trump_word2vec_features

December 11, 2019

https://github.com/QuantCS109/TrumpTweets/blob/master/notebooks_features/trump_word2vec_features.ipy

[1]:
```
import torch

import sys
sys.path.append('..') #to add top-level to path
sys.path.append('../modules') #to add top-level to path

from modules.project_helper import TweetData

import numpy as np
import pandas as pd
import os
```

[2]:
```
tweet_data = TweetData()
```

[3]:
```
daily_tweets = tweet_data.clean_tweets #[pd.to_datetime(tweet_data.clean_tweets.
 ↪after4_date)
                                       #<= pd.to_datetime(daily_df.index[-1])]
daily_tweets.after4_date = pd.to_datetime(daily_tweets.after4_date)
```

[4]:
```
daily_tweets.head()
```

[4]:
```
                                                                tweets  \
timestamp
2019-11-17 19:57:12-06:00  tell jennifer williams whoever that is to read…
2019-11-17 19:56:02-06:00
2019-11-17 19:49:47-06:00  paul krugman of has been wrong about me from t…
2019-11-17 19:47:32-06:00                    schiff is a corrupt politician
2019-11-17 19:30:09-06:00  blew the nasty amp obnoxious chris wallace wil…

                                        timestamp after4_date
timestamp
2019-11-17 19:57:12-06:00 2019-11-17 19:57:12-06:00  2019-11-18
2019-11-17 19:56:02-06:00 2019-11-17 19:56:02-06:00  2019-11-18
2019-11-17 19:49:47-06:00 2019-11-17 19:49:47-06:00  2019-11-18
2019-11-17 19:47:32-06:00 2019-11-17 19:47:32-06:00  2019-11-18
2019-11-17 19:30:09-06:00 2019-11-17 19:30:09-06:00  2019-11-18
```

[5]:
```
model = torch.load('../models/tweet_embeddings/trump_rnn_1911.net')
```

[6]:
```
embeddings = model.in_embed.weight.to('cpu').data.numpy()
```

[7]:
```
embeddings.shape
```

[7]:
```
(4574, 100)
```

[8]:
```
arches = os.listdir('../models/tweet_embeddings')
arches[0:5]
```

[8]:
```
['trump_rnn_1702.net',
 'trump_rnn_1703.net',
 'trump_rnn_1908.net',
 'trump_rnn_1707.net',
 'trump_rnn_1712.net']
```

[9]:
```
embedding_list = []
for arch in arches:
    model = torch.load('../models/tweet_embeddings/{}'.format(arch))
    embedding_list.append(model.in_embed.weight.to('cpu').data.numpy())
```

[10]:
```
embedding_list[0].shape
```

[10]:
```
(2576, 100)
```

[11]:
```
data = TweetData('../data/intermediate_data/trump_archive_ts/
 ↪trump_archive_db_1911.csv')
words = data.words
vocab_to_int, int_to_vocab = data.vocab_to_int, data.int_to_vocab
int_words = data.int_words
```

[12]:
```
embedding_list[-1]
```

[12]:
```
array([[-0.07983071,  0.13481925, -0.3339518 , …,  0.07891949,
        -0.03683575, -0.06613905],
       [-0.4210904 ,  0.02706584, -0.1898706 , …, -0.17446454,
        -0.08358873, -0.03138035],
       [-0.22094908, -0.20108747, -0.04167196, …,  0.11216594,
         0.00401473,  0.15296404],
       …,
       [-0.38425526, -0.3069386 ,  0.7174769 , …, -0.4966805 ,
        -0.22396143, -0.90884376],
       [-1.4828042 ,  0.38311642,  0.23013367, …,  0.69838697,
         0.10621499,  0.49422497],
       [-0.72725123,  0.19940047, -0.78007627, …,  0.00616149,
```

```
        0.37083137, -0.43986282]], dtype=float32)
```

[13]:
```
daily_tweets.after4_date
```

[13]:
```
timestamp
2019-11-17 19:57:12-06:00   2019-11-18
2019-11-17 19:56:02-06:00   2019-11-18
2019-11-17 19:49:47-06:00   2019-11-18
2019-11-17 19:47:32-06:00   2019-11-18
2019-11-17 19:30:09-06:00   2019-11-18
                                  …
2009-05-12 14:07:28-05:00   2009-05-12
2009-05-08 20:40:15-05:00   2009-05-09
2009-05-08 13:38:08-05:00   2009-05-08
2009-05-05 01:00:10-05:00   2009-05-05
2009-05-04 18:54:02-05:00   2009-05-05
Name: after4_date, Length: 28813, dtype: datetime64[ns]
```

[14]:
```
tweet_embeddings = {}
dtws = daily_tweets.tweets[daily_tweets.after4_date >= pd.
 ↪to_datetime('1-1-2017')]
tweet_embeddings_np = np.zeros([dtws.shape[0],100])

for i, tweet in enumerate(dtws):
    embed_sum = np.zeros(100)
    tw = tweet.split()
    wd_count = 0
    for word in tw:
        try:
            if vocab_to_int[word] >= 25:
                embed_sum += embedding_list[-1][vocab_to_int[word]]
                wd_count += 1

        except:
            pass
    if wd_count > 0:
        tweet_embeddings[i] = embed_sum/wd_count
        tweet_embeddings_np[i,:] = embed_sum/wd_count
    else:
        tweet_embeddings[i] = embed_sum
        tweet_embeddings_np[i,:] = embed_sum
```

[15]:
```
tweet_embeddings_np.shape
```

[15]:
```
(9582, 100)
```

[16]:
```
#pd.DataFrame(tweet_embeddings_np).to_csv('../data/intermediate_data/
↪tweet_embeddings.csv')
```

[17]:
```
tw = 1000

embedding = model.in_embed
embed_vectors = embedding.weight
magnitudes = embed_vectors.pow(2).sum(dim=1).sqrt().unsqueeze(0)
tweet_after = daily_tweets.tweets[daily_tweets.after4_date >= pd.
↪to_datetime('1-1-2017')]
valid_vector = torch.FloatTensor(tweet_embeddings[tw])
wds = (torch.matmul(valid_vector, embed_vectors.t()) / torch.
↪FloatTensor(magnitudes)).topk(10)
for wd in list(wds[1][0].numpy()):
    if wd> 25 :
        print(int_to_vocab[wd])
tweet_after[tw]
```

```
collusion
democrats
friday
president
up
do
has
```

[17]:
```
'democrats wrote to the ukrainian government in may urging it to continue
investigations into president donald trumps alleged collusion with russia in the
presidential campaign collusion later found not to exist '
```

[18]:
```
wds
```

[18]:
```
torch.return_types.topk(
values=tensor([[0.6764, 0.6728, 0.6672, 0.6502, 0.6453, 0.6393, 0.6383, 0.6304,
0.6291,
         0.6283]], grad_fn=<TopkBackward>),
indices=tensor([[   8,  203,   75, 1083,    6,   44,    0,   73,   47,   30]]))
```

[64]:
```
tw = 750

embedding = model.in_embed
embed_vectors = torch.FloatTensor(tweet_embeddings_np)
magnitudes = embed_vectors.pow(2).sum(dim=1).sqrt().unsqueeze(0)
tweet_after = daily_tweets.tweets[daily_tweets.after4_date >= pd.
↪to_datetime('1-1-2017')]
```

```
tweet_after_date = daily_tweets.timestamp[daily_tweets.after4_date >= pd.
 ↪to_datetime('1-1-2017')]
valid_vector = torch.FloatTensor(tweet_embeddings[tw])
wds = (torch.matmul(valid_vector, embed_vectors.t()) / torch.
 ↪FloatTensor(magnitudes))
wds[wds != wds] = 0
for i in wds.topk(10).indices.numpy():
    for tweet_range in i:
        print(tweet_after_date[tweet_range], '||', tweet_after[tweet_range])
        print('')
```

```
2019-10-06 22:03:03-05:00 || democrat lawyer is same for both whistleblowers all
support obama and crooked hillary witch hunt

2018-07-29 19:35:14-05:00 || there is no collusion the robert mueller rigged
witch hunt headed now by increased from including an obama white house lawyer
angry democrats was started by a fraudulent dossier paid for by crooked hillary
and the dnc therefore the witch hunt is an illegal scam

2018-09-13 12:49:12-05:00 || this was done by the democrats in order to make me
look as bad as possible when i was successfully raising billions of dollars to
help rebuild puerto rico if a person died for any reason like old age just add
them onto the list bad politics i love puerto rico

2019-06-19 14:18:11-05:00 || the dems are very unhappy with the mueller report
so after almost years they want a redo or do over this is extreme presidential
harassment they gave crooked hillarys people complete immunity yet now they
bring back hope hicks why arent the dems looking at the

2019-07-24 02:29:14-05:00 || so robert mueller has now asked for his long time
never trumper lawyer to sit beside him and help with answers whats this all
about his lawyer represented the basement server guy who got off free in the
crooked hillary case this should not be allowed rigged witch hunt

2019-02-07 11:26:48-06:00 || the dems and their committees are going nuts the
republicans never did this to president obama there would be no time left to run
government i hear other committee heads will do the same thing even stealing
people who work at white house a continuation of witch hunt

2018-12-24 15:23:22-06:00 || for all of the sympathizers out there of brett
mcgurk remember he was the obama appointee who was responsible for loading up
airplanes with billion dollars in cash amp sending it to iran as part of the
horrific iran nuclear deal now terminated approved by little bob corker

2019-05-20 11:20:54-05:00 || was very good and highly professional to deal with
and if for any reason i didnt like them i would have gone elsewhere there was
always plenty of money around and banks to choose from they would be very happy
```

```
to take my money fake news

2018-12-13 13:34:14-06:00 || guilty even on a civil bases those charges were
just agreed to by him in order to embarrass the president and get a much reduced
prison sentence which he did including the fact that his family was temporarily
let off the hook as a lawyer michael has great liability to me

2018-11-26 01:59:13-06:00 || did a phony story about child separation when they
know we had the exact same policy as the obama administration in fact a picture
of children in jails was used by other fake media to show how bad cruel we are
but it was in during o years obama separated
```

[20]:
```
tweet_after[800]
```

[20]:
```
'my daughter ivanka will be on tonight on at p m following the great at enjoy '
```

[21]:
```
tweet_after[9335]
```

[21]:
```
'my daughter ivanka has been treated so unfairly by she is a great person always
pushing me to do the right thing terrible '
```

[22]:
```
tweet_after[4955]
```

[22]:
```
'will be interviewed tonight by on at p m eastern enjoy '
```

[23]:
```
tweet_after[2434]
```

[23]:
```
'i will be interviewed live tonight by on p m enjoy '
```