df=skills_df.reset_index()
skills_df = skills_df.rename(columns = {'index':'skill_id'})
skills_df.head()
```

	skill_id	skill_name
0	0	java
1	1	spark
2	2	pytorch
3	3	sql
4	4	php

```
In [629]:
#find skill id based on skill name
job_new_df= pd.merge(job_new_df, skills_df[['skill_name','skill_id']], on="skill_name", how="left")
job_new_df=job_new_df.drop("skill_name",axis=1)
job_new_df["skill_id"]= job_new_df["skill_id"].fillna(0)
job_new_df["skill_id"] = job_new_df["skill_id"].astype(int)
```

```
In [680]:
#cleaned data frame for jobs
job_new_df["job_post_expiry"]=job_new_df["job_post_expiry"].astype('datetime64[ns]')
type(job_new_df.job_post_expiry[0])

pandas._libs.tslibs.timestamps.Timestamp
```

```
In [ ]:
```

Fetch data for courses

```
In [631]:
courses_df = pd.read_csv('C:\MIS_NEU\Assignments\DMDD\MakeMyCareer\Assignment4\courses_data.csv' , keep_default_na=F

courses_df.drop(columns = courses_df.columns[0], axis = 1, inplace= True)
courses_df = courses_df.dropna(how='all')
textCleaning(courses_df)
```

```
In [632]:
#fetch skill ids based on skill name from extracted skill
courses_m_df = pd.DataFrame(columns=["title","type","url","description","ratings","recent_views","enrolled","skills"]
c=0

for j in courses_df.index:
    for o in skills_df.index:
        skill = skills_df.loc[o, 'skill_name']
        if(skill in courses_df.loc[j, 'description']):
            courses_m_df.loc[c]=courses_df.loc[j]
            courses_m_df.loc[c,"skills"]=o
            c+=1
```

```
In [633]:
#courses_new_df = courses_m_df.copy()
#for i in courses_m_df.index:
#    for j in skills_df.index:
#        if(courses_new_df.loc[i, "skills"]==skills_df.loc[j, "skill_name"]):
#            courses_new_df.loc[i, "skills"]=j
#courses_new_df.head()
```

Database creation from Data frames -

Creating tables for Jobs, Courses, Skills and Location:

```
In [634]:
!pip install sqlalchemy

Requirement already satisfied: sqlalchemy in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (1.4.45)
Requirement already satisfied: greenlet!=0.4.17 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from sqlalchemy) (2.0.1)
Requirement already satisfied: importlib-metadata in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from sqlalchemy) (3.10.0)
Requirement already satisfied: typing-extensions>=3.6.4 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from importlib-metadata->sqlalchemy) (3.7.4.3)
Requirement already satisfied: zipp>=0.5 in c:\users\shabina\.conda\envs\python-cvcourse\lib\site-packages (from importlib-metadata->sqlalchemy) (3.4.1)
```

```
In [736]:
from sqlalchemy import create_engine
from sqlalchemy.sql import text

db_data = 'mysql:mysqldb://' + 'root' + ':' + 'admin' + '@' + 'localhost' + ':3306/' \
          + 'mmc_backup' + '?charset=utf8mb4'
engine = create_engine(db_data)
```

```
In [ ]:

fnew_df=job_new_df.copy()
fnew_df=fnew_df.drop("job_id",axis=1, errors='ignore')
fnew_df.reset_index(inplace=True)
fnew_df = fnew_df.rename(columns = {'index':'job_id'})
fnew_df = fnew_df.rename(columns = {'description':'job_description'})
fnew_df = fnew_df.rename(columns = {'skill_id':'skill'})
fnew_df.job_id = pd.to_numeric(fnew_df.job_id)
fnew_df.skill_id = pd.to_numeric(fnew_df.skill)

#fnew_df = fnew_df.replace(np.nan, None)
#fnew_df.to_sql(con=conn, name='jobs', index=False, if_exists='replace')
#fnew_df.to_sql('jobs', engine, if_exists='append', index=False)
#try:
connection = pymysql.connect(host='localhost',
                             user='root',
                             password='admin',
                             db='mmc_backup')

cursor=connection.cursor()
for index in fnew_df.index:
    statement = "INSERT INTO jobs (job_id, job_title, industry, company, job_description, job_locality, job_region,
    params = (fnew_df.loc[index,'job_id'], fnew_df.loc[index,'job_title'],
              fnew_df.loc[index,'industry'], fnew_df.loc[index,'company'],
              fnew_df.loc[index,'job_description'], fnew_df.loc[index,'job_locality'],
              fnew_df.loc[index,'job_region'], fnew_df.loc[index,'job_post_expiry'],
              fnew_df.loc[index,'education_requirements'], fnew_df.loc[index,'base_salary_min'],
              fnew_df.loc[index,'base_salary_max'], fnew_df.loc[index,'base_salary_currency'],
              fnew_df.loc[index,'url'], fnew_df.loc[index,'skill']
              )

    try:
        cursor.execute(statement, params)
    except:
        connection.rollback()
    else:
        connection.commit()

cursor.close()
```


In [706]:

```
locality_df=locality_df.drop("location_id",axis=1, errors='ignore')
locality_df.reset_index(inplace=True)
locality_df = locality_df.rename(columns = {'index':'location_id'})
#locality_df = locality_df.replace("", np.nan)
#locality_df.to_sql('locations',con=conn,index=False, if_exists='replace')
#try:
connection = pymysql.connect(host='localhost',
                             user='root',
                             password='admin',
                             db='mmc_backup')

cursor=connection.cursor()
for index in locality_df.index:
    statement = "INSERT INTO locations(location_id, job_id, job_locality, job_region) VALUES(%s,%s,%s,%s) "
    params = (locality_df.loc[index,'location_id'], locality_df.loc[index,'job_id'], locality_df.loc[index,'job_loca
try:
    cursor.execute(statement, params)
except:
    connection.rollback()
else:
    connection.commit()
cursor.close()
```

course_id	title	type	url	course_description	ratings	recent_views	enrc
0	0	Accounting for Mergers and Acquisitions: Foun...	Course	https://www.coursera.org/learn/accounting-for-...	course aim assisting interpreting financial ac...	8,851 recent views	
1	1	Assisting Public Sector Decision Makers With ...	Course	https://www.coursera.org/learn/assist-public-s...	develop data analysis skill support public sec...	7,734 recent views	
2	2	Assisting Public Sector Decision Makers With ...	Course	https://www.coursera.org/learn/assist-public-s...	develop data analysis skill support public sec...	7,734 recent views	
3	3	AtenciÃ³n prehospitalaria del ictus agudo y s...	Course	https://www.coursera.org/learn/ictus-agudo-esc...	el ictus e una emergencia mÃ¡ dica tiempo depen...	231 ratings 3,616 recent views	4,257 alrea enrol
4	4	AtenciÃ³n prehospitalaria del ictus agudo y s...	Course	https://www.coursera.org/learn/ictus-agudo-esc...	el ictus e una emergencia mÃ¡ dica tiempo depen...	231 ratings 3,616 recent views	4,257 alrea enrol
...
1000	1000	Preparing for Google Cloud Certification: Clou...	Professional Certificate	https://www.coursera.org/professional-certific...	tensorflowbigquerygoogle cloud platformcloud c...	70 ratings 1,736 recent views	
1001	1001	Soutien des TI de Google Professional Certific...	Professional Certificate	https://www.coursera.org/professional-certific...	protocoles de rÃª seuuinfonuagiquedÃ bogagealgo...	6 ratings 4,509 recent views	
1002	1002	Soutien des TI de Google Professional Certific...	Professional Certificate	https://www.coursera.org/professional-certific...	protocoles de rÃª seuuinfonuagiquedÃ bogagealgo...	6 ratings 4,509 recent views	
1003	1003	Soutien des TI de Google Professional Certific...	Professional Certificate	https://www.coursera.org/professional-certific...	protocoles de rÃª seuuinfonuagiquedÃ bogagealgo...	6 ratings 4,509 recent views	
1004	1004	IBM Machine Learning Professional Certificate ...	Professional Certificate	https://www.coursera.org/professional-certific...	artificial intelligence ai machine learningfea...	934 ratings 42,204 recent views	5,668 alrea enrol

1005 rows x 9 columns

```

In [ ]:

courses_m_df=courses_m_df.drop("course_id",axis=1, errors='ignore')
courses_m_df.reset_index(inplace=True)
courses_m_df = courses_m_df.rename(columns = {'index':'course_id'})
courses_m_df = courses_m_df.rename(columns = {'title':'course_title'})
courses_m_df = courses_m_df.rename(columns = {'description':'course_description'})
#courses_m_df = courses_m_df.replace("", np.nan)
#courses_m_df.to_sql('courses',con=conn,index=False, if_exists='replace')
#try:
connection = pymysql.connect(host='localhost',
                             user='root',
                             password='admin',
                             db='mmc_backup')

cursor=connection.cursor()
for index in courses_m_df.index:
    statement = "INSERT INTO courses(course_id, course_title, course_description, url, ratings, recent_views, enroll
    params = (courses_m_df.loc[index, 'course_id'], courses_m_df.loc[index, 'course_title'],
              courses_m_df.loc[index, 'course_description'], courses_m_df.loc[index, 'url'],
              courses_m_df.loc[index, 'ratings'], courses_m_df.loc[index, 'recent_views'],
              courses_m_df.loc[index, 'enrolled'], courses_m_df.loc[index, 'skills'] )

    try:
        cursor.execute(statement, params)
    except:
        connection.rollback()
    else:
        connection.commit()
cursor.close()

```

```

In [ ]:

skills_df=skills_df.drop("skill_id",axis=1, errors='ignore')
skills_df.reset_index(inplace=True)
skills_df = skills_df.rename(columns = {'index':'skill_id'})
#skills_df = skills_df.replace("", np.nan)
#skills_df.to_sql(con=conn, name='skills', index=False, if_exists='replace')

connection = pymysql.connect(host='localhost',
                             user='root',
                             password='admin',
                             db='mmc_backup')

cursor=connection.cursor()
for index in skills_df.index:
    statement = "INSERT INTO skills(skill_id, skill_name) VALUES(%s,%s) "
    params = (skills_df.loc[index, 'skill_id'], skills_df.loc[index, 'skill_name'] )
    try:
        cursor.execute(statement, params)
    except:
        connection.rollback()
    else:
        connection.commit()

```

```

In [735]:

engine.dispose()
connection.close()

```

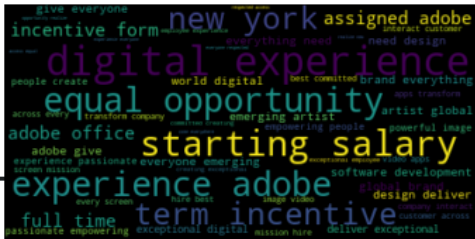
In []:

```
!cp "/content/MakeMyCareer.db" "/content/drive/MyDrive/MakeMyCareer.sql"
```

Visualizations, Data Quality and Checks

```
In [640]:
```

```
## Visualize data
jobs_list = aggregate.job_title.unique().tolist()
for j in range(0,5):
    job=jobs_list[j]
    # Start with one review:
    text = aggregate[aggregate.job_title == job].iloc[0].description
    # Create and generate a word cloud image:
    wordcloud = WordCloud().generate(text)
    # Display the generated image:
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis("off")
    plt.show()
```

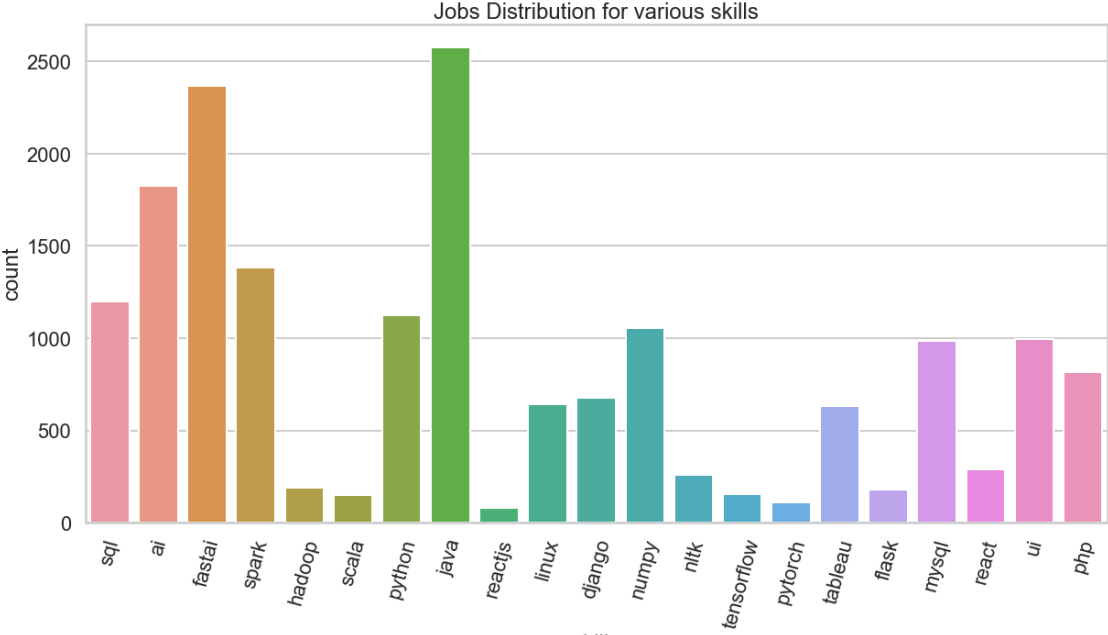


```
In [652]:  
  
sns.set_context("poster")  
sns.set_style("whitegrid")  
  
input_data = visualize_df.groupby(["industry"],as_index=False).agg(jobs = ('job_title', 'count'))  
input_data  
  
fig, ax = plt.subplots(figsize=(10, 30))  
ax = sns.countplot(y='industry', data=visualize_df)  
  
ax.set_title('Jobs Distribution through various industries')  
ax.set_xticklabels(ax.get_xticklabels(), rotation=75)  
fig.show()  
  
C:\Users\Shabina\.conda\envs\python-cvcourse\lib\site-packages\matplotlib\figure.py:457: UserWarning: matplotlib is currently using a non-  
GUI backend, so cannot show the figure  
  "matplotlib is currently using a non-GUI backend, "
```

```
input_data = fnew_df.groupby('industry').agg(jobs = ('job_title', 'count'))
input_data

fig, ax = plt.subplots(figsize=(10, 7))
ax = sns.countplot(x='skills', data=input_data)

ax.set_title('Jobs Distribution through various industries')
ax.set_xticklabels(ax.get_xticklabels()[:7])
fig.show()
```



```
fnew_df.describe()
```

	job_id	skill_id
count	17688.000000	17688.000000
mean	8843.500000	7.682214
std	5106.230116	5.609333
min	0.000000	0.000000
25%	4421.750000	3.000000
50%	8843.500000	7.000000
75%	13265.250000	12.000000
max	17687.000000	20.000000

In [644]:

```
fnew_df['base_salary_max'].isnull().sum()
```

0

In [645]:

```
fnew_df.shape
```

(17688, 14)

In [646]:

```
fnew_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17688 entries, 0 to 17687
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype  
---  -
 0   job_id                17688 non-null  int64  
 1   job_title             17688 non-null  object  
 2   industry              17688 non-null  object  
 3   company               17688 non-null  object  
 4   description            17688 non-null  object  
 5   job_locality          17688 non-null  object  
 6   job_region            17688 non-null  object  
 7   job_post_expiry       17688 non-null  object  
 8   education_requirements 17688 non-null  object  
 9   base_salary_min       17688 non-null  object  
10   base_salary_max       17688 non-null  object  
11   base_salary_currency  17688 non-null  object  
12   url                   17688 non-null  object  
13   skill_id              17688 non-null  int32  
dtypes: int32(1), int64(1), object(12)
memory usage: 1.8+ MB
```

In [647]:

```
courses_m_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1005 entries, 0 to 1004
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype  
---  -
 0   title           1005 non-null  object  
 1   type            1005 non-null  object  
 2   url             1005 non-null  object  
 3   description     1005 non-null  object  
 4   ratings         1005 non-null  object  
 5   recent_views   1005 non-null  object  
 6   enrolled       1005 non-null  object  
 7   skills         1005 non-null  object  
dtypes: object(8)
memory usage: 110.7+ KB
```

```
from google.colab import files
from google.colab import auth
from google.colab import drive
auth.authenticate_user()
drive.mount('/content/drive')
```

Mounted at /content/drive

```
!pip install pip
```

```
import pip
```

```
for each in ["requests","beautifulsoup4"]:
    pip.main(['install', each])
```

```
import requests
from bs4 import BeautifulSoup
```

```
↳ Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pip in /usr/local/lib/python3.8/dist-packages (21.1.3)
WARNING: pip is being invoked by an old script wrapper. This will fail in a future version of pip.
Please see https://github.com/pypa/pip/issues/5599 for advice on fixing the underlying issue.
To avoid this problem you can invoke Python with '-m pip' instead of running pip directly.
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: requests in /usr/local/lib/python3.8/dist-packages (2.23.0)
Requirement already satisfied: urllib3!=1.25.0,!1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.8/dist-packages (from requests) (1)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.8/dist-packages (from requests) (2022.9.24)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.8/dist-packages (from requests) (2.10)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.8/dist-packages (from requests) (3.0.4)
WARNING: pip is being invoked by an old script wrapper. This will fail in a future version of pip.
Please see https://github.com/pypa/pip/issues/5599 for advice on fixing the underlying issue.
To avoid this problem you can invoke Python with '-m pip' instead of running pip directly.
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.8/dist-packages (4.6.3)
```

```
#scraping indeed for positions according to candidate profile
```

```
import datetime, time
```

```
!pip install wget
```

```
import wget
```

```
import csv, re
```

```
import logging
```

```
import pandas as pd
```

```
import sqlite3
```

```
import json
```

```
from pytz import timezone
```

```
from requests.auth import HTTPBasicAuth
```

```
import requests
```

```
from requests.exceptions import HTTPError
```

```
import traceback
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting wget
  Downloading wget-3.2.zip (10 kB)
Building wheels for collected packages: wget
  Building wheel for wget (setup.py) ... done
  Created wheel for wget: filename=wget-3.2-py3-none-any.whl size=9674 sha256=a9e31dbfa5af116e92b5058b1c74b8ca3722c9bb922671efe0948
  Stored in directory: /root/.cache/pip/wheels/bd/a8/c3/3cf2c14a1837a4e04bd98631724e81f33f462d86a1d895fae0
Successfully built wget
Installing collected packages: wget
Successfully installed wget-3.2
NumExpr defaulting to 2 threads.
```

```
client_key = "77zk7frida7k3"
```

```
client_secret = "CCISHQyTUjP3xqvZ"
```

```
job_positions = ["data%20scientist%20intern","software%20engineering%20intern","Devops%20intern","Software%20Engineer","Data%20Analyst",
headers = {'Accept': 'application/json'}
auth = HTTPBasicAuth(client_key, client_secret)
job_ur_template = "https://www.linkedin.com/jobs/view/"
job_df = pd.DataFrame(columns=["job_title","industry","company","description","job_locality","job_region","skills","job_post_expiry","edu
```

```
list_of_links = []
try:
```

```

s = requests.Session()

for position in job_positions:
    for i in range(0,500,25):
        #for i in range(0,1):
            linkedin_url = 'https://www.linkedin.com/jobs-guest/jobs/api/seeMoreJobPostings/search?keywords='+position+\
                '&geoId=103644278&trk=public_jobs_jobs-search-bar_search-submit&position=8&pageNum=0&currentJobId=32578844338'

            time.sleep(1)
            html_text = s.get(linkedin_url, headers=headers).text
            soup = BeautifulSoup(html_text, "html.parser")
            count=0
            for meat in soup.find_all('a'):
                if (meat.get('href') != None):
                    if(job_url_template in meat.get('href')):
                        list_of_links.append(meat.get('href'))
                        count+=1

            print(len(list_of_links))
except HTTPError as http_err:
    print(f'HTTP error occurred: {http_err}')
except Exception as err:
    print(f'Other error occurred: {err}')

2997

job_data = []
print(len(list_of_links))
for i in range(0,len(list_of_links)):
    job_req = requests.Session()
    job_url=list_of_links[i]
    try:
        job_text = job_req.get(job_url, headers=headers, auth=auth).text
        job_soup = BeautifulSoup(job_text, "html.parser")
        job_desc_soup = BeautifulSoup(job_text, "xml")
        job_details_content = job_soup.find('script', {"type": "application/ld+json"})
        job_desc = job_desc_soup.find('div', class_='description__text')

        if(job_details_content and job_details_content.text):
            job_json = json.loads(job_details_content.text)
            job_details = [job_json.get('title'), job_json.get('industry')]

            if(job_json.get('identifier')):
                job_details.append(job_json.get('identifier').get('name'))
            else:
                job_details.append(None)

            if(job_desc):
                job_details.append(job_desc.text.replace("Show more", " ").replace("Show less", " "))
            else:
                job_details.append(None)

            if('jobLocation' in job_json and 'address' in job_json.get('jobLocation')):
                job_details.extend( [job_json.get('jobLocation').get('address').get('addressLocality'),
                                    job_json.get('jobLocation').get('address').get('addressRegion')])
            else:
                job_details.extend([None,None])

            job_details.extend( [job_json.get('skills'), job_json.get('validThrough')])

            if('educationRequirements' in job_json):
                job_details.append(job_json.get('educationRequirements').get('credentialCategory'))
            else:
                job_details.append(None)

            if('baseSalary' in job_json and 'value' in job_json.get('baseSalary')):
                job_details.extend([job_json.get('baseSalary').get('value').get('minValue'),
                                    job_json.get('baseSalary').get('value').get('maxValue'), job_json.get('baseSalary').get('currency')])
            else:
                job_details.extend([None,None,None])
            job_details.append(job_url)

            job_data.append(job_details)
            time.sleep(1.25)
        else:
            print(job_url)
    except Exception as e:
        logging.error(traceback.format_exc())

2997
https://www.linkedin.com/jobs/view/data-scientist-at-pomercy-3384652177?refId=rB2A0c0hyuwFqg3vO5ORPg%3D%3D&trackingId=0hQTD%2B%2F
https://www.linkedin.com/jobs/view/data-scientist-full-time-at-bardess-group-ltd-3389875542?refId=BV5TqViF4IDtWAggaunTqw%3D%3D&tr
https://www.linkedin.com/jobs/view/data-scientist-at-r-d-partners-3384656781?refId=BV5TqViF4IDtWAggaunTqw%3D%3D&trackingId=Waevn1

```

<https://www.linkedin.com/jobs/view/data-scientist-at-hireresources-3384151389?refId=BV5TqViF4IDtWaggaunTqw%3D%3D&trackingId=0Sn60>

<https://www.linkedin.com/jobs/view/python-data-scientist-at-brains-workgroup-3395566644?refId=P85ov0M%2B4Tnt8o8jAr4Hkg%3D%3D&trac>

<https://www.linkedin.com/jobs/view/machine-vision-engineer-at-w-h-leary-co-inc-3395534646?refId=L61B5qqe00RMJ%2B1vm525ew%3D%3D&tr>

<https://www.linkedin.com/jobs/view/data-scientist-cost-analyst-at-quest-consulting-inc-3382643925?refId=Oczps%2BwTp6REK0TrY%2F%2F>

<https://www.linkedin.com/jobs/view/software-engineer-summer-intern-2023-at-staples-3385013751?refId=Cmp0u4wgS53ilCG8%2BjFFVA%3D%3>

<https://www.linkedin.com/jobs/view/software-engineer-internship-jan-march-at-arivo-acceptance-3368897330?refId=OpGQZkdVelss5bY6Su>

<https://www.linkedin.com/jobs/view/software-engineer-intern-at-cubrc-3361176547?refId=EcV%2BkC%2FypOgXJTiZf0xjzA%3D%3D&trackingId>

<https://www.linkedin.com/jobs/view/aws-devops-admin-at-avance-consulting-3385121117?refId=pe60mf5Q54iZc0B5lFBxyg%3D%3D&trackingId>

<https://www.linkedin.com/jobs/view/cloud-engineer-at-finlocker-3395490904?refId=pe60mf5Q54iZc0B5lFBxyg%3D%3D&trackingId=2FXJhPts>

<https://www.linkedin.com/jobs/view/sql-developer-intern-at-custom-computer-specialists-3394347972?refId=pe60mf5Q54iZc0B5lFBxyg%3D>

<https://www.linkedin.com/jobs/view/cloud-operations-engineer-at-veterans-sourcing-group-llc-3395533898?refId=JxdTA84TvwjvFB%2FYd0>

<https://www.linkedin.com/jobs/view/cloud-security-engineer-at-trapp-technology-3394077063?refId=jo1ISWkv1p8ilGbsOf6RIO%3D%3D&trac>

<https://www.linkedin.com/jobs/view/it-applications-engineer-at-globe-life-3395492896?refId=jo1ISWkv1p8ilGbsOf6RIO%3D%3D&trackingI>

<https://www.linkedin.com/jobs/view/it-applications-engineer-at-globe-life-3395492896?refId=g5H6dRPMbQwGF3a262Rbs0%3D%3D&trackingI>

<https://www.linkedin.com/jobs/view/cloud-infrastructure-engineer-at-opspro-llc-3395456417?refId=L2Czj2o5Xte0IEkQdoEH%2Fw%3D%3D&tr>

<https://www.linkedin.com/jobs/view/cloud-infrastructure-engineer-at-opspro-llc-3395456417?refId=ad0tevDvRjyzHOKUBk00Vw%3D%3D&trac>

<https://www.linkedin.com/jobs/view/systems-engineer-at-sja-solutions-3395494595?refId=g1lBMWFF9G0GxkhhbyGxpww%3D%3D&trackingId=hB2>

<https://www.linkedin.com/jobs/view/junior-software-developer-at-cagenix-3395457398?refId=mo0%2F6yPLN4ZG1nVXOt2ag%3D%3D&trackingI>

<https://www.linkedin.com/jobs/view/associate-cloud-engineer-at-umass-chan-medical-school-3395594737?refId=FNdgciUMnqyQSiN3YaVmw%3>

<https://www.linkedin.com/jobs/view/jr-software-engineer-at-tellabs-3382702753?refId=vuvnGNCKmIbUdtpKqghpw%3D%3D&trackingId=BlA6B>

<https://www.linkedin.com/jobs/view/software-engineer-at-foxconn-industrial-internet-3394335327?refId=S8Jfn%2F2Fx4iWtqhGCFUdiC%3D%3>

<https://www.linkedin.com/jobs/view/software-engineer-at-productive-robotics-llc-3385802893?refId=S8Jfn%2F2Fx4iWtqhGCFUdiC%3D%3D&t>

<https://www.linkedin.com/jobs/view/software-engineer-at-lightriver-companies-3382651454?refId=4mrE20BT0mPd%2B13C090lV%3D%3D&trac>

<https://www.linkedin.com/jobs/view/software-engineer-at-tribalco-3395458254?refId=4mrE20BT0mPd%2B13C090lV%3D%3D&trackingId=TSWwY>

<https://www.linkedin.com/jobs/view/software-engineer-at-nagrastar-3384330203?refId=4mrE20BT0mPd%2B13C090lV%3D%3D&trackingId=Gvnl>

<https://www.linkedin.com/jobs/view/software-engineer-at-impact-consulting-solutions-inc-3382908427?refId=4mrE20BT0mPd%2B13C090lV>

<https://www.linkedin.com/jobs/view/software-engineer-c%23-net-at-darbytek-338982452?refId=4mrE20BT0mPd%2B13C090lV%3D%3D&trackin>

<https://www.linkedin.com/jobs/view/software-engineer-at-rsi-visuals-3394061760?refId=3JgqkNSezkdaMS9v0DjCsw%3D%3D&trackingId=u5h%3>

<https://www.linkedin.com/jobs/view/software-engineer-at-enhance-it-3390256067?refId=3JgqkNSezkdaMS9v0DjCsw%3D%3D&trackingId=oZS6>

<https://www.linkedin.com/jobs/view/software-engineer-at-omnitech-inc-3395460187?refId=3JgqkNSezkdaMS9v0DjCsw%3D%3D&trackingId=wi0>

<https://www.linkedin.com/jobs/view/software-engineer-at-omnitech-inc-3395460187?refId=8VkvW0wRwPAzZdW2ukIkA%3D%3D&trackingId=x0P>

<https://www.linkedin.com/jobs/view/software-developer-at-information-technology-3395496313?refId=8VkvW0wRwPAzZdW2ukIkA%3D%3D&tra>

<https://www.linkedin.com/jobs/view/software-developer-backend-at-certdrive-3395491498?refId=FarYanI9SHt0s6y6yWAZ9A%3D%3D&trackin>

<https://www.linkedin.com/jobs/view/back-end-software-engineer-at-healium-3394334470?refId=hSxhJ59lq7%2FgKh27Alh0o0%3D%3D&tracking>

<https://www.linkedin.com/jobs/view/software-engineer-backend-application-engineer-at-ashael-tek-solutions-llc-3395451531?refId=Wa>

<https://www.linkedin.com/jobs/view/software-application-engineer-hybrid-at-devonway-3394406604?refId=X6G0LnyaIcyW6Rigrt0n8A%3D%3D>

<https://www.linkedin.com/jobs/view/full-stack-software-engineer-at-avirtek-3390024233?refId=y8eWg4vLp87YvH3sJx8Z6A%3D%3D&tracking>

<https://www.linkedin.com/jobs/view/software-engineer-at-staff-experts-llc-3387515528?refId=%2BYRhVyXvOscyqLY073yU10%3D%3D&trackin>

<https://www.linkedin.com/jobs/view/software-engineer-at-staff-experts-llc-3387515528?refId=BFKPsVU15SraRzM46DT6Tw%3D%3D&trackingI>

<https://www.linkedin.com/jobs/view/java-developer-software-engineer-at-enhance-it-3388668654?refId=IN4W7cvHIB61AslS1jJptg%3D%3D&t>

<https://www.linkedin.com/jobs/view/software-development-engineer-at-the-job-connection-inc-3395538518?refId=IN4W7cvHIB61AslS1jJpt>

<https://www.linkedin.com/jobs/view/software-developer-at-quality-manufacturing-systems-inc-qmsi-3395513146?refId=4b5ZgyV7iGfeex3g>

<https://www.linkedin.com/jobs/view/c%23-software-engineer-at-black-swan-search-3385772333?refId=4b5ZgyV7iGfeex3g2jBPaa%3D%3D&trac>

<https://www.linkedin.com/jobs/view/software-engineer-front-end-at-ashael-tek-solutions-llc-3395455101?refId=Z9gJzrghI0vJjpj0KKKcS>

<https://www.linkedin.com/jobs/view/data-analyst-at-surge-staffing-3395492666?refId=DkwbLn8n%2BjJxNA9Fwzy60%3D%3D&trackingId=dx1r>

<https://www.linkedin.com/jobs/view/data-analyst-at-bluewave-resource-partners-3385812728?refId=I77hvCsPIvaa%2FyldXa5iW0%3D%3D&tra>

<https://www.linkedin.com/jobs/view/data-analyst-entry-level-at-flexon-technologies-inc-3387396862?refId=I77hvCsPIvaa%2FyldXa5iW0%3>

<https://www.linkedin.com/jobs/view/entry-level-data-analyst-at-asta-crs-inc-3395541204?refId=I77hvCsPIvaa%2FyldXa5iW0%3D%3D&track>

<https://www.linkedin.com/jobs/view/data-analyst-at-paradigm-3394407665?refId=ISVef0cGc0sDcnTeIXaYw%3D%3D&trackingId=095kX4r4NOIV>

<https://www.linkedin.com/jobs/view/data-analyst-i-at-tech-providers-inc-3388709415?refId=GHeTi8J4W6vY4UMDAJ80g%3D%3D&trackingId>

<https://www.linkedin.com/jobs/view/data-analyst-remote-at-aston-carter-3395514800?refId=H7HCBsvZ0c5JGwM%2FsaI78A%3D%3D&trackingId>

```
job_df = pd.DataFrame(job_data,
                        columns=["job_title", "industry", "company", "description", "job_locality", "job_region", "skills", "job_post_expiry", "e

job_df
```

```

    job_title    industry    company    description
0      Data Scientist Internship - OR (US)    Insurance    Asurion    \n\n\n 2022-75202Asurion Data Science I...
1      Intern - Data Scientist    Medical Equipment Manufacturing    Dexcom    \n\n\nAbout DexcomDexcom, Inc. empowers people...
2      Data Science, Graduate Internship    Retail    Lowe's Companies, Inc.    \n\n\nLowe's Summer Internship Program Overvie...
3      Data Scientist Internship - ML (US)    Insurance    Asurion    \n\n\n 2022-75205Asurion Data Science I...

job_df['description']

0      \n\n\n      2022-75202Asurion Data Science I...
1      \n\n\nAbout DexcomDexcom, Inc. empowers people...
2      \n\n\nLowe's Summer Internship Program Overvie...
3      \n\n\n      2022-75205Asurion Data Science I...
4      \n\n\n      We're a leader in the pet care i...
...
2930    \n\n\nDescriptionData EngineerWe are currently...
2931    \n\n\nJob DescriptionPerform Data Engineering ...
2932    \n\n\nJob DescriptionData model development an...
2933    \n\n\nAbout NeudesicPassion for technology dri...
2934    \n\n\n      Evaluating existing data systems...
Name: description, Length: 2935, dtype: object

print(job_data)

IOPub data rate exceeded.
The notebook server will temporarily stop sending output
to the client in order to avoid crashing it.
To change this limit, set the config variable
`--NotebookApp.iopub_data_rate_limit`.

Current values:
NotebookApp.iopub_data_rate_limit=1000000.0 (bytes/sec)
NotebookApp.rate_limit_window=3.0 (secs)

job_df = pd.DataFrame(job_data,
                       columns=["job_title", "industry", "company", "description", "job_locality", "job_region", "skills", "job_post_expiry", "e
outfile="jobs_data"
outfile=outfile+'.csv'
print(outfile)
job_df.to_csv(outfile, sep=',', encoding='utf-8')

job_df
```

```
jobs_data.csv
```

	job_title	industry	company	description
0	Data Scientist Internship - OR (US)	Insurance	Asurion	\n\n\n 2022-75202Asurion Data Science I...
1	Intern - Data Scientist	Medical Equipment Manufacturing	Dexcom	\n\n\nAbout DexcomDexcom, Inc. empowers people...
2	Data Science, Graduate Internship	Retail	Lowe's Companies, Inc.	\n\n\nLowe's Summer Internship Program Overvie...
3	Data Scientist Internship - ML (US)	Insurance	Asurion	\n\n\n 2022-75205Asurion Data Science I...
4	Data Science	Manufacturing	Nestlé	\n\n\n We're a leader

```
!cp "/content/jobs_data.csv" "/content/drive/MyDrive/jobs_data.csv"
print("Uploaded")

Uploaded
```

2930	Engineer (AWS, Python, SQL)	Staffing and Recruiting	Cambay Healthcare, LLC	\n\n\nDescriptionData EngineerWe are currently...
2931	Data Engineer	Software Development	Diverse Lynx	\n\n\nJob DescriptionPerform Data Engineering ...
	Data	Software		\n\n\nJob DescriptionData

```
import tweepy #Library required for Twitter API
from tweepy.auth import OAuthHandler
```

```
import datetime, time
!pip install wget
import wget
```

```
import csv, re
import logging
```

```
import pandas as pd
import sqlite3
```

```
from pytz import timezone
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting wget
  Downloading wget-3.2.zip (10 kB)
Building wheels for collected packages: wget
  Building wheel for wget (setup.py) ... done
  Created wheel for wget: filename=wget-3.2-py3-none-any.whl size=9674 sha256=8ea6625d0b31e88d36c942ed5e8541852273c4c73d5b406c04e9a
  Stored in directory: /root/.cache/pip/wheels/bd/a8/c3/3cf2c14a1837a4e04bd98631724e81f33f462d86a1d895fae0
Successfully built wget
Installing collected packages: wget
Successfully installed wget-3.2
```

```
!pip install pip
```

```
import pip
package = 'tweepy' #Just replace the package name with any package to install it.
pip.main(['install',package])
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pip in /usr/local/lib/python3.7/dist-packages (21.1.3)
WARNING: pip is being invoked by an old script wrapper. This will fail in a future version of pip.
Please see https://github.com/pypa/pip/issues/5599 for advice on fixing the underlying issue.
To avoid this problem you can invoke Python with '-m pip' instead of running pip directly.
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: tweepy in /usr/local/lib/python3.7/dist-packages (3.10.0)
Requirement already satisfied: six>=1.10.0 in /usr/local/lib/python3.7/dist-packages (from tweepy) (1.15.0)
Requirement already satisfied: requests[socks]>=2.11.1 in /usr/local/lib/python3.7/dist-packages (from tweepy) (2.23.0)
Requirement already satisfied: requests-oauthlib>=0.7.0 in /usr/local/lib/python3.7/dist-packages (from tweepy) (1.3.1)
Requirement already satisfied: oauthlib>=3.0.0 in /usr/local/lib/python3.7/dist-packages (from requests-oauthlib>=0.7.0->tweepy) (3.0.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests[socks]>=2.11.1->tweepy) (2021.10.8)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests[socks]>=2.11.1->tweepy) (2.10)
Requirement already satisfied: urllib3!=1.25.0,!<1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-packages (from requests[socks]>=2.11.1->tweepy) (1.25.11)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests[socks]>=2.11.1->tweepy) (3.0.2)
Requirement already satisfied: PySocks!=1.5.7,>=1.5.6 in /usr/local/lib/python3.7/dist-packages (from requests[socks]>=2.11.1->tweepy) (1.7.1)
```

Initializing credentials and Data Frames to hold the data.

```
consumer_key = "N9UT39Ye0SUEk4DY0SCUhUste"
consumer_secret = "660kzn3kiviV5Q23TPm96F00tth0yu2SHqSBTGrnvjuzttI97h"
access_key = "1590826381532798976-jqrLEziKmsFDF8sKZpriSGWDY04unh"
access_secret = "r0fmPQvqxkDD0da0ivipkta2u6Z4FvphSJPtUyooAgvJ"
```

```
#Creating an empty dataframe to store the information
tweets = pd.DataFrame(columns=["id", "created_at", "user", "text", "hashtags", "user_mentions", "media_url", "urls", "location", "retweetec
users = pd.DataFrame(columns=["id", "user_id", "created_at", "screen_name", "name", "url", "profile_image_url_https", "statuses_count"])
```

```
urls_df = pd.DataFrame(columns=["url", "tweet_ids"])
hashtags_df = pd.DataFrame(columns=["start_index", "end_index", "text", "tweet_ids"])
user_mentions_df = pd.DataFrame(columns=["id", "screen_name"])
media_df = pd.DataFrame(columns=["id", "url", "type", "tweet_ids"])
jobs_df = pd.DataFrame(columns=["job_title", "description", "url", "poster", "posted_at", "tweet_ids"])
```

Fetching current datetime and date in the past with a difference of 14 days

```
eastern = timezone('US/Eastern')
lh = datetime.datetime.now()
last_hour = lh.astimezone(eastern) - datetime.timedelta(days=14)

#getting tweets since today
since_tweets = last_hour.strftime("%Y-%m-%d")
```

```

print('Authentication OK')

auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_key, access_secret)

api = tweepy.API(auth,wait_on_rate_limit=True)

try:
    api.verify_credentials()
    print("Authentication OK")
except:
    print("Error during authentication")

    Authentication OK

```

Downloading data from Twitter based on hashtag relevant to job recommendation.

```

api = tweepy.API(auth,wait_on_rate_limit=True)
tweets_data = [] #initialize master list to hold our ready tweets
user_data = [] #initialize master list to hold users

media_urls_data = []
media = []
urls_data = []
urls = []
hashtags_data = []
hashtags = []
user_mentions_data = []
user_mentions = []
#keywords=['#jobalert', '#JobSearch', "#job", "#hiring"]
#new_search = " OR ".join(keywords)+" #jobsearch -filter:retweets"
new_search="#jobalert"
num_tweets=100
cnt=0
job_positions = ["ENGINEER", "DEVELOPER", "OFFICER", "NURSE", "CLERK", "SUPERVISOR"]
job_data = []
temp=[]

tweet_set = tweepy.Cursor(api.search,q=new_search,count=100, #The q variable holds the hashtag
                           lang="en",
                           since=since_tweets).items()

try:
    for tweet in tweet_set:
        for position in job_positions:
            if(position in tweet.text):
                job=[]
                job.append(position)
                job.append(tweet.text.encode("utf-8"))
                cnt=cnt+1
                user=tweet.user
                user_data.append([user.id, user.created_at, user.screen_name, user.name, user.url, user.profile_image_url_https, user.status])
                if(tweet.entities.get('media',[])) :
                    for m in tweet.entities.get('media',[]):
                        media_urls_data.append([m["id"], m["media_url"], m["url"], m["type"], tweet.id])
                        media.append(str(m["id"]))
                if(tweet.entities.get('urls',[])):
                    urlStr=[]
                    for u in tweet.entities.get('urls'):
                        urls_data.append([u["url"], tweet.id])
                        urls.append(u["url"])
                        urlStr.append(u["url"])
                    job.append("".join(urlStr))
                else:
                    job.append(None)
                if(tweet.entities.get('hashtags',[])) :
                    for ht in tweet.entities.get('hashtags'):
                        hashtags_data.append([ht["indices"][0], ht["indices"][1],ht["text"], tweet.id])
                        hashtags.append(ht["text"])
                if(tweet.entities.get('user_mentions',[])) :
                    for user in tweet.entities.get('user_mentions'):
                        user_data.append([user["id"], None,user["screen_name"], None, None, None, None])
                        user_mentions.append(str(user["id"]))
                tweets_data.append([tweet.id, tweet.created_at, tweet.user.id, tweet.text.encode("utf-8"),
                                    "", ".join(hashtags)", "", ".join(user_mentions)", "", ".join(media)",
                                    "", ".join(urls)", tweet.user.location, tweet.retweeted, tweet.retweet_count,
                                    tweet.favorite_count])
                job.append(tweet.user.name)
                job.append(tweet.created_at)
                job.append(tweet.id)

```



```

        job_data.append(job)
        continue
    except BaseException as e:
        print('BaseException',str(e)) # print the error code obtained from twitter
        time.sleep(5)

tweets_df = pd.DataFrame(tweets_data,
                          columns = ["tweet_id","created_at", "user_id", "text", "hashtags", "user_mentions", "media_url","url", "location"])
tweets_df["tweet_id"] = pd.to_numeric(tweets_df["tweet_id"])
tweets_df['created_at'] = pd.to_datetime(tweets_df['created_at'])
tweets_df["retweeted_status"] = tweets_df['retweeted_status'].astype('bool')
tweets_df["retweet_count"] = pd.to_numeric(tweets_df["retweet_count"])
tweets_df["favorite_count"] = pd.to_numeric(tweets_df["favorite_count"])
tweets_df

```

	tweet_id	created_at	user_id	text	hashtags
0	1591558499686707200	2022-11-12 22:28:10	1491776087684104193	b'RT @careersingov: Don't miss this...	
1	1591558499627958272	2022-11-12 22:28:10	1491776087684104193	b'RT @careersingov: .@SanBenitoCounty is #hiring...	
2	1591558499544092672	2022-11-12 22:28:10	1491776087684104193	b'RT @careersingov: Could this job be yours? @...	
3	1591552924198608897	2022-11-12 22:06:01	525005120	b'Could this job be yours? @SanBenitoCounty is...	
4	1591551183335952386	2022-11-12 21:59:06	40591485	b'RT @careersingov: .@SanBenitoCounty is #hiring...	
...
91	1588587645449428997	2022-11-04 17:43:03	525005120	b'Great job! @YorkCountySCGov is #hiring a REG...	
92	1588572198071111680	2022-11-04 16:41:40	1491776087684104193	b'RT @careersingov: Don't miss this...	
93	1588571813826760706	2022-11-04 16:40:09	40591485	b'RT @careersingov: Don't miss this...	
94	1588571792028962817	2022-11-04 16:40:04	525005120	b'Don't miss this job opportunity! ...	
95	1588571792028962817	2022-11-04 16:40:04	525005120	b'@CityofHesperia is #hiring a	

```

jobs_df = pd.DataFrame(job_data,
                        columns = ["job_title", "description", "url", "poster", "posted_at","tweet_ids"] )

jobs_df['posted_at'] = pd.to_datetime(jobs_df['posted_at'])
jobs_df

```

```

    job_title      description      url      poster      pos
0      NURSE      b'RT @careersingov: Don't miss this...'      None      Jobs via Tweet      2
1  SUPERVISOR      b'RT @careersingov: @SanBenitoCounty is #hiring...'      None      Jobs via Tweet      2
2  SUPERVISOR      b'RT @careersingov: Could this job be yours? @...'      None      Jobs via Tweet      2
3  SUPERVISOR      b'Could this job be yours? @...'      https://t.co/...      CareersInGovernment      2

user_df = pd.DataFrame(user_data,
                        columns=["user_id", "created_at", "screen_name", "name", "url", "profile_image_url_https", "statuses_count"])

user_df["user_id"] = pd.to_numeric(user_df["user_id"])
user_df["created_at"] = pd.to_datetime(user_df["created_at"])
user_df["statuses_count"] = pd.to_numeric(user_df["statuses_count"])
user_df
```

	user_id	created_at	screen_name	name	
0	1491776087684104193	2022-02-10 14:08:51	jobsviatweet	Jobs via Tweet	https://t
1	525005120	NaT	careersingov	None	
2	1171080641133236225	NaT	SanBenitoCounty	None	
3	1491776087684104193	2022-02-10 14:08:51	jobsviatweet	Jobs via Tweet	https://t
4	525005120	NaT	careersingov	None	
...	
207	41657673	NaT	KansasCity	None	
208	525005120	2012-03-15 03:52:59	careersingov	CareersInGovernment	https://t.c
209	41657673	NaT	KansasCity	None	
210	525005120	2012-03-15 03:52:59	careersingov	CareersInGovernment	https://t.c

```

hashtags_df = pd.DataFrame(hashtags_data,
                            columns = ["start_index", "end_index", "text", "tweet_ids"])

hashtags_df["start_index"] = pd.to_numeric(hashtags_df["start_index"])
hashtags_df["end_index"] = pd.to_numeric(hashtags_df["end_index"])
hashtags_df
```

	start_index	end_index	text	tweet_ids
0	59	66	hiring	1591558499686707200
1	39	46	hiring	1591558499627958272
2	63	70	hiring	1591558499544092672
3	45	52	hiring	1591552924198608897
4	39	46	hiring	1591551183335952386
...
90	31	38	hiring	1588587645449428997
91	66	73	hiring	1588572198071111680
92	66	73	hiring	1588571813826760706
93	48	55	hiring	1588571792028962817
94	20	27	hiring	1588547130301595648

95 rows x 4 columns

```

media_df = pd.DataFrame(media_urls_data,
                        columns = ["id", "media_url", "url", "type", "tweet_ids"])
```

```
media_df["id"]= pd.to_numeric(media_df["id"])
media_df
```

id	media_url	url	type	tweet_ids
----	-----------	-----	------	-----------

```
urls_df = pd.DataFrame(urls_data,
                        columns = ["url", "tweet_ids"])
```

```
urls_df
```

	url	tweet_ids
0	https://t.co/ugdjFewaYD	1591552924198608897
1	https://t.co/hr7oUFU3H9	1591551161554837504
2	https://t.co/qSjYNC1DdD	1591549399402008579
3	https://t.co/FDOZBGpxkF	1591461323451535360
4	https://t.co/NxEwkaqhyR	1591404658106552321
...
57	https://t.co/vQN2t6QMdE	1588628155912404992
58	https://t.co/Jp8N7ZNvdl	1588605260527984641
59	https://t.co/8PmyUsGBzZ	1588587645449428997
60	https://t.co/Pol0dJybL0	1588571792028962817
61	https://t.co/l72eOtK7dj	1588547130301595648

62 rows × 2 columns

Save Data Frames to SQL database in different tables.

```
conn = sqlite3.connect('MakeMyCareer.db')
cur = conn.cursor()
```

```
create_tweets_query="CREATE TABLE tweets(tweet_id int PRIMARY KEY NOT NULL UNIQUE, \
                        created_at DATE NOT NULL, user_id int NOT NULL, text VARCHAR(255) NOT NULL, hashtags VARCHAR, user_mentions VARCHAR, \
                        location VARCHAR(255), retweeted_status BOOL NOT NULL, retweet_count int NOT NULL, favorite_count int, \
                        FOREIGN KEY (user_mentions) REFERENCES user_mentions (VARCHAR) );"
```

```
cur.execute("DROP TABLE IF EXISTS tweets;")
cur.execute(create_tweets_query)
```

```
tweets_df.to_sql('tweets',con=conn,index=False, if_exists='replace')
```

```
create_users_query="CREATE TABLE users(user_id int PRIMARY KEY NOT NULL UNIQUE,created_at DATE NOT NULL,screen_name VARCHAR(255) NOT NULL, \
                        name VARCHAR(255) NOT NULL,url VARCHAR ,profile_image_url_https VARCHAR, statuses_count);"
```

```
cur.execute("DROP TABLE IF EXISTS users;")
cur.execute(create_users_query)
```

```
#user_df.drop(columns = user_df.columns[0], axis = 1, inplace= True)
user_df.to_sql('users',con=conn,index=False, if_exists='replace')
```

```
create_hashtags_query="CREATE TABLE hashtags(hashtag_id int PRIMARY KEY NOT NULL UNIQUE,text VARCHAR(255),tweet_ids VARCHAR, \
                        FOREIGN KEY (tweet_ids) REFERENCES tweets (id) );"
```

```
cur.execute("DROP TABLE IF EXISTS hashtags;")
cur.execute(create_hashtags_query)
```

```
hashtags_df.to_sql('hashtags',con=conn,index=False, if_exists='replace')
```

```
create_media_query="CREATE TABLE media_urls(media_id int PRIMARY KEY NOT NULL UNIQUE,media_url VARCHAR,url VARCHAR, \
                        type VARCHAR(255), tweet_ids VARCHAR, FOREIGN KEY (tweet_ids) REFERENCES tweets (tweet_id) );"
```

```
cur.execute("DROP TABLE IF EXISTS media_urls;")
cur.execute(create_media_query)
```

```
media_df.to_sql('media_urls',con=conn,index=False, if_exists='replace')
```

```
create_urls_query="CREATE TABLE urls(url VARCHAR PRIMARY KEY NOT NULL UNIQUE, \
                        tweet_ids VARCHAR, FOREIGN KEY (tweet_ids) REFERENCES tweets (tweet_id) );"
```

```
cur.execute("DROP TABLE IF EXISTS urls;")
```

```

cur.execute(create_urls_query)

urls_df.to_sql('urls',con=conn,index=False, if_exists='replace')

create_jobs_query="CREATE TABLE jobs(job_title VARCHAR(255), description VARCHAR,\
                                url VARCHAR PRIMARY KEY NOT NULL UNIQUE,poster VARCHAR ,posted_at VARCHAR, \
                                tweet_ids VARCHAR, FOREIGN KEY (tweet_ids) REFERENCES tweets (tweet_id) );"

cur.execute("DROP TABLE IF EXISTS jobs;")
cur.execute(create_jobs_query)

jobs_df.to_sql('jobs',con=conn,index=False, if_exists='replace')

def run_query(query):
    return pd.read_sql(query,conn)

```

SQL queries to express the below questions:

- What user posted this tweet?
- When did the user post this tweet?
- What tweets have this user posted in the past 24 hours?
- How many tweets have this user posted in the past 24 hours?
- When did this user join Twitter?
- What keywords/ hashtags are popular?
- What tweets are popular?

```

list_of_tables = ["tweets","users","hashtags","media_urls","urls","jobs"]
table=list_of_tables[0]
select_query="SELECT * FROM "+table+" LIMIT 5"

query_q1 = "SELECT DISTINCT users.user_id, users.screen_name, users.name \
            FROM users INNER JOIN tweets ON users.user_id=tweets.user_id \
            where tweets.tweet_id = '1591558499627958272' "
query_q2 = "SELECT DISTINCT tweets.created_at \
            FROM users INNER JOIN tweets ON users.user_id=tweets.user_id \
            where tweets.tweet_id = '1591558499627958272' "
query_q3 = "SELECT DISTINCT tweets.tweet_id,tweets.text,users.user_id \
            FROM tweets INNER JOIN users ON users.user_id=tweets.user_id \
            WHERE tweets.user_id=1491776087684104193 AND \
            tweets.created_at>'2022-11-11 22:06:01'"
query_q4 = "SELECT users.user_id,COUNT(tweets.tweet_id) as number_of_tweets \
            FROM tweets INNER JOIN users ON users.user_id=tweets.user_id \
            WHERE tweets.user_id=1491776087684104193 AND \
            tweets.created_at>'2022-11-11 22:06:01'"
query_q5 = "SELECT DISTINCT users.user_id,users.created_at FROM users WHERE users.user_id=1491776087684104193"
query_q6 = "SELECT DISTINCT hashtags.text, tweets.retweet_count FROM hashtags \
            INNER JOIN tweets on tweets.tweet_id=hashtags.tweet_ids \
            WHERE hashtags.retweet_count > 3"
query_q7 = "SELECT tweet_id, text, retweet_count FROM tweets ORDER BY retweet_count DESC"

run_query(query_q6)

```

	text	retweet_count
0	jobopening	4
1	hiring	4

▼ Career Recommendation System : Use Cases

1. Use Case: User can look for opening for their target job position Description: User can look for opening for a position named "Engineer"

Actor: User

Precondition: User should have a valid target position name

Steps:

Actor action: User request for list of job openings for his target position.

System Responses: If the position exists, the system will return a list of job openings posted.

Post Condition: List of job openings suggested

Alternate Path: The user request is not correct and system throws an error

Error: User information is incorrect

2. Use Case: User can look for openings posted by their dream company handle Description: Search for job posts posted by a particular user

Actor: User

Precondition: User should have a company name user is target

Steps:

Actor action: User request for list of job openings for his target position.

System Responses: If the company has posted job openings, the system will return the list.

Post Condition: List of job openings suggested

Alternate Path: The user request is not correct and system throws an error

Error: User information is incorrect

3. Use Case: User can look for openings posted within last 5 days and for a particular position Description: Search for job posts posted within last 5 days

Actor: User

Precondition: User should have a valid target position name

Steps:

Actor action: User request for list of job openings for his target position.

System Responses: The system will return a list of job posts.

Post Condition: List of job openings suggested

Alternate Path: The user request is not correct and system throws an error

Error: User information is incorrect

4. Use Case: User can assess which job positions are more in demand Description: Search for job posts for different job positions

Actor: User

Precondition: User should have a valid target position name

Steps:

Actor action: User request for list of job openings for his target position.

System Responses: The system will return a count of job posts for a position.

Post Condition: List of job openings suggested

Alternate Path: The user request is not correct and system throws an error

Error: User information is incorrect

5. Use Case: User can assess which companies are posting more jobs

Description: Search for job posts for job positions by different companies

Actor: User

Precondition: User should have a valid target position name and target company

Steps:

Actor action: User request for list of job openings for his target position posted by company.

System Responses: The system will return a count of job posts posted by company handle.

Post Condition: Count of job openings suggested

Alternate Path: The user request is not correct and system throws an error

Error: User information is incorrect

```
#Use Case: User can look for opening for their target job position
use_case_1 = "SELECT jobs.job_title,jobs.description, jobs.poster, jobs.posted_at FROM jobs INNER JOIN tweets ON tweets.tweet_id=jobs.tw
```

```
#Use Case: User can look for openings posted by their dream company
use_case_2 = "SELECT jobs.job_title,jobs.description, jobs.poster, jobs.posted_at \
FROM jobs WHERE poster='CareersInGovernment'"
```

```
#Use Case: User can look for openings posted within last 5 days and for a particular position
use_case_3 = "SELECT jobs.job_title,jobs.description, jobs.poster, jobs.posted_at \
FROM jobs WHERE job_title='NURSE' and posted_at>'2022-11-11 22:28:10'"
```

```
#Use Case: User can assess which job positions are more in demand
```

```
use_case_4 = "SELECT jobs.job_title,COUNT(jobs.job_title) as number_of_postings,jobs.description, jobs.poster, jobs.posted_at \
FROM jobs GROUP BY job_title ORDER BY number_of_postings DESC"

#Use Case: User can assess which companies are posting more jobs
use_case_5 = "SELECT jobs.job_title,COUNT(jobs.job_title) as number_of_postings,jobs.description, jobs.poster, jobs.posted_at \
FROM jobs GROUP BY poster"

run_query(use_case_4)
```

	job_title	number_of_postings	description	poster	posted
0	SUPERVISOR	36	b'.@CityofHesperia is #hiring a MAINTENANCE CR...	CareersInGovernment	2022 15:04
1	CLERK	20	b'ADMIN CLERK \n\nOrganization: Department of ...	Jobshaven	2022 04:36
2	OFFICER	15	b'Don't miss this job	CareersInGovernment	2022

