Imports

Data Management

In [1]:
```
import datetime
import numpy as np
import pandas as pd
```

Analysis and Cleaning

In [2]:
```
import string
import re

from gensim.parsing.preprocessing import preprocess_string, strip_tags, strip_punctuati
                                        strip_multiple_whitespaces, strip_numeric, \
                                        remove_stopwords, strip_short
from gensim.models import Word2Vec
```

Learning

In [3]:
```
from sklearn import cluster
from sklearn import metrics
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
```

Visualization

In [4]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

Data Analysis & Cleanup

In [5]:
```
fake = pd.read_csv('datasets/Fake.csv')
true = pd.read_csv('datasets/True.csv')
print('False Sample')
display(fake.sample(10))
print('\n\n\n\n')
print('True Sample')
display(true.sample(10))
```

```
False Sample
```

	title	text	subject	date
7062	Ted Cruz Arrives In The Bronx Only To Put Dow...	You ve honestly got to be a special kind of st...	News	April 6, 2016
15006	Who do you think won the CNBC GOP Presidential...	Please help us determine who real grassroots c...	politics	Oct 29, 2015
19697	BREAKING EMAIL LEAK: “Bernie needs to be groun...	Hey Bernie how s that whole Queen of Wall Stre...	left-news	Nov 1, 2016

	title	text	subject	date
12088	MTV Releases Racist ‘Hey Fellow White Guy’s Vi...		politics	Dec 21, 2016
2087	Trump’s SCOTUS Nominee Has DISTURBING History...	We all knew that Donald Trump would be a night...	News	March 20, 2017
14379	WAKE UP AMERICA! SOMALI CANDIDATES IN MINNESOT...	While our eyes are on the invasion of Europe, ...	politics	Mar 2, 2016
8905	Ted Cruz Reveals Paranoid Fantasy Involving O...	Republican presidential hopeful Ted Cruz stunn...	News	January 7, 2016
14050	BEYONCE DOUBLES DOWN...Debuts #LEMONADE, Another...	Most of the world will be obsessed with Beyonc...	politics	Apr 24, 2016
2603	Trump Just Met With Airline Execs, RUINS Meet...	Donald Trump has turned yet another meeting in...	News	February 9, 2017
15462	SHAKEDOWN AL SHARPTON MEETS WITH GM TO PRESSUR...	I m a mom who grew up in Romeo, MI., the same ...	politics	Jul 15, 2015

```
True Sample
```

	title	text	subject	date
5634	Trump says Pence will lead voter fraud panel	WEST PALM BEACH, Fla. (Reuters) - President Do...	politicsNews	February 5, 2017
12132	Thai tour guide arrested for inappropriate beh...	BANGKOK (Reuters) - Thai authorities have arre...	worldnews	December 16, 2017
7223	Schumer, McConnell elected top leaders in Senate	WASHINGTON (Reuters) - Democratic U.S. senator...	politicsNews	November 16, 2016
10816	Foreclosure crisis snarls Clinton, Sanders' ef...	(Reuters) - Democratic presidential hopefuls B...	politicsNews	February 14, 2016
20199	As North Korea girds for latest sanctions, eco...	DANDONG, China (Reuters) - The United Nations ...	worldnews	September 13, 2017
247	Aide tries to refocus tax debate after Trump's...	WASHINGTON (Reuters) - President Donald Trump’...	politicsNews	December 8, 2017
8880	House No. 2 Republican says still questions Cl...	(Reuters) - U.S. House Majority Leader Kevin M...	politicsNews	July 5, 2016
19677	Rescued migrants say lucky to dodge Libyan coa...	ABOARD AQUARIUS RESCUE SHIP (Reuters) - Migran...	worldnews	September 19, 2017
8555	Trump ally Christie calls criticisms of slain ...	WASHINGTON (Reuters) - Chris Christie, a close...	politicsNews	August 2, 2016
3883	Trump review of Wall Street rules to be done i...	NEW YORK/WASHINGTON (Reuters) - The U.S. gover...	politicsNews	May 8, 2017

Getting rid of unwanted strings

In [6]:
```
cleansed_data = []
for data in true.text:
    if "@realDonaldTrump : - " in data:
        cleansed_data.append(data.split("@realDonaldTrump : - ")[1])
    elif "(Reuters) -" in data:
        cleansed_data.append(data.split("(Reuters) - ")[1])
    else:
        cleansed_data.append(data)

true["text"] = cleansed_data
display(true.head(10))
```

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	The head of a conservative Republican faction ...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	Transgender people will be allowed for the fir...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	The special counsel investigation of links bet...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	Trump campaign adviser George Papadopoulos tol...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	President Donald Trump called on the U.S. Post...	politicsNews	December 29, 2017
5	White House, Congress prepare for talks on spe...	The White House said on Friday it was set to k...	politicsNews	December 29, 2017
6	Trump says Russia probe will be fair, but time...	President Donald Trump said on Thursday he bel...	politicsNews	December 29, 2017
7	Factbox: Trump on Twitter (Dec 29) - Approval ...	While the Fake News loves to talk about my so-...	politicsNews	December 29, 2017
8	Trump on Twitter (Dec 28) - Global Warming	Together, we are MAKING AMERICA GREAT AGAIN! b...	politicsNews	December 29, 2017
9	Alabama official to certify Senator-elect Jone...	Alabama Secretary of State John Merrill said h...	politicsNews	December 28, 2017

Joining title and text

In [7]:
```
fake['Sentences'] = fake['title'] + ' ' + fake['text']
true['Sentences'] = true['title'] + ' ' + true['text']
```

In [8]:
```
true.head()
```

Out[8]:

	title	text	subject	date	Sentences
0	As U.S. budget fight looms, Republicans flip t...	The head of a conservative Republican faction ...	politicsNews	December 31, 2017	As U.S. budget fight looms, Republicans flip t...
1	U.S. military to accept transgender recruits o...	Transgender people will be allowed for the fir...	politicsNews	December 29, 2017	U.S. military to accept transgender recruits o...

	title	text	subject	date	Sentences
2	Senior U.S. Republican senator: 'Let Mr. Muell...	The special counsel investigation of links bet...	politicsNews	December 31, 2017	Senior U.S. Republican senator: 'Let Mr. Muell...
3	FBI Russia probe helped by Australian diplomat...	Trump campaign adviser George Papadopoulos tol...	politicsNews	December 30, 2017	FBI Russia probe helped by Australian diplomat...
4	Trump wants Postal Service to charge 'much mor...	President Donald Trump called on the U.S. Post...	politicsNews	December 29, 2017	Trump wants Postal Service to charge 'much mor...

Adding Labels, concatenating and mixing

In [9]:

```
fake['Label'] = 0
true['Label'] = 1

final_data = pd.concat([fake, true])

final_data = final_data.sample(frac=1, random_state=42).reset_index(drop=True)
```

Droping uneeded columns

In [10]:

```
final_data = final_data.drop(['title', 'text', 'subject', 'date'], axis = 1)
display(final_data.head(10))
```

	Sentences	Label
0	Ben Stein Calls Out 9th Circuit Court: Committ...	0
1	Trump drops Steve Bannon from National Securit...	1
2	Puerto Rico expects U.S. to lift Jones Act shi...	1
3	OOPS: Trump Just Accidentally Confirmed He Le...	0
4	Donald Trump heads for Scotland to reopen a go...	1
5	Paul Ryan Responds To Dem’s Sit-In On Gun Con...	0
6	AWESOME! DIAMOND AND SILK Rip Into The Press: ...	0
7	STAND UP AND CHEER! UKIP Party Leader SLAMS Ge...	0
8	North Korea shows no sign it is serious about ...	1
9	Trump signals willingness to raise U.S. minimu...	1

Processing Sentences

Function

In [11]:

```
def remove_URL(s):
    regex = re.compile(r'https?://\S+|www\.\S+|bit\.ly\S+')
    return regex.sub(r'',s)
```

List of functions

```
In [12]:  CUSTOM_FILTERS = [lambda x: x.lower(), strip_tags, remove_URL, strip_punctuation, strip
```

Useful info

```
In [13]:  words_broken_up = [preprocess_string(sentence, CUSTOM_FILTERS) for sentence in final_da
```

```
In [14]:  processed_data = [word for word in words_broken_up if len(word) > 0]
```

```
In [15]:  processed_labels = [label for num, label in enumerate(final_data.Label) if len(words_br
```

Word2Vec

```
In [16]:  model = Word2Vec(processed_data, min_count=1)
          display(model.wv.most_similar("country"))
```

```
[('nation', 0.8211482763290405),
 ('america', 0.6449085474014282),
 ('countries', 0.6018150448799133),
 ('europe', 0.5721968412399292),
 ('dealmaker”', 0.5451688766479492),
 ('world', 0.524400532245636),
 ('especially', 0.514208197593689),
 ('means', 0.4974181056022644),
 ('path', 0.48721593618392944),
 ('fear', 0.4780063033103943)]
```

Sentence Vectors

```
In [17]:  def return_vector(model_made, x):
              try:
                  return model_made[x]
              except:
                  return np.zeros(100)


          def sentence_vector(model_made, sentence):
              word_vectors = list(map(lambda x: return_vector(model_made, x), sentence))
              return np.average(word_vectors, axis=0).tolist()
```

```
In [18]:  X = np.array([sentence_vector(model, data) for data in processed_data])
```

```
C:\Users\jarc1\anaconda3\envs\env1\lib\site-packages\ipykernel_launcher.py:3: Deprecatio
nWarning: Call to deprecated `__getitem__` (Method will be removed in 4.0.0, use self.w
v.__getitem__() instead).
  This is separate from the ipykernel package so we can avoid doing imports until
```

Clustering

```
In [19]:  kmeans = cluster.KMeans(n_clusters=2, verbose=1)
          clustered = kmeans.fit_predict(X)
```

```
Initialization complete
Iteration 0, inertia 797034.5814754908
Iteration 1, inertia 469822.3581141416
Iteration 2, inertia 454721.90205345955
Iteration 3, inertia 449163.4764854106
Iteration 4, inertia 446957.490341186
Iteration 5, inertia 445725.57276875945
Iteration 6, inertia 444988.7387002231
Iteration 7, inertia 444631.13049879455
Iteration 8, inertia 444471.74890456063
Iteration 9, inertia 444414.9960367619
Iteration 10, inertia 444396.94775127666
Iteration 11, inertia 444389.32258432487
Iteration 12, inertia 444386.69616800034
Iteration 13, inertia 444385.9518836059
Converged at iteration 13: center shift 1.1116875513575477e-05 within tolerance 1.182505
9075582397e-05
Initialization complete
Iteration 0, inertia 711966.5527862577
Iteration 1, inertia 490386.9343802891
Iteration 2, inertia 482759.41313339345
Iteration 3, inertia 470492.40758340835
Iteration 4, inertia 455541.67887200473
Iteration 5, inertia 447028.1790759409
Iteration 6, inertia 444889.9942452808
Iteration 7, inertia 444477.8944834342
Iteration 8, inertia 444404.1155266255
Iteration 9, inertia 444389.5861153842
Iteration 10, inertia 444386.29967509233
Iteration 11, inertia 444385.43223583774
Converged at iteration 11: center shift 8.901074107577207e-06 within tolerance 1.1825059
075582397e-05
Initialization complete
Iteration 0, inertia 825712.9212451687
Iteration 1, inertia 461575.27426523034
Iteration 2, inertia 453998.2295087943
Iteration 3, inertia 450794.97270186845
Iteration 4, inertia 448724.2315331823
Iteration 5, inertia 447027.3103016334
Iteration 6, inertia 445814.41699653835
Iteration 7, inertia 445040.81112924375
Iteration 8, inertia 444660.32879998954
Iteration 9, inertia 444487.4494418016
Iteration 10, inertia 444421.86736424617
Iteration 11, inertia 444399.38657099917
Iteration 12, inertia 444390.1253206167
Iteration 13, inertia 444386.8843754947
Iteration 14, inertia 444386.0570644636
Iteration 15, inertia 444385.56608521333
Converged at iteration 15: center shift 8.273156374409398e-06 within tolerance 1.1825059
075582397e-05
Initialization complete
Iteration 0, inertia 791204.3731917946
Iteration 1, inertia 450996.8197895801
Iteration 2, inertia 444696.13676217856
Iteration 3, inertia 444434.002323673
Iteration 4, inertia 444394.9245029976
Iteration 5, inertia 444387.3940248714
Iteration 6, inertia 444385.49764435616
Iteration 7, inertia 444385.1345608201
Converged at iteration 7: center shift 1.4312123107181263e-06 within tolerance 1.1825059
075582397e-05
Initialization complete
Iteration 0, inertia 905450.9006095811
Iteration 1, inertia 505321.57567987265
Iteration 2, inertia 475245.2414401457
Iteration 3, inertia 454575.00193936453
Iteration 4, inertia 447527.52548588597
Iteration 5, inertia 445412.3574518756
Iteration 6, inertia 444716.16315272235
Iteration 7, inertia 444489.5553116242
Iteration 8, inertia 444419.7704118113
Iteration 9, inertia 444395.33215859224
Iteration 10, inertia 444387.8616300548
Iteration 11, inertia 444385.8582270308
Iteration 12, inertia 444385.229078487
Converged at iteration 12: center shift 3.5331823936080315e-06 within tolerance 1.182505
9075582397e-05
Initialization complete
Iteration 0, inertia 724871.5527820098
Iteration 1, inertia 467918.0160870421
Iteration 2, inertia 455579.4469178993
Iteration 3, inertia 450626.18249213864
Iteration 4, inertia 448245.8391944835
Iteration 5, inertia 446594.0979387968
Iteration 6, inertia 445507.7116950821
Iteration 7, inertia 444881.3397365774
Iteration 8, inertia 444585.90624431643
Iteration 9, inertia 444459.7136675668
Iteration 10, inertia 444410.1741781191
Iteration 11, inertia 444394.7664575611
Iteration 12, inertia 444388.5576711203
Iteration 13, inertia 444386.472051852
Iteration 14, inertia 444385.8279104046
Converged at iteration 14: center shift 1.0231099111160025e-05 within tolerance 1.182505
9075582397e-05
Initialization complete
Iteration 0, inertia 680945.6979138054
Iteration 1, inertia 515696.56706046854
Iteration 2, inertia 502546.44463076873
Iteration 3, inertia 491263.26452488446
Iteration 4, inertia 478339.039400209
Iteration 5, inertia 463467.00576245127
Iteration 6, inertia 452849.7290338717
Iteration 7, inertia 447212.62070170604
Iteration 8, inertia 445157.23316283955
Iteration 9, inertia 444580.32976048597
Iteration 10, inertia 444427.39076911024
Iteration 11, inertia 444396.77404235577
Iteration 12, inertia 444388.6103638619
Iteration 13, inertia 444386.40478678915
Iteration 14, inertia 444385.80248727393
Converged at iteration 14: center shift 8.882905886104413e-06 within tolerance 1.1825059
075582397e-05
Initialization complete
Iteration 0, inertia 844676.8213299435
Iteration 1, inertia 460604.5681773221
Iteration 2, inertia 448732.38236676104
Iteration 3, inertia 445591.69256136386
Iteration 4, inertia 444755.5736675169
Iteration 5, inertia 444508.89584607
Iteration 6, inertia 444425.3434537398
Iteration 7, inertia 444398.3566692798
Iteration 8, inertia 444390.2759723376
Iteration 9, inertia 444386.94104570657
Iteration 10, inertia 444386.09537966654
Iteration 11, inertia 444385.5881598441
Converged at iteration 11: center shift 9.370973252894358e-06 within tolerance 1.1825059
075582397e-05
Initialization complete
Iteration 0, inertia 816824.6983104716
Iteration 1, inertia 504373.7435899613
Iteration 2, inertia 481634.06468190363
Iteration 3, inertia 449903.3085471271
Iteration 4, inertia 444976.30144724133
Iteration 5, inertia 444482.2367799041
Iteration 6, inertia 444402.962778053
Iteration 7, inertia 444389.43754668144
Iteration 8, inertia 444386.5221587359
Iteration 9, inertia 444385.65789591684
Converged at iteration 9: center shift 1.0973834484185811e-05 within tolerance 1.1825059
075582397e-05
Initialization complete
Iteration 0, inertia 671045.2665509919
Iteration 1, inertia 483858.43822997354
Iteration 2, inertia 459766.83609074814
Iteration 3, inertia 448724.78830022574
Iteration 4, inertia 445776.9687904379
Iteration 5, inertia 444879.7323503956
Iteration 6, inertia 444567.88943177625
Iteration 7, inertia 444453.0618109822
Iteration 8, inertia 444407.8809879115
Iteration 9, inertia 444393.6776780456
Iteration 10, inertia 444388.29665103334
Iteration 11, inertia 444386.39396845835
Iteration 12, inertia 444385.7805568746
Converged at iteration 12: center shift 9.358810224088597e-06 within tolerance 1.1825059
075582397e-05
```

In [20]:
```
testing_df = pd.DataFrame({'Sentence': processed_data, 'Labels': processed_labels, 'Pre
display(testing_df.head(20))
```

	Sentence	Labels	Prediction
0	[ben, stein, calls, circuit, court, committed,...	0	0
1	[trump, drops, steve, bannon, national, securi...	1	1
2	[puerto, rico, expects, lift, jones, act, ship...	1	1
3	[oops, trump, accidentally, confirmed, leaked,...	0	0
4	[donald, trump, heads, scotland, reopen, golf,...	1	0
5	[paul, ryan, responds, dem’s, sit, gun, contro...	0	0
6	[awesome, diamond, silk, rip, press, “we, don’...	0	0
7	[stand, cheer, ukip, party, leader, slams, ger...	0	1
8	[north, korea, shows, sign, talking, official,...	1	1
9	[trump, signals, willingness, raise, minimum, ...	1	0
10	[new, jersey, christie, mulls, run, lead, repu...	1	0
11	[where’s, hillary, clinton, spotted, dining]	0	0
12	[france, germany, want, iran, reverse, ballist...	1	1
13	[aide, commission, head, tweets, picture, whit...	1	1
14	[trump, issues, warning, man, army”, “could, i...	0	1
15	[gives, laos, extra, million, help, clear, une...	1	1
16	[judge, declares, baby, “illegal”, prevent, “e...	0	0

	Sentence	Labels	Prediction
17	[paul, ryan, takes, monumentally, humiliating,...	0	0
18	[republicans, dine, trump, try, railroad, come...	0	0
19	[house, panel, offers, alternative, retirement...	1	1

Validating

In [21]:

```
testing_df['accuracy'] = np.logical_not(np.logical_xor(testing_df['Labels'], testing_df
assertion = np.sum(testing_df.accuracy)/np.sum(len(testing_df.accuracy))*100

print('Data classificated correctly: ', assertion, '%')
```

```
Data classificated correctly:  87.39557575352536 %
```

Visualization

Prinicpal Component Analysis (PCA)

In [22]:

```
pca = PCA(n_components=2)
pca_result = pca.fit_transform(X)

PCA_df = pd.DataFrame(pca_result)
PCA_df['cluster'] = clustered
PCA_df.columns = ['x1','x2','cluster']
```

T-Distributed Stochastic Neighbor Embedding (TSNE)

In [23]:

```
tsne = TSNE(n_components=2)
tsne_result = tsne.fit_transform(pca_result)

TSNE_df = pd.DataFrame(tsne_result)
TSNE_df['cluster'] = clustered
TSNE_df.columns = ['x1','x2','cluster']
```

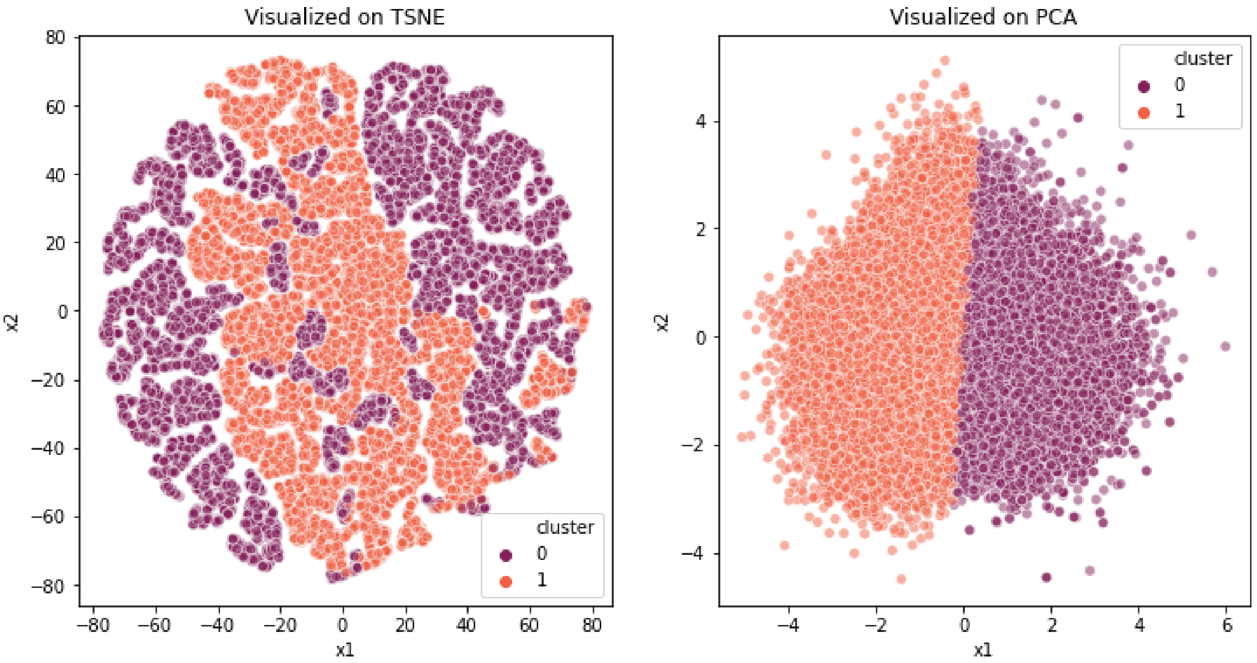
Plots

In [24]:

```
fig, ax = plt.subplots(1, 2, figsize=(12,6))
sns.scatterplot(data=PCA_df,x='x1',y='x2',hue='cluster',legend="full",alpha=0.5,ax=ax[1
sns.scatterplot(data=TSNE_df,x='x1',y='x2',hue='cluster',legend="full",alpha=0.5,ax=ax[
ax[0].set_title('Visualized on TSNE')
ax[1].set_title('Visualized on PCA')
```

Out[24]:

```
Text(0.5, 1.0, 'Visualized on PCA')
```



Custom new tests

Testing with fake news generated from https://www.thefakenewsgenerator.com/

Onion

In [25]:
```
onion_data = "Flint Residents Learn To Harness Superpowers, But Trump Gets Away Again T

# Preprocess article
onion_data = preprocess_string(onion_data, CUSTOM_FILTERS)

# Get sentence vector
onion_data = sentence_vector(model, onion_data)

# Get prediction
kmeans.predict(np.array([onion_data]))
```

```
C:\Users\jarc1\anaconda3\envs\env1\lib\site-packages\ipykernel_launcher.py:3: Deprecatio
nWarning: Call to deprecated `__getitem__` (Method will be removed in 4.0.0, use self.w
v.__getitem__() instead).
  This is separate from the ipykernel package so we can avoid doing imports until
```

Out[25]:
```
array([0])
```

News from BBC

In [26]:
```
bbc_data = "Nasa Mars 2020 Mission's MiMi Aung on women in space Next year, Nasa will s

# Preprocess article
bbc_data = preprocess_string(bbc_data, CUSTOM_FILTERS)

# Get sentence vector
bbc_data = sentence_vector(model, bbc_data)

# Get prediction
kmeans.predict(np.array([bbc_data]))
```

```
C:\Users\jarc1\anaconda3\envs\env1\lib\site-packages\ipykernel_launcher.py:3: Deprecatio
nWarning: Call to deprecated `__getitem__` (Method will be removed in 4.0.0, use self.w
v.__getitem__() instead).
  This is separate from the ipykernel package so we can avoid doing imports until
```

```
Out[26]: array([1])
```