Text Clustering

In this project we will perform a text classification on the health tweets text using kmeans clustering. The health tweets provide a large dataset to test many types of text clustering, from kmeans to kmedians, to LSA and LDA

The code, along with the files necessary and versions of packages in this instance can be found on this repo: [https://github.com/Benjamin-Siebold/MSDS-682-Text-Analytics (https://github.com/Benjamin-Siebold/MSDS-682-Text-Analytics)](https://github.com/Benjamin-Siebold/MSDS-682-Text-Analytics)

In [1]:
```
import time
from glob import iglob
import spacy
import nltk
import string
import re

import numpy as np
import pandas as pd
pd.options.display.max_rows = 999
from matplotlib import pyplot as plt
%matplotlib inline

#from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans

nlp = spacy.load('en_core_web_lg')
stopwords = nltk.corpus.stopwords.words('english')
np.random.seed(50)
```

1 - Load data and prepare text

The first step in this analysis is to load the data in with iglob, and clean the tweets to lemmatize, change words to lowercase, and remove stopwords in order to cluster the data most accurately. Because the amount of tweets (62,000) is so large, the cleaning function and clustering can take an extremely long time. One option to this time taken is to multiprocess the cleaning portion of the tweets by either pooling them, or creating a managed list in the multiprocessing package. Another option is to take a sample of the tweets for analysis. The latter was chosen for this project.

In [2]:
```
news_files = []
for file in list(iglob('Health-Tweets/*.txt')):
    news_files.append(pd.read_csv(file, sep= '|', header=None, error_bad_li

news_df = pd.concat(news_files)
```

```
b'Skipping line 846: expected 3 fields, saw 4\nSkipping line 904: expecte
d 3 fields, saw 4\nSkipping line 914: expected 3 fields, saw 4\nSkipping
line 1264: expected 3 fields, saw 4\nSkipping line 1269: expected 3 field
s, saw 4\nSkipping line 1293: expected 3 fields, saw 4\nSkipping line 134
8: expected 3 fields, saw 4\nSkipping line 1430: expected 3 fields, saw 4
\nSkipping line 1486: expected 3 fields, saw 4\nSkipping line 1710: expec
ted 3 fields, saw 4\nSkipping line 2699: expected 3 fields, saw 4\nSkippi
ng line 2728: expected 3 fields, saw 4\nSkipping line 3000: expected 3 fi
elds, saw 4\n'
b'Skipping line 4015: expected 3 fields, saw 4\nSkipping line 6118: expec
ted 3 fields, saw 4\nSkipping line 6354: expected 3 fields, saw 4\nSkippi
ng line 6429: expected 3 fields, saw 4\nSkipping line 6528: expected 3 fi
elds, saw 4\nSkipping line 6930: expected 3 fields, saw 4\nSkipping line
6944: expected 3 fields, saw 4\nSkipping line 6948: expected 3 fields, sa
w 4\nSkipping line 6949: expected 3 fields, saw 4\nSkipping line 6951: ex
pected 3 fields, saw 5\nSkipping line 6954: expected 3 fields, saw 4\nSki
pping line 6956: expected 3 fields, saw 4\nSkipping line 6965: expected 3
fields, saw 4\nSkipping line 6984: expected 3 fields, saw 4\nSkipping lin
e 6988: expected 3 fields, saw 4\nSkipping line 7020: expected 3 fields,
```

In [3]:
```
news_df.columns = ['id','date','tweet']
```

In [4]:
```
news_df
```

Out[4]:

	id	date	tweet
0	586266687948881921	Thu Apr 09 20:37:25 +0000 2015	Drugs need careful monitoring for expiry dates...
1	586266687017771008	Thu Apr 09 20:37:25 +0000 2015	Sabra hummus recalled in U.S. http://www.cbc.c...
2	586266685495214080	Thu Apr 09 20:37:24 +0000 2015	U.S. sperm bank sued by Canadian couple didn't...
3	586226316820623360	Thu Apr 09 17:57:00 +0000 2015	Manitoba pharmacists want clampdown on Tylenol...
4	586164344452354048	Thu Apr 09 13:50:44 +0000 2015	Mom of 7 'spooked' by vaccinations reverses st...
...	...	...	...
3194	106106376224378880	Tue Aug 23 20:51:46 +0000 2011	Mainstay Meds Often Cut Off Accidentally After...
3195	106106374735400960	Tue Aug 23 20:51:45 +0000 2011	Injectable Psoriasis Drugs May Not Hike Heart ...
3196	106106373275787264	Tue Aug 23 20:51:45 +0000 2011	Certain Foods Said to Help Lower Bad Cholester...
3197	106106371736485889	Tue Aug 23 20:51:45 +0000 2011	Boys Mature Sexually Earlier Than Ever Before:...
3198	106073015590203393	Tue Aug 23 18:39:12 +0000 2011	Mental Illness Affects Women, Men Differently,...

62817 rows × 3 columns

In [5]:
```
stopwords = set(stopwords + ['RT', 'Health', 'Healthcare', 'health', 'healt
```

In [6]:
```
def clean_text(docs):

    #print('remove https')
    docs = [re.sub(r'http\S+', '', doc).rstrip() for doc in docs]
    docs = [re.sub(r'@\S+', '', doc) for doc in docs]

    print('removing punc')
    table = str.maketrans({key: None for key in string.punctuation + string
    clean_docs = [d.translate(table) for d in docs]

    print('nlp')
    nlp_docs = [nlp(doc) for doc in clean_docs]

    print('lemmatize')
    lemm_docs = [[w.lower_ if w.lemma_ == '-PRON-'
                   else w.lemma_ for w in doc]
                 for doc in nlp_docs]

    print('remove stopwords')
    lemm_docs = [[lemma for lemma in doc if lemma not in stopwords] for doc

    print('combine words')
    cleaned_docs = [' '.join(word) for word in lemm_docs]

    return cleaned_docs
```

In [7]:
```
tweet_sample = news_df['tweet'].sample(5000)
```

In [8]:
```
cleaned_tweets = clean_text(tweet_sample)
```

```
removing punc
nlp
lemmatize
remove stopwords
combine words
```

In [9]:
```
tweets_no_special = []
for r in cleaned_tweets:
    tweets_no_special.append(r.encode('ascii','ignore').decode('ascii'))
```

2 - Tranform Data and apply model

Now that the data is cleaned, we transform the data using TfidfVectorizer, and apply a KMeans model to the data. Once that is done, we check the score, and then create an iteration of models in order to build a plot to determine which amount of clusters is the most optimal. This spot occurs where the data starts to flatten out for the sum of squared distances between points and their centers, or by plotting the derivative of the within sum of squares (wss) to see where it flattens.

In [10]:
```
vectorizer = TfidfVectorizer(min_df=100)
features = vectorizer.fit_transform(tweets_no_special)
type(features)
```

Out[10]:
```
scipy.sparse.csr.csr_matrix
```

In [11]:
```
features.shape
```

Out[11]:
```
(5000, 27)
```

In [12]:
```
features = features.todense()
```

In [13]:
```
model = KMeans(n_clusters=10, random_state=50, n_jobs=-1)
```

In [14]:
```
model.fit(features)
```

Out[14]:
```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=10, n_init=10, n_jobs=-1, precompute_distances='auto',
       random_state=50, tol=0.0001, verbose=0)
```

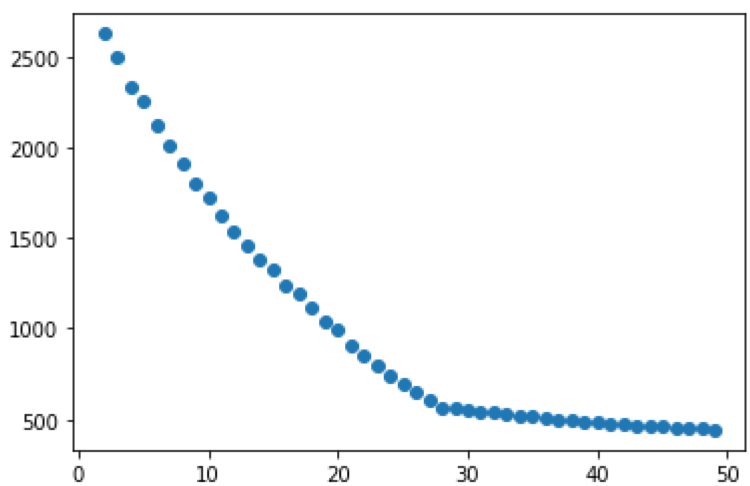
In [15]:
```
model.score(features)
```

Out[15]:
```
-1721.261590397848
```

In [16]:
```
wss = []
for n in range(2, 50):
    model = KMeans(n_clusters=n, random_state=50, n_jobs=-1)
    model.fit(features)
    wss.append(-model.score(features))
```

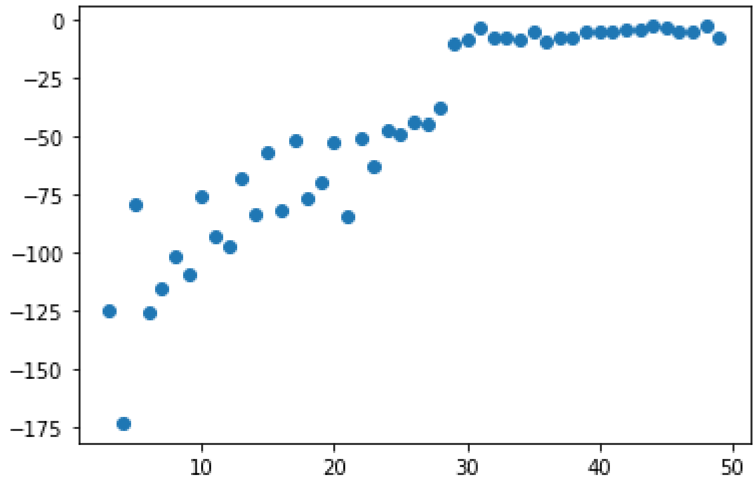
In [17]:
```
plt.scatter(range(2,50),wss)
```

Out[17]:
```
<matplotlib.collections.PathCollection at 0x1a47135810>
```



In [18]:
```
plt.scatter(range(3,50), np.diff(wss))
```

Out[18]:
```
<matplotlib.collections.PathCollection at 0x1a47135710>
```



3 - Apply best model

From the plots above, we can see for the most part the best model to apply to this data has 28 clusters. With this known, we create a model, and cluster the data accordingly, predicting our features, and then tying each tweet back to a cluster.

In [19]:
```
model = KMeans(n_clusters=28, random_state=50, n_jobs=-1)
model.fit(features)
cluster_labels = model.predict(features)
```

In [20]:
```
np.bincount(cluster_labels)
```

Out[20]:
```
array([ 110, 2072,  123,  103,  114,  162,  129,  109,  100,  102,  103,
        138,  105,  124,  102,   90,  101,  105,  108,   79,  111,  116,
        120,   94,  105,   87,   92,   96])
```

In [21]:
```
tweet_df = pd.DataFrame(np.array(tweets_no_special),columns=['tweet'])
clusters_df = pd.DataFrame(cluster_labels,columns=['prediction'])
```

In [22]:
```
tweets_clustered = pd.concat([tweet_df,clusters_df], axis=1,join='inner')
tweets_clustered
```

...

In [23]:
```
tweets_clustered.groupby(['prediction']).head(5).sort_values(by=['predictio
```

Out[23]:

	tweet	prediction
153	food eat long life	0
140	well trendy Food Kosher Passover	0
197	young kid food allergy may learn helplessness	0
265	FDA propose strict new safety rule animal food	0
12	feel bloated lil gassy top gasproduce food avoid	0
0	abduction survivor humanfactor	1
3	Boston Hospitals Share lesson Marathon bombing	1
1	manage chronic condition workplace	1
4	Pfizer buy two Baxter vaccine mln	1
2	Amish seek Measles shot Ohio Outbreak sicken	1
149	get pilates see interested try	2
155	get Obamacare insurance	2
158	Fun way get Fit without gym	2
14	remember its hard get Ebola even airplane	2
108	Saskatchewan nurse order get flu shot wear mask	2
75	maker pom wonderful pomegranate juice allow ma...	3
198	hate diet Heres burn calorie lose pound make...	3
202	today getfit tip eat fat make fat eat enough...	3
8	Regrets Outlook May make sunny old age	3
174	make tonight ingredient sweet Potato amp Black...	3
182	Fight Seasonal Affective Disorder New Tracker ...	4
227	New IVF use timelapse image boost live birth	4
288	sick Drawn New Coverage	4
55	new approach HIV promise	4
19	New York Governor lay Ebola rule Home quarantine	4
88	predict Ebola case week	5
163	CDC confirm first Ebola case diagnose United S...	5
59	shortage engineer sanitation expert may slow f...	5
69	homemade ebola protection suit wait sick mom...	5
22	North Carolina monitoring person return Liberi...	5
116	thank Hope join us today ET HealthTalk	6
18	US childhood obesity turn corner Rates presc...	6
131	US voter May prefer LowPitched Male voice	6

	tweet	prediction
45	cost birth US high world think cost baby	6
72	hello thank invite us discuss Xtreme eat ext...	6
98	well help Smokers quit start First Place	7
35	well ask well Genetic Testing breast Cancer	7
81	well brisk walk well	7
25	well training dog sniff cancer	7
49	basic public well use resource experimental ...	7
21	end culture patient deference towards NHS prof...	8
173	NHS face tricky winter	8
205	VIDEO NHS plan recipe disaster	8
120	major incident risk new normal NHS think a...	8
162	Incl patient datum access wld ask NHS Gdnh...	8
327	girl commit dating violence often boy study show	9
67	study examine Efficacy taxis sugary drink	9
142	Red Meat Can Unhealthy Study suggest	9
86	VIDEO Alzheimers insight DNA study	9
96	senior mental wellbeing affect live say ital...	9
87	breathe easier spring allergy season good Wors...	10
133	paracetamol no good back pain	10
47	drink day may good everyone	10
39	Smokeout Day good Ways quit	10
9	People ill accident happen job amp Ill good ...	10
114	Wild Insurance rate hike May settle filing Show	11
43	early school start time may tie teen drive acc...	11
84	sunscreen may slow skin age use daily	11
91	woman Dialysis May experience Sexual Problems ...	11
128	Canola oilenriched diet may benefit people dia...	11
99	go glutenfree mean give fave healthy carb sh...	12
347	enter healthy contest chance win close K p...	12
201	use avocado make virtually meal little bit hea...	12
121	obese kid Head Start get healthy year	12
215	world healthy meal scientist come Supperdinner...	12
6	Everyone take deep breath chill relax say ne...	13
186	regulation place narrow focus organisation wro...	13
161	Moms Tykes Should eat More fish low Mercury sa...	13

	tweet	prediction
31	Theres link MMR vaccine autism say Dr Anne S...	13
118	maternal death fall say	13
92	Q risk factor develop diabetes HealthTalk	14
60	confused dustup new statin guideline risk calc...	14
138	VIDEO Rural GPs surgery risk	14
11	Common Plastics Chemical may boost diabetes risk	14
117	Syria Doctors risk life Juggle ethic	14
113	antismoking campaign CDC help	15
241	athlete weird Rituals actually help win	15
76	Texas Groups Promote Insurance Exchange help s...	15
213	snack help lose weight burn fat build muscle	15
112	combine vaccine may help eradicate polio	15
103	Courage Unmasked turn symbol cancer torture art	16
34	prostate cancer chance rise vitamin e selenium...	16
51	big drop colon cancer fuel push get More Peopl...	16
119	face breast Cancer	16
52	black woman less likely get breast cancer di...	16
29	US probe Sanofi blockbuster drug plavix	17
44	lack drug datum extreme concern	17
89	Indias Ranbaxy share fall US FDA revoke approv...	17
65	UK cost agency reject british company gw canna...	17
85	FDA slam Ranbaxy time cover negative test dr...	17
271	Twitter week New Followers mention K menti...	18
281	way reap Red Wines benefit without drink via	18
329	WorkLife Balance dangerous via	18
284	Measles party California Via remind pox ...	18
58	blame third cup coffee gene via	18
316	US Ebola patient initially turn away may expos...	19
518	cancer patient fight parent right die	19
361	cancer patient Canada get weak dose chemo drug	19
20	federal government lose appeal stop medical ma...	19
56	social networking site connect multiple sclero...	19
146	Texas hospital consult Emory Hospital Atlanta ...	20
27	hospital bad patient injury rate lose Me...	20
28	fame US hospital pay M MD secretly film woman	20

	tweet	prediction
37	want hospital w medical error one w hotel perk	20
82	video hospital prepare winter crisis	20
148	doctor urge caution payment drug maker make Heres	21
111	ban new smoker call doctor	21
16	Can doctor really Can demand Extra front	21
252	doctor characteristic may influence prostate c...	21
154	think free drug sample get doctor save money t...	21
7	regular binge drinking may raise risk type d...	22
54	emotion may improve ability recall memories st...	22
236	look find especially hunt methane	22
83	Heroin overdose cure exist Can user find	22
57	eat late lunch may thwart weightloss effort ne...	22
349	Homeopathy work australian expert say	23
115	afraid eat night fear weight gain common diet ...	23
17	popular week start day antidepressant end wish...	23
185	like work clinical psychologist area high soci...	23
177	engineer turn table e coli put bacteria work...	23
225	new report estimate number annual ER visit inv...	24
15	report More hospital Face Medicare crackdo...	24
93	Ethiopia WVa Community Workers help close Acce...	24
36	Medicare Social Security Report thing wat...	24
270	big bill surprise ER patient even innetwork ho...	24
94	Paradigm shift medical model care person mod...	25
104	plenty option care law study	25
95	candidate position care Obama Romney	25
97	well Futile care Lifes end	25
38	video Emergency care system confusing	25
246	childrens heart surgery unit safe	26
10	obesity lead heart disease type diabetes ill...	26
68	heart matter treat Disease instead person	26
66	Sepsis plus heart Rhythm disorder link Stroke ...	26
207	dementia care cost treat heart disease cancer	26
240	study Medicare dump random drug plan assignmen...	27
325	regular walking routine could prevent depressi...	27
328	much tv could damage sperm production new stud...	27

	tweet	prediction
342	bank mobile want manage way Could agree	27
260	cup GREENTEA day could help lose twice much ...	27

4 - Analyze data

The last step in the analysis is to look into what kind of clustering took place of the tweets. We once again take a sample of the 5,000 tweets to see if we can find a general understanding for how the clusters were done. From above, it can be seen there are still potential flaws with the clustering. For example cluster 6 has almost now relation between the examples, but US is found multiple times, suggesting either that cluster is focused on the US location and it potentially would've been beneficial to remove additional stopwords.