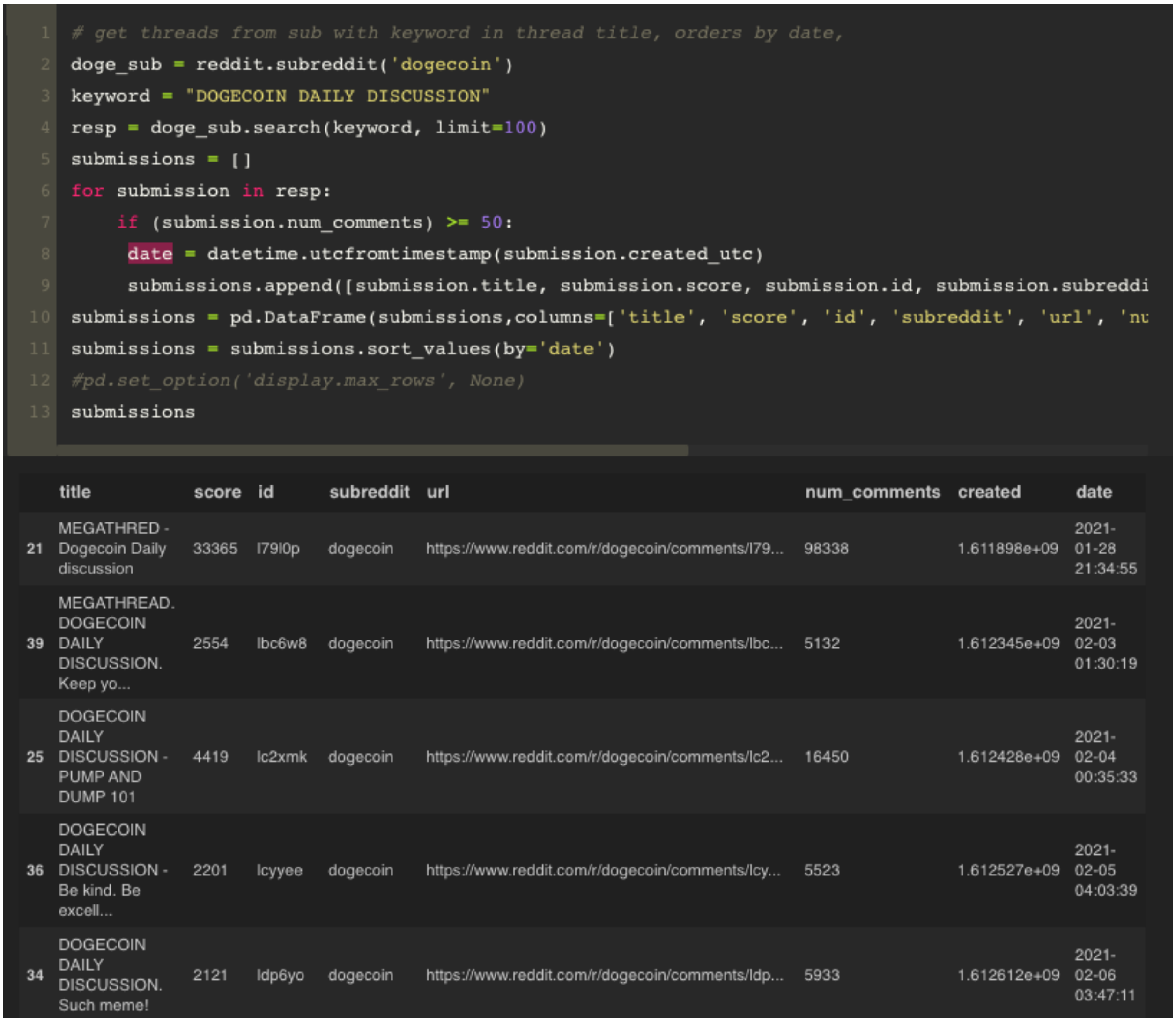
**subReddit Extraction Process using praw, pmaw, pushshift.io**
[notebook link]
install and import dependencies praw, pmaw
set-up PRAW (Reddit API) client ID and client secret using Oauth2 [link]

using PRAW Reddit API, connect to dogecoin subreddit to extract threads containing keywords “DOGECOIN DAILY DISCUSSION”, select chosen date range, set parameters for extraction of thread id’s with >50 comments.
extract and append necessary values for threads to new dataframe, sort by date

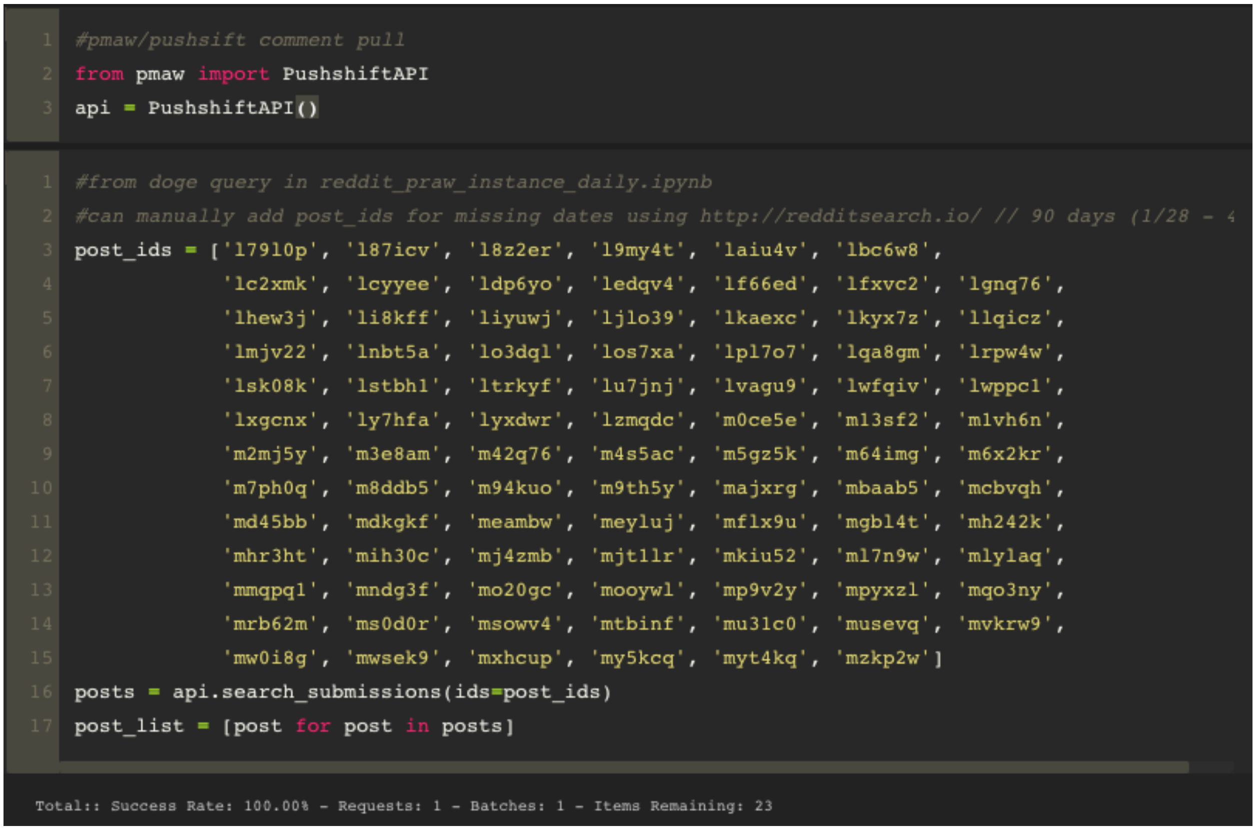


```
# get threads from sub with keyword in thread title, orders by date,
doge_sub = reddit.subreddit('dogecoin')
keyword = "DOGECOIN DAILY DISCUSSION"
resp = doge_sub.search(keyword, limit=100)
submissions = []
for submission in resp:
    if (submission.num_comments) >= 50:
     date = datetime.utcfromtimestamp(submission.created_utc)
     submissions.append([submission.title, submission.score, submission.id, submission.subreddi
submissions = pd.DataFrame(submissions,columns=['title', 'score', 'id', 'subreddit', 'url', 'nu
submissions = submissions.sort_values(by='date')
#pd.set_option('display.max_rows', None)
submissions
```

| | title | score | id | subreddit | url | num_comments | created | date |
|---|---|---|---|---|---|---|---|---|
| 21 | MEGATHRED - Dogecoin Daily discussion | 33365 | l79l0p | dogecoin | https://www.reddit.com/r/dogecoin/comments/l79... | 98338 | 1.611898e+09 | 2021-01-28 21:34:55 |
| 39 | MEGATHREAD. DOGECOIN DAILY DISCUSSION. Keep yo... | 2554 | lbc6w8 | dogecoin | https://www.reddit.com/r/dogecoin/comments/lbc... | 5132 | 1.612345e+09 | 2021-02-03 01:30:19 |
| 25 | DOGECOIN DAILY DISCUSSION - PUMP AND DUMP 101 | 4419 | lc2xmk | dogecoin | https://www.reddit.com/r/dogecoin/comments/lc2... | 16450 | 1.612428e+09 | 2021-02-04 00:35:33 |
| 36 | DOGECOIN DAILY DISCUSSION - Be kind. Be excell... | 2201 | lcyyee | dogecoin | https://www.reddit.com/r/dogecoin/comments/lcy... | 5523 | 1.612527e+09 | 2021-02-05 04:03:39 |
| 34 | DOGECOIN DAILY DISCUSSION. Such meme! | 2121 | ldp6yo | dogecoin | https://www.reddit.com/r/dogecoin/comments/ldp... | 5933 | 1.612612e+09 | 2021-02-06 03:47:11 |

Make note of missing dates within date range for thread id extraction
for missing thread id’s use http://redditsearch.io/ to search for threads containing the most comments for each of the missing dates, record thread id
add thread id to list of post_ids

Using PMAW, a third-party wrapper, and Pushshift.io a third-party Reddit API that makes available non-extractable Reddit API data (i.e. batch comments for multiple thread id's) check to ensure availability of thread id's

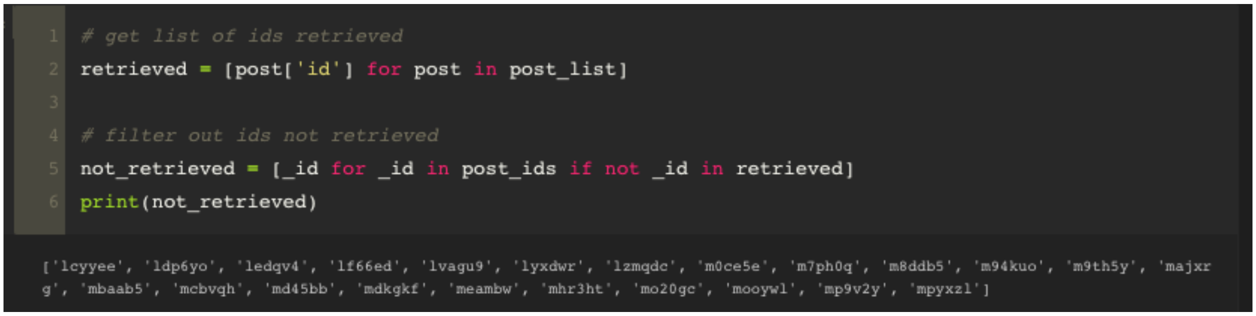


```
#pmaw/pushsift comment pull
from pmaw import PushshiftAPI
api = PushshiftAPI()
```

```
#from doge query in reddit_praw_instance_daily.ipynb
#can manually add post_ids for missing dates using http://redditsearch.io/ // 90 days (1/28 - 4
post_ids = ['l79l0p', 'l87icv', 'l8z2er', 'l9my4t', 'laiu4v', 'lbc6w8',
            'lc2xmk', 'lcyyee', 'ldp6yo', 'ledqv4', 'lf66ed', 'lfxvc2', 'lgnq76',
            'lhew3j', 'li8kff', 'liyuwj', 'ljlo39', 'lkaexc', 'lkyx7z', 'llqicz',
            'lmjv22', 'lnbt5a', 'lo3dql', 'los7xa', 'lpl7o7', 'lqa8gm', 'lrpw4w',
            'lsk08k', 'lstbhl', 'ltrkyf', 'lu7jnj', 'lvagu9', 'lwfqiv', 'lwppcl',
            'lxgcnx', 'ly7hfa', 'lyxdwr', 'lzmqdc', 'm0ce5e', 'm13sf2', 'm1vh6n',
            'm2mj5y', 'm3e8am', 'm42q76', 'm4s5ac', 'm5gz5k', 'm64img', 'm6x2kr',
            'm7ph0q', 'm8ddb5', 'm94kuo', 'm9th5y', 'majxrg', 'mbaab5', 'mcbvqh',
            'md45bb', 'mdkgkf', 'meambw', 'meyluj', 'mflx9u', 'mgbl4t', 'mh242k',
            'mhr3ht', 'mih30c', 'mj4zmb', 'mjtllr', 'mkiu52', 'ml7n9w', 'mlylaq',
            'mmqpql', 'mndg3f', 'mo20gc', 'mooywl', 'mp9v2y', 'mpyxzl', 'mqo3ny',
            'mrb62m', 'ms0d0r', 'msowv4', 'mtbinf', 'mu31c0', 'musevq', 'mvkrw9',
            'mw0i8g', 'mwsek9', 'mxhcup', 'my5kcq', 'myt4kq', 'mzkp2w']
posts = api.search_submissions(ids=post_ids)
post_list = [post for post in posts]
```

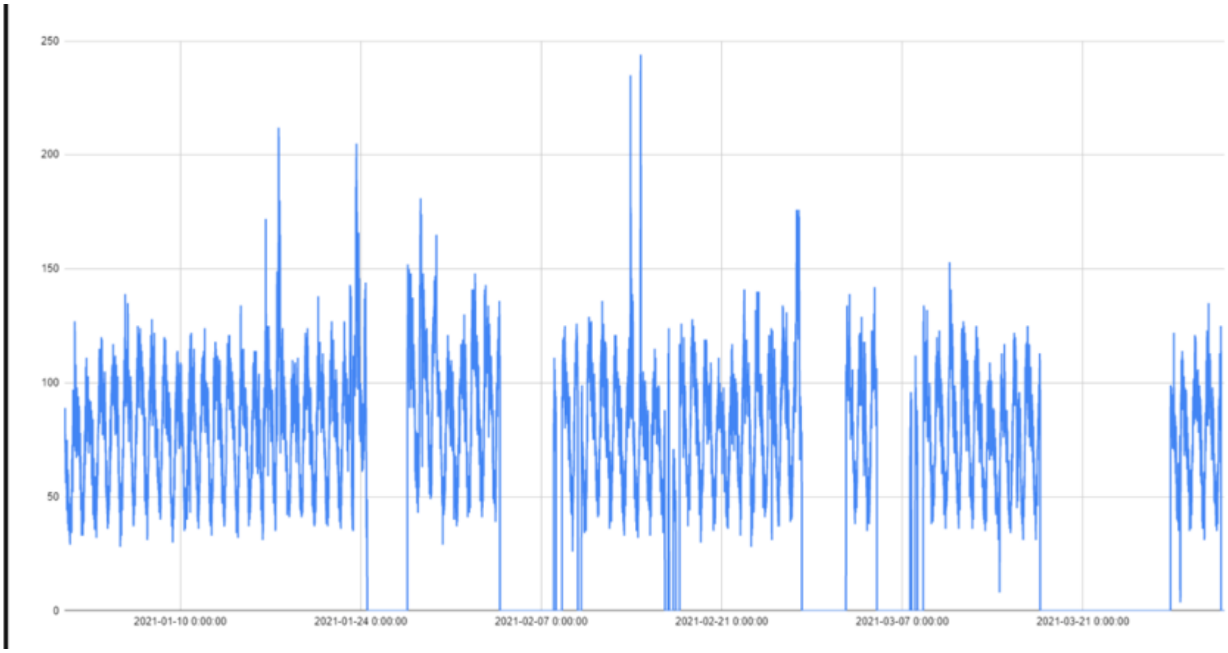
```
Total:: Success Rate: 100.00% - Requests: 1 - Batches: 1 - Items Remaining: 23
```

print list of missing thread id's from Pushshift.io (for use later)



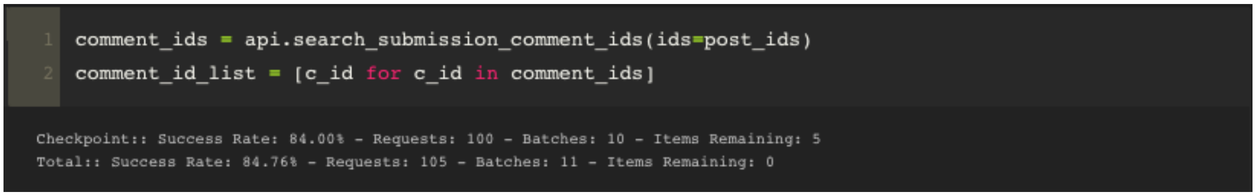
```
# get list of ids retrieved
retrieved = [post['id'] for post in post_list]

# filter out ids not retrieved
not_retrieved = [_id for _id in post_ids if not _id in retrieved]
print(not_retrieved)
```

```
['lcyyee', 'ldp6yo', 'ledqv4', 'lf66ed', 'lvagu9', 'lyxdwr', 'lzmqdc', 'm0ce5e', 'm7ph0q', 'm8ddb5', 'm94kuo', 'm9th5y', 'majxr
g', 'mbaab5', 'mcbvqh', 'md45bb', 'mdkgkf', 'meambw', 'mhr3ht', 'mo20gc', 'mooywl', 'mp9v2y', 'mpyxzl']
```



Data gaps in Pushshift.io for selected date range

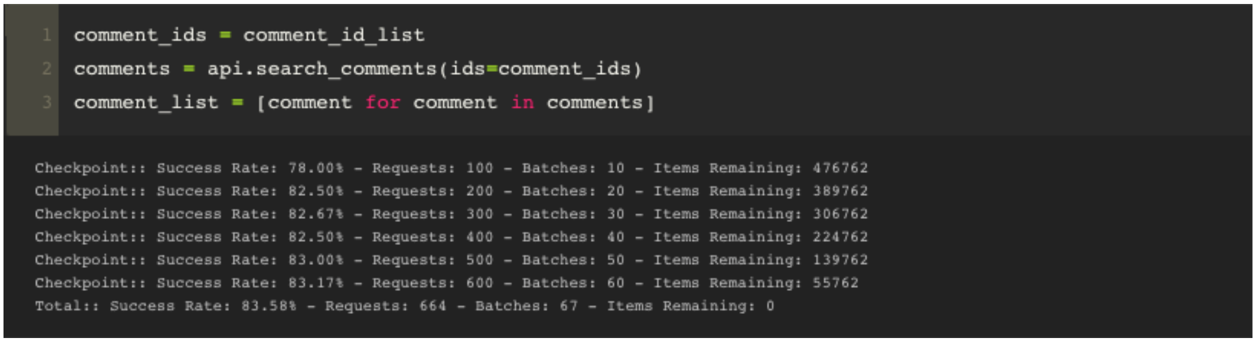
for each of the available thread id's found using PMAW, extract comments using comment id's found in thread id's.



```
comment_ids = api.search_submission_comment_ids(ids=post_ids)
comment_id_list = [c_id for c_id in comment_ids]
```

```
Checkpoint:: Success Rate: 84.00% - Requests: 100 - Batches: 10 - Items Remaining: 5
Total:: Success Rate: 84.76% - Requests: 105 - Batches: 11 - Items Remaining: 0
```

For each comment id, extract comment information using pmaw batch extraction and set to dataframe



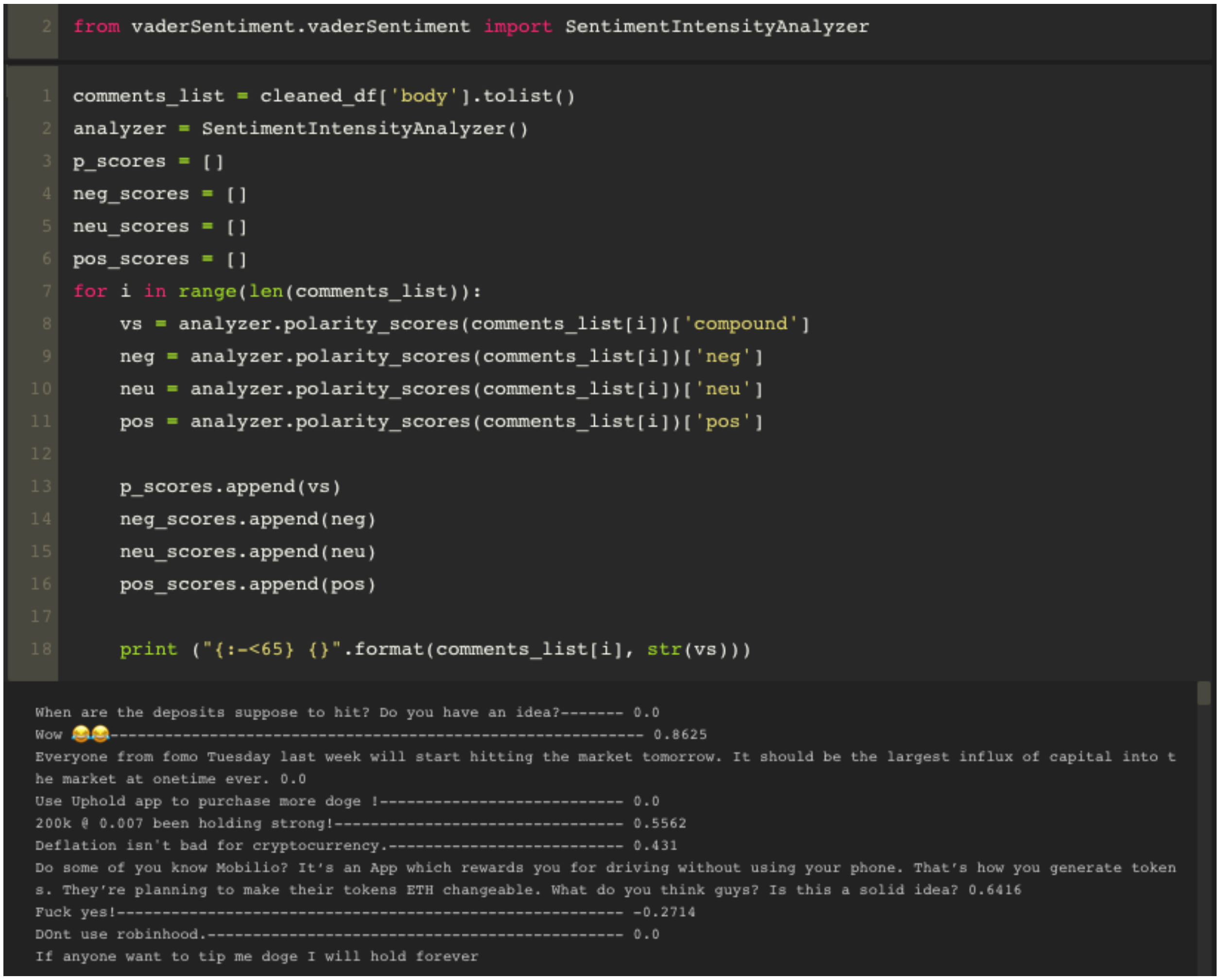
```
comment_ids = comment_id_list
comments = api.search_comments(ids=comment_ids)
comment_list = [comment for comment in comments]
```

```
Checkpoint:: Success Rate: 78.00% - Requests: 100 - Batches: 10 - Items Remaining: 476762
Checkpoint:: Success Rate: 82.50% - Requests: 200 - Batches: 20 - Items Remaining: 389762
Checkpoint:: Success Rate: 82.67% - Requests: 300 - Batches: 30 - Items Remaining: 306762
Checkpoint:: Success Rate: 82.50% - Requests: 400 - Batches: 40 - Items Remaining: 224762
Checkpoint:: Success Rate: 83.00% - Requests: 500 - Batches: 50 - Items Remaining: 139762
Checkpoint:: Success Rate: 83.17% - Requests: 600 - Batches: 60 - Items Remaining: 55762
Total:: Success Rate: 83.58% - Requests: 664 - Batches: 67 - Items Remaining: 0
```

Filter out deleted comments from dataframe

Convert extracted utc column to datetime

Run comments through VADER Analyzer
[1]

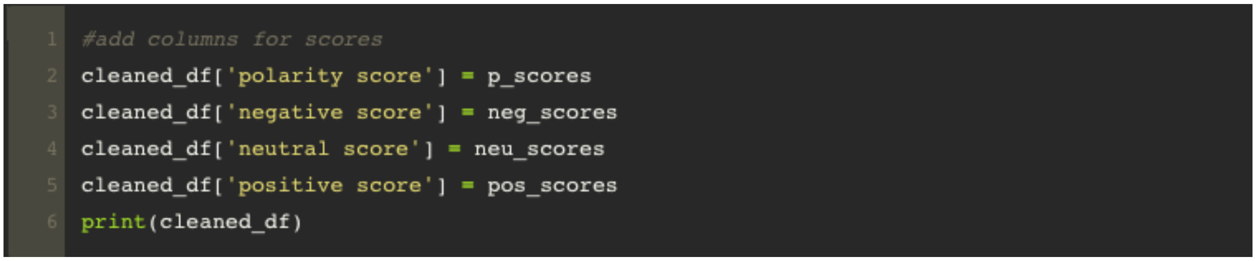


```python
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
```

```python
comments_list = cleaned_df['body'].tolist()
analyzer = SentimentIntensityAnalyzer()
p_scores = []
neg_scores = []
neu_scores = []
pos_scores = []
for i in range(len(comments_list)):
    vs = analyzer.polarity_scores(comments_list[i])['compound']
    neg = analyzer.polarity_scores(comments_list[i])['neg']
    neu = analyzer.polarity_scores(comments_list[i])['neu']
    pos = analyzer.polarity_scores(comments_list[i])['pos']

    p_scores.append(vs)
    neg_scores.append(neg)
    neu_scores.append(neu)
    pos_scores.append(pos)

    print ("{:-<65} {}".format(comments_list[i], str(vs)))
```

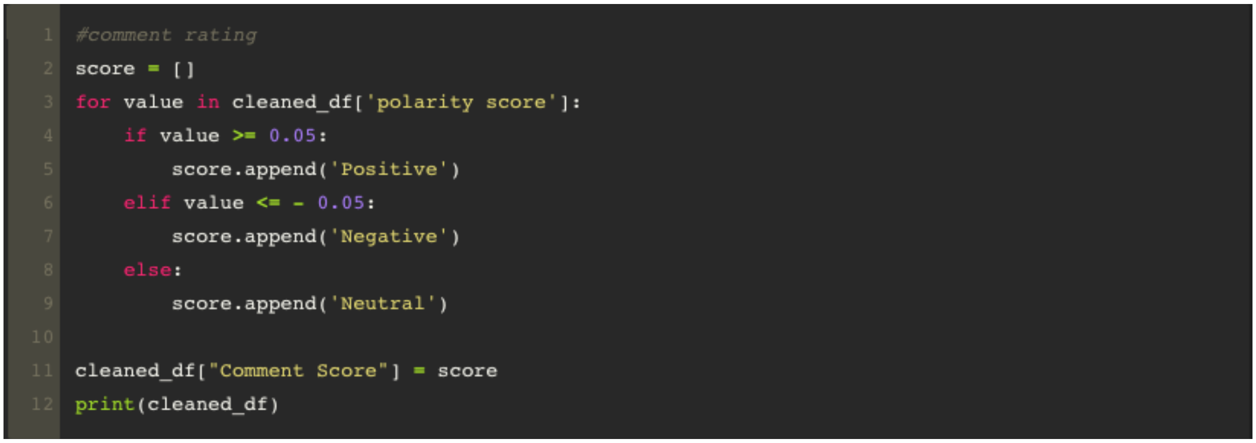
```
When are the deposits suppose to hit? Do you have an idea?------- 0.0
Wow 😂😂----------------------------------------------------------- 0.8625
Everyone from fomo Tuesday last week will start hitting the market tomorrow. It should be the largest influx of capital into t
he market at onetime ever. 0.0
Use Uphold app to purchase more doge !--------------------------- 0.0
200k @ 0.007 been holding strong!-------------------------------- 0.5562
Deflation isn't bad for cryptocurrency.-------------------------- 0.431
Do some of you know Mobilio? It’s an App which rewards you for driving without using your phone. That’s how you generate token
s. They’re planning to make their tokens ETH changeable. What do you think guys? Is this a solid idea? 0.6416
Fuck yes!-------------------------------------------------------- -0.2714
DOnt use robinhood.---------------------------------------------- 0.0
If anyone want to tip me doge I will hold forever
```

Create columns in dataframe for extracted scores
[2]



```python
#add columns for scores
cleaned_df['polarity score'] = p_scores
cleaned_df['negative score'] = neg_scores
cleaned_df['neutral score'] = neu_scores
cleaned_df['positive score'] = pos_scores
print(cleaned_df)
```

Using VADER determined compound/polarity score for each comment, set value ranges to an overall Comment Score of Positive, Negative, or Neutral
Append Comment Score column to dataframe
[3]



```python
#comment rating
score = []
for value in cleaned_df['polarity score']:
    if value >= 0.05:
        score.append('Positive')
    elif value <= - 0.05:
        score.append('Negative')
    else:
        score.append('Neutral')

cleaned_df["Comment Score"] = score
print(cleaned_df)
```

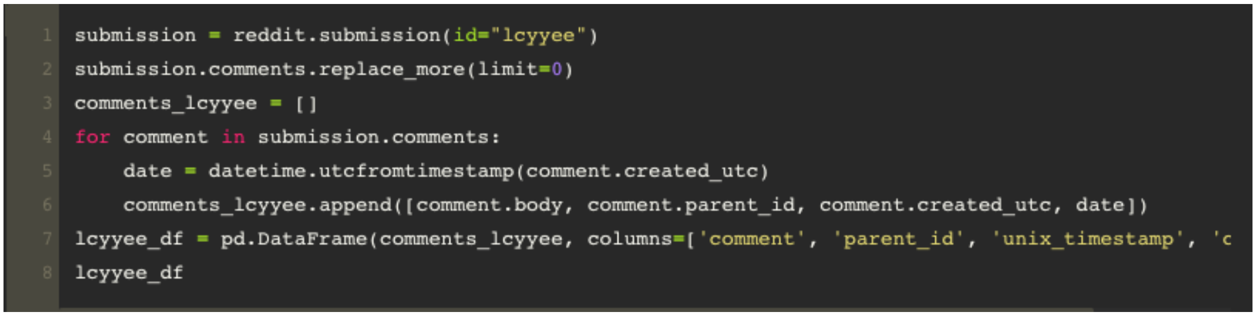
Export created dataframe to csv

**For missing thread ids**

[notebook link]

Create new script with PRAW import, set up client id, client secret

Run each of the missing thread id's through PRAW request for comment extraction



```python
submission = reddit.submission(id="lcyyee")
submission.comments.replace_more(limit=0)
comments_lcyyee = []
for comment in submission.comments:
    date = datetime.utcfromtimestamp(comment.created_utc)
    comments_lcyyee.append([comment.body, comment.parent_id, comment.created_utc, date])
lcyyee_df = pd.DataFrame(comments_lcyyee, columns=['comment', 'parent_id', 'unix_timestamp', 'c
lcyyee_df
```

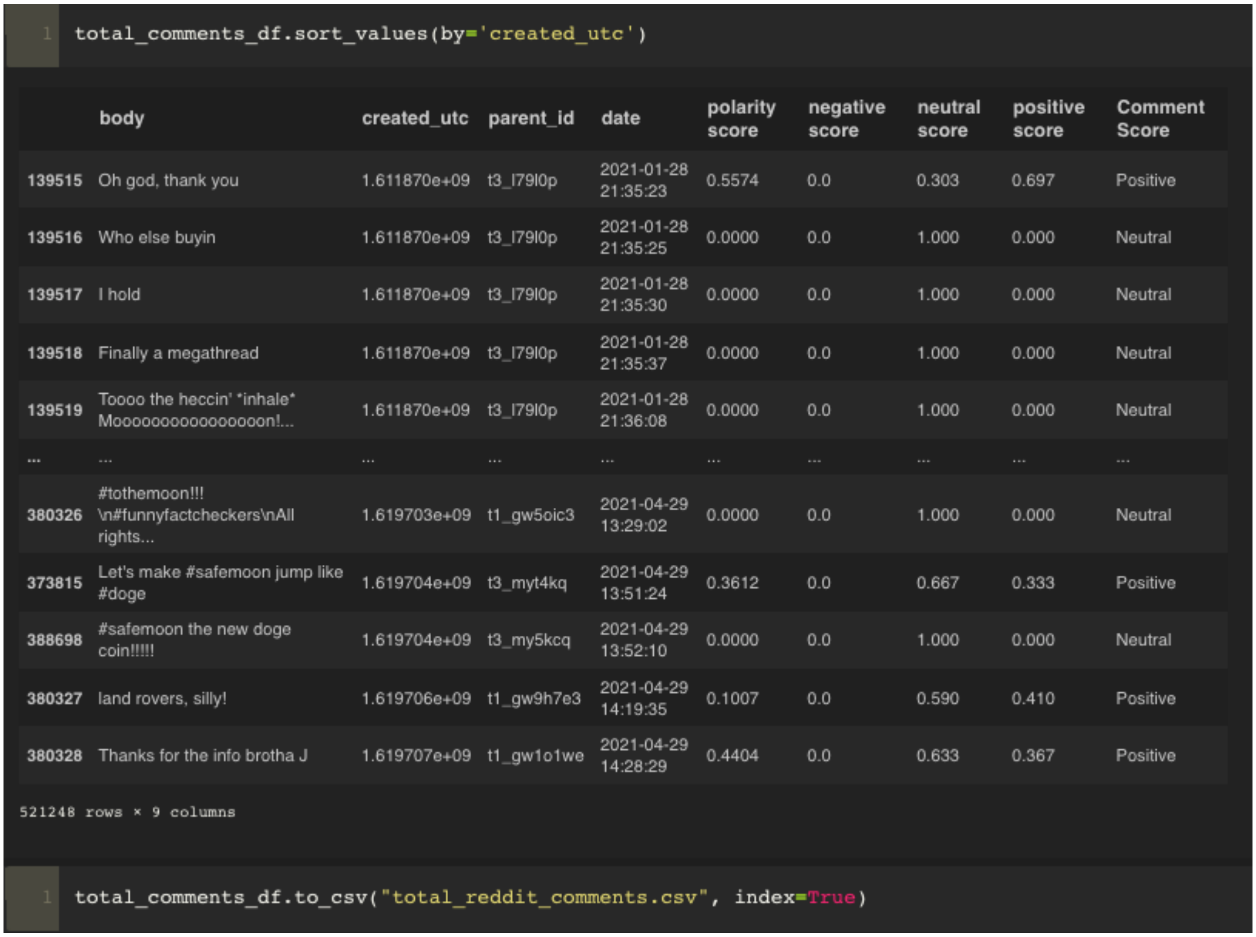
After running each submission, combine each submission dataframe and run through VADER Analyzer (steps 1 - 3)

Import csv from previous extraction and combine the two dataframes

Note: because comment extraction process was split between two API's due to missing thread submissions, the column headers of the two completed dataframes will differ, be sure to rename column headers accordingly before combining the two analyzed dataframes

Drop unnecessary columns that will not be used in final analysis, charts, or plots and sort values for final dataframe by date

Export combined dataframe for final csv



```python
total_comments_df.sort_values(by='created_utc')
```

| | body | created_utc | parent_id | date | polarity score | negative score | neutral score | positive score | Comment Score |
|---|---|---|---|---|---|---|---|---|---|
| **139515** | Oh god, thank you | 1.611870e+09 | t3_l79l0p | 2021-01-28 21:35:23 | 0.5574 | 0.0 | 0.303 | 0.697 | Positive |
| **139516** | Who else buyin | 1.611870e+09 | t3_l79l0p | 2021-01-28 21:35:25 | 0.0000 | 0.0 | 1.000 | 0.000 | Neutral |
| **139517** | I hold | 1.611870e+09 | t3_l79l0p | 2021-01-28 21:35:30 | 0.0000 | 0.0 | 1.000 | 0.000 | Neutral |
| **139518** | Finally a megathread | 1.611870e+09 | t3_l79l0p | 2021-01-28 21:35:37 | 0.0000 | 0.0 | 1.000 | 0.000 | Neutral |
| **139519** | Toooo the heccin' *inhale* Mooooooooooooooooon!... | 1.611870e+09 | t3_l79l0p | 2021-01-28 21:36:08 | 0.0000 | 0.0 | 1.000 | 0.000 | Neutral |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **380326** | #tothemoon!!! \n#funnyfactcheckers\nAll rights... | 1.619703e+09 | t1_gw5oic3 | 2021-04-29 13:29:02 | 0.0000 | 0.0 | 1.000 | 0.000 | Neutral |
| **373815** | Let's make #safemoon jump like #doge | 1.619704e+09 | t3_myt4kq | 2021-04-29 13:51:24 | 0.3612 | 0.0 | 0.667 | 0.333 | Positive |
| **388698** | #safemoon the new doge coin!!!!! | 1.619704e+09 | t3_my5kcq | 2021-04-29 13:52:10 | 0.0000 | 0.0 | 1.000 | 0.000 | Neutral |
| **380327** | land rovers, silly! | 1.619706e+09 | t1_gw9h7e3 | 2021-04-29 14:19:35 | 0.1007 | 0.0 | 0.590 | 0.410 | Positive |
| **380328** | Thanks for the info brotha J | 1.619707e+09 | t1_gw1o1we | 2021-04-29 14:28:29 | 0.4404 | 0.0 | 0.633 | 0.367 | Positive |

```
521248 rows × 9 columns
```

```python
total_comments_df.to_csv("total_reddit_comments.csv", index=True)
```