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_
punctuation, \
                                        strip_multiple_whitespaces, strip_nume
ric, \
                                        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
6472	Sarah Palin’s Pick To Replace Paul Ryan Wants...	The media has been telling us for months now t...	News	May 8, 2016
2020	WATCH: Hell Freezes As Fox News Host Calls Re...	Things just went from bad to worse for the Rep...	News	March 24, 2017
4341	Trump Has Been BRIBING Government Officials F...	While the amount of Trump s ahem charitable gi...	News	October 6, 2016
9431	LEFTIST MEDIA EXPOSES Democrat Party For Ignor...	Hollywood producer kingpin and mega-Democrat P...	politics	Nov 14, 2017
21110	WATCH HILLARY SQUIRM When Mainstream Media Ask...	Hillary was too busy to be bothered with makin...	left-news	Jan 17, 2016
18089	WATCH SOLAR ECLIPSE LIVE HERE	If you don t have the proper glasses to watch ...	left-news	Aug 21, 2017
14445	FEMALE VETERAN Slays Hillary In a Few Short Mi...	This is great! We love it when young women get...	politics	Feb 20, 2016
13141	4 THINGS THE MEDIA WON’T TELL YOU About “Oppre...	San Fransisco 49er s quarterback Colin Kaepern...	politics	Aug 29, 2016
6951	Poverty Kills As The Rich Live Longer Than Ev...	The Republican war on poor people has been kil...	News	April 12, 2016
4647	Stephen King Compares Trump To Cthulhu, An Of...	Donald Trump is so deplorable that even Cthulh...	News	September 13, 2016

```
True Sample
```

	title	text	subject	date
12262	Colombia urgently crafting law to allow crime ...	BOGOTA (Reuters) - Colombia s government is ur...	worldnews	December 14, 2017
19423	China urges North Korea not to go further in a...	UNITED NATIONS (Reuters) - China s foreign min...	worldnews	September 21, 2017
2113	Trump fires adviser Bannon	WASHINGTON/HAGERSTOWN, Md. (Reuters) - Preside...	politicsNews	August 18, 2017
10922	U.S. health official: Widespread Zika vaccine ...	(Reuters) - U.S. health officials said on Mond...	politicsNews	February 8, 2016
14605	China says 'dual suspension' proposal still be...	BEIJING (Reuters) - China said on Thursday a ...	worldnews	November 16, 2017
16827	Malta offers 1 million-euro reward to find jou...	VALLETTA (Reuters) - Malta s government said o...	worldnews	October 21, 2017
8464	Talk of shifting funds away from Trump prematu...	WASHINGTON (Reuters) - A senior official with ...	politicsNews	August 14, 2016
13132	Britain's Hammond says 'very confident' of dea...	BRUSSELS (Reuters) - Britain s Chancellor of t...	worldnews	December 5, 2017
3109	U.S. State Department questions Gulf motives o...	WASHINGTON (Reuters) - The U.S. State Departme...	politicsNews	June 20, 2017
20762	UK's Prince George starts school, pregnant mum...	LONDON (Reuters) - Britain s Prince George, th...	worldnews	September 7, 2017

Getting rid of unwanted strings

In [6]:

```python
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
true.head(10)
```

Out[6]:

	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]:

```python
fake['Sentences'] = fake['title'] + ' ' + fake['text']
true['Sentences'] = true['title'] + ' ' + true['text']
```

Adding Labels, concatenating and mixing

In [8]:

```
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 [9]:

```
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 [10]:

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

List of functions

In [11]:

```
CUSTOM_FILTERS = [lambda x: x.lower(), strip_tags, remove_URL, strip_punctuati
on, strip_multiple_whitespaces, strip_numeric, remove_stopwords, strip_short]
```

Useful info

In [12]:

```
words_broken_up = [preprocess_string(sentence, CUSTOM_FILTERS) for sentence in 
final_data.Sentences]
```

In [13]:

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

In [14]:

```
processed_labels = [label for num, label in enumerate(final_data.Label) if len
(words_broken_up[num]) > 0]
```

Word2Vec

In [15]:

```
model = Word2Vec(processed_data, min_count=1)
display(model.wv.most_similar("country"))
```

```
[('nation', 0.8124138116836548),
 ('america', 0.6545636057853699),
 ('europe', 0.5747597813606262),
 ('countries', 0.5584433674812317),
 ('especially', 0.49935755133628845),
 ('prosperous', 0.49221035838127136),
 ('feel', 0.48947253823280334),
 ('abroad', 0.4890715479850769),
 ('world', 0.48643702268600464),
 ('africa', 0.47567102313041687)]
```

Sentence Vectors

In [16]:

```
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 [17]:

```
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: 
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 u
ntil
```

Clustering

In [18]:

```
kmeans = cluster.KMeans(n_clusters=2, verbose=1)
clustered = kmeans.fit_predict(X)
```

```
Initialization complete
Iteration 0, inertia 646358.9154381026
Iteration 1, inertia 479174.78568014095
Iteration 2, inertia 458252.62133920245
Iteration 3, inertia 448388.6043523719
Iteration 4, inertia 445796.8233863477
Iteration 5, inertia 445046.4458195295
Iteration 6, inertia 444859.6258746112
Iteration 7, inertia 444805.0153343606
Iteration 8, inertia 444786.6924731565
Iteration 9, inertia 444780.8224541747
Iteration 10, inertia 444779.2221854268
Iteration 11, inertia 444778.68996250944
Converged at iteration 11: center shift 3.206914606224222e-06 within toleranc
e 1.1845714228159858e-05
Initialization complete
Iteration 0, inertia 777618.5267570995
Iteration 1, inertia 483221.46217111824
Iteration 2, inertia 446974.8537673994
Iteration 3, inertia 445279.44100356207
Iteration 4, inertia 444931.5301610576
Iteration 5, inertia 444825.1707570919
Iteration 6, inertia 444791.54453206
Iteration 7, inertia 444782.70562524284
Iteration 8, inertia 444779.7002658901
Iteration 9, inertia 444778.8951731812
Converged at iteration 9: center shift 9.24403774375695e-06 within tolerance
1.1845714228159858e-05
Initialization complete
Iteration 0, inertia 801453.0786591835
Iteration 1, inertia 479167.1504049558
Iteration 2, inertia 462373.9288628715
Iteration 3, inertia 454742.17593144375
Iteration 4, inertia 451189.2718656833
Iteration 5, inertia 449049.9312260796
Iteration 6, inertia 447354.2327471069
Iteration 7, inertia 446163.40909162036
Iteration 8, inertia 445418.6696321503
Iteration 9, inertia 445042.28246655065
Iteration 10, inertia 444877.044654669
Iteration 11, inertia 444815.1421684565
Iteration 12, inertia 444791.45314770495
Iteration 13, inertia 444782.700704972
Iteration 14, inertia 444780.33104440593
Iteration 15, inertia 444779.39913793234
Iteration 16, inertia 444778.9690788849
Converged at iteration 16: center shift 4.6724187682457e-06 within tolerance
1.1845714228159858e-05
Initialization complete
Iteration 0, inertia 737269.3330422463
Iteration 1, inertia 494445.3759406277
Iteration 2, inertia 468649.33454727236
Iteration 3, inertia 451743.1982069422
Iteration 4, inertia 446575.1375122909
Iteration 5, inertia 445243.80141140526
Iteration 6, inertia 444893.6872132383
Iteration 7, inertia 444810.33289645193
Iteration 8, inertia 444788.46201054624
Iteration 9, inertia 444781.92596435954
Iteration 10, inertia 444780.12057111785
Iteration 11, inertia 444779.2685245906
Converged at iteration 11: center shift 9.742140536847953e-06 within toleranc
e 1.1845714228159858e-05
Initialization complete
Iteration 0, inertia 706817.8949163846
Iteration 1, inertia 492886.8336924054
Iteration 2, inertia 482637.2751833688
Iteration 3, inertia 470465.4558249034
Iteration 4, inertia 457926.3743569074
Iteration 5, inertia 450345.0890976133
Iteration 6, inertia 446787.9236901671
Iteration 7, inertia 445549.4525579802
Iteration 8, inertia 445035.86700172024
Iteration 9, inertia 444860.5050350835
Iteration 10, inertia 444807.99246080674
Iteration 11, inertia 444787.6967792908
Iteration 12, inertia 444781.78717467794
Iteration 13, inertia 444780.12057111785
Iteration 14, inertia 444779.2685245906
Converged at iteration 14: center shift 9.742140536847865e-06 within toleranc
e 1.1845714228159858e-05
Initialization complete
Iteration 0, inertia 660475.649880208
Iteration 1, inertia 461244.72224639804
Iteration 2, inertia 451914.707224504
Iteration 3, inertia 449015.3937244078
Iteration 4, inertia 447220.51708811184
Iteration 5, inertia 446048.27482923324
Iteration 6, inertia 445361.54385017767
Iteration 7, inertia 445022.77162806713
Iteration 8, inertia 444869.3238140668
Iteration 9, inertia 444813.39428276476
Iteration 10, inertia 444791.15363943984
Iteration 11, inertia 444782.62095077144
Iteration 12, inertia 444780.33104440593
Iteration 13, inertia 444779.39913793234
Iteration 14, inertia 444778.9690788849
Converged at iteration 14: center shift 4.6724187682457495e-06 within toleran
ce 1.1845714228159858e-05
Initialization complete
Iteration 0, inertia 719903.5877453176
Iteration 1, inertia 472047.3707503147
Iteration 2, inertia 453833.0926722607
Iteration 3, inertia 448838.8397467963
Iteration 4, inertia 447016.64549836627
Iteration 5, inertia 445930.8211350542
Iteration 6, inertia 445286.1974751019
Iteration 7, inertia 444983.442628219
Iteration 8, inertia 444854.2507036427
Iteration 9, inertia 444806.9522977306
Iteration 10, inertia 444787.2741860823
Iteration 11, inertia 444781.7824512382
Iteration 12, inertia 444780.01520727226
Iteration 13, inertia 444779.2493469654
```

```
Converged at iteration 13: center shift 1.0098529681401642e-05 within toleran
ce 1.1845714228159858e-05
Initialization complete
Iteration 0, inertia 654237.111012707
Iteration 1, inertia 463075.3758536617
Iteration 2, inertia 454997.07844064763
Iteration 3, inertia 451348.4545639733
Iteration 4, inertia 449182.1405942238
Iteration 5, inertia 447458.2360687532
Iteration 6, inertia 446226.4230954077
Iteration 7, inertia 445451.3215352411
Iteration 8, inertia 445056.76831236336
Iteration 9, inertia 444882.1355827968
Iteration 10, inertia 444816.8543294063
Iteration 11, inertia 444792.66860920895
Iteration 12, inertia 444782.92359972117
Iteration 13, inertia 444780.36309790326
Iteration 14, inertia 444779.3991379324
Iteration 15, inertia 444778.9690788849
Converged at iteration 15: center shift 4.6724187682458325e-06 within toleran
ce 1.1845714228159858e-05
Initialization complete
Iteration 0, inertia 966741.7035354939
Iteration 1, inertia 463370.11344668834
Iteration 2, inertia 450648.3110408531
Iteration 3, inertia 447642.8563213908
Iteration 4, inertia 446171.25763937586
Iteration 5, inertia 445382.1958523393
Iteration 6, inertia 445028.96686383383
Iteration 7, inertia 444872.37115436397
Iteration 8, inertia 444814.0901491148
Iteration 9, inertia 444791.20274607884
Iteration 10, inertia 444782.62095077144
Iteration 11, inertia 444780.33104440593
Iteration 12, inertia 444779.3991379324
Iteration 13, inertia 444778.9690788849
Converged at iteration 13: center shift 4.672418768245696e-06 within toleranc
e 1.1845714228159858e-05
Initialization complete
Iteration 0, inertia 716499.9240820269
Iteration 1, inertia 495409.94500878465
Iteration 2, inertia 459110.05129976297
Iteration 3, inertia 447598.44153984386
Iteration 4, inertia 445266.374874223
Iteration 5, inertia 444862.8224159512
Iteration 6, inertia 444791.8575828882
Iteration 7, inertia 444780.75859850564
Iteration 8, inertia 444778.7835116862
Converged at iteration 8: center shift 9.316624248558061e-06 within tolerance
1.1845714228159858e-05
```

In [19]:
```
testing_df = pd.DataFrame({'Sentence': processed_data, 'Labels': processed_lab
els, 'Prediction': clustered})
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
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 [20]:
```
testing_df['assertion'] = np.logical_not(np.logical_xor(testing_df['Labels'],
testing_df['Prediction']))
assertion = np.sum(testing_df.assertion)/np.sum(len(testing_df.assertion))*100

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

```
Data classificated correctly:  87.3465659738466 %
```

Visualization

Prinicpal Component Analysis (PCA)

In [21]:
```
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)

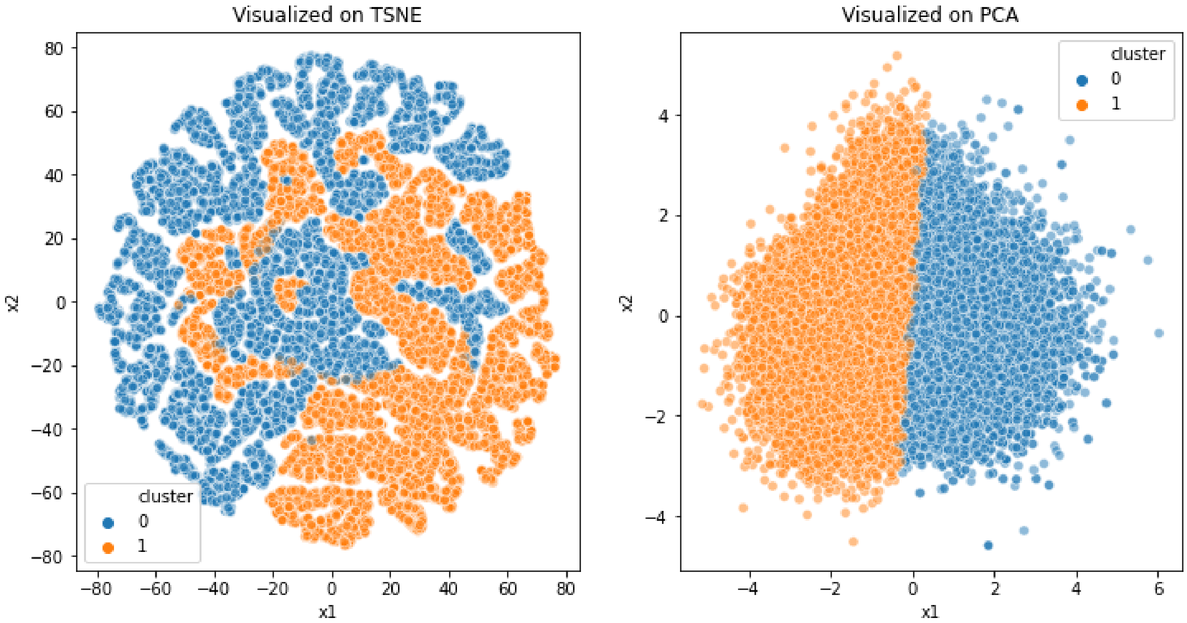
In [22]:
```
tsne = TSNE(n_components=2)
tsne_result = tsne.fit_transform(pca_result)

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

Plots

In [23]:
```
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[0])
ax[0].set_title('Visualized on TSNE')
ax[1].set_title('Visualized on PCA')
```

Out[23]:
```
Text(0.5, 1.0, 'Visualized on PCA')
```



Custom new tests

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

Onion

In [24]:

```
onion_data = "Flint Residents Learn To Harness Superpowers, But Trump Gets Awa
y Again They developed superpowers after years of drinking from a lead-poisone
d water supply. But just having incredible abilities doesn't make them superhe
roes. Not yet. Donald Trump faced off against the superpowered civilians but h
e got away before they could catch him"

# 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:
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 u
ntil
```

Out[24]:

```
array([0])
```

News from BBC

In [25]:

```
bbc_data = "Nasa Mars 2020 Mission's MiMi Aung on women in space Next year, Na
sa will send a mission to Mars. The woman in charge of making the helicopter t
hat will be sent there - which is set to become the first aircraft to fly on a
nother planet - is MiMi Aung. At 16, MiMi travelled alone from Myanmar to the
 US for access to education. She is now one of the lead engineers at Nasa. We
 find out what it's like being a woman in space exploration, and why her mum i
s her biggest inspiration."

# 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 u
ntil
```

Out[25]:

```
array([1])
```