Course Project Wallstreet Bets Scrape

Authors: Joel Quek (SG)

In [30]:

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

import random

import json
import csv
```

In [31]:

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

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

In [32]:

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

```
1667627085.067862
```

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

Function

In [33]:

```
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/wallstreetbets

In [34]:

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

In [35]:

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

print(wallstreetbets.shape)

wallstreetbets.head(3)
```

```
(2500, 83)
```

Out[35]:

	all_awardings	allow_live_comments	author	author_flair_css_class	author_flair_richte
0	[]	False	Pro-Gambler99	None	
1	[]	False	Fit_One4445	None	
2	[]	False	Plastic-Ad-2191	None	

In [36]:

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

Out[36]:

```
1667308264
```

In [37]:

```
wallstreetbets2 = PushShift5000('wallstreetbets', 500, 1667308264)
```

In [38]:

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

Out[38]:

```
(3000, 83)
```

In [39]:

```
wallstreetbets2.iloc[wallstreetbets2.shape[0]-1]['created_utc']
```

Out[39]:

```
1667234194
```

In [40]:

```
wallstreetbets3 = PushShift5000('wallstreetbets', 500, 1667234194)
```

In [41]:

```
wallstreetbets=pd.concat([wallstreetbets,wallstreetbets3],ignore_index=True)
wallstreetbets.shape
```

Out[41]:

```
(3749, 83)
```

In [42]:

```
wallstreetbets3.iloc[wallstreetbets3.shape[0]-1]['created_utc']
```

Out[42]:

```
1667081029
```

In [43]:

```
wallstreetbets4 = PushShift5000('wallstreetbets', 500, 1667081029)
```

In [44]:

```
wallstreetbets=pd.concat([wallstreetbets,wallstreetbets4],ignore_index=True)
wallstreetbets.shape
```

Out[44]:

```
(4498, 83)
```

In [45]:

```
wallstreetbets4.iloc[wallstreetbets4.shape[0]-1]['created_utc']
```

Out[45]:

```
1666977111
```

In [46]:

```
wallstreetbets5 = PushShift5000('wallstreetbets', 500, 1666977111)
```

In [47]:

```
wallstreetbets=pd.concat([wallstreetbets,wallstreetbets5],ignore_index=True)
wallstreetbets.shape
```

Out[47]:

```
(4998, 83)
```

In [48]:

```
wallstreetbets5.iloc[wallstreetbets5.shape[0]-1]['created_utc']
```

Out[48]:

```
1666922376
```

In [51]:

```
wallstreetbets6 = PushShift5000('wallstreetbets', 500, 1666922376)
```

In [52]:

```
wallstreetbets=pd.concat([wallstreetbets,wallstreetbets6],ignore_index=True)
wallstreetbets.shape
```

Out[52]:

```
(5498, 83)
```

In [53]:

```
wallstreetbets6.iloc[wallstreetbets6.shape[0]-1]['created_utc']
```

Out[53]:

```
1666890339
```

In [55]:

```
wallstreetbets7 = PushShift5000('wallstreetbets', 500, 1666890339)
```

In [56]:

```
wallstreetbets=pd.concat([wallstreetbets,wallstreetbets7],ignore_index=True)
wallstreetbets.shape
```

Out[56]:

```
(5998, 83)
```

In [57]:

```
wallstreetbets7.iloc[wallstreetbets6.shape[0]-1]['created_utc']
```

Out[57]:

```
1666828956
```

In [58]:

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