Submission Gathering

April 6, 2022

1 Setup

[1]:
```
# !pip install gensim
```

[2]:
```
# !pip install python-Levenshtein
```

[3]:
```
# !pip install praw
```

[4]:
```
# !pip install psaw
```

[5]:
```
# !pip install p
```

[6]:
```
import gensim
import pandas as pd
import praw
from psaw import PushshiftAPI
import datetime as dt
import matplotlib.pyplot as plt
```

1.1 API

PRAW Reddit API wrapper

[8]:
```
p_reddit = praw.Reddit(client_id='[redacted]',
                       client_secret='[redacted]',
                       username='[redacted]',
                       password='[redacted]',
                       user_agent='[redacted]')

p_subreddit = p_reddit.subreddit('leagueoflegends')
```

PushshiftAPI

[9]:
```
ps_api = PushshiftAPI()

subr = 'leagueoflegends'
fi = ['id','created_utc', 'url', 'title']
```

1.2 Gathering Reddit posts, or "submissions"

[10]:
```
def get_submissions(subreddit, start_time, end_time, filters):
    if(len(filters) == 0):
        filters = ['id', 'author', 'created_utc','domain', 'url','title',␣
↪'num_comments']
                    #We set by default some useful columns

    posts = list(ps_api.search_submissions(
        subreddit=subreddit,    #Subreddit we want to audit
        after=start_time,       #Start date
        before=end_time,        #End date
        filter=filters         #Column names we want to retrieve
    ))          ##Max number of posts

    return pd.DataFrame(posts)
```

Gathering submissions by period

[ ]:
```
st = '2021-06-01'
en = '2021-06-15'

june_1_2_submissions = get_submissions(subr, st, en, fi)

st = '2021-06-16'
en = '2021-06-30'

june_2_2_submissions = get_submissions(subr, st, en, fi)

st = '2021-07-01'
en = '2021-07-15'

july_1_2_submissions = get_submissions(subr, st, en, fi)

st = '2021-07-16'
en = '2021-07-31'

july_2_2_submissions = get_submissions(subr, st, en, fi)

st = '2021-08-01'
en = '2021-08-15'

aug_1_2_submissions = get_submissions(subr, st, en, fi)

st = '2021-08-16'
en = '2021-08-31'

aug_2_2_submissions = get_submissions(subr, st, en, fi)
```

Send to CSV

[ ]:
```
june_1_2_submissions.to_csv('june_1.csv')
june_2_2_submissions.to_csv('june_2.csv')
july_1_2_submissions.to_csv('july_1.csv')
july_2_2_submissions.to_csv('july_2.csv')
aug_1_2_submissions.to_csv('aug_1.csv')
aug_2_2_submissions.to_csv('aug_2.csv')
```

1.3 Cleaning functions

[39]:
```
def valid_url(u):
    """
    Tell us whether subreddit submissions url is valid

    This function takes a url from a row of the submissions dataframe
    and tells us whether the domain is a valid type of submission for
    later text processing. Namely, that the domain is either
    'https://www.reddit.com/r/leagueoflegends' or
    'https://v.redd.it'.

    Parameters
    ----------
    u : str
        url from the subreddit submissions

    Returns
    -------
    bool
        whether ``u`` is in valid domain

    Example
    -------
    >>> valid_url('https://www.reddit.com/r/leagueoflegends/s82nf81')
    True
    """
    return 'https://www.reddit.com/r/leagueoflegends' in u or 'https://v.redd.
 ↪it' in u

def get_valid_i(dframe):
    """
    Obtains indices of submissions that have valid url domains

    Taking a dataframe whose entries are submissions pulled from the
    pushshift reddit api, we obtain a list of indices of entries
    whose url domains are valid.
```

```
    Parameters
    ----------
    dframe : Pandas DataFrame
        Dataframe of leagueoflegends subreddit submissions and other
        results obtained from pushshift

    Returns
    -------
    list
        List of indices whose url domains are valid
    """
    temp = list()
    # print('type dframe:\t', type(dframe))

    for i in range(dframe.shape[0]):
        u = str(dframe.iloc[i]['url'])
        if valid_url(u):
            temp.append(i)

    return temp

def clean_table(df):
    """
    Take a dataframe of leagueoflegends subreddit submissions and
    other results from pushshift and remove the results whose url
    domains are not valid.

    Parameters
    ----------
    df : Pandas DataFrame
        Dataframe of leagueoflegends subreddit submissions and other
        results obtained from pushshift

    Returns
    -------
    Pandas DataFrame
    """
    valid_i = get_valid_i(df)

    df = df.iloc[valid_i]

    return df
```

[29]:
```
st = '2021-09-01'
en = '2021-09-30'

sept_submissions = get_submissions(subr, st, en, fi)
```

```
C:\Users\aKost\anaconda3\lib\site-packages\psaw\PushshiftAPI.py:192:
UserWarning: Got non 200 code 429
  warnings.warn("Got non 200 code %s" % response.status_code)
C:\Users\aKost\anaconda3\lib\site-packages\psaw\PushshiftAPI.py:180:
UserWarning: Unable to connect to pushshift.io. Retrying after backoff.
  warnings.warn("Unable to connect to pushshift.io. Retrying after backoff.")
```

[32]:
```
st = '2021-10-01'
en = '2021-10-07'

oct_1_4_submissions = get_submissions(subr, st, en, fi)
```

```
C:\Users\aKost\anaconda3\lib\site-packages\psaw\PushshiftAPI.py:192:
UserWarning: Got non 200 code 429
  warnings.warn("Got non 200 code %s" % response.status_code)
C:\Users\aKost\anaconda3\lib\site-packages\psaw\PushshiftAPI.py:180:
UserWarning: Unable to connect to pushshift.io. Retrying after backoff.
  warnings.warn("Unable to connect to pushshift.io. Retrying after backoff.")
```

[33]:
```
st = '2021-10-08'
en = '2021-10-14'

oct_2_4_submissions = get_submissions(subr, st, en, fi)
```

[34]:
```
st = '2021-10-15'
en = '2021-10-21'

oct_3_4_submissions = get_submissions(subr, st, en, fi)
```

[35]:
```
st = '2021-10-22'
en = '2021-10-31'

oct_4_4_submissions = get_submissions(subr, st, en, fi)
```

[36]:
```
st = '2021-11-01'
en = '2021-11-30'

november_submissions = get_submissions(subr, st, en, fi)
```

[40]:
```
#sept_submissions = pd.read_csv('sept.csv')
#sept_submissions = clean_table(sept_submissions)
#sept_submissions.to_csv('sept.csv')
#
#oct_1_4_submissions = pd.read_csv('oct_1_4.csv')
#oct_1_4_submissions = clean_table(oct_1_4_submissions)
#oct_1_4_submissions.to_csv('oct_1_4.csv')
#
```

```
#oct_2_4_submissions = pd.read_csv('oct_2_4.csv')
#oct_2_4_submissions = clean_table(oct_2_4_submissions)
#oct_2_4_submissions.to_csv('oct_2_4.csv')
#
#oct_3_4_submissions = pd.read_csv('oct_3_4.csv')
#oct_3_4_submissions = clean_table(oct_3_4_submissions)
#oct_3_4_submissions.to_csv('oct_3_4.csv')
#
#oct_4_4_submissions = pd.read_csv('oct_4_4.csv')
#oct_4_4_submissions = clean_table(oct_4_4_submissions)
#oct_4_4_submissions.to_csv('oct_4_4.csv')
#
#november_submissions = pd.read_csv('november.csv')
#november_submissions = clean_table(november_submissions)
#november_submissions.to_csv('november.csv')
```

[41]:
```
st = '2021-12-01'
en = '2021-12-31'

decemeber_submissions = get_submissions(subr, st, en, fi)

decemeber_submissions = clean_table(decemeber_submissions)

decemeber_submissions.to_csv('decemeber.csv')
```

```
C:\Users\aKost\anaconda3\lib\site-packages\psaw\PushshiftAPI.py:192:
UserWarning: Got non 200 code 429
  warnings.warn("Got non 200 code %s" % response.status_code)
C:\Users\aKost\anaconda3\lib\site-packages\psaw\PushshiftAPI.py:180:
UserWarning: Unable to connect to pushshift.io. Retrying after backoff.
  warnings.warn("Unable to connect to pushshift.io. Retrying after backoff.")
```

[42]:
```
st = '2022-01-01'
en = '2022-01-31'

january_submissions = get_submissions(subr, st, en, fi)

january_submissions = clean_table(january_submissions)

january_submissions.to_csv('january.csv')
```

[43]:
```
st = '2022-02-01'
en = '2022-02-28'

february_submissions = get_submissions(subr, st, en, fi)

february_submissions = clean_table(february_submissions)
```

```
february_submissions.to_csv('feb.csv')
```

[44]:
```
st = '2022-03-01'
en = '2022-03-31'

march_submissions = get_submissions(subr, st, en, fi)

march_submissions = clean_table(march_submissions)

march_submissions.to_csv('march.csv')
```

2 Creating a master table

Concatenation of all tables

[48]:
```
files = [
    'june_2.csv',
    'july_1.csv',
    'july_2.csv',
    'aug_1.csv',
    'aug_2.csv',
    'sept.csv',
    'oct_1_4.csv',
    'oct_2_4.csv',
    'oct_3_4.csv',
    'oct_4_4.csv',
    'november.csv',
    'decemeber.csv',
    'january.csv',
    'feb.csv',
    'march.csv'
]
```

[54]:
```
mega_table = pd.read_csv('june_1.csv')

for f in files:
    temp = pd.read_csv(f)
    mega_table = pd.concat([mega_table, temp], axis=0)
```

[55]:
```
mega_table
```

[55]:

	Unnamed: 0	Unnamed: 0.1	created_utc	id	\
0	0	0.0	1623715013	o00m44	
1	1	1.0	1623714840	o00k4w	
2	2	2.0	1623714789	o00jk8	
3	3	3.0	1623714743	o00j2e	

```
4              4           4.0   1623714562  o00gyo
…          …           …          …     …
9549       13055           NaN   1646093044  t3txtq
9550       13056           NaN   1646092994  t3tx4z
9551       13057           NaN   1646092983  t3twyu
9552       13058           NaN   1646092952  t3twhr
9553       13059           NaN   1646092817  t3tuqh

                                                  title  \
0                                  I dont cs when im fed
1     Totally planned prediction 2021 outplay that i…
2     UCAM Esports vs MAD Lions Madrid - LEAGUE OF L…
3                  One hella lucky play on my le blanc
4                        A digital painting of Wolf!!
…                                                    …
9549                                The Double Steal
9550                    Frequent crashes on Windows 11.
9551  Tarzaned leaving game after dying lvl 1, flame…
9552  Do not criticize russian aggression or Vladimi…
9553                                Lunar Revel Pass

                                                   url       created  \
0     https://www.reddit.com/r/leagueoflegends/comme…  1.623740e+09
1                       https://v.redd.it/iqthzskakb571  1.623740e+09
2     https://youtube.com/watch?v=wfHM4nX2z-A&amp;fe…  1.623740e+09
3                       https://v.redd.it/c9puqrqyjb571  1.623740e+09
4                    https://i.redd.it/73ln16dhjb571.jpg  1.623740e+09
…                                                    …           …
9549                    https://v.redd.it/f11af9f4ynk81  1.646118e+09
9550  https://www.reddit.com/r/leagueoflegends/comme…  1.646118e+09
9551                    https://v.redd.it/tpem6nm0ynk81  1.646118e+09
9552  https://www.reddit.com/r/leagueoflegends/comme…  1.646118e+09
9553  https://www.reddit.com/r/leagueoflegends/comme…  1.646118e+09

                                                    d_
0     {'created_utc': 1623715013, 'id': 'o00m44', 't…
1     {'created_utc': 1623714840, 'id': 'o00k4w', 't…
2     {'created_utc': 1623714789, 'id': 'o00jk8', 't…
3     {'created_utc': 1623714743, 'id': 'o00j2e', 't…
4     {'created_utc': 1623714562, 'id': 'o00gyo', 't…
…                                                   …
9549  {'created_utc': 1646093044, 'id': 't3txtq', 't…
9550  {'created_utc': 1646092994, 'id': 't3tx4z', 't…
9551  {'created_utc': 1646092983, 'id': 't3twyu', 't…
9552  {'created_utc': 1646092952, 'id': 't3twhr', 't…
9553  {'created_utc': 1646092817, 'id': 't3tuqh', 't…
```

```
[143567 rows x 8 columns]
```

[60]:
```
mega_table = mega_table.drop(columns=['Unnamed: 0','Unnamed: 0.1'])
```

[61]:
```
mega_table
```

[61]:
```
       created_utc      id                                                title  \
0       1623715013  o00m44                                I dont cs when im fed
1       1623714840  o00k4w  Totally planned prediction 2021 outplay that i…
2       1623714789  o00jk8  UCAM Esports vs MAD Lions Madrid - LEAGUE OF L…
3       1623714743  o00j2e                One hella lucky play on my le blanc
4       1623714562  o00gyo                       A digital painting of Wolf!!
…              …     …                                                   …
9549    1646093044  t3txtq                                   The Double Steal
9550    1646092994  t3tx4z                    Frequent crashes on Windows 11.
9551    1646092983  t3twyu  Tarzaned leaving game after dying lvl 1, flame…
9552    1646092952  t3twhr  Do not criticize russian aggression or Vladimi…
9553    1646092817  t3tuqh                                   Lunar Revel Pass

                                                    url       created  \
0     https://www.reddit.com/r/leagueoflegends/comme…  1.623740e+09
1                       https://v.redd.it/iqthzskakb571  1.623740e+09
2     https://youtube.com/watch?v=wfHM4nX2z-A&amp;fe…  1.623740e+09
3                       https://v.redd.it/c9puqrqyjb571  1.623740e+09
4                    https://i.redd.it/73ln16dhjb571.jpg  1.623740e+09
…                                                    …           …
9549                    https://v.redd.it/f11af9f4ynk81  1.646118e+09
9550  https://www.reddit.com/r/leagueoflegends/comme…  1.646118e+09
9551                    https://v.redd.it/tpem6nm0ynk81  1.646118e+09
9552  https://www.reddit.com/r/leagueoflegends/comme…  1.646118e+09
9553  https://www.reddit.com/r/leagueoflegends/comme…  1.646118e+09

                                                    d_
0     {'created_utc': 1623715013, 'id': 'o00m44', 't…
1     {'created_utc': 1623714840, 'id': 'o00k4w', 't…
2     {'created_utc': 1623714789, 'id': 'o00jk8', 't…
3     {'created_utc': 1623714743, 'id': 'o00j2e', 't…
4     {'created_utc': 1623714562, 'id': 'o00gyo', 't…
…                                                   …
9549  {'created_utc': 1646093044, 'id': 't3txtq', 't…
9550  {'created_utc': 1646092994, 'id': 't3tx4z', 't…
9551  {'created_utc': 1646092983, 'id': 't3twyu', 't…
9552  {'created_utc': 1646092952, 'id': 't3twhr', 't…
9553  {'created_utc': 1646092817, 'id': 't3tuqh', 't…

[143567 rows x 6 columns]
```

Save table to CSV and check shape

[62]:
```
mega_table.to_csv('june_to_march_postings.csv')
```

[63]:
```
mega_table.shape
```

[63]: (143567, 6)

3 Final Comment

We have 143,567 Reddit posts from which to gather comments, these comments will be used to implement Word2Vec featurization to act in our binary classification model.