Course Project Reddit Scrape

Authors: Joel Quek (SG)

Web Scraping for r/investing

In [1]:

```
import requests
import pandas as pd
import time
from datetime import datetime

import random

import json
import csv
```

In [2]:

```
url = 'https://api.pushshift.io/reddit/search/submission'
```

Function to Pull 5000 rows of data

Current Unix Timestamp [Epoch Converter] (https://www.epochconverter.com/)

In [3]:

```
presentDate = datetime.now()
unix_timestamp = datetime.timestamp(presentDate)
print(unix_timestamp)
```

```
1667611970.297744
```

https://www.epochconverter.com/ (https://www.epochconverter.com/)

Function

In [4]:

```
def PushShift5000(sub, size, present):
    url = 'https://api.pushshift.io/reddit/search/submission'
    #---------------------------------------------------------------
    params ={
        'subreddit': str(sub),
        'size': int(size),
        'before': int(present)
    }
    res = requests.get(url,params)
    data=res.json()
    posts = data['data']
    df=pd.DataFrame(posts)
    max_size = df.shape[0]-1
    #---------------------------------------------------------------
    while df.shape[0] < size:
        params2 ={
            'subreddit': str(sub),
            'size': int(size),
            'before': posts[max_size]['created_utc']
        }
        res2 = requests.get(url,params2)
        data2=res2.json()
        posts = data2['data']
        df2=pd.DataFrame(posts)
        df=pd.concat([df,df2],ignore_index=True)
    return df
```

r/investing

In [11]:

```
investing = PushShift5000('investing', 4000, unix_timestamp)
```

In [12]:

```
investing.head()
```

Out[12]:

	all_awardings	allow_live_comments	author	author_flair_css_class	author_flair
0	[]	False	HomeInvading	None	
1	[]	False	ocean-airseashell10	None	
2	[]	False	ocean-airseashell10	None	
3	[]	False	iamjokingiamserious	None	
4	[]	False	jamesterryburke01	None	

5 rows × 74 columns

In [13]:

```
investing.shape[0]
```

Out[13]:

```
4000
```

In [14]:

```
investing.iloc[investing.shape[0]-1]['created_utc']
```

Out[14]:

```
1662586585
```

In [15]:

```
investing2 = PushShift5000('investing', 1000, investing.iloc[investing.shape[0]-1]['creat
```

In [16]:

```
investing2.shape[0]
```

Out[16]:

```
1248
```

In [17]:

```
investing=pd.concat([investing,investing2],ignore_index=True)
investing.shape
```

Out[17]:

```
(5248, 74)
```

In [55]:

```
investing3 = PushShift5000('investing', 1000, investing2.iloc[investing2.shape[0]-1]['cre
```

In [56]:

```
investing=pd.concat([investing,investing3],ignore_index=True)
investing.shape
```

Out[56]:

```
(6248, 74)
```

In [60]:

```
investing4 = PushShift5000('investing', 1000, investing3.iloc[investing3.shape[0]-1]['cre
```

In [61]:

```
investing=pd.concat([investing,investing4],ignore_index=True)
investing.shape
```

Out[61]:

```
(7248, 74)
```

In [63]:

```
investing5 = PushShift5000('investing', 500, investing4.iloc[investing4.shape[0]-1]['crea
```

In [64]:

```
investing=pd.concat([investing,investing5],ignore_index=True)
investing.shape
```

Out[64]:

```
(7995, 74)
```

In [66]:

```
investing[['subreddit', 'author', 'selftext', 'title']].head(10)
```

Out[66]:

	subreddit	author	selftext	title
0	investing	HomeInvading	Hey guys, I’m a 22 year old male, I grew up wi...	Help a young man out would ya?
1	investing	ocean-airseashell10	[removed]	Treasury bonds is it a good idea to buy
2	investing	ocean-airseashell10	[removed]	How to buy treasury bonds? Is treasury’s direc...
3	investing	iamjokingiamserious	[removed]	Early Exercise of Stock Options
4	investing	jamesterryburke01	Hello Redditors 👋 \n\nI work as a Investment C...	Alternative Investments -
5	investing	James_OrangeRiver	[removed]	Alternative Investing - Interested to hear the...
6	investing	woke4evainfinity	[removed]	Best way to invest $250k? what do rich people do?
7	investing	Whistleblower793	Back around 2005ish, I remember reading an ad ...	Can someone explain why “your house is not a r...
8	investing	Lonely_Possibility92	[removed]	silly question for saving bond
9	investing	FlounderFair9656	[removed]	What is the next wallstreetbets?

In [67]:

```
investing.to_csv('datasets/investing.csv')
```

In [68]:

```
investing.iloc[investing.shape[0]-1]['created_utc']
```

Out[68]:

```
1657271926
```

In [21]:

```
datetime.now(int(1660670168))
```

```
---------------------------------------------------------------------------
-
TypeError                                 Traceback (most recent call las
t)
c:\Users\redoc\OneDrive\Desktop\DSI-Roughpaper\0. project_3 (DO THIS)\redd
it-scrape.ipynb Cell 22 in <cell line: 1>()
----> <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI
-Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#X30sZmlsZ
Q%3D%3D?line=0'>1</a> datetime.now(int(1660670168))

TypeError: tzinfo argument must be None or of a tzinfo subclass, not type
'int'
```

r/stockmarket

In [27]:

```
stockmarket = PushShift5000('StockMarket', 3000, unix_timestamp)
```

In [28]:

```
stockmarket.shape
```

Out[28]:

```
(3000, 80)
```

In [29]:

```
stockmarket.iloc[2499]['created_utc']
```

Out[29]:

```
1664718418
```

In [31]:

```
stockmarket2 = PushShift5000('StockMarket', 2000, stockmarket.iloc[stockmarket.shape[0]-1
```

In [32]:

```
stockmarket=pd.concat([stockmarket,stockmarket2],ignore_index=True)
stockmarket.shape
```

Out[32]:

```
(5248, 80)
```

In [33]:

```
stockmarket.iloc[0]
```

Out[33]:

```
all_awardings                    []
allow_live_comments           False
author                     zitrored
author_flair_css_class         None
author_flair_richtext            []
                             ...   
author_flair_text_color         NaN
gallery_data                    NaN
media_metadata                  NaN
author_cakeday                  NaN
banned_by                       NaN
Name: 0, Length: 80, dtype: object
```

In [36]:

```
stockmarket.iloc[3000]
```

Out[36]:

```
all_awardings                              []
allow_live_comments                     False
author                     Obvious-Expert-007
author_flair_css_class                   None
author_flair_richtext                      []
                                  ...         
author_flair_text_color                   NaN
gallery_data                              NaN
media_metadata                            NaN
author_cakeday                            NaN
banned_by                                 NaN
Name: 3000, Length: 80, dtype: object
```

In [69]:

```
stockmarket3 = PushShift5000('StockMarket', 1000, stockmarket2.iloc[stockmarket2.shape[0]
```

In [70]:

```
stockmarket=pd.concat([stockmarket,stockmarket3],ignore_index=True)
stockmarket.shape
```

Out[70]:

```
(6494, 80)
```

In [72]:

```
stockmarket4 = PushShift5000('StockMarket', 1000, stockmarket3.iloc[stockmarket3.shape[0]
```

In [73]:

```
stockmarket=pd.concat([stockmarket,stockmarket4],ignore_index=True)
stockmarket.shape
```

Out[73]:

```
(7494, 80)
```

In [77]:

```
# stockmarket5 = PushShift5000('StockMarket', 500, stockmarket4.iloc[stockmarket4.shape[
```

```
---------------------------------------------------------------------------
-
IndexError                                Traceback (most recent call las
t)
c:\Users\redoc\OneDrive\Desktop\DSI-Roughpaper\0. project_3 (DO THIS)\redd
it-scrape.ipynb Cell 41 in <cell line: 1>()
----> <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI
-Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y102sZmls
ZQ%3D%3D?line=0'>1</a> stockmarket5 = PushShift5000('StockMarket', 500, st
ockmarket4.iloc[stockmarket4.shape[0]-1]['created_utc'])

c:\Users\redoc\OneDrive\Desktop\DSI-Roughpaper\0. project_3 (DO THIS)\redd
it-scrape.ipynb Cell 41 in PushShift5000(sub, size, present)
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y102sZmlsZ
Q%3D%3D?line=13'>14</a> #--------------------------------------------------
-----------------
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y102sZmlsZ
Q%3D%3D?line=14'>15</a> while df.shape[0] < size:
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y102sZmlsZ
Q%3D%3D?line=15'>16</a>     params2 ={
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y102sZmlsZ
Q%3D%3D?line=16'>17</a>         'subreddit': str(sub),
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y102sZmlsZ
Q%3D%3D?line=17'>18</a>         'size': int(size),
---> <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y102sZmlsZ
Q%3D%3D?line=18'>19</a>         'before': posts[max_size]['created_utc']
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y102sZmlsZ
Q%3D%3D?line=19'>20</a>     }
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y102sZmlsZ
Q%3D%3D?line=20'>21</a>     res2 = requests.get(url,params2)
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y102sZmlsZ
Q%3D%3D?line=21'>22</a>     data2=res2.json()

IndexError: list index out of range
```

In [ ]:

```
#stockmarket=pd.concat([stockmarket,stockmarket5],ignore_index=True)
#stockmarket.shape
```

In [78]:

```
stockmarket.to_csv('datasets/stockmarket.csv')
```

In [79]:

```
stockmarket.iloc[stockmarket.shape[0]-1]['created_utc']
```

Out[79]:

```
1657678438
```

r/wallstreetbets

In [45]:

```
wallstreetbets = PushShift5000('wallstreetbets', 2500, unix_timestamp)
```

In [49]:

```
pd.set_option('display.max_columns', None)

print(wallstreetbets.shape)

wallstreetbets.head()
```

```
(2500, 83)
```

Out[49]:

	all_awardings	allow_live_comments	author	author_flair_css_class	author_flair_
0	[]	False	Icy_Finance_23	None	
1	[]	False	JohnnyFaang	None	
2	[]	False	JohnnyFaang	None	
3	[]	False	Scaldie-B	None	
4	[]	False	barackosamathe3rd	None	

In [53]:

```
wallstreetbets.iloc[wallstreetbets.shape[0]-1]['created_utc']
```

Out[53]:

```
1667300528
```

In [50]:

```
wallstreetbets2 = PushShift5000('wallstreetbets', 500, wallstreetbets.iloc[wallstreetbets
```

```
---------------------------------------------------------------------------
-
IndexError                                Traceback (most recent call las
t)
c:\Users\redoc\OneDrive\Desktop\DSI-Roughpaper\0. project_3 (DO THIS)\redd
it-scrape.ipynb Cell 36 in <cell line: 1>()
----> <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI
-Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#X50sZmlsZ
Q%3D%3D?line=0'>1</a> wallstreetbets2 = PushShift5000('wallstreetbets', 50
0, wallstreetbets.iloc[wallstreetbets.shape[0]-1]['created_utc'])

c:\Users\redoc\OneDrive\Desktop\DSI-Roughpaper\0. project_3 (DO THIS)\redd
it-scrape.ipynb Cell 36 in PushShift5000(sub, size, present)
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#X50sZmlsZ
Q%3D%3D?line=13'>14</a> #--------------------------------------------------
-----------------
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#X50sZmlsZ
Q%3D%3D?line=14'>15</a> while df.shape[0] < size:
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#X50sZmlsZ
Q%3D%3D?line=15'>16</a>     params2 ={
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#X50sZmlsZ
Q%3D%3D?line=16'>17</a>         'subreddit': str(sub),
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#X50sZmlsZ
Q%3D%3D?line=17'>18</a>         'size': int(size),
---> <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#X50sZmlsZ
Q%3D%3D?line=18'>19</a>         'before': posts[max_size]['created_utc']
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#X50sZmlsZ
Q%3D%3D?line=19'>20</a>     }
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#X50sZmlsZ
Q%3D%3D?line=20'>21</a>     res2 = requests.get(url,params2)
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#X50sZmlsZ
Q%3D%3D?line=21'>22</a>     data2=res2.json()

IndexError: list index out of range
```

In [80]:

```
wallstreetbets2 = PushShift5000('wallstreetbets', 2500, 1667300528)
```

```
---------------------------------------------------------------------------
-
IndexError                                Traceback (most recent call las
t)
c:\Users\redoc\OneDrive\Desktop\DSI-Roughpaper\0. project_3 (DO THIS)\redd
it-scrape.ipynb Cell 50 in <cell line: 1>()
----> <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI
-Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y104sZmls
ZQ%3D%3D?line=0'>1</a> wallstreetbets2 = PushShift5000('wallstreetbets', 2
500, 1667300528)

c:\Users\redoc\OneDrive\Desktop\DSI-Roughpaper\0. project_3 (DO THIS)\redd
it-scrape.ipynb Cell 50 in PushShift5000(sub, size, present)
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y104sZmlsZ
Q%3D%3D?line=13'>14</a> #--------------------------------------------------
-----------------
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y104sZmlsZ
Q%3D%3D?line=14'>15</a> while df.shape[0] < size:
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y104sZmlsZ
Q%3D%3D?line=15'>16</a>     params2 ={
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y104sZmlsZ
Q%3D%3D?line=16'>17</a>         'subreddit': str(sub),
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y104sZmlsZ
Q%3D%3D?line=17'>18</a>         'size': int(size),
---> <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y104sZmlsZ
Q%3D%3D?line=18'>19</a>         'before': posts[max_size]['created_utc']
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y104sZmlsZ
Q%3D%3D?line=19'>20</a>     }
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y104sZmlsZ
Q%3D%3D?line=20'>21</a>     res2 = requests.get(url,params2)
     <a href='vscode-notebook-cell:/c%3A/Users/redoc/OneDrive/Desktop/DSI-
Roughpaper/0.%20project_3%20%28DO%20THIS%29/reddit-scrape.ipynb#Y104sZmlsZ
Q%3D%3D?line=21'>22</a>     data2=res2.json()

IndexError: list index out of range
```

In [ ]:

```
wallstreetbets=pd.concat([wallstreetbets,wallstreetbets2],ignore_index=True)
wallstreetbets.shape
```

In [42]:

```
wallstreetbets.shape[0]-1
```

Out[42]:

```
2499
```

In [43]:

```
wallstreetbets.to_csv('datasets/wallstreetbets.csv')
```

In [ ]:

```
# wallstreetbets2 = PushShift5000('wallstreetbets', 500, wallstreetbets.iloc[wallstreetbe
```