**subReddit Extraction Process using praw, pmaw, pushshift.io**

[[notebook link](https://github.com/cspence001/crypto_sentiment_analysis/blob/main/extraction_notebooks/pushshift_doge_sub.ipynb)]

* install and import dependencies praw, pmaw in notebook
* set-up PRAW (Reddit API) client ID and client secret using Oauth2 [[link](https://github.com/reddit-archive/reddit/wiki/OAuth2)]
* 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

Text

Description automatically generated

* [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

Graphical user interface, text

Description automatically generated

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

Text

Description automatically generated

Chart

Description automatically generated

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.

Graphical user interface, text

Description automatically generated

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

Text

Description automatically generated with low confidence

* Filter out deleted comments from dataframe
* Convert extracted utc column to datetime

[1] Run comments through VADER Analyzer

Text

Description automatically generated

[2] Create columns in dataframe for extracted scores

Text

Description automatically generated with medium confidence

[3] 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

Text

Description automatically generated

* Export created dataframe to csv

**For missing thread ids**

[[notebook link](https://github.com/cspence001/crypto_sentiment_analysis/blob/main/extraction_notebooks/missing_submissions.ipynb)]

* 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 as follows:Text

  Description automatically generated
* After running for 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 the comment extraction process was split between two API’s, 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 :

Graphical user interface, text

Description automatically generated