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
6487	Experts Warn This Trump Plan Would Cause 'Glo...	Reality TV star and presumptive Republican pre...	News	May 7, 2016
12899	THIS ONE VERY IMPORTANT THING THE MEDIA MISSED...		politics	Sep 27, 2016
12162	UNREAL! HILLARY CLINTON Greeted By Sobbing Fem...	After the Reid event, Hillary Clinton greeted ...	politics	Dec 12, 2016

	title	text	subject	date
15966	BREAKING: DNC STAFFERS May Have Sold Sensitive...	Well, well, well The DNC IT scandal could beco...	Government News	Aug 20, 2017
16447	HOW WOULD YOU DEFEND YOUR FAMILY FROM VIOLENT ...	Run, yes run, to the nearest gun store and buy...	Government News	Aug 17, 2016
16096	TUCKER CARLSON Confronts Nasty Activist: “I’m ...	Tucker takes on the co-director of Popular Res...	Government News	May 16, 2017
19772	WATCH HILLARY LAUGH When Trump Mentions Gays W...	Hillary s such a champion of gay rights that h...	left-news	Oct 20, 2016
13870	WATCH: TRANSEXUAL MICHELLE OBAMA LOOK-ALIKE Ki...	So it begins the suing of Americans who don t ...	politics	May 20, 2016
15971	CRAZED PROTESTERS Pull Down Confederate Statue...	Crazed lunatic fringe elements of America have...	Government News	Aug 14, 2017
1476	Trump Tweeted Nine Times To Deleted Russian T...	Trump s long record of tweet-before-think acti...	News	May 13, 2017

```
True Sample
```

	title	text	subject	date
21341	Brexit bill gets bigger as euro strengthens	LONDON (Reuters) - As Britain s pound declines...	worldnews	August 25, 2017
6099	McCain proposes $7.5 billion of new U.S. mili...	WASHINGTON (Reuters) - The head of the U.S. Se...	politicsNews	January 24, 2017
15362	EU, U.S. affirm Lebanon support, diverging fro...	BEIRUT (Reuters) - The European Union and the ...	worldnews	November 8, 2017
3260	Sessions says he will discuss Comey with Senat...	WASHINGTON (Reuters) - U.S. Attorney General J...	politicsNews	June 10, 2017
3195	Democratic lawmakers sue Trump over foreign st...	WASHINGTON (Reuters) - More than 190 Democrati...	politicsNews	June 14, 2017
16758	Trump's tougher stance could backfire by boost...	ANKARA/LONDON (Reuters) - Iran s elite Islamic...	worldnews	October 23, 2017
20573	U.S. calls for U.N. Security Council vote on N...	UNITED NATIONS (Reuters) - The United States o...	worldnews	September 9, 2017
16401	Tillerson says Iraq must resist Iran influence	GENEVA (Reuters) - U.S. Secretary of State Rex...	worldnews	October 26, 2017
847	Green groups sue for access to U.S. monument d...	NEW YORK (Reuters) - The Sierra Club and five ...	politicsNews	November 3, 2017
14312	Trump declares North Korea state sponsor of te...	WASHINGTON (Reuters) - President Donald Trump ...	worldnews	November 20, 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 [11]:
```
true.head()
```

Out[11]:

	title	text	subject	date	Sentences	Label
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
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...	1

	title	text	subject	date	Sentences	Label
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...	1
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...	1
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...	1

Adding Labels, concatenating and mixing

In [16]:
```
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 [17]:
```
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 [18]:
```
def remove_URL(s):
    regex = re.compile(r'https?://\S+|www\.\S+|bit\.ly\S+')
    return regex.sub(r'',s)
```

List of functions

In [19]:
```
CUSTOM_FILTERS = [lambda x: x.lower(), strip_tags, remove_URL, strip_punctuation, strip
```

Useful info

In [20]:
```
words_broken_up = [preprocess_string(sentence, CUSTOM_FILTERS) for sentence in final_da
```

In [22]:
```
processed_data = [word for word in words_broken_up if len(word) > 0]
```

In [23]:
```
processed_labels = [label for num, label in enumerate(final_data.Label) if len(words_br
```

Word2Vec

In [26]:
```
model = Word2Vec(processed_data, min_count=1)
display(model.wv.most_similar("country"))
```

```
[('nation', 0.8215814828872681),
 ('america', 0.6285639405250549),
 ('europe', 0.5672596096992493),
 ('countries', 0.5627425909042358),
 ('world', 0.5163451433181763),
 ('especially', 0.5064713358879089),
 ('planet', 0.4862319231033325),
 ('americans', 0.4859480559825897),
 ('communities', 0.47674286365509033),
 ('abroad', 0.474148690700531)]
```

Sentence Vectors

In [27]:
```
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 [43]:
```
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 [31]:
```
kmeans = cluster.KMeans(n_clusters=2, verbose=1)
clustered = kmeans.fit_predict(X)
```

```
Initialization complete
Iteration 0, inertia 810999.4389091282
Iteration 1, inertia 450533.04123004584
Iteration 2, inertia 447901.15583913843
Iteration 3, inertia 446432.5693528905
Iteration 4, inertia 445502.06354076695
Iteration 5, inertia 444952.61006822385
Iteration 6, inertia 444712.7573555605
Iteration 7, inertia 444604.7876254867
Iteration 8, inertia 444566.7221354316
Iteration 9, inertia 444552.29224350274
Iteration 10, inertia 444547.1037764706
Iteration 11, inertia 444544.76187169494
Iteration 12, inertia 444543.95941159996
Converged at iteration 12: center shift 8.681443044054806e-06 within tolerance 1.1832936
542979685e-05
Initialization complete
Iteration 0, inertia 640079.5136113721
Iteration 1, inertia 456489.96046381263
Iteration 2, inertia 449549.99662131275
Iteration 3, inertia 447001.4671568593
Iteration 4, inertia 445770.91196222743
Iteration 5, inertia 445072.46004522755
Iteration 6, inertia 444761.3843881492
Iteration 7, inertia 444623.59421036293
Iteration 8, inertia 444573.8587853524
Iteration 9, inertia 444555.41696768923
Iteration 10, inertia 444547.9007779977
Iteration 11, inertia 444545.1992245343
Iteration 12, inertia 444544.0389473401
Converged at iteration 12: center shift 9.234284272638666e-06 within tolerance 1.1832936
542979685e-05
Initialization complete
Iteration 0, inertia 866645.7818922837
Iteration 1, inertia 475986.5931123498
Iteration 2, inertia 449809.4370914807
Iteration 3, inertia 445935.6242199342
Iteration 4, inertia 444981.2218729469
Iteration 5, inertia 444667.76920299296
Iteration 6, inertia 444581.59515653097
Iteration 7, inertia 444556.4117497593
Iteration 8, inertia 444548.0916478003
Iteration 9, inertia 444545.305272086
Iteration 10, inertia 444544.0432231083
Converged at iteration 10: center shift 1.0879895709704006e-05 within tolerance 1.183293
6542979685e-05
Initialization complete
Iteration 0, inertia 897470.635838007
Iteration 1, inertia 468845.4303673573
Iteration 2, inertia 452501.95589275274
Iteration 3, inertia 446116.6266681562
Iteration 4, inertia 444847.52574705257
Iteration 5, inertia 444600.2056659988
Iteration 6, inertia 444554.01762863446
Iteration 7, inertia 444545.78283916303
Iteration 8, inertia 444544.101945055
Iteration 9, inertia 444543.52998593036
Converged at iteration 9: center shift 4.492427668763844e-06 within tolerance 1.18329365
42979685e-05
Initialization complete
Iteration 0, inertia 825994.6905506456
Iteration 1, inertia 445533.75602303556
```

```
Iteration 2, inertia 444637.3739362086
Iteration 3, inertia 444568.88756037597
Iteration 4, inertia 444551.8583009382
Iteration 5, inertia 444546.44382127613
Iteration 6, inertia 444544.4790412817
Iteration 7, inertia 444543.7891246144
Converged at iteration 7: center shift 8.057952656825175e-06 within tolerance 1.18329365
42979685e-05
Initialization complete
Iteration 0, inertia 800178.8751888698
Iteration 1, inertia 456626.17319965997
Iteration 2, inertia 445650.6754039017
Iteration 3, inertia 444682.9572888273
Iteration 4, inertia 444569.84971194784
Iteration 5, inertia 444550.96168459574
Iteration 6, inertia 444545.9386075608
Iteration 7, inertia 444544.3159024447
Iteration 8, inertia 444543.71225588775
Converged at iteration 8: center shift 7.42671023213798e-06 within tolerance 1.183293654
2979685e-05
Initialization complete
Iteration 0, inertia 733680.8842540112
Iteration 1, inertia 444937.375430951
Iteration 2, inertia 444649.28465905844
Iteration 3, inertia 444579.6959780805
Iteration 4, inertia 444557.7670667637
Iteration 5, inertia 444548.8896119019
Iteration 6, inertia 444545.6325064063
Iteration 7, inertia 444544.17711115465
Converged at iteration 7: center shift 1.1461042908515297e-05 within tolerance 1.1832936
542979685e-05
Initialization complete
Iteration 0, inertia 845459.1156135998
Iteration 1, inertia 477552.9677891203
Iteration 2, inertia 461284.46058569424
Iteration 3, inertia 454218.3948804996
Iteration 4, inertia 450866.9318772187
Iteration 5, inertia 448774.0591500787
Iteration 6, inertia 447109.0708748334
Iteration 7, inertia 445917.3981073505
Iteration 8, inertia 445167.25834394235
Iteration 9, inertia 444800.52954013407
Iteration 10, inertia 444643.07314299355
Iteration 11, inertia 444581.0577296819
Iteration 12, inertia 444558.7770699155
Iteration 13, inertia 444549.29745104787
Iteration 14, inertia 444545.8805175102
Iteration 15, inertia 444544.2767239068
Iteration 16, inertia 444543.7466628344
Converged at iteration 16: center shift 6.3734655217966885e-06 within tolerance 1.183293
6542979685e-05
Initialization complete
Iteration 0, inertia 693915.7883272534
Iteration 1, inertia 455778.00850657275
Iteration 2, inertia 451364.1563567481
Iteration 3, inertia 449007.33425923885
Iteration 4, inertia 447267.7608871586
Iteration 5, inertia 446013.51676272816
Iteration 6, inertia 445209.09275955934
Iteration 7, inertia 444817.5804643324
Iteration 8, inertia 444650.29418101325
Iteration 9, inertia 444583.19176331867
Iteration 10, inertia 444559.1438957858
Iteration 11, inertia 444549.42732426134
Iteration 12, inertia 444545.8980840112
```

```
Iteration 13, inertia 444544.2767239068
Iteration 14, inertia 444543.7466628344
Converged at iteration 14: center shift 6.373465521796564e-06 within tolerance 1.1832936
542979685e-05
Initialization complete
Iteration 0, inertia 908575.0674408921
Iteration 1, inertia 481546.27845967776
Iteration 2, inertia 465867.4235719301
Iteration 3, inertia 451807.80886073894
Iteration 4, inertia 446021.62358087784
Iteration 5, inertia 444799.8900896346
Iteration 6, inertia 444593.825957866
Iteration 7, inertia 444553.2017334842
Iteration 8, inertia 444546.3048587095
Iteration 9, inertia 444544.4316738958
Iteration 10, inertia 444543.625122391
Converged at iteration 10: center shift 6.6001819932360276e-06 within tolerance 1.183293
6542979685e-05
```

In [32]:
```
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	1
1	[trump, drops, steve, bannon, national, securi...	1	0
2	[puerto, rico, expects, lift, jones, act, ship...	1	0
3	[oops, trump, accidentally, confirmed, leaked,...	0	1
4	[donald, trump, heads, scotland, reopen, golf,...	1	1
5	[paul, ryan, responds, dem’s, sit, gun, contro...	0	1
6	[awesome, diamond, silk, rip, press, “we, don’...	0	1
7	[stand, cheer, ukip, party, leader, slams, ger...	0	0
8	[north, korea, shows, sign, talking, official,...	1	0
9	[trump, signals, willingness, raise, minimum, ...	1	1
10	[new, jersey, christie, mulls, run, lead, repu...	1	1
11	[where’s, hillary, clinton, spotted, dining]	0	1
12	[france, germany, want, iran, reverse, ballist...	1	0
13	[aide, commission, head, tweets, picture, whit...	1	0
14	[trump, issues, warning, man, army”, “could, i...	0	0
15	[gives, laos, extra, million, help, clear, une...	1	0
16	[judge, declares, baby, “illegal”, prevent, “e...	0	1
17	[paul, ryan, takes, monumentally, humiliating,...	0	1
18	[republicans, dine, trump, try, railroad, come...	0	1
19	[house, panel, offers, alternative, retirement...	1	0

Validating

In [34]:
```
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:  12.620018267281516 %
```

Visualization

Prinicpal Component Analysis (PCA)

In [35]:
```
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)

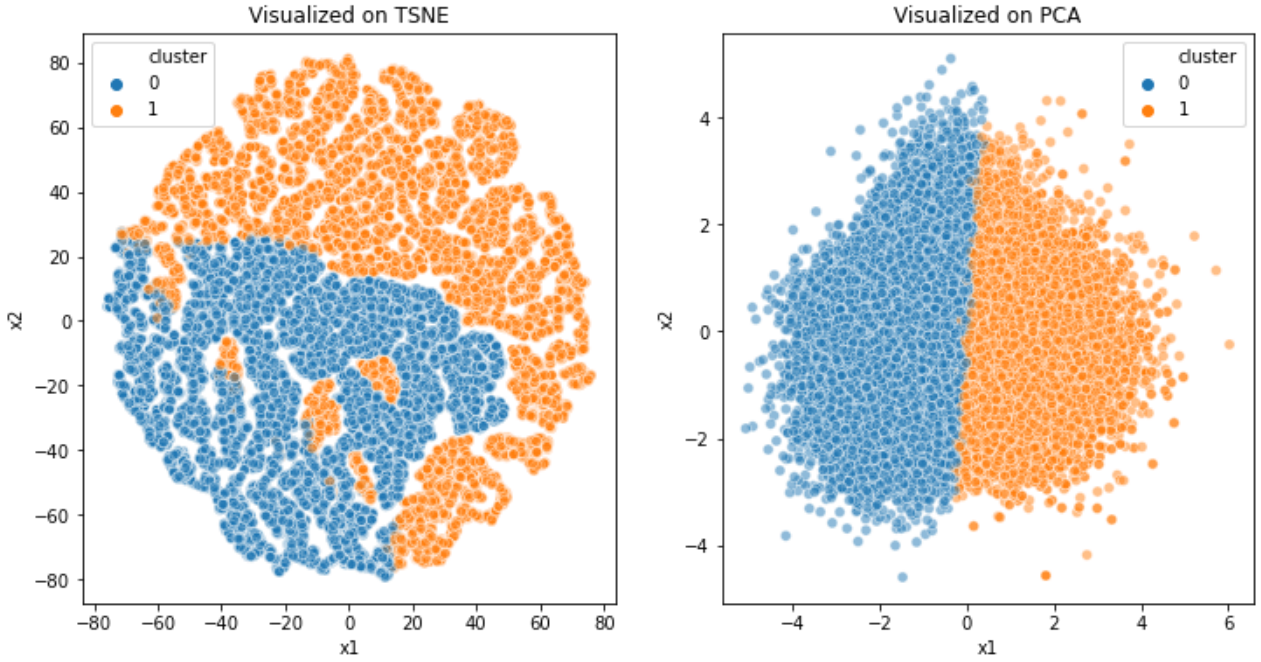
In [36]:
```
tsne = TSNE(n_components=2)
tsne_result = tsne.fit_transform(pca_result)

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

Plots

In [40]:
```
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[40]:
```
Text(0.5, 1.0, 'Visualized on PCA')
```



Custom new tests

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

Onion

In [41]:
```
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[41]:
```
array([1])
```

News from BBC

In [42]:
```
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: DeprecationWarning: Call to deprecated `__getitem__` (Method will be removed in 4.0.0, use self.wv.__getitem__() instead).
  This is separate from the ipykernel package so we can avoid doing imports until
```

Out[42]:
```
array([0])
```