Welcome to The Carpentries Etherpad!

This pad is synchronized as you type, so that everyone viewing this page sees the same text. This allows you to collaborate seamlessly on documents.

Use of this service is restricted to members of The Carpentries community; this is not for general purpose use (for that, try https://etherpad.wikimedia.org).

Users are expected to follow our code of conduct: https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html

---

Harvesting Twitter Data with Twarc

Instructors: Jon, Kristi, Amanda

Helpers: Amelia, Jingwei

Workshop Page: https://ucsbcarpentry.github.io/2022-05-26-ucsb-twarc/

Zoom Meeting: ucsb.zoom.us/j/88525068176?pwd=ieEd9iGjdjBQpJPMwJLQrFTWUp2Az0.1

JupyterLab Instance: http://carpentryworkshop.lsit.ucsb.edu/

Twitter Developer Account: https://developer.twitter.com/en

Daily Feedback survey: https://forms.gle/aH7WqZBkETfzkNrF9

==========

Post Workshop Resources

twarc2 docs: https://twarc-project.readthedocs.io/en/latest/twarc2_en_us/

DocNow: vhttp://www.docnow.io/
Textblob doc: https://textblob.readthedocs.io/en/dev/quickstart.html?highlight=sentiment#sentiment-analysis

Jupyter Notebook: https://github.com/UCSBCarpentry/twitter-with-twarc/blob/gh-pages/code/TwarcWorkshop.ipynb
==========

Name | Industry or UCSB Departmental Affiliation | Weekend plans?
jonjab. | ucsb library. |. none.
Faithe J. Day | Center for Black Studies Research - UCSB | Blog Writing
Amanda | UCSB Library | cat caring
Chloe | Linguistics Department (but want to transition to industry | catching up on work
Kristi | UCSB Library | none
Arnav | Computer Science | Work
Thao | Computer Engineering | Take grad pictures
Dina | Education | Galaxy game
Bryan | MCDB | Study
Amy | ARIT | Live Music
Lei | Econ | Sleep and do hw

What obscure terms have you encounterd on twitter _____
dynamic shelves (library twitter?)
#invertbutt - There was an entire week on bug twitter about bug butts
Obscure terms: sock puppets, botnets, orbiting, subtweeting - maybe not obscure more twitter lingo

Code

```
! pip install twarc_csv
! pip install emojis
```

import the following:

```
import pandas as pd
import twarc_csv
import textblob
import nltk
import os
import emojis
```

From Kristi:
```
!twarc2 flatten 'raw_data/ucsb_timeline.jsonl' > 'output_data/ucsb_timeline_flat.jsonl'
!twarc2 csv 'output_data/ucsb_timeline_flat.jsonl' 'output_data/ucsb_timeline.csv'
```

```
ucsb_timeline_df = pd.read_csv("output_data/ucsb_timeline.csv")

#outputs first 5 lines in dataframe
ucsb_timeline_df.head()

#catsofinstagram
```

Ethics/textblob Code:

```
hashtagcats_df = pd.read_csv("output_data/hashtagcats.csv")
list_tweets = hashtagcats_df['text'].tolist()
string_tweets = ' '.join(list_tweets)
library_blob = TextBlob(string_tweets)
library_freq = library_blob.word_countslibrary_sorted_freq =
sorted(library_freq.items(), key = lambda kv: kv[1], reverse = True)

from nltk.corpus import stopwords
# sw_nltk is our dictionary of stopwordssw_nltk = stopwords.words('english',
'japanese')

cats_str_stopped = [word for word in cats_string.split()
                    if word.lower() not in sw_nltk]
cats_words_stopped = " ".join(cats_str_stopped)

!python -m textblob.download_corpora
nltk.download('stopwords')

^ do this is you haven't added the stopwords


cats_blob_stopped = textblob.TextBlob(cats_words_stopped)
cats_blob_stopped_freq = cats_blob_stopped.word_counts
cats_blob_stopped_sorted_freq = sorted(cats_blob_stopped_freq.items(),
                        key = lambda kv: kv[1],
                        reverse = True)
cats_blob_stopped_sorted_freq[1:50]
```

Kristi's Output:

```
[('catsofinstagram', 496),
 ('cats', 289),
 ('cat', 131),
 ('’', 122),
 ('catsontwitter', 116),
 ('catsoftwitter', 110),
 ('family', 86),
 ('s', 77),
 ('sweet', 68),
 ('❀❀❀', 68),
 ('catlovers', 67),
```

('catstagram', 61),
('weekend', 56),
('he', 55),
('♥\\n', 53),
('t.co/6rxu3rirdo', 53),
('friday', 51),
('😻', 50),
('eating', 49),
('😻', 49),
('treats', 47),
('day', 47),
('long', 47),
('gigi_the_explorer_cat', 47),
('looking', 47),
('furrrtastic', 47),
(' ', 47),
('t.co/bw0yjcnb03', 47),
('furryfritz', 46),
('i', 46),
('could', 46),
('listen', 46),
('crunchy', 46),
('t.co/szim1yzxr9', 46),
('kitten', 44),
('meow', 38),
('love', 37),
('petsofinstagram', 35),
('catlover', 34),
('ねこ', 33),
('that', 33),
('babies', 33),
('two', 31),
('n\\n', 31),
('fresh', 31),
('kittycat', 31),
('spring', 30),
('getting', 30),
('猫のいる暮らし', 30)]

Capitol Riots

riots_dehydrated_df = pd.read_csv("raw_data/dehydratedCapitolRiotTweets.txt")
len(riots_dehydrated_df)

^Check your pandas, mine is imported as pd

Search & Stream

```
!twarc2 search --limit 500 "(#catsofinstagram #cute)" > "raw_data/hashtag_cats500.jsonl"
```