INFO 6210-DATA MANAGEMENT AND DATABASE DESIGN

ABSTRACT

In this project, we aim at developing a jobs database using tools like MySQL and python. This jobs database consists of the recently open job positions for top 5 banks in the USA. Data was extracted from multiple sources like the individual career website of each bank, glassdoor and twitter.

INTRODUCTION

This is an unprecedented situation mankind has been facing and is certainly challenging. Economies have taken a huge hit and amidst all of this, job search is another aspect for all the people who had been laid off as a result of this pandemic. This database consists of recent job postings (posted in the duration of Feb-April) from top 5 banks in the US.

BUILDING THE LIST OF COMPANIES

For extracting the biggest 5 banking institutions in the US, we decided to scrape data from Wikipedia. We scraped relevant information like the headquarters, Total assets and market capitalization. The output looks like this.

	A	B	C	D	E
1	Rank	Bank name	Headquarters locatio	Total assets (billions	Market capitalizatior
2	1	JPMorgan Chase	New York City	$2,687	$327
3	2	Bank of America	Charlotte, North Carc	$2,434	$301
4	3	Citigroup	New York City	$1,951	$174
5	4	Wells Fargo	San Francisco, Calif	$1,927	$273
6	5	Goldman Sachs	New York City	$992	$87

Code: -

Importing libraries

```
from bs4 import BeautifulSoup
import requests
import pandas as pd

```

Web scraping the essential information

```
URL= "https://en.wikipedia.org/wiki/List_of_largest_banks_in_the_United_States"
#url for getting list
```

```
r=requests.get(URL)
soup=BeautifulSoup(r.content, "html.parser")
table=soup.find('table')  #print(table)
rows = table.find_all('tr')
columns = [v.text.replace('\n','') for v in rows[0].find_all('th')]
  #for row in table.findAll('tr'):
    #for column in row.findAll('td'):
  #print(columns)
df = pd.DataFrame(columns=columns)
for i in range(1,6):
  tds = rows[i].findAll('td')

  if len(tds) == 4:
    values = [tds[0].text.replace('\n',''), tds[1].text.replace('\n',''),'',tds[2].text.replace('\n',''),tds[3].text.replace
  else:
    values = [td.text.replace('\n','') for td in tds]

     #print(values)
  df = df.append(pd.Series(values, index=columns),ignore_index=True)
  print(df)
  df.to_csv(r'wiki_list.csv',index=False, header=True)


#C:\Users\Rohan\Desktop\NEU\Term 2\DMDD\Final project\wiki list\

fe06dfcd94abefcae8a25b363331
```

We also scraped some more information about these companies and the code, and the scraped information are as follows: -

BofA and JP Morgan Chase

258

BofA	JPMC
Charlotte, North Carolina, U.S.	New York, NY, U.S.
Kenneth D. Lewis, Chairman, President & C	Jamie Dimon, Chairman, President & CEO
Banking	Finance and Insurance
Financial Services	Financial Services
$117.01 billion USD	US$99.9 billion
$21.13 billion USD	US$14.4 billion
203,425	174,360
www.bankofamerica.com www.bofa.com	www.jpmorganchase.com
	www.jpmorgan.com
	www.chase.com
Bank of America (NYSE: BAC TYO: 8648) is th	JPMorgan Chase (NYSE: JPM) is one of the oldest financial ser
Public (NYSE: BAC TYO: 8648)	Public (NYSE: JPM, TYO: 8634)
(as "Bank of Italy") San Francisco, CA (1928)	1799

Code: -

```
[ ]  bofa_jpmc_url= 'https://www.diffen.com/difference/Bank_of_America_vs_JPMorgan_Chase'
```

```
[ ]  bofa_jpmc_soup = getSoup(bofa_jpmc_url)
```

```
[ ]  table = bofa_jpmc_soup.find('table')
     result = []
     for row in table.findAll('tr'):
       result.append([at.string for at in row.findAll('td')])
     result = result[3:]
     print(result)
```

```
[ ]
     bofa_jpmc_df = pd.DataFrame( columns = ['BofA', 'JPMC'], index= ['Token', 'Headquarters', 'Industry', 'Products', 'Revenue
```

```
[ ]  bofa_jpmc_df.head()
```

Wells Fargo and Goldman Sachs: -

Wells	Goldman
The Goldman Sachs Group, Inc., or si	Wells Fargo & Co. (NYSE: WFC) is a diversified f
Public (NYSE: GS)	Public (NYSE: WFC)
1869	New York, New York, USA (March 18, 1852)
New York, NY	420 Montgomery, San Francisco, California, USA
Lloyd Blankfein, Chairman & CEO	
Gary Cohn, President & COO	
Jon Winkelried, President and COO	
John S. Weinberg, Vice Chairman	
David A. Viniar, CFO	
Edward C. Forst, CAO	Richard Kovacevich,
Gregory K. Palm, General Counsel	Chairman
Esta E. Stecher, General Counsel	John Stumpf,
Kevin W.	President and CEO
Investment Banking	Retail Banking
	Insurance
	Investments
	Mortgages
	Consumer Finance

Code: -

```
[ ]  wells_gold_url = 'https://www.diffen.com/difference/Goldman_Sachs_vs_Wells_Fargo'
```

```
[ ]  wells_gold_soup = getSoup(wells_gold_url)
     #print(soup)
     table = wells_gold_soup.find('table')
     result = []
     for row in table.findAll('tr'):
       result.append([at.string for at in row.findAll('td')])
     result = result[3:]
     print(result)
```

```
[ ]  wells_gold_df = pd.DataFrame( columns = ['Wells', 'Goldman'], index= ['Introduction', 'Type', 'Founded', 'Headquarters','H
```

```
[ ]  wells_gold_df.head()
```

Citi: -

Citi	
New York City, USA	
Sir Win Bischoff, Chairman	
Financial services	
Consumer Banking	
Corporate Banking	
Stockbroking	
Investment Banking	
Global Wealth Management	
Investment Research	
Private Equity	
Structured Products	
US $146.7 billion	
US $21.538 billion	
332,000	
Let's get it done.	
www.citigroup.com	
Citigroup Inc. (NYSE: C), operating as Citi, is	
Public (NYSE: C)	
New York City, USA (1812)	

```
[ ]  bofa_citi_url = 'https://www.diffen.com/difference/Bank_of_America_vs_Citigroup'
```

```
[ ]  bofa_citi_soup = getSoup(bofa_citi_url)
     table = bofa_citi_soup.find('table')
     result = []
     for row in table.findAll('tr'):
       result.append([at.string for at in row.findAll('td')])
     result = result[3:]
     print(result)
```

```
[['Charlotte, North Carolina, U.S.', 'New York City, USA'], ['Kenneth D. Lewis, Chairman, President & CEO  Amy W. Brinkley,
```

```
[ ]  bofa_citi_df = pd.DataFrame( columns = ['BofA', 'Citi'], index= ['Headquarters', 'Key People', 'Industry', 'Products', 'Re
```

```
[ ]  bofa_citi_df.head()
```

SCRAPING JOB POSTINGS FOR THESE 5 BANKS.

Now with information on these 5 banks, the next obvious step was to scrape job postings from their careers section. Again, using the same approach of Python, in this section we will be discussing the process of scraping the recent job postings from these top 5 banks.

We decided to scrape the Job posting, location and date posted. Example of the job postings and the code associated with the scraping look like this:-

Bank of America

```
[ ]  bofa_url = "https://careers.bankofamerica.com/en-us/job-search?ref=search&search=jobsByCityCountry&city=Azusa&state=Calif
```

```
[ ]  bofa_soup = getSoup(bofa_url)
     #print(soup)
     job = bofa_soup.find('div', attrs={'aem-wrap--job-search-results-listing section'})
     #print(job)
     job_name = []
     for row in job.findAll('a'):
       job_name.append(row.text)
```

```
[ ]  locations = []
     dates = []
     for location in bofa_soup.findAll('div', attrs={'class':'job-search-tile__detail'}):
       dates.append((location.text.split("\n")[1]).split()[1])
       locations.append(location.text.split("\n")[2])
```

```
[ ]  print(dates)
```

```
[ ]  def getJobsDf(jobs, loc, dates):
       df = pd.DataFrame({'Job Name':jobs, 'Location':loc, 'Date Posted': dates})
       df.head()
       return(df)
```

```
[ ]  bofa_df = getJobsDf(job_name, locations, dates)
```

```
[ ]  bofa_df.head()
```

CSV file looks like this: -

Job Name	Location	Date Posted
Merchant Services C	Azusa, California	04/13/2020
Merchant Services C	Multiple Locations	04/16/2020
Merchant Services C	Multiple Locations	04/20/2020
Merchant Services C	Multiple Locations	04/13/2020
Merchant Services C	Multiple Locations	04/21/2020
Client Service Repre	Rosemead, Californi	04/08/2020
Client Service Repre	Whittier, California	04/10/2020
Client Service Repre	Whittier, California	04/10/2020
Client Service Repre	Chino Hills, Californi	04/20/2020
Merchant Services C	Multiple Locations	04/13/2020
Relationship Banker	Monterey Park, Califo	04/08/2020
Merrill Lynch Financ	Multiple Locations	04/23/2020
Quantitative Operati	Multiple Locations	04/23/2020

Citibank

Code: -

```
[ ]  citi_url= "https://jobs.citi.com/search-jobs"
```

```
[ ]  citi_soup = getSoup(citi_url)


     job_name = []
     locations = []
     dates = []
     #to get all li
     job_list = citi_soup.find('section', attrs = {'id':'search-results-list'})
     for row in job_list.findAll('li'):
       job_name.append(row.h2.text)
       locations.append(row.find('span', attrs = {'class': 'job-location'}).text)
       dates.append(row.find('span', attrs = {'class': 'job-date-posted'}).text)
```

```
[ ]  citi_df = getJobsDf(job_name, locations, dates)
```

CSV File: -

Job Name	Location	Date posted
Vice President, Capi	New York, New York,	04/23/2020
Vice President - Tree	New York, New York,	04/23/2020
Service Rep 5	Tuxtla Gutiérrez, Me	04/23/2020
Solutions Sales Sr. G	Miami, Florida, Unite	04/23/2020
SME Relationship M	Multiple Locations	04/23/2020
Pue Huauchinango I	Huauchinango, Mex	04/23/2020
Senior Global Progra	New York, New York,	04/23/2020
Promotor de Tarjetas	Saltillo, Mexico	04/23/2020
Independent Operati	Belfast, United Kingc	04/23/2020
Full Stack developer	Tampa, Florida, Unit	04/23/2020
EJECUTIVO/A FINA	Villahermosa, Mexic	04/23/2020
Digital Data Collecti	Multiple Locations	04/23/2020
Cmpl AML Executior	Tampa, Florida, Unit	04/23/2020
Cash and Trade Proc	Tampa, Florida, Unit	04/23/2020
Containers & Kubern	Irving, Texas, United	04/23/2020

JP Morgan Chase

Code :-

```
[ ] import re
```

```
[ ] jpmc_url ="https://jobs.jpmorganchase.com/ListJobs/ByCountry/US/"
```

Code Text

```
[ ] jpmc_soup = getSoup(jpmc_url)
    #print(jpmc_soup)# , 'style':'display: block;'
    table = jpmc_soup.find('table')
    #print(table)

    job_name = []
    locations = []
    dates = []
    for row in table.findAll('tr'):
      #print(row)
      job = row.find('td', {'class': 'coloriginaljobtitle'})
      if job != None:
        job_name.append(job.a.text)
      location = row.find('td', {'class':'colstate'})
      city = row.find('td',{'class':'colcity'})
      if location != None:
        locations.append(city.text.strip() + ", " +location.text.strip())
```

```
      date_post = row.find('td',{'class':'colpostedon'}) #colpostedon
      if date_post != None:
        dates.append(date_post.text.strip())
```

```
[ ] job_name
```

```
[ ] #df name
    jpmc_df = getJobsDf(job_name, locations, dates)
```

CSV File: -

Job Title	Location	Date Posted
CIB F&BM - Web De	Brooklyn, NY	4-24-2020
Global Supplier Ser	Plano, TX	4-24-2020
Global Supplier Ser	Newark, DE	4-24-2020
Global Supplier Ser	Columbus, OH	4-24-2020
CIB F&BM - DPS Op	Brooklyn, NY	4-24-2020
Finance Control Mar	Brooklyn, NY	4-9-2020
Content Analyst-U.S.	Jersey City, NJ	4-23-2020
Consumer and Comr	Harrisburg, NC	4-23-2020
Consumer and Comr	Charlotte, NC	4-23-2020
Consumer and Comr	Harrisburg, NC	4-23-2020
Consumer and Comr	Charlotte, NC	4-23-2020
Relationship Banker	Charlotte, NC	4-23-2020
Consumer and Comr	Charlotte, NC	4-23-2020
Consumer and Comr	Charlotte, NC	4-23-2020
Consumer and Comr	Charlotte, NC	4-23-2020
Consumer and Comr	Charlotte, NC	4-23-2020
Consumer and Comr	Gastonia, NC	4-23-2020

Wells Fargo

Code: -

```
[ ]  wells_url='https://employment.wellsfargo.com/psc/PSEA/APPLICANT_NW/HRMS/c/HRS_HRAM_FL.HRS_CG_SEARCH_FL.GBL?Page=HRS_APP_SC
```

```
[ ]  wells_soup = getSoup(wells_url)

     #job_search = soup.find('div', {'title':'Search Results List', 'class':'ps_grid-list'})

     #job_search = soup.find('div', {'id':"win0divHRS_AGNT_RSLT_I$grid$0" })
     job_list = wells_soup.find('ul', {'class':'ps_grid-body'})
     #
     print(job_list)
```

```
[ ]  import re
     #in job search find li
     job_name = []
     locations = []
     dates = []
     for row in wells_soup.findAll('li', {'id': re.compile('^HRS_AGNT')}):

       job = row.find('span', {'class':'ps_box-value'})
       job_name.append(job.text)
       location = row.find('span', {'id': re.compile('^LOCATION')})
       locations.append(location.text)
       date_post = row.find('span', {'id': re.compile('^SCH_OPENED')})
       dates.append(date_post.text)
```

```
[ ]  wells_df = getJobsDf(job_name, locations, dates)
```

```
[ ]  wells_df.head()
```

CSV File: -

Job Title	Location	Date Posted
ECMO QA Analyst 2	IA-West Des Moines	4/24/2020
IDQ/PowerCenter Developer (Infor	NJ-Summit	4/24/2020
Operational Risk Management Tech	Multiple	4/24/2020
Tech Initiative Supprt Coord 3	TX-San Antonio	4/24/2020
Technology Manager 4 - Cisco VoIP	Multiple	4/24/2020
Account Resolution Specialist 2 - Ov	OR-Hillsboro	4/23/2020
Account Resolution Specialist 3 - Ov	OR-Hillsboro	4/23/2020
Administrative Assistant 4	TX-Westlake	4/23/2020
Analytic Consultant 3 - Home Lendi	Multiple	4/23/2020
Analytic Consultant 5 - Home Lendi	Multiple	4/23/2020
Analytic Consultant 6	Multiple	4/23/2020

Note: - The CSV files attached are not all the jobs scraped. Just some of the examples scraped.

Goldman Sachs

Goldman had a very limited number of jobs and hence we decided to just add them to our database on MySQL.

GETTING DATA FROM TWITTER.

Twitter is a great medium to talk about the media relevance of these 5 companies and what these companies are talking about. We extracted information like user, date, text, favorite count, hashtags and user location.

General Function :-

```
[ ]  from twython import Twython
     import json
```

```
[ ]  def getTaggedTweets(hashtag):
       """tutorial for accessing twitter api using twython
       link - https://stackabuse.com/accessing-the-twitter-api-with-python/"""

       #api details
       # Enter your keys/secrets as strings in the following fields
       credentials = {}
       credentials['CONSUMER_KEY'] = 'GJrKx4QiGP4VBCvWzwzL9Jc3s'
       credentials['CONSUMER_SECRET'] = 'SRrnCvSvJ75wilwE62wrj22hQGA8rJrNtQJEFNemORgSLQNTqY'
       credentials['ACCESS_TOKEN'] = '1246732645284728832-I6ReqO8tMqdASeGMK62op2Ygq7RXXj'
       credentials['ACCESS_SECRET'] = 'mGeWISnxFe04t0rqtgql7QloBPLUgnjeazlodmD1VHNeJ'


       # Instantiate an object
       python_tweets = Twython(credentials['CONSUMER_KEY'], credentials['CONSUMER_SECRET'])
       # Create our query
       query = {'q': hashtag,
               'result_type': 'mixed',
               'count': 100,
               'lang': 'en'
               }
```

```
       # Search tweets
       dict_ = {'user': [], 'date': [], 'text': [], 'favorite_count': [], 'hashtags' : [], 'user_loc' :[]}
       for status in python_tweets.search(**query)['statuses']:

           dict_['user'].append(status['user']['screen_name'])
           dict_['date'].append(status['created_at'])
           dict_['text'].append(status['text'])
           dict_['favorite_count'].append(status['favorite_count'])
           dict_['hashtags'].append([hashtag['text'] for hashtag in status['entities']['hashtags']])
           dict_['user_loc'].append(status['user']['location'])
           # Structure data in a pandas DataFrame for easier manipulation
       df = pd.DataFrame(dict_)
       return(df)
```

BofA Twitter Data

Code: -

```
[ ]  bofa_tweet = getTaggedTweets('#bankofamerica')
```

```
[ ]  bofa_tweet.head()
```

```
[ ]  bofa_tweet.to_csv(r'bofa tweets.csv',index=False, header=True)
```

CSV file: -

user	date	text	favorite_count	hashtags	user_loc
financebrokerag	Fri Apr 24 06:42:01 +	Gold prices dropped	0	[]	British Virgin Islands
slowdownwandwa1	Fri Apr 24 06:16:15 +	#bankofamerica #horrible. They are r @BankofAmerica ke	0	['bankofamerica', 'horrible']	
GlobalDividend	Fri Apr 24 06:16:11 +	Company: Bank of A Dividend: 0,18 USD Period: 2020 3 mont Dividend yield: 3,28 Record date: 5.06.2(Ex-... https://t.co/CP(	0	[]	
DrThomasPaul	Fri Apr 24 04:55:05 +	RT @PLTC_PastLive https://t.co/60bqrPB(	0	['pollution', 'GMO', 'b	Los Angeles . New Y
AJJohns40444296	Fri Apr 24 04:34:26 +	RT @lola666babe: # https://t.co/29Mv4dj1	0	['bankofamerica']	Lorain, OH
jennifercreador	Fri Apr 24 04:22:07 +	@Anomalous__ The	1	[]	Scottsdale, AZ
jeff91597805	Fri Apr 24 02:42:05 +	Banks stop taking "fe	0	[]	
dracobate	Fri Apr 24 01:59:40 +	@GavinNewsom con	1	['debt']	Los Angeles, CA
publiusfed1	Fri Apr 24 01:33:58 +	DO NOT use #Banko	0	['BankofAmerica']	

Citibank Twitter Data: -

Code

```
[ ]  # Structure data in a pandas DataFrame for easier manipulation
     citi_tweets = getTaggedTweets("#CitiGroup")
```

```
[ ]  citi_tweets.head(5)
```

CSV File: -

user	date	text	favorite_count	hashtags	user_loc
peacepumpkinpic	Thu Apr 23 16:34:40	RT @peacepumpkin	0	[]	Toronto, Canada
Tickeron	Thu Apr 23 16:09:45	$C enters an Uptrend	0	[]	Sunnyvale, CA
SamSharplesMT	Thu Apr 23 14:58:50	RT @flyingstocksman	0	['Citigroup', 'America	Montana, USA
flyingstocksman	Thu Apr 23 14:31:27	@carlquintanilla @R	2	['Citigroup']	Manchester, England
MindMoneyMedicl	Thu Apr 23 00:29:31	Look at some nice pe	1	['C', 'CitiGroup']	
LuxuryDecorDeb	Wed Apr 22 22:08:31	Let's not forget he wa	1	['Citigroup', 'Obama']	East Coast
LuxuryDecorDeb	Wed Apr 22 22:08:03	@impulsivewoman L	1	['Citigroup']	East Coast
Bartz70Tim	Wed Apr 22 10:19:24	RT @SchuermannCh	0	['JPMorgan', 'BofA', 'Barclays', 'Citigroup',	
SchuermannChris	Wed Apr 22 07:14:33	A US lawsuit accuse	0	['JPMorgan', 'BofA', '	Rhine area Germany
GlobalDividend	Wed Apr 22 05:58:02	Company: Citigroup Dividend: 0,51 USD Period: 2020 3 mont Dividend yield: 4,62 Record date: 4.05.20 Ex-divide... https://t.c	0	[]	
AusBeyondCoal	Wed Apr 22 04:05:00	International finance	1	['citigroup', 'coal']	Australia
HottestStockNow	Wed Apr 22 00:46:11	$C Time to Reward $	0	['Citigroup']	Las Vegas, NV
Market_Screener	Tue Apr 21 21:18:04	Citigroup : Remarks b	0	['Citigroup', 'Stock', 'MarketScreener']	
WRFoxKidsNation	Tue Apr 21 21:10:46	RT @amazonwatch:	0	['Citigroup']	Dayton, OH

JP Morgan Chase Twitter Data: -

Code: -

```
[ ]  # Structure data in a pandas DataFrame for easier manipulation
     jpmc_tweets = getTaggedTweets("#jpmorganchase")
```

```
[ ]  jpmc_tweets.head(5)
```

CSV File: -

	user	date	text	favorite_count	hashtags	user_loc
0	FordWMaverick	Fri Apr 24 03:55:09 +	RT @lloydkaufman:	0	[]	Columbus Ohio
1	Lynchian_Dream	Fri Apr 24 03:27:39 +	RT @lloydkaufman:	0	[]	
2	lloydkaufman	Fri Apr 24 03:25:42 +	@choptopmoseley @	2	[]	Tromaville,New York
3	bencaroncreates	Fri Apr 24 00:24:42 +	I've been committed With chase #JPMorganChase is	1	['JPMorganChase']	Los Angeles, CA
4	AndreasBoos	Thu Apr 23 19:32:19	How big banks helpe // #JPMorganChase	0	['JPMorganChase', 'C	USA
5	AllTheFunInOne	Thu Apr 23 18:46:55	I applied for a small Help the Animals Pl.	1	[]	Chicago, Illinois USA

Wells Fargo Twitter Data: -

Code: -

```
[ ]  # Structure data in a pandas DataFrame for easier manipulation
     wells_tweets = getTaggedTweets("#wellsfargo")
```

```
[ ]  wells_tweets.head(5)
```

CSV File: -

	user	date	text	favorite_count	hashtags	user_loc
0	DickBurls	Fri Apr 24 07:01:41 +	RT @osudaltaco: #w	0	['wellsfargo']	
1	205_575_9595	Fri Apr 24 06:47:28 +	It's getting Partly Clo	0	[]	
2	DonaldHoffler	Fri Apr 24 03:56:56 +	Wells Fargo, America	0	[]	
3	FordWMaverick	Fri Apr 24 03:55:09 +	RT @lloydkaufman:	0	[]	Columbus Ohio
4	bizjournals	Fri Apr 24 03:45:13 +	The second round of	1	['PPP', 'WellsFargo']	
5	Lynchian_Dream	Fri Apr 24 03:27:39 +	RT @lloydkaufman:	0	[]	
6	lloydkaufman	Fri Apr 24 03:25:42 +	@choptopmoseley @	2	[]	Tromaville,New York
7	jdm_crypto_boy	Fri Apr 24 00:47:32 +	THE MOVE TO #ISO #Sibos #Swift #Wells #Fintech #Chainlink https://t.co/WdqratHl	5	['ISO2022', 'Sibos', 'S	the moon
8	TardWatcher	Fri Apr 24 00:09:14 +	@WellsFargo Shame	0	['WellsFargo', 'DoTheRightThing']	

Goldman Sachs Twitter Data

Code: -

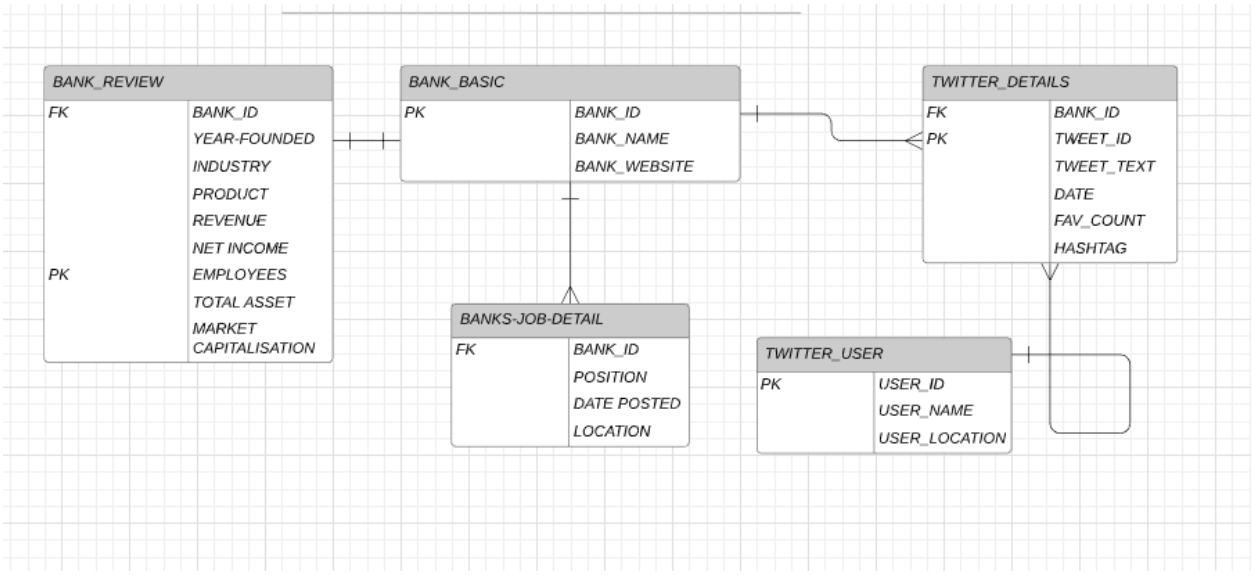
```
[ ]  # Structure data in a pandas DataFrame for easier manipulation
     goldman_tweets = getTaggedTweets('#goldmansachs')
```

```
[ ]  goldman_tweets.head(5)
```

CSV File: -

user	date	text	favorite_count	hashtags	user_loc
WestJournalism	Fri Apr 24 02:00:32 +	Kelly took a big stan	17	['2020election', 'Ariz	Phoenix, AZ
_FinancialWorld	Fri Apr 24 08:01:37 +	Goldman CEO Solor	0	['banks', 'competition	London, UK
mihaidumitres12	Fri Apr 24 07:36:12 +	RT @Kevin_Cage_: Just a reminder of we No Hype https://t.co/wc2Gi8jh RT's are appreciated #XRPCommunity...	0	['XRPCommunity']	
cherobeam141	Fri Apr 24 07:22:09 +	RT @Kevin_Cage_: Just a reminder of we No Hype https://t.co/wc2Gi8jh RT's are appreciated #XRPCommunity...	0	['XRPCommunity']	
lmnysf	Fri Apr 24 07:05:19 +	@OJPgov @TheJust	0	['AGBarr', 'GoldmanSachs', 'Corruption']	
krilliklip	Fri Apr 24 07:02:15 +	#Gold & FOMO	0	['Gold', 'UBS', 'Goldn	Planet #Earth, most
lmnysf	Fri Apr 24 07:01:13 +	@OJPgov @TheJust	0	['AGBarr', 'GoldmanSachs', '1MDB']	

ER MODEL



BUILDING THE DATABASE ON MYSQL

We used MySQL to create the database. Using the insert command and importing the csv file which contains the data scraped, the database looked like this:-

Bank_ID	Name	Bank Website
1	Bank of America	https://www.bankofamerica.com/
2	JP Morgan Chase	https://www.jpmorgan.com/country/US/en/jpm...
3	Wells Fargo	https://www.wellsfargo.com/
4	Goldman Sachs	https://www.goldmansachs.com/index.html
5	Citigroup	https://www.citigroup.com/citi/

Bank_ID : Unique digit for the bank

Name: Name of bank

Bank-Website: website of the bank

Bank_ID	Job Serial Number	Position	Date Posted	Location
5	630	Cmpl AML Execution Int Asc Ast	4/23/20	Tampa, Florida, Unit...
5	640	Cash and Trade Proc Analyst2	4/23/20	Tampa, Florida, Unit...
5	650	Containers & Kubernetes Engineer â€" Logging ...	4/23/20	Irving, Texas, Unite...
2	2001	CIB F&BM - Web Developer - Associate	4/24/2020	Brooklyn, NY
2	2002	Global Supplier Services - Digital Invoice Platfor...	4/24/2020	
2	2003	Global Supplier Services - Digital Invoice Platfor...	4/24/2020	Newark, DE
2	2004	Global Supplier Services - Digital Invoice Platfor...	4/24/2020	Columbus, OH
2	2005	CIB F&BM - DPS Operations Finance Manager - ...	4/24/2020	Brooklyn, NY
2	2006	Finance Control Management - External Reporti...	4/9/2020	Brooklyn, NY
2	2007	Content Analyst-U.S. Wealth Management- Jer...	4/23/2020	Jersey City, NJ
3	3001	ECMO QA Analyst 2	4/24/20	IA-West Des Moines
3	3002	IDQ/Pwer Center Dev.	4/24/20	NJ-Summit
4	4001	Summer Analyst Internship	NA	
4	4002	Digital Innovators Research Program	NA	Cities
4	4003	Summer Associate Program	NA	Multiple
4	4004	New Analyst Program	NA	Multiple
4	4005	New Associate Program	NA	Multiple
4	4006	Chile Intern Program	NA	Santiago
4	4007	Brazil Intern Program	NA	Sao Paolo
4	4008	Mexico Intern Program	NA	Mexico City

2	2007	Content Analyst-U.S. Wealth Management- Jer...	4/23/2020	Jersey City, NJ
2	2008	Consumer and Community Banking - Market Exp...	4/23/2020	Harrisburg, NC
2	2009	Consumer and Community Banking - Market Exp...	4/23/2020	Charlotte, NC
2	2010	Consumer and Community Banking - Market Exp...	4/23/2020	Harrisburg, NC
2	2011	Consumer and Community Banking - Market Exp...	4/23/2020	Charlotte, NC
2	2012	Relationship Banker (Market Expansion) Hwy 73...	4/23/2020	Charlotte, NC
2	2013	Consumer and Community Banking - Market Exp...	4/23/2020	Charlotte, NC
2	2014	Consumer and Community Banking - Market Exp...	4/23/2020	Charlotte, NC
2	2015	Consumer and Community Banking - Market Exp...	4/23/2020	Charlotte, NC
2	2016	Consumer and Community Banking - Market Exp...	4/23/2020	Charlotte, NC

Bank_ID : Unique Bank ID

Job_ID: unique Job_ID (created by us)

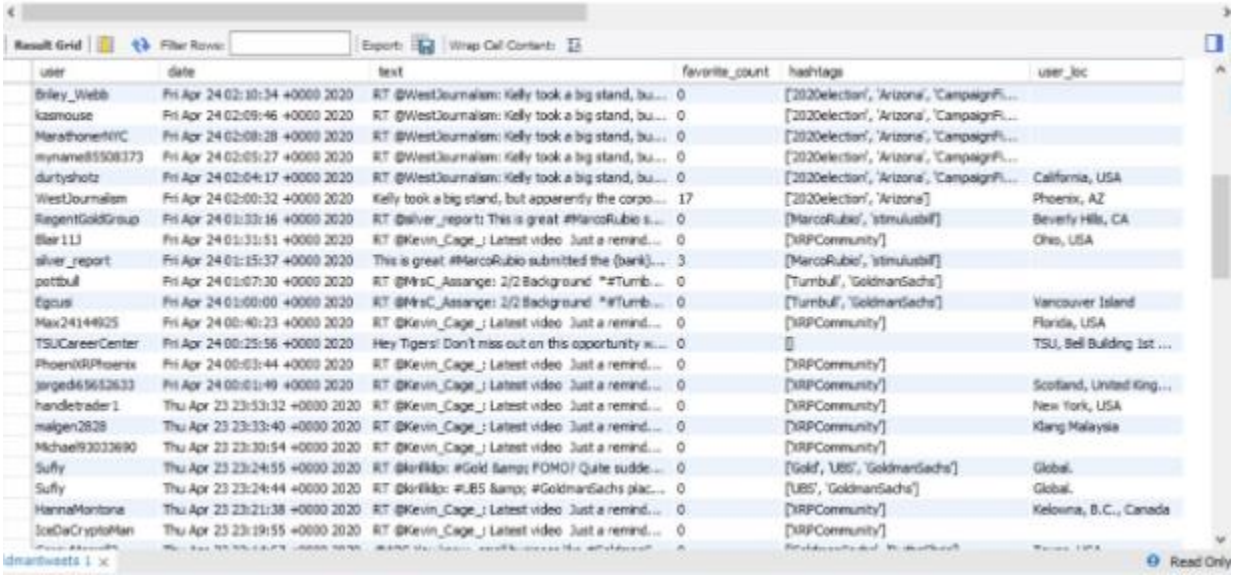
Position: Title of the job.

Date Posted: Date on which the job was posted.

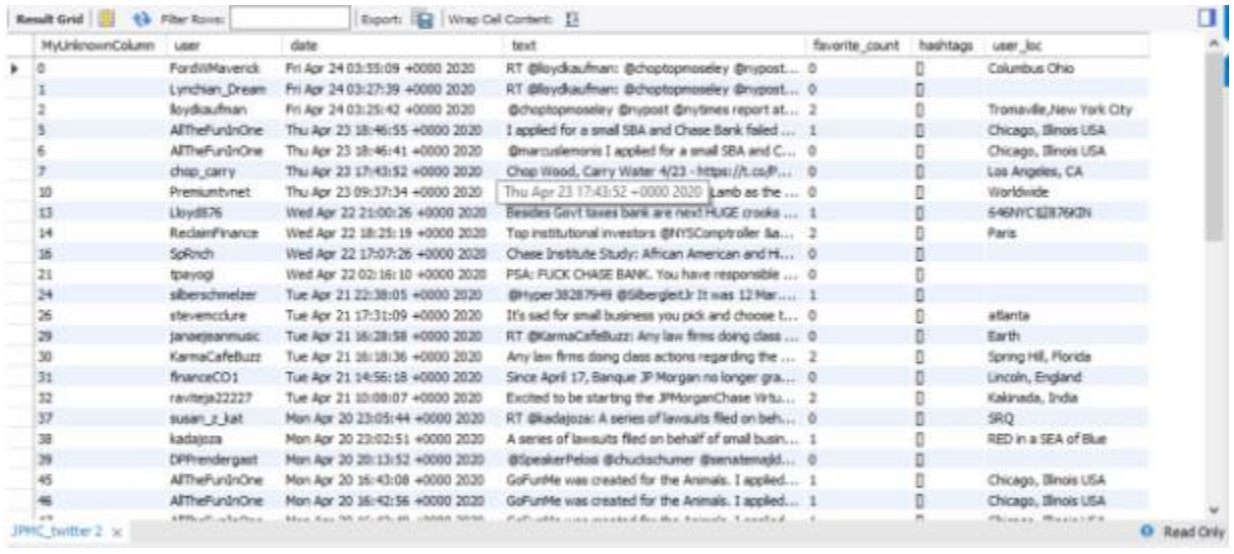
Location: Location of job

These are some of the screenshots of the bank job details look like.

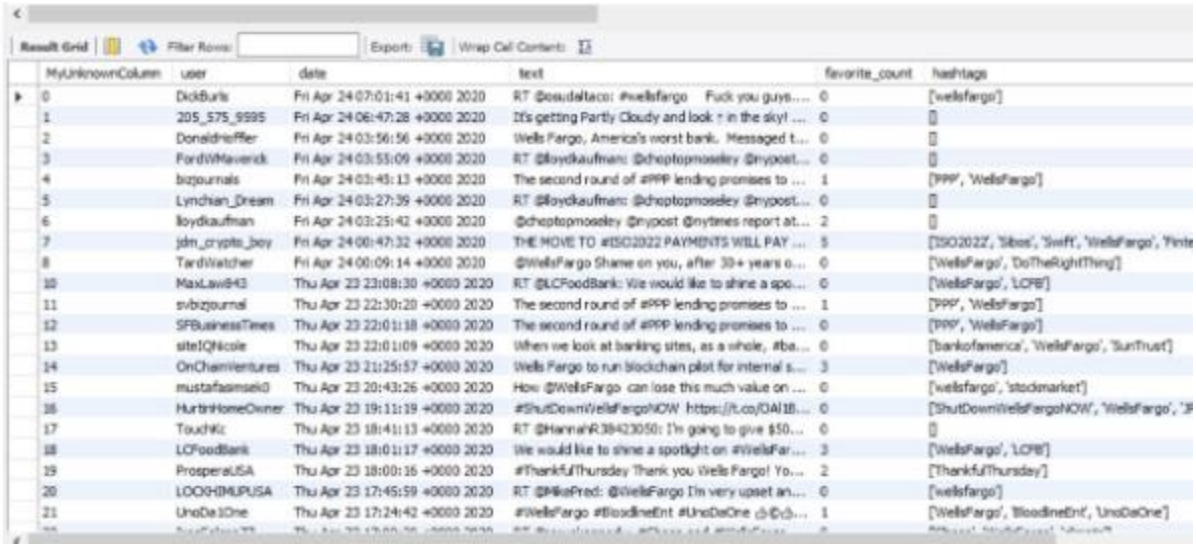
user	date	text	favorite_count	hashtags	user_loc
financebrokerag	Fri Apr 24 06:42:01 +0000 2020	Gold prices dropped on Tuesday. The Bank of A...	0	[]	British Virgin Islands
GlobalDividend	Fri Apr 24 06:16:11 +0000 2020	Company: Bank of America Dividend: 0,18 USD ...	0	[]	
jennifercreador	Fri Apr 24 04:22:07 +0000 2020	@Anomalous__ They also didn't close their bran...	1	[]	Scottsdale, AZ
jeff91597805	Fri Apr 24 02:42:05 +0000 2020	Banks stop taking "fees" for processing ppp fun...	0	[]	
millionssold	Thu Apr 23 16:02:44 +0000 2020	RT @Reenies: SBA did not prioritize any busines...	0	[]	Melbourne, FL
Tickeron	Thu Apr 23 15:01:45 +0000 2020	$BAC enters a Downtrend because Momentum I...	0	[]	Sunnyvale, CA
Reenies	Thu Apr 23 14:21:54 +0000 2020	SBA did not prioritize any businesses, banks ma...	3	[]	Florida
JimJachetta	Thu Apr 23 14:03:26 +0000 2020	@marcuslemonis Thank you. You are a patriot. ...	1	[]	Los Angeles, CA
cashgiveskayla	Thu Apr 23 13:00:10 +0000 2020	Invest today and profit today!!! DM now to get...	1	[]	United States
gaoconnor	Thu Apr 23 12:37:14 +0000 2020	@ElisabethKuhns @BankofAmerica @sheribrons...	1	[]	
PeaceAtelier	Thu Apr 23 12:19:08 +0000 2020	"[A]lleging banks processed clients with larger lo...	0	[]	
onlinecheck	Thu Apr 23 08:11:25 +0000 2020	SBA has released limited data showing granular ...	0	[]	Essex, Connecticut
LongLyndi	Thu Apr 23 05:55:40 +0000 2020	2 weeks after I uploaded my docs-and almost a ...	1	[]	
sukiyaki_1	Thu Apr 23 04:21:08 +0000 2020	@MarisaKabas Love his accent!! He needs a Hol...	1	[]	
kavitadaiya	Thu Apr 23 02:22:38 +0000 2020	@ChrisMurphyCT I don't understand why @Spe...	0	[]	
christo02366670	Thu Apr 23 01:30:13 +0000 2020	@sageleader What are you basing your down...	0	[]	Long Beach, CA
shannon_gapuz	Thu Apr 23 00:26:38 +0000 2020	@queenselassie1 @BankofAmerica @gpconserv...	0	[]	Stockton, CA
shrifilm	Wed Apr 22 22:38:44 +0000 2020	.@BankofAmerica customer service rep was not...	0	[]	Internet
OLAGUNJUOLA...	Wed Apr 22 21:28:18 +0000 2020	@JR_Howell_JR @twmays1974 @BofA_Help @...	0	[]	Palo Alto, CA
Lloyd876	Wed Apr 22 21:00:26 +0000 2020	Besides Govt taxes bank are next HUGE crooks ...	1	[]	646NYC\u2876KIN
AutoCaviar	Wed Apr 22 20:27:04 +0000 2020	@TDBank_US worst banking experience of my li...	2	[]	



user	date	text	favorite_count	hashtags	user_loc
Briley_Webb	Fri Apr 24 02:10:34 +0000 2020	RT @WestJournalism: Kelly took a big stand, bu...	0	['2020election', 'Arizona', 'CampaignFi...	
kasmouse	Fri Apr 24 02:09:46 +0000 2020	RT @WestJournalism: Kelly took a big stand, bu...	0	['2020election', 'Arizona', 'CampaignFi...	
MarathonerNYC	Fri Apr 24 02:08:28 +0000 2020	RT @WestJournalism: Kelly took a big stand, bu...	0	['2020election', 'Arizona', 'CampaignFi...	
myname85508373	Fri Apr 24 02:05:27 +0000 2020	RT @WestJournalism: Kelly took a big stand, bu...	0	['2020election', 'Arizona', 'CampaignFi...	
dirtyshotz	Fri Apr 24 02:04:17 +0000 2020	RT @WestJournalism: Kelly took a big stand, bu...	0	['2020election', 'Arizona', 'CampaignFi...	California, USA
WestJournalism	Fri Apr 24 02:00:32 +0000 2020	Kelly took a big stand, but apparently the corpo...	17	['2020election', 'Arizona']	Phoenix, AZ
RegentGoldGroup	Fri Apr 24 01:33:16 +0000 2020	RT @silver_report: This is great #MarcoRubio s...	0	['MarcoRubio', 'stimulusbill']	Beverly Hills, CA
Blair11J	Fri Apr 24 01:31:51 +0000 2020	RT @Kevin_Cage_: Latest video Just a remind...	0	['XRPCommunity']	Ohio, USA
silver_report	Fri Apr 24 01:15:37 +0000 2020	This is great #MarcoRubio submitted the (bank)...	3	['MarcoRubio', 'stimulusbill']	
pottbull	Fri Apr 24 01:07:30 +0000 2020	RT @MrsC_Assange: 2/2 Background *#Turnb...	0	['Turnbull', 'GoldmanSachs']	
Egcusi	Fri Apr 24 01:00:00 +0000 2020	RT @MrsC_Assange: 2/2 Background *#Turnb...	0	['Turnbull', 'GoldmanSachs']	Vancouver Island
Max24144925	Fri Apr 24 00:40:23 +0000 2020	RT @Kevin_Cage_: Latest video Just a remind...	0	['XRPCommunity']	Florida, USA
TSUCareerCenter	Fri Apr 24 00:25:56 +0000 2020	Hey Tigers! Don't miss out on this opportunity w...	0	[]	TSU, Bell Building 1st ...
PhoenixRPhoenix	Fri Apr 24 00:03:44 +0000 2020	RT @Kevin_Cage_: Latest video Just a remind...	0	['XRPCommunity']	
jorged65652633	Fri Apr 24 00:01:49 +0000 2020	RT @Kevin_Cage_: Latest video Just a remind...	0	['XRPCommunity']	Scotland, United King...
handletrader1	Thu Apr 23 23:53:32 +0000 2020	RT @Kevin_Cage_: Latest video Just a remind...	0	['XRPCommunity']	New York, USA
malgen2828	Thu Apr 23 23:33:40 +0000 2020	RT @Kevin_Cage_: Latest video Just a remind...	0	['XRPCommunity']	Klang Malaysia
Michael93033690	Thu Apr 23 23:30:54 +0000 2020	RT @Kevin_Cage_: Latest video Just a remind...	0	['XRPCommunity']	
Sufiy	Thu Apr 23 23:24:55 +0000 2020	RT @kirilldp: #Gold & FOMO? Quite sudde...	0	['Gold', 'UBS', 'GoldmanSachs']	Global.
Sufiy	Thu Apr 23 23:24:44 +0000 2020	RT @kirilldp: #UBS & #GoldmanSachs plac...	0	['UBS', 'GoldmanSachs']	Global.
HannaMontana	Thu Apr 23 23:21:38 +0000 2020	RT @Kevin_Cage_: Latest video Just a remind...	0	['XRPCommunity']	Kelowna, B.C., Canada
IceDaCryptoMan	Thu Apr 23 23:19:55 +0000 2020	RT @Kevin_Cage_: Latest video Just a remind...	0	['XRPCommunity']	



MyUnknownColumn	user	date	text	favorite_count	hashtags	user_loc
0	FordWMaverick	Fri Apr 24 03:55:09 +0000 2020	RT @lloydkaufman: @choptopmoseley @nypost...	0	[]	Columbus Ohio
1	Lynchian_Dream	Fri Apr 24 03:27:39 +0000 2020	RT @lloydkaufman: @choptopmoseley @nypost...	0	[]	
2	lloydkaufman	Fri Apr 24 03:25:42 +0000 2020	@choptopmoseley @nypost @nytimes report at...	2	[]	Tromaville,New York City
5	AllTheFun1nOne	Thu Apr 23 18:46:55 +0000 2020	I applied for a small SBA and Chase Bank failed ...	1	[]	Chicago, Illinois USA
6	AllTheFun1nOne	Thu Apr 23 18:46:41 +0000 2020	@marcuslemonis I applied for a small SBA and C...	0	[]	Chicago, Illinois USA
7	chop_carry	Thu Apr 23 17:43:52 +0000 2020	Chop Wood, Carry Water 4/23 - https://t.co/P...	0	[]	Los Angeles, CA
10	Premiumtvnet	Thu Apr 23 09:37:34 +0000 2020	Thu Apr 23 17:43:52 +0000 2020 Lamb as the ...	0	[]	Worldwide
13	Lloyd876	Wed Apr 22 21:00:26 +0000 2020	Besides Govt taxes bank are next HUGE crooks ...	1	[]	646NYC\u2876KIN
14	ReclaimFinance	Wed Apr 22 18:25:19 +0000 2020	Top institutional investors @NYSComptroller &a...	2	[]	Paris
16	SpRnch	Wed Apr 22 17:07:26 +0000 2020	Chase Institute Study: African American and Hi...	0	[]	
21	tpayogi	Wed Apr 22 02:16:10 +0000 2020	PSA: FUCK CHASE BANK. You have responsible ...	0	[]	
24	silberschmelzer	Tue Apr 21 22:38:05 +0000 2020	@Hyper38287949 @SilbergietJr It was 12 Mar....	1	[]	
26	stevemcclure	Tue Apr 21 17:31:09 +0000 2020	It's sad for small business you pick and choose t...	0	[]	atlanta
29	janaejeanmusic	Tue Apr 21 16:28:58 +0000 2020	RT @KarmaCafeBuzz: Any law firms doing class ...	0	[]	Earth
30	KarmaCafeBuzz	Tue Apr 21 16:18:36 +0000 2020	Any law firms doing class actions regarding the ...	2	[]	Spring Hill, Florida
31	financeCO1	Tue Apr 21 14:56:18 +0000 2020	Since April 17, Banque JP Morgan no longer gra...	0	[]	Lincoln, England
32	raviteja22227	Tue Apr 21 10:08:07 +0000 2020	Excited to be starting the JPMorganChase Virtu...	2	[]	Kakinada, India
37	susan_z_kat	Mon Apr 20 23:05:44 +0000 2020	RT @kadajoza: A series of lawsuits filed on beh...	0	[]	SRQ
38	kadajoza	Mon Apr 20 23:02:51 +0000 2020	A series of lawsuits filed on behalf of small busin...	1	[]	RED in a SEA of Blue
39	DPPrendergast	Mon Apr 20 20:13:52 +0000 2020	@SpeakerPelosi @chuckschumer @senatemajld...	0	[]	
45	AllTheFun1nOne	Mon Apr 20 16:43:08 +0000 2020	GoFundMe was created for the Animals. I applied...	1	[]	Chicago, Illinois USA
46	AllTheFun1nOne	Mon Apr 20 16:42:56 +0000 2020	GoFundMe was created for the Animals. I applied...	1	[]	Chicago, Illinois USA



	MyUnknownColumn	user	date	text	favorite_count	hashtags
▸	0	DickBurls	Fri Apr 24 07:01:41 +0000 2020	RT @osudeltaco: #wellsfargo Fuck you guys....	0	['wellsfargo']
	1	205_575_9595	Fri Apr 24 06:47:28 +0000 2020	It's getting Partly Cloudy and look ↑ in the sky! ...	0	[]
	2	DonaldrieffJr	Fri Apr 24 03:56:56 +0000 2020	Wells Fargo, America's worst bank. Messaged t...	0	[]
	3	FordVMaverick	Fri Apr 24 03:55:09 +0000 2020	RT @lloydkaufman: @cheptopmoseley @nypost...	0	[]
	4	bizjournals	Fri Apr 24 03:45:13 +0000 2020	The second round of #PPP lending promises to ...	1	['PPP', 'WellsFargo']
	5	Lynchian_Dream	Fri Apr 24 03:27:39 +0000 2020	RT @lloydkaufman: @cheptopmoseley @nypost...	0	[]
	6	lloydkaufman	Fri Apr 24 03:25:42 +0000 2020	@cheptopmoseley @nypost @nytimes report at...	2	[]
	7	jdm_crypto_boy	Fri Apr 24 00:47:32 +0000 2020	THE MOVE TO #ISO2022 PAYMENTS WILL PAY ...	5	['ISO2022', 'Sibos', 'Swift', 'WellsFargo', 'Finte
	8	TardWatcher	Fri Apr 24 00:09:14 +0000 2020	@WellsFargo Shame on you, after 30+ years o...	0	['WellsFargo', 'DoTheRightThing']
	10	MaxLaw843	Thu Apr 23 23:08:30 +0000 2020	RT @LCFoodBank: We would like to shine a spo...	0	['WellsFargo', 'LCFB']
	11	svbizjournal	Thu Apr 23 22:30:20 +0000 2020	The second round of #PPP lending promises to ...	1	['PPP', 'WellsFargo']
	12	SFBusinessTimes	Thu Apr 23 22:01:18 +0000 2020	The second round of #PPP lending promises to ...	0	['PPP', 'WellsFargo']
	13	siteIQNicole	Thu Apr 23 22:01:09 +0000 2020	When we look at banking sites, as a whole, #ba...	0	['bankofamerica', 'WellsFargo', 'SunTrust']
	14	OnChainVentures	Thu Apr 23 21:25:57 +0000 2020	Wells Fargo to run blockchain pilot for internal s...	3	['WellsFargo']
	15	mustafaaimseki0	Thu Apr 23 20:43:26 +0000 2020	How @WellsFargo can lose this much value on ...	0	['wellsfargo', 'stockmarket']
	16	HurtinHomeOwner	Thu Apr 23 19:11:19 +0000 2020	#ShutDownWellsFargoNOW https://t.co/OAI1B...	0	['ShutDownWellsFargoNOW', 'WellsFargo', '#
	17	TouchKc	Thu Apr 23 18:41:13 +0000 2020	RT @HannahR38423050: I'm going to give $50...	0	[]
	18	LCFoodBank	Thu Apr 23 18:01:17 +0000 2020	We would like to shine a spotlight on #WellsFar...	3	['WellsFargo', 'LCFB']
	19	ProsperaUSA	Thu Apr 23 18:00:16 +0000 2020	#ThankfulThursday Thank you Wells Fargo! Yo...	2	['ThankfulThursday']
	20	LOCKHIMUPUSA	Thu Apr 23 17:45:59 +0000 2020	RT @MikePred: @WellsFargo I'm very upset an...	0	['wellsfargo']
	21	UnoDe1One	Thu Apr 23 17:24:42 +0000 2020	#WellsFargo #BloodlineEnt #UnoDeOne ◷©◷...	1	['WellsFargo', 'BloodlineEnt', 'UnoDeOne']

The twitter table looked like this.

User: the one who tweeted

Date: When the user tweeted.

Text: What he tweets

User_loc: location of the user

Bank Review: -

bank_id	year_founded	industry	product	revenue(billions)	net_income(billions)	employees	total asset(billions)
4	1869	Investment Banking	Commerical	37	10	38300	992
2	1877	Finance and Insurance	Financial Services	100	14	174360	2687
1	1904	Banking	Financial Services	117	21	203425	2434
5	1998	Equity,fixed income	Banking	73	18	204000	1951
3	1852	Retail Banking	Financial Services	86	5	258700	1927

total asset(billions)	market_capitalisation(billion
992	87
2687	327
2434	301
1951	174
1927	273

GENERATING USER CASES

This database created, is a consolidated database of all the jobs by top 5 banks in the banking industry. The database can be sliced according to the location, title and in what service and products the user wants to work.

Any person analyzing the jobs can estimate the trends like what the banks are tweeting, number of jobs over time and how the banks history has been with KPI likes net income, total assets and market capitalization.

OPTIMIZATION OF DATABASE

While working on the database, we tried to keep the data as consistent as possible, Missing values were checked and all the issues were hence resolved.

Following the rules of normalization, we tried normalizing the data down till 3NF.

REPORT

Overview of the code: -

For scraping:-

Requests to access the data

Pandas to create the dataset

BeautifulSoup for scraping the data

NumPy for reshaping the data

Twitter Data: -

Twitter API for the data

Requests to request the data

Pandas to create the dataframe

Twython for working with twitter API

For creating the database :-

We used MySQL to create the database.

LucidChart for creating the ERD Diagram.

CONCLUSION

Working on this project gave an all-round sense of database management. From web scraping to extracting data from twitter API, relevant information was extracted to build this consolidated data set. Using MySQL to build this dataset was a challenging aspect.

Over the duration, we got knowledge of how inconsistency can cause issues, understood how websites are designed in various manners and understood various other aspects of database management systems.

CONTRIBUTION

Own Contribution: 60%

Professor: 30%

External: 10%

CITATIONS

https://www.diffen.com/difference/Bank_of_America_vs_JPMorgan_Chase

https://www.diffen.com/difference/Goldman_Sachs_vs_Wells_Fargo

https://www.diffen.com/difference/Bank_of_America_vs_Citigroup

https://www.goldmansachs.com/careers/students/programs/index.html

https://jobs.citi.com/search-jobs?glat=38.9071998596191&glon=-77.0369033813477

https://careers.jpmorgan.com/us/en/students/programs?search=&tags=location__Americas__UnitedStatesofAmerica

https://employment.wellsfargo.com/psc/PSEA/APPLICANT_NW/HRMS/c/HRS_HRAM_FL.HRS_CG_SEARCH_FL.GBL?Page=HRS_APP_SCHJOB_FL&Action=U

https://careers.bankofamerica.com/en-us/job-search?ref=search&search=jobsByCityState&city=herndon&state=virginia&country=united%20states&searchstring=herndon,%20virginia&start=0&rows=10

https://developer.twitter.com/en/docs/api-reference-index

https://github.com/nikbearbrown/INFO_6210/blob/master/Project/Jobs_DB_Project/Jobs_DB_Project.pdf

https://youtu.be/cdAYhLSXDj4

https://stackabuse.com/accessing-the-twitter-api-with-python/

LICENSE

SUBMITTED BY: -

Rohan Saji George

https://github.com/Rohan-George5

Yash Mahansaria

https://github.com/yashmahansaria