topic_modelling

December 11, 2019

1 Topic Modelling

(https://github.com/QuantCS109/TrumpTweets/blob/master/notebooks_features/topic_modelling.ipynb)

[3]:
```
import sys
sys.path.append('..') #to add top-level to path


import numpy as np
from sklearn import preprocessing
from sklearn.cluster import KMeans
import pandas as pd
from modules.project_helper import TweetData
```

1.0.1 This notebook uses the Word2Vec features created in Trump_Word2Vec and in trump_word2vec_features

Creating cluster model with 25 clusters.

[4]:
```
num_clusters = 25

tweet_data = TweetData()
topics_df = tweet_data.clean_tweets[tweet_data.clean_tweets.after4_date >= pd.
 ↪to_datetime('1-1-2017')]

emb = pd.read_csv('../data/intermediate_data/tweet_embeddings.csv',index_col=0)
X_Norm = preprocessing.normalize(np.array(emb))
kmeans = KMeans(n_clusters=num_clusters, random_state=0).fit(X_Norm)

topics_df['topic'] = kmeans.predict(X_Norm)

topics_df.groupby('topic').agg('count')

topics_analysis = pd.DataFrame()
topics_analysis['tweet_list'] = topics_df.tweets.str.split(' ')
topics_analysis['topic'] = topics_df['topic']
topics_analysis_melt = topics_analysis.explode('tweet_list')
```

```
topics_analysis_agg = topics_analysis_melt.assign(topic_count=1).
 ↪groupby(['tweet_list','topic']).agg('count').reset_index()
all_count = topics_analysis_melt.groupby('tweet_list').agg(all_count=pd.
 ↪NamedAgg('topic','count'))
topics_analysis_joined = topics_analysis_agg.join(all_count,on='tweet_list')
topics_analysis_joined['prop'] = topics_analysis_joined.topic_count/␣
 ↪topics_analysis_joined.all_count
```

```
c:\users\gufra\.virtualenvs\trump_tweets-t_tuxmg9\lib\site-
packages\ipykernel_launcher.py:10: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  # Remove the CWD from sys.path while we load stuff.
```

2 Aggregating topics at day level

We want to know how many times has trump tweeted about each 'topic' in a given day.

[153]:
```
featues_df = topics_df\
        .groupby(['after4_date','topic'])['tweets']\
        .agg('count')\
        .reset_index()\
        .pivot(index='after4_date',columns='topic',values='tweets')\
        .fillna(0)\
        .apply(lambda x: x/sum(x),axis=1)
featues_df.index.name = 'date'
featues_df.head()
```

[153]:

topic	0	1	2	3	4	5	6	7	…	15	16	17	18	19	20	21
date									…							
2017-01-01	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0	…	0.000000	0.000000	0.5	0.0	0.0	0.0	0.0
2017-01-02	0.0	1.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0	…	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
2017-01-03	0.0	0.222222	0.0	0.111111	0.222222	0.000000	0.0	0.0	…	0.000000	0.111111	0.0	0.0	0.0	0.0	0.0
2017-01-04	0.0	0.214286	0.0	0.214286	0.142857	0.000000	0.0	0.0	…	0.071429	0.071429	0.0	0.0	0.0	0.0	0.0
2017-01-05	0.0	0.000000	0.0	0.000000	0.333333	0.166667	0.0	0.0	…	0.166667	0.166667	0.0	0.0	0.0	0.0	0.0

```
topic        22   23   24
date
2017-01-01  0.0  0.0  0.0
2017-01-02  0.0  0.0  0.0
2017-01-03  0.0  0.0  0.0
2017-01-04  0.0  0.0  0.0
2017-01-05  0.0  0.0  0.0

[5 rows x 25 columns]
```

[149]:
```
features_df.to_csv(f'../data/features/topic_features_clusters={num_clusters}.
↪csv')
```

2.1 Topic Analysis

The code below prints a random sample of 10 tweets for each cluster.

[154]:
```
def sample_print(df,count=5):
    np.random.seed(seed=0)
    samples = np.random.choice(len(df),10,replace=False)
    if (len(df)>0):
        for i in samples:
            print(df.tweets[i])
            print('')
    else:
        print('No tweets in cluster')
```

[155]:
```
for cluster in range(num_clusters):
    print(f'Cluster {cluster}')
    print('')
    sample_print(topics_df[topics_df.topic==cluster])
  ␣
↪print('--------------------------------------------------------------------------------')
    print('')
```

```
Cluster 0

congratulations

congratulations to the tigers full ceremony

congratulations to our new cia director gina haspel

congratulations to the great jerry west
```

congratulations to the class of

congratulations to dana rohrabacher on his big california win we are proud of you dana

big day for israel congratulations

a great win for brooks congratulations to a great champion

it was my great honor to deliver the at the congratulations to the class of

congratulations to the philadelphia eagles on a great super bowl victory

---

Cluster 1

mitch get back to work and put repeal amp replace tax reform amp cuts and a great infrastructure bill on my desk for signing you can do it

it was a great day for the united states of america this is a great plan that is a repeal amp replace of obamacare

i suggest that we add more dollars to healthcare and make it the best anywhere obamacare is dead the republicans will do much better

house republicans could easily pass a bill on strong border security but remember it still has to pass in the senate and for that we need democrat votes and all they do is resist they want open borders and don t care about crime need more republicans to win in november

it is my opinion that many of the leaks coming out of the white house are fabricated lies made up by the media

the woodward book is a joke just another assault against me in a barrage of assaults using now disproven unnamed and anonymous sources many have already come forward to say the quotes by them like the book are fiction dems can t stand losing i ll write the real book

ainsley earnhardt a truly great person just wrote a wonderful book the light within me which is doing really well she is very special and so is her new book bring it to number one

all signs are that business is looking really good for next year only to be helped further by our tax cut bill will be a great year for companies and jobs stock market is poised for another year of success

today we announced vital new actions that we are taking to help former inmates

find a job live a crime free life and succeed beyond their dreams

prime minister of japan has been working with me to help balance out the one sided trade with japan these are some of the investments they are making in our country just the beginning

---

Cluster 2

Cluster 3

thanks to republican leadership america is winning again and america is being respected again all over the world because we are finally putting america first

talks with mexico will resume tomorrow with the understanding that if no agreement is reached tariffs at the level will begin on monday with monthly increases as per schedule the higher the tariffs go the higher the number of companies that will move back to the usa

we salute our great american flag we put our hands on our hearts for the pledge of allegiance and we all proudly stand for the national anthem

a few of the many clips of john mccain talking about repealing amp replacing ocare my oh my has he changed complete turn from years of talk

today it was my great honor to sign the largest tax cuts and reform in the history of our country full remarks

kim strassel of the wsj just said after reviewing the dumb comey memos you got to ask what was the purpose of the special counsel there s no there there dan henninger of the wsj said memos would show that this would be one of the weakest obstruction cases ever brought

thank you general mazloum for your kind words and courage please extend my warmest regards to the kurdish people i look forward to seeing you soon

today i am also directing the department of justice to propose legislation ensuring that those who commit hate crimes and mass murders face the death penalty and that this capital punishment be delivered quickly decisively and without years of needless delay

this month we celebrate the contributions of asian americans amp pacific islanders that enrich our nation

law enforcement has been strongly notified to watch closely for any illegal voting which may take place in tuesday s election or early voting anyone caught will be subject to the maximum criminal penalties allowed by law thank you

---

Cluster 4

excellent jobs numbers just released and i have only just begun many job stifling regulations continue to fall movement back to usa

fraudulently and illegally inserted his made up amp twisted words into my call with the ukrainian president to make it look like i did something very wrong he then boldly read those words to congress and millions of people defaming amp libeling me he must resign from congress

very proud of my executive order which will allow greatly expanded access and far lower costs for healthcare millions of people benefit

congressman keith rothfus continues to do a great job for the people of pennsylvania keith is strong on crime the border and our second amendment loves our military and our vets he has my total endorsement

because the ban was lifted by a judge many very bad and dangerous people may be pouring into our country a terrible decision

kentucky get out today and vote for and the entire republican ticket mississippi get out today and vote for and the entire republican ticket polls are open for a few more hours find your location below

as i predicted all along obamacare has been struck down as an unconstitutional disaster now congress must pass a strong law that provides great healthcare and protects pre existing conditions mitch and nancy get it done

happy canada day to all of the great people of canada and to your prime minister and my new found friend

our record economy would crash just like in if any of those clowns became president

it has been years and donald trump hasn t done anything wrong donald trump hasn t done a single thing of which he has been accused a witch hunt like no other

---

Cluster 5

today i signed the veterans our heroes choice program extension amp improvement act the watch

conquests how brave he was and it was all a lie he cried like a baby and begged for forgiveness like a child now he judges collusion

crooked hillary clinton blames everybody and every thing but herself for her election loss she lost the debates and lost her direction

this is the best time ever to look for a job james freeman of wsj

when someone gets nominated overwhelmingly and then wins the election as he did then he gets to set the national agenda the press is just outrages this story is the most irresponsible thing i ve ever seen i agree they also lose too much

with eleven republican candidates running in georgia on tuesday for congress a runoff will be a win vote r for lower taxes amp safety

such a total miscarriage of justice in san francisco

reports are there was indeed at least one fbi representative implanted for political purposes into my campaign for president it took place very early on and long before the phony russia hoax became a hot fake news story if true all time biggest political scandal

i hearby demand a second investigation after schumer of pelosi for her close ties to russia and lying about it

please join me with your thoughts and prayers for both aviators their families and our incredible

---

Cluster 6

god bless the people of el paso texas god bless the people of dayton ohio

the passage of the accountability and whistleblower protection act is great news for veterans i lo

the whistleblower has disappeared where is the whistleblower

the truth about impeachment

manufacturing openings hires rise to highest levels of the recovery

will be at the womens u s open today

finally great news at the border

border wall prototypes underway

just arrived in wisconsin to help two great people and

this is the real and only story

Cluster 7

thank you france

thank you

thank you

thank you

thank you kentucky

thank you

thank you and

thank you to nbc for the correction

thank you

thank you nicole

Cluster 8

the border or large sections of the border next week this would be so easy for mexico to do but they just take our money and talk besides we lose so much money with them especially when you add in drug trafficking etc that the border closing would be a good thing

the new york times and a third rate reporter named maggie haberman known as a crooked h flunkie who i don t speak to and have nothing to do with are going out of their way to destroy michael cohen and his relationship with me in the hope that he will flip they use

isn t it a shame that someone can write an article or book totally make up stories and form a picture of a person that is literally the exact opposite of the fact and get away with it without retribution or cost don t know why washington politicians don t change libel laws

as we celebrate lgbt pride month and recognize the outstanding contributions lgbt people have made to our great nation let us also stand in solidarity with the many lgbt people who live in dozens of countries worldwide that punish imprison or even execute individuals

democrats are frantic to throw something else at the president that s why you saw those subpoenas it s ridiculous just because your still upset over an election that happened years ago you should not be allowed to ruin people s lives like this lara trump

congressman lee zeldin is doing a fantastic job in d c tough and smart he loves our country and will always be there to do the right thing he has my complete and total endorsement

yesterday was the radical left democrats big impeachment day they worked so hard to make it something really big and special but had one problem almost nobody showed up the media admits low turnout for anti trump rallies all around the country people are

the people of germany are turning against their leadership as migration is rocking the already tenuous berlin coalition crime in germany is way up big mistake made all over europe in allowing millions of people in who have so strongly and violently changed their culture

sleepy joe biden just admitted he worked with segregationists and separately has already been very plain about the fact that he will be substantially raising everyone s taxes if he becomes president ridiculously all democrats want to

substantially raise taxes

sorry can t let them into our country if too crowded tell them not to come to usa and tell the dems to fix the loopholes problem solved

---

Cluster 9

the democrats in the southwest part of virginia have been abandoned by their party republican ed gillespie will never let you down

days goes very quickly negotiations with democrats will start immediately will not be easy to make a deal both parties very dug in the case for national security has been greatly enhanced by what has been happening at the border amp through dialogue we will build the wall

today we celebrate the lives and achievements of americans with down syndrome and i will always stand with these wonderful families and together we will always stand for life

i was just informed by marillyn hewson ceo of lockheed martin of her decision to keep the sikorsky helicopter plant in coatesville pennsylvania open and humming we are very proud of pennsylvania and the people who work there

heading to joint base andrews on with prime minister shinz earlier today

the democrat congresswomen have been spewing some of the most vile hateful and disgusting things ever said by a politician in the house or senate amp yet they get a free pass and a big embrace from the democrat party horrible anti israel anti usa pro terrorist amp public

we have been working every day to deliver for america s farmers just as they work every day to deliver for us

our prayers are with those affected by the flooding in japan we commend the rescue efforts and offer condolences to all who were injured or lost loved ones

loser terrorists must be dealt with in a much tougher manner the internet is their main recruitment tool which we must cut off amp use better

will be giving a full pardon to dinesh d souza today he was treated very unfairly by our government

---

Cluster 10

thank you to wayne allyn root for the very nice words president trump is the greatest president for jews and for israel in the history of the world not just america he is the best president for israel in the history of the world and the jewish people in israel love him

fbi top lawyer confirms unusual steps they relied on the clinton campaign s fake amp unverified dossier which is illegal that has corrupted them that has enabled them to gather evidence by unconstitutional means and that s what they did to the president judge napolitano

join me in las vegas nevada at pm for a make america great again rally tickets

one my way to new mexico see you all shortly at the

my deepest sympathies to the families and friends of those involved in the terrible boat accident which just took place in missouri such a tragedy such a great loss may god be with you all

many lawyers and top law firms want to represent me in the russia case don t believe the fake news narrative that it is hard to find a lawyer who wants to take this on fame amp fortune will never be turned down by a lawyer though some are conflicted problem is that a new

with one yes vote in hospital amp very positive signs from alaska and two others mccain is out we have the hcare vote but not for friday

dems totally control the u s senate many great republican bills will never pass like kates law and complete healthcare get smart

california has been forced to cancel the massive bullet train project after having spent and wasted many billions of dollars they owe the federal government three and a half billion dollars we want that money back now whole project is a green disaster

thank you hopefully this is just the beginning of a massive story of injustice and treason you will never learn this from the corrupt lamestream media who get pulitzer prizes for reporting the story totally wrong the ones who report it right get only respect

---

Cluster 11

bayer ag has pledged to add u s jobs and investments after meeting with president elect donald trump the latest in a string

russia started their anti us campaign in long before i announced that i would run for president the results of the election were not impacted the trump

campaign did nothing wrong no collusion

the united states has an billion dollar yearly trade deficit because of our very stupid trade deals and policies our jobs and wealth are being given to other countries that have taken advantage of us for years they laugh at what fools our leaders have been no more

a big scandal at news they got caught using really gruesome fake footage of the turks bombing in syria a real disgrace tomorrow they will ask softball questions to sleepy joe biden s son hunter like why did ukraine amp china pay you millions when you knew nothing payoff

the democrats have become known as the do nothing party

the paris agreement isn t working out so well for paris protests and riots all over france people do not want to pay large sums of money much to third world countries that are questionably run in order to maybe protect the environment chanting we want trump love france

president donald j trump is following through on his promise to cut burdensome red tape and unleash the american economy read more

sadly past administrations have allowed china to get so far ahead of fair and balanced trade that it has become a great burden to the american taxpayer as president i can no longer allow this to happen in the spirit of achieving fair trade we must balance this very

the do nothing dems have nothing but time

if we bought billion dollars of agriculture from our farmers far more than china buys now we would have more than billion dollars left over for new infrastructure healthcare or anything else china would greatly slow down and we would automatically speed up

---

Cluster 12

thank you governor phil bryant it was my great honor to be there

israel saudi arabia and the middle east were great trying hard for peace doing well heading to vatican amp pope then and

pledge to america s workers

this is really an incredible time for our nation we are respected again

my great honor

it was an honor to meet with republic of rwanda president paul kagame this morning in davos switzerland many great discussions

the greatest scam in the history of american politics

presidential executive order on promoting agriculture and rural prosperity in americaexecutive order

so beautiful show this picture to the nfl players who still kneel

america is open for business

Cluster 13

we will never forget

comey gave strozk his marching orders mueller is comey s best friend witch hunt

wow they got caught end the witch hunt now

so much happening with the now discredited witch hunt this total hoax will be studied for years

just a continuation of the witch hunt

the witch hunt continues

just out unemployment is broken in the meantime witch hunt

witch hunt

the mueller witch hunt is completely over

as the witch hunt continues

Cluster 14

just got back only to hear of a last minute change allowing a never trumper attorney to help robert mueller with his testimony before congress tomorrow what a disgrace to our system never heard of this before very unfair should not be allowed a rigged witch hunt

leaving el paso for the white house what great people i met there and in dayton

ohio the fake news worked overtime trying to disparage me and the two trips but it just didn t work the love respect amp enthusiasm were there for all to see they have been through so much sad

jesse watters the only thing trump obstructed was hillary getting to the white house so true

thank you to kurt volker u s envoy to ukraine who said in his congressional testimony just released you asked what conversations did i have about that quid pro quo et cetra none because i didn t know there was a quid pro quo witch hunt

separating families at the border is the fault of bad legislation passed by the democrats border security laws should be changed but the dems can t get their act together started the wall

the fake news media has lost tremendous credibility with its corrupt coverage of the illegal democrat witch hunt of your all time favorite duly elected president me t v ratings of cnn amp msnbc tanked last night after seeing the mueller report statement up big

so much fake news about what is going on in the white house very calm and calculated with a big focus on open and fair trade with china the coming north korea meeting and of course the vicious gas attack in syria feels great to have bolton amp larry k on board i we are

when will the fake news media start asking democrats if they are ok with the hiring of christopher steele a foreign agent paid for by crooked hillary and the dnc to dig up dirt and write a phony dossier against the presidential candidate of the opposing party

congrats to the senate for taking the first step to now its onto the house

the russia hoax continues now its ads on facebook what about the totally biased and dishonest media coverage in favor of crooked hillary

---

Cluster 15

despite the fact that the mueller report was composed by trump haters and angry democrats who had unlimited funds and human resources the end result was no collusion no obstruction amazing

working hard to get the olympics for the united states l a stay tuned

the trump administration will be announcing the new secretary of the interior next week

great bilateral meetings at lys e palace w president the friendship between our two nations and ourselves is unbreakable

correct a total scam

its hard to read the failing new york times or the amazon washington post because every story opinion even if should be positive is bad

trump recognized russian meddling many times

thank you jon working hard

oppressive regimes cannot endure forever and the day will come when the iranian people will face a choice the world is watching

they are trying to stop me because i am fighting for you

Cluster 16

congratulations to legislators in new jersey for not passing taxes that would have driven large numbers of high end taxpayers out of the state many were planning to leave amp will now be staying new york amp others should start changing their thought process on taxes fast

everyone very excited about the new deal with mexico

great news as a result of our tax cuts amp jobs act

thank you to everyone who joined me at the yesterday together we are making america great again

congratulations to paul ryan kevin mccarthy kevin brady steve scalise cathy mcmorris rodgers and all great house republicans who voted in favor of cutting your taxes

my lawyers should sue the democrats and shifty adam schiff for fraud

good great meeting in the oval office tonight with the nra

senator luther strange has gone up a lot in the polls since i endorsed him a month ago now a close runoff he will be great in d c

h b holders in the united states can rest assured that changes are soon coming which will bring both simplicity and certainty to your stay including a potential path to citizenship we want to encourage talented and highly skilled people to pursue career options in the u s

today it was my great honor to visit the martin luther king jr memorial with mike pence in honor of

---

Cluster 17

happy birthday

happy mother s day

happy birthday to our melania

happy anniversary

happy birthday to our great united states

happy columbus day

happy

happy birthday to a truly great champion and person

happy

happy birthday america

Cluster 18

presidential harassment

jobs jobs jobs

unemployment filings are at their lowest level in over years great news for workers and jobs jobs jobs

just arrived in wisconsin to discuss jobs jobs jobs

jobs jobs jobs

jobs jobs jobs

jobs jobs jobs

presidential harassment it should never be allowed to happen again

presidential harassment

jobs jobs jobs unemployment claims have fallen to a year low together we are making the economy great again

---

Cluster 19

thank you houston texas get out and

welcome to the amir sabah al ahmed al jaber al sabah of kuwait joint press conference coming up soon

approval rating in the republican party thank you

consumer confidence hits highest level since december read more

beautiful evening in mesa arizona with great patriots thank you

just stopped in alaska and said hello to our great troops

trump approval hits

thank you bill say hello to our great veterans

were all thinking of you

consumer confidence in february highest since november

Cluster 20

make america great again

beautiful make america great again rally in lebanon ohio thank you watch here

together we are making america great again

in nashville tennessee lets make america great again

making america great again

make america great again

make america great again

make america great again

make america great again

make america great again

Cluster 21

fake news the enemy of the people

too bad a large portion of the media refuses to report the lies and corruption having to do with the rigged witch hunt but that is why we call them fake news

the only collusion with russia was with the democrats so now they are looking at my tweets along with million other people the rigged witch hunt continues how stupid and unfair to our country and so the fake news doesn t waste my time with dumb questions no

this decision but in the end it will be best for all concerned as president i will always be there to help new york and the great people of new york it will always have a special place in my heart

years ago today the national security council met for the first time great history of advising presidents then amp now thanks nsc staff

with a president and federal government that wants our wonerful city and state to flourish and thrive i love new york

there is great anger in our country caused in part by inaccurate and even fraudulent reporting of the news the fake news media the true enemy of the people must stop the open amp obvious hostility amp report the news accurately amp fairly that will do much to put out the flame

stock market up points today also great jobs numbers

the reason sarah sanders does not go to the podium much anymore is that the press covers her so rudely amp inaccurately in particular certain members of the press i told her not to bother the word gets out anyway most will never cover us fairly amp hence the term fake news

with all of the fake and made up news out there iran can have no idea what is actually going on

---

Cluster 22

joining tonight at pme on enjoy

will be interviewed by on at pm tonight enjoy

it is very possible that those sources dont exsist but are made up by fake news writers is the enemy

my son donald will be interviewed by tonight at p m he is a great person who loves our country

i will be interviewed on by tomorrow from a m to a m enjoy

i will be interviewed by laura ingraham tonight at p m on

an extended interview from the super bowl with airs tonight at p m enjoy

the failing new york times

i will be interviewed by on at enjoy

will be interviewed on tonight at p m

Cluster 23

thank you cleveland ohio

i finally agree with

the usa loves india

robert will do a great job for our vets we also recently won choice

read the transcript it is perfect

longest bull run in the history of the stock market congratulations america

great job thom

thank you to and all first responders

trump imitation syndrome is afflicting the president s liberal enemies thank you

funding bill

```
--------------------------------------------------------------------------------

Cluster 24



␣
↪--------------------------------------------------------------------------------

       ValueError                                Traceback (most recent call␣
↪last)

       <ipython-input-155-980a536c3e16> in <module>
         2     print(f'Cluster {cluster}')
         3     print('')
   ----> 4     sample_print(topics_df[topics_df.topic==cluster])
         5     ␣
↪print('--------------------------------------------------------------------------------')
         6     print('')


       <ipython-input-154-6d2377205a1a> in sample_print(df, count)
         1 def sample_print(df,count=5):
         2     np.random.seed(seed=0)
   ----> 3     samples = np.random.choice(len(df),10,replace=False)
         4     if (len(df)>0):
         5         for i in samples:


       mtrand.pyx in numpy.random.mtrand.RandomState.choice()


       ValueError: Cannot take a larger sample than population when␣
↪'replace=False'
```