```python
In [1]:  # importing libraries
         import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import seaborn as sns
         import nltk
         from nltk.corpus import stopwords
         from nltk.tokenize import word_tokenize
         from nltk.stem import WordNetLemmatizer
         from sklearn import cluster
         from sklearn.feature_extraction.text import TfidfVectorizer
         from sklearn.preprocessing import StandardScaler
         from sklearn.decomposition import PCA
         from sklearn.manifold import TSNE
         from sklearn import metrics
         from sklearn.model_selection import GridSearchCV
         from sklearn.neighbors import NearestNeighbors
         import string
         import re
         from gensim.parsing.preprocessing import preprocess_string, strip_tags, strip_punctuatio
         from gensim.models import Word2Vec
         from pylab import savefig
         from collections import Counter
         def warn(*args, **kwargs):
             pass
         import warnings
         warnings.warn = warn
         warnings.filterwarnings('ignore')
```

```python
In [2]:  # Functions

         def most_frequent(List):
             most_freq_start = []
             most_freq_mid = []
             occurence_count = Counter(List).most_common(1000)
             occurence_count_start = occurence_count[0:99]
             occurence_count_mid = occurence_count[500:599]
             for inst_start, inst_mid in zip( occurence_count_start, occurence_count_mid):
                 if (len(inst_start[0]) >= 5 and len(most_freq_start) <= 10):
                     most_freq_start.append(inst_start[0])
                 if (len(inst_mid[0]) >= 5 and len(most_freq_mid) <= 10):
                     most_freq_mid.append(inst_mid[0])
                 if (len(most_freq_start) > 9 and len(most_freq_mid) > 9):
                     break
             return most_freq_start, most_freq_mid

         def accuracy(dataset, label, prediction):
             correct = 0
             incorrect = 0
             for index, row in dataset.iterrows():
                 if row[label] == row[prediction]:
                     correct += 1
                 else:
                     incorrect += 1

             return (correct / (correct + incorrect) * 100)
```

```python
In [3]:  # Import Data
         filepath = "data\WELFake_Dataset.csv\WELFake_Dataset.csv"

         data = pd.read_csv(filepath)
```

```
In [4]:  # Data Cleaning
         cols = [x for x in data.columns if x in ['title', 'text', 'label']]
         raw_data = data[cols]
```

```
In [5]:  # Initial EDA
         raw_data.head()
```

Out[5]:

	title	text	label
0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	No comment is expected from Barack Obama Membe...	1
1	NaN	Did they post their votes for Hillary already?	1
2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	Now, most of the demonstrators gathered last ...	1
3	Bobby Jindal, raised Hindu, uses story of Chri...	A dozen politically active pastors came here f...	0
4	SATAN 2: Russia unvelis an image of its terrif...	The RS-28 Sarmat missile, dubbed Satan 2, will...	1

```
In [6]:  plot = plt.axes()
         plot.set_title('Fake vs Real News Count')
         sns.countplot(raw_data['label'])
         plot.set_xlabel('label')
         plot.set_xticklabels(['Fake(0)','Real(1)'])
```

Out[6]: [Text(0, 0, 'Fake(0)'), Text(1, 0, 'Real(1)')]
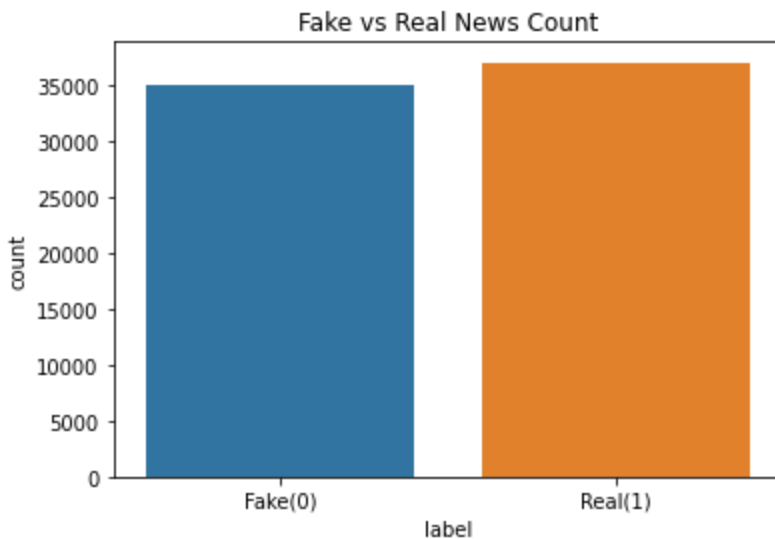

```
In [7]:  raw_data.isnull().sum()
```

Out[7]: title   558
         text   39
         label   0
         dtype: int64

```
In [8]:  raw_data.dropna(inplace=True)
```

```
In [9]:  final_data = raw_data.copy()
         final_data['text'] = raw_data['title'] + " " + raw_data['text']
         final_data.drop(['title'], axis=1, inplace=True)
         final_data.head()
```

Out[9]:

	text	label
0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	1

2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	1
3	Bobby Jindal, raised Hindu, uses story of Chri...	0
4	SATAN 2: Russia unvelis an image of its terrif...	1
5	About Time! Christian Group Sues Amazon and SP...	1

In [13]:
```python
from nltk.corpus import stopwords
from wordcloud import WordCloud

wordcloud = WordCloud(width = 800, height = 800,
            background_color ='white',
            stopwords = stopwords.words('english'),
            min_font_size = 10).generate(" ".join(final_data[final_data['label'] ==

# plot the word cloud for fake news data
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)
plt.show()
```


In [15]:
```python
from wordcloud import WordCloud

wordcloud = WordCloud(width = 800, height = 800,
            background_color ='white',
```

```
                        stopwords = stopwords.words('english'),
                        min_font_size = 10).generate(" ".join(final_data[final_data['label'] ==

# plot the WordCloud image for genuine news data
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)
plt.show()
```



```
In [16]:  lemmatizer = WordNetLemmatizer()
          nltk.download('stopwords')
          nltk.download('punkt')
          nltk.download('wordnet')
          nltk.download('omw-1.4')

          corpus = []
          for i in range(0,len(final_data)):
              message = re.sub('[^a-zA-Z0-9]', ' ', final_data['text'].iloc[i])
              message = message.lower()
              message = word_tokenize(message)
              message = [lemmatizer.lemmatize(w) for w in message if not w in stopwords.words('eng
              message = ' '.join(message)
              corpus.append(message)
```

```
[nltk_data] Downloading package stopwords to C:\Users\Liam's
[nltk_data]     Computer\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [17]:
```python
tfidf_v = TfidfVectorizer(max_features=1000)
X = tfidf_v.fit_transform(corpus).toarray()
```

In [18]:
```python
scaler = StandardScaler()
X = scaler.fit_transform(X)
X.shape
```

Out[18]:
```
(71537, 1000)
```

In [20]:
```python
Kmeans_two = cluster.KMeans(n_clusters=2,init='k-means++',max_iter=500,verbose=True,rand
clustered_two = Kmeans_two.fit_predict(X)
```

```
Initialization complete
Iteration 0, inertia 126907378.6682678.
Iteration 1, inertia 71393971.80174512.
Iteration 2, inertia 71210064.51525897.
Iteration 3, inertia 71116660.17064306.
Iteration 4, inertia 71054145.72680879.
Iteration 5, inertia 71009743.47583708.
Iteration 6, inertia 70979301.68345556.
Iteration 7, inertia 70961413.36761042.
Iteration 8, inertia 70951670.50327478.
Iteration 9, inertia 70946864.65941939.
Iteration 10, inertia 70944218.98051375.
Iteration 11, inertia 70942640.82491356.
Iteration 12, inertia 70941618.0541618.
Iteration 13, inertia 70940979.85584612.
Iteration 14, inertia 70940582.6102931.
Iteration 15, inertia 70940340.1700507.
Iteration 16, inertia 70940159.55415925.
Iteration 17, inertia 70940009.61082678.
Iteration 18, inertia 70939903.43891843.
Iteration 19, inertia 70939847.05255662.
Iteration 20, inertia 70939796.71910559.
Iteration 21, inertia 70939748.92557037.
Iteration 22, inertia 70939712.0029668.
Iteration 23, inertia 70939688.54810639.
Iteration 24, inertia 70939671.37678596.
Iteration 25, inertia 70939658.59936193.
Iteration 26, inertia 70939644.30980809.
Iteration 27, inertia 70939632.59871891.
Iteration 28, inertia 70939625.82302088.
Iteration 29, inertia 70939619.68782094.
Converged at iteration 29: center shift 7.063882217052981e-05 within tolerance 0.0001000
0000000000223.
Initialization complete
Iteration 0, inertia 113018754.64627728.
Iteration 1, inertia 71444701.68760891.
Iteration 2, inertia 71389223.94163036.
Iteration 3, inertia 71230716.31143019.
Iteration 4, inertia 71041349.68651347.
Iteration 5, inertia 70975547.27076502.
Iteration 6, inertia 70954819.2012967.
Iteration 7, inertia 70947293.0064354.
Iteration 8, inertia 70943790.28473523.
```

```
Iteration 9, inertia 70942087.10068129.
Iteration 10, inertia 70941150.95710817.
Iteration 11, inertia 70940583.57783408.
Iteration 12, inertia 70940294.918616.
Iteration 13, inertia 70940112.71921828.
Iteration 14, inertia 70939960.1963909.
Iteration 15, inertia 70939875.14093682.
Iteration 16, inertia 70939821.68710682.
Iteration 17, inertia 70939769.13114269.
Iteration 18, inertia 70939726.59215048.
Iteration 19, inertia 70939698.734638.
Iteration 20, inertia 70939678.40365902.
Iteration 21, inertia 70939662.59711349.
Iteration 22, inertia 70939652.83072646.
Iteration 23, inertia 70939636.58598904.
Iteration 24, inertia 70939629.41783582.
Iteration 25, inertia 70939623.2142228.
Iteration 26, inertia 70939618.66217393.
Converged at iteration 26: center shift 7.325008708742591e-05 within tolerance 0.0001000
0000000000223.
Initialization complete
Iteration 0, inertia 112762321.05626398.
Iteration 1, inertia 71278547.75490151.
Iteration 2, inertia 71157511.83061016.
Iteration 3, inertia 71112849.2163999.
Iteration 4, inertia 71082610.53422195.
Iteration 5, inertia 71054371.22181453.
Iteration 6, inertia 71035947.80386528.
Iteration 7, inertia 71024507.5197763.
Iteration 8, inertia 71012795.86248703.
Iteration 9, inertia 70996853.85852094.
Iteration 10, inertia 70979159.88314964.
Iteration 11, inertia 70965058.70506093.
Iteration 12, inertia 70957170.34982494.
Iteration 13, inertia 70952768.87000386.
Iteration 14, inertia 70950041.2578285.
Iteration 15, inertia 70947919.8900221.
Iteration 16, inertia 70946471.63100281.
Iteration 17, inertia 70945476.73220016.
Iteration 18, inertia 70944675.85344349.
Iteration 19, inertia 70944016.99878797.
Iteration 20, inertia 70943529.21886519.
Iteration 21, inertia 70943135.22985706.
Iteration 22, inertia 70942732.38962817.
Iteration 23, inertia 70942331.0172832.
Iteration 24, inertia 70941910.22054645.
Iteration 25, inertia 70941519.71393809.
Iteration 26, inertia 70941169.5565074.
Iteration 27, inertia 70940937.97931373.
Iteration 28, inertia 70940735.20040855.
Iteration 29, inertia 70940555.69921368.
Iteration 30, inertia 70940403.48649488.
Iteration 31, inertia 70940304.34984492.
Iteration 32, inertia 70940209.97498292.
Iteration 33, inertia 70940131.23880106.
Iteration 34, inertia 70940067.27718341.
Iteration 35, inertia 70940021.00180022.
Iteration 36, inertia 70939984.96444507.
Iteration 37, inertia 70939949.47499633.
Iteration 38, inertia 70939901.88600594.
Iteration 39, inertia 70939846.76541801.
Iteration 40, inertia 70939788.65226126.
Iteration 41, inertia 70939743.60423847.
Iteration 42, inertia 70939699.5981079.
Iteration 43, inertia 70939668.37103558.
Iteration 44, inertia 70939648.55839719.
```

```
Iteration 45, inertia 70939630.5153456.
Iteration 46, inertia 70939620.3791443.
Iteration 47, inertia 70939614.47695231.
Converged at iteration 47: center shift 9.724628479080752e-05 within tolerance 0.0001000
0000000000223.
Initialization complete
Iteration 0, inertia 117742312.57251918.
Iteration 1, inertia 71374172.50352204.
Iteration 2, inertia 71277709.79146034.
Iteration 3, inertia 71248237.23310892.
Iteration 4, inertia 71218380.09436089.
Iteration 5, inertia 71168145.02943736.
Iteration 6, inertia 71124101.04733402.
Iteration 7, inertia 71098270.99266176.
Iteration 8, inertia 71084453.84069178.
Iteration 9, inertia 71076928.77503173.
Iteration 10, inertia 71070534.62352452.
Iteration 11, inertia 71062376.84625646.
Iteration 12, inertia 71049764.26084417.
Iteration 13, inertia 71030624.76154512.
Iteration 14, inertia 71005443.87698856.
Iteration 15, inertia 70981047.40015002.
Iteration 16, inertia 70962733.14967862.
Iteration 17, inertia 70951555.88286576.
Iteration 18, inertia 70946148.25961415.
Iteration 19, inertia 70943283.24177015.
Iteration 20, inertia 70941784.67877412.
Iteration 21, inertia 70941003.75792311.
Iteration 22, inertia 70940588.22188482.
Iteration 23, inertia 70940320.30017833.
Iteration 24, inertia 70940117.16933128.
Iteration 25, inertia 70939977.62746483.
Iteration 26, inertia 70939885.30653809.
Iteration 27, inertia 70939830.78183815.
Iteration 28, inertia 70939776.61509503.
Iteration 29, inertia 70939738.68196456.
Iteration 30, inertia 70939706.54805525.
Iteration 31, inertia 70939685.29476452.
Iteration 32, inertia 70939668.21722597.
Iteration 33, inertia 70939655.06145199.
Iteration 34, inertia 70939642.45314093.
Iteration 35, inertia 70939629.9171657.
Iteration 36, inertia 70939623.21962921.
Converged at iteration 36: center shift 8.55093157563338e-05 within tolerance 0.00010000
000000000223.
Initialization complete
Iteration 0, inertia 120881155.63810182.
Iteration 1, inertia 71416581.42428283.
Iteration 2, inertia 71366101.42478657.
Iteration 3, inertia 71307383.75421074.
Iteration 4, inertia 71263834.55820017.
Iteration 5, inertia 71194710.2367394.
Iteration 6, inertia 71094086.91866249.
Iteration 7, inertia 71046360.88578676.
Iteration 8, inertia 71032449.94094141.
Iteration 9, inertia 71023015.3152013.
Iteration 10, inertia 71011016.63037428.
Iteration 11, inertia 70994393.28154883.
Iteration 12, inertia 70976545.83383621.
Iteration 13, inertia 70963749.69477676.
Iteration 14, inertia 70956769.37113275.
Iteration 15, inertia 70952532.70882738.
Iteration 16, inertia 70949936.83561985.
Iteration 17, inertia 70947884.0549.
Iteration 18, inertia 70946468.92321178.
Iteration 19, inertia 70945475.62877336.
```

```
Iteration 20, inertia 70944682.64303757.
Iteration 21, inertia 70944028.17217189.
Iteration 22, inertia 70943533.92417537.
Iteration 23, inertia 70943143.76968716.
Iteration 24, inertia 70942740.82980248.
Iteration 25, inertia 70942341.9053883.
Iteration 26, inertia 70941911.95345874.
Iteration 27, inertia 70941519.35984299.
Iteration 28, inertia 70941171.28590415.
Iteration 29, inertia 70940938.4194293.
Iteration 30, inertia 70940734.5071546.
Iteration 31, inertia 70940553.75423944.
Iteration 32, inertia 70940401.80456835.
Iteration 33, inertia 70940303.80154513.
Iteration 34, inertia 70940208.98552069.
Iteration 35, inertia 70940129.03664835.
Iteration 36, inertia 70940065.75694975.
Iteration 37, inertia 70940020.4134427.
Iteration 38, inertia 70939984.34966445.
Iteration 39, inertia 70939949.47499633.
Iteration 40, inertia 70939901.88600594.
Iteration 41, inertia 70939846.76541801.
Iteration 42, inertia 70939788.65226126.
Iteration 43, inertia 70939743.60423847.
Iteration 44, inertia 70939699.5981079.
Iteration 45, inertia 70939668.37103558.
Iteration 46, inertia 70939648.55839719.
Iteration 47, inertia 70939630.5153456.
Iteration 48, inertia 70939620.3791443.
Iteration 49, inertia 70939614.47695231.
Converged at iteration 49: center shift 9.724628479080763e-05 within tolerance 0.0001000
0000000000223.
Initialization complete
Iteration 0, inertia 136436786.51022908.
Iteration 1, inertia 71281784.98309095.
Iteration 2, inertia 71122707.8081944.
Iteration 3, inertia 71067216.11856696.
Iteration 4, inertia 71030770.0831398.
Iteration 5, inertia 71000389.29470243.
Iteration 6, inertia 70978970.97097859.
Iteration 7, inertia 70965718.35386476.
Iteration 8, inertia 70958363.07921338.
Iteration 9, inertia 70954236.84578255.
Iteration 10, inertia 70950474.73975077.
Iteration 11, inertia 70948305.11971653.
Iteration 12, inertia 70946824.13299246.
Iteration 13, inertia 70945658.70616972.
Iteration 14, inertia 70944802.24852714.
Iteration 15, inertia 70944110.72404456.
Iteration 16, inertia 70943567.1217458.
Iteration 17, inertia 70943182.50561449.
Iteration 18, inertia 70942775.73716843.
Iteration 19, inertia 70942364.13779704.
Iteration 20, inertia 70941949.75436454.
Iteration 21, inertia 70941560.614674.
Iteration 22, inertia 70941208.39178784.
Iteration 23, inertia 70940973.23975092.
Iteration 24, inertia 70940772.40000318.
Iteration 25, inertia 70940586.56903763.
Iteration 26, inertia 70940430.13201399.
Iteration 27, inertia 70940322.19309637.
Iteration 28, inertia 70940234.44661741.
Iteration 29, inertia 70940160.72591765.
Iteration 30, inertia 70940089.79095127.
Iteration 31, inertia 70940036.61242048.
Iteration 32, inertia 70940003.39126721.
```

```
Iteration 33, inertia 70939968.23878312.
Iteration 34, inertia 70939928.48453036.
Iteration 35, inertia 70939876.54276863.
Iteration 36, inertia 70939822.33268991.
Iteration 37, inertia 70939764.39114237.
Iteration 38, inertia 70939721.47769146.
Iteration 39, inertia 70939682.97564648.
Iteration 40, inertia 70939656.18622087.
Iteration 41, inertia 70939638.08268017.
Iteration 42, inertia 70939623.7179258.
Iteration 43, inertia 70939616.9091.
Iteration 44, inertia 70939612.06736994.
Converged at iteration 44: center shift 7.643618963285351e-05 within tolerance 0.0001000
0000000000223.
Initialization complete
Iteration 0, inertia 121261474.59189206.
Iteration 1, inertia 71354231.64692473.
Iteration 2, inertia 71267767.4108159.
Iteration 3, inertia 71216730.18542942.
Iteration 4, inertia 71169519.64076395.
Iteration 5, inertia 71118965.47411087.
Iteration 6, inertia 71077639.53587398.
Iteration 7, inertia 71044182.15825486.
Iteration 8, inertia 71011576.90381688.
Iteration 9, inertia 70985001.06087737.
Iteration 10, inertia 70968141.58410715.
Iteration 11, inertia 70958844.02087569.
Iteration 12, inertia 70953050.36000644.
Iteration 13, inertia 70949134.32672061.
Iteration 14, inertia 70947202.73683721.
Iteration 15, inertia 70945821.36061576.
Iteration 16, inertia 70944836.54897425.
Iteration 17, inertia 70944108.3259157.
Iteration 18, inertia 70943545.14862224.
Iteration 19, inertia 70943145.25664663.
Iteration 20, inertia 70942732.95745453.
Iteration 21, inertia 70942308.33055209.
Iteration 22, inertia 70941894.48776698.
Iteration 23, inertia 70941505.03507732.
Iteration 24, inertia 70941156.71462491.
Iteration 25, inertia 70940923.97644728.
Iteration 26, inertia 70940730.98252636.
Iteration 27, inertia 70940550.97307266.
Iteration 28, inertia 70940395.02912425.
Iteration 29, inertia 70940297.75363004.
Iteration 30, inertia 70940205.68631332.
Iteration 31, inertia 70940126.17891671.
Iteration 32, inertia 70940062.47451809.
Iteration 33, inertia 70940018.83012474.
Iteration 34, inertia 70939982.03736113.
Iteration 35, inertia 70939948.32323517.
Iteration 36, inertia 70939899.2046187.
Iteration 37, inertia 70939844.69585094.
Iteration 38, inertia 70939787.26842971.
Iteration 39, inertia 70939742.4285446.
Iteration 40, inertia 70939699.44984105.
Iteration 41, inertia 70939668.55988288.
Iteration 42, inertia 70939648.79807144.
Iteration 43, inertia 70939630.90157627.
Iteration 44, inertia 70939619.86692187.
Converged at iteration 44: center shift 9.501057077260497e-05 within tolerance 0.0001000
0000000000223.
Initialization complete
Iteration 0, inertia 129956436.85143813.
Iteration 1, inertia 71503597.14389494.
Iteration 2, inertia 71453105.86408398.
```

```
Iteration 3, inertia 71268249.98540254.
Iteration 4, inertia 71096045.07691893.
Iteration 5, inertia 71018583.94169419.
Iteration 6, inertia 70979167.24098.
Iteration 7, inertia 70959822.30611652.
Iteration 8, inertia 70950622.58318782.
Iteration 9, inertia 70946366.73815441.
Iteration 10, inertia 70943900.3624204.
Iteration 11, inertia 70942418.26076795.
Iteration 12, inertia 70941429.06532913.
Iteration 13, inertia 70940857.61742139.
Iteration 14, inertia 70940503.21665166.
Iteration 15, inertia 70940282.33208202.
Iteration 16, inertia 70940126.79119325.
Iteration 17, inertia 70939980.78606537.
Iteration 18, inertia 70939889.67868865.
Iteration 19, inertia 70939836.47123924.
Iteration 20, inertia 70939786.27926236.
Iteration 21, inertia 70939742.24809968.
Iteration 22, inertia 70939709.36764167.
Iteration 23, inertia 70939689.63184811.
Iteration 24, inertia 70939671.38781556.
Iteration 25, inertia 70939659.00746845.
Iteration 26, inertia 70939644.8539691.
Iteration 27, inertia 70939632.00434653.
Iteration 28, inertia 70939625.53870091.
Iteration 29, inertia 70939619.65504794.
Converged at iteration 29: center shift 5.7998193826831344e-05 within tolerance 0.000100
00000000000223.
Initialization complete
Iteration 0, inertia 110987947.12610132.
Iteration 1, inertia 71424834.99960089.
Iteration 2, inertia 71229042.69821979.
Iteration 3, inertia 71072693.40629669.
Iteration 4, inertia 71031225.4239119.
Iteration 5, inertia 71014094.68028308.
Iteration 6, inertia 70996695.56560768.
Iteration 7, inertia 70978226.66681284.
Iteration 8, inertia 70964359.40895978.
Iteration 9, inertia 70956716.37851958.
Iteration 10, inertia 70952517.69926855.
Iteration 11, inertia 70949769.77249955.
Iteration 12, inertia 70947725.5246126.
Iteration 13, inertia 70946320.69747865.
Iteration 14, inertia 70945344.04331832.
Iteration 15, inertia 70944544.26454884.
Iteration 16, inertia 70943913.26108064.
Iteration 17, inertia 70943446.30420293.
Iteration 18, inertia 70943063.41415381.
Iteration 19, inertia 70942658.43509449.
Iteration 20, inertia 70942266.7023793.
Iteration 21, inertia 70941849.55933715.
Iteration 22, inertia 70941461.64293979.
Iteration 23, inertia 70941112.64541443.
Iteration 24, inertia 70940895.35240367.
Iteration 25, inertia 70940708.37156208.
Iteration 26, inertia 70940534.78010015.
Iteration 27, inertia 70940385.76838143.
Iteration 28, inertia 70940293.03571877.
Iteration 29, inertia 70940205.01399828.
Iteration 30, inertia 70940125.51501125.
Iteration 31, inertia 70940061.61516745.
Iteration 32, inertia 70940017.42802343.
Iteration 33, inertia 70939980.38478905.
Iteration 34, inertia 70939947.47632554.
Iteration 35, inertia 70939897.85545343.
```

```
Iteration 36, inertia 70939844.2519016.
Iteration 37, inertia 70939785.67889042.
Iteration 38, inertia 70939742.43857117.
Iteration 39, inertia 70939699.3967315.
Iteration 40, inertia 70939668.6693642.
Iteration 41, inertia 70939648.7422919.
Iteration 42, inertia 70939630.43204105.
Iteration 43, inertia 70939619.9463916.
Iteration 44, inertia 70939614.6191473.
Iteration 45, inertia 70939610.15694748.
Converged at iteration 45: center shift 6.157037681295185e-05 within tolerance 0.0001000
0000000000223.
Initialization complete
Iteration 0, inertia 128711058.26859924.
Iteration 1, inertia 71503420.52809727.
Iteration 2, inertia 71465610.11933064.
Iteration 3, inertia 71423377.85479641.
Iteration 4, inertia 71326947.44984414.
Iteration 5, inertia 71222320.18018176.
Iteration 6, inertia 71178532.05199039.
Iteration 7, inertia 71156433.4727684.
Iteration 8, inertia 71140234.00326912.
Iteration 9, inertia 71126332.52366413.
Iteration 10, inertia 71113141.2666147.
Iteration 11, inertia 71101679.11271968.
Iteration 12, inertia 71092668.71357445.
Iteration 13, inertia 71086458.51935452.
Iteration 14, inertia 71082165.88636464.
Iteration 15, inertia 71079296.03149122.
Iteration 16, inertia 71077032.71227688.
Iteration 17, inertia 71074952.8768694.
Iteration 18, inertia 71073149.26802817.
Iteration 19, inertia 71071266.9559493.
Iteration 20, inertia 71069094.33908692.
Iteration 21, inertia 71066591.69192056.
Iteration 22, inertia 71063396.73292239.
Iteration 23, inertia 71059010.90220417.
Iteration 24, inertia 71052520.22839351.
Iteration 25, inertia 71042257.70832488.
Iteration 26, inertia 71027331.558768.
Iteration 27, inertia 71008058.37359935.
Iteration 28, inertia 70987809.40718749.
Iteration 29, inertia 70971888.21058336.
Iteration 30, inertia 70962040.5480084.
Iteration 31, inertia 70956193.02141646.
Iteration 32, inertia 70951757.65200883.
Iteration 33, inertia 70948999.80275838.
Iteration 34, inertia 70947305.04820284.
Iteration 35, inertia 70946017.54505576.
Iteration 36, inertia 70945063.46367586.
Iteration 37, inertia 70944301.68765804.
Iteration 38, inertia 70943702.69724438.
Iteration 39, inertia 70943301.77645776.
Iteration 40, inertia 70942891.43614936.
Iteration 41, inertia 70942498.74425854.
Iteration 42, inertia 70942099.05736901.
Iteration 43, inertia 70941699.07325824.
Iteration 44, inertia 70941313.42218782.
Iteration 45, inertia 70941036.93227005.
Iteration 46, inertia 70940825.18913159.
Iteration 47, inertia 70940649.97838385.
Iteration 48, inertia 70940483.76871668.
Iteration 49, inertia 70940354.66831695.
Iteration 50, inertia 70940262.51931463.
Iteration 51, inertia 70940187.58391742.
Iteration 52, inertia 70940112.00520435.
```

```
Iteration 53, inertia 70940052.31191027.
Iteration 54, inertia 70940012.62785694.
Iteration 55, inertia 70939977.96226554.
Iteration 56, inertia 70939941.70384717.
Iteration 57, inertia 70939892.4953284.
Iteration 58, inertia 70939834.06541738.
Iteration 59, inertia 70939773.89183164.
Iteration 60, inertia 70939729.55087207.
Iteration 61, inertia 70939689.6500127.
Iteration 62, inertia 70939662.5411342.
Iteration 63, inertia 70939645.38163638.
Iteration 64, inertia 70939627.6951457.
Iteration 65, inertia 70939619.05406956.
Iteration 66, inertia 70939614.10648862.
Converged at iteration 66: center shift 8.011498678636532e-05 within tolerance 0.0001000
0000000000223.
```

In [21]:
```python
dataset_predict = final_data.copy()
dataset_predict['prediction_two'] = clustered_two
dataset_predict.head(10)
```

Out[21]:

	text	label	prediction_two
0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	1	1
2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	1	0
3	Bobby Jindal, raised Hindu, uses story of Chri...	0	1
4	SATAN 2: Russia unvelis an image of its terrif...	1	0
5	About Time! Christian Group Sues Amazon and SP...	1	0
6	DR BEN CARSON TARGETED BY THE IRS: "I never ha...	1	0
7	HOUSE INTEL CHAIR On Trump-Russia Fake Story: ...	1	0
8	Sports Bar Owner Bans NFL Games...Will Show Only...	1	1
9	Latest Pipeline Leak Underscores Dangers Of Da...	1	0
10	GOP Senator Just Smacked Down The Most Puncha...	1	1

In [22]:
```python
print(str(accuracy(dataset_predict, 'label', 'prediction_two')) + '% Articles Correctly
```
```
69.40184799474397% Articles Correctly Clustered
```

In [23]:
```python
pca = PCA(n_components=2)
pca_result = pca.fit_transform(X)
pca_result.shape
```

Out[23]:
```
(71537, 2)
```

In [24]:
```python
file_name = 'pca_result.csv'
pd.DataFrame(pca_result).to_csv(file_name)
```

In [25]:
```python
Kmeans_PCA_two = cluster.KMeans(n_clusters=2,init='k-means++',max_iter=500,verbose=True,
clustered_PCA_two = Kmeans_PCA_two.fit_predict(pca_result)
```
```
Initialization complete
Iteration 0, inertia 1256623.7427207571.
Iteration 1, inertia 1129793.5988176877.
Iteration 2, inertia 1074926.6707693024.
Iteration 3, inertia 1040971.480994144.
Iteration 4, inertia 1021188.7673884124.
Iteration 5, inertia 1009599.2288710562.
```

```
Iteration 6, inertia 1003534.9791816946.
Iteration 7, inertia 1000367.2456550621.
Iteration 8, inertia 998426.7541044132.
Iteration 9, inertia 997373.9007164704.
Iteration 10, inertia 996814.711179529.
Iteration 11, inertia 996481.2912968348.
Iteration 12, inertia 996238.9682565404.
Iteration 13, inertia 996081.4271664933.
Iteration 14, inertia 996005.3424902244.
Converged at iteration 14: center shift 0.0007320040155624746 within tolerance 0.0011096
20681239686.
Initialization complete
Iteration 0, inertia 1304880.672803718.
Iteration 1, inertia 1072199.5429208768.
Iteration 2, inertia 1020987.6841006003.
Iteration 3, inertia 1005049.7286695084.
Iteration 4, inertia 1000042.4638434208.
Iteration 5, inertia 998175.7710396963.
Iteration 6, inertia 997322.9769128986.
Iteration 7, inertia 996817.3935140179.
Iteration 8, inertia 996481.0517477762.
Iteration 9, inertia 996299.5732225708.
Iteration 10, inertia 996178.7065796647.
Iteration 11, inertia 996078.16698548.
Iteration 12, inertia 995999.6178416157.
Converged at iteration 12: center shift 0.001036178254424796 within tolerance 0.00110962
0681239686.
Initialization complete
Iteration 0, inertia 2042567.0567252696.
Iteration 1, inertia 1086294.8996133583.
Iteration 2, inertia 1064085.8319141264.
Iteration 3, inertia 1042387.9766222172.
Iteration 4, inertia 1027691.0144473348.
Iteration 5, inertia 1018310.0854289633.
Iteration 6, inertia 1011139.7118973644.
Iteration 7, inertia 1006825.5616484187.
Iteration 8, inertia 1004078.4263898855.
Iteration 9, inertia 1002205.91561114.
Iteration 10, inertia 1000816.7665555256.
Iteration 11, inertia 999649.7196851522.
Iteration 12, inertia 998645.7638783893.
Iteration 13, inertia 997929.6734302558.
Iteration 14, inertia 997358.9980903446.
Iteration 15, inertia 996860.4272202271.
Iteration 16, inertia 996516.229786485.
Iteration 17, inertia 996323.4610716987.
Iteration 18, inertia 996197.5818336988.
Iteration 19, inertia 996092.1015933478.
Iteration 20, inertia 996012.1556083161.
Converged at iteration 20: center shift 0.0009153761765912562 within tolerance 0.0011096
20681239686.
Initialization complete
Iteration 0, inertia 1482225.6331779098.
Iteration 1, inertia 1021054.8916429228.
Iteration 2, inertia 1003460.5865211894.
Iteration 3, inertia 998166.1095833484.
Iteration 4, inertia 996591.4191010187.
Iteration 5, inertia 996106.3172211477.
Iteration 6, inertia 995940.8191926253.
Converged at iteration 6: center shift 0.001063836365397435 within tolerance 0.00110962
0681239686.
Initialization complete
Iteration 0, inertia 1479913.834170286.
Iteration 1, inertia 1068945.8856988098.
Iteration 2, inertia 1032700.8204807538.
Iteration 3, inertia 1015014.5586317453.
```

```
Iteration 4, inertia 1005867.5207771492.
Iteration 5, inertia 1001581.2958835439.
Iteration 6, inertia 999088.3055112746.
Iteration 7, inertia 997752.4057606203.
Iteration 8, inertia 997006.5748199269.
Iteration 9, inertia 996616.8103843316.
Iteration 10, inertia 996342.617233049.
Iteration 11, inertia 996139.5958250453.
Iteration 12, inertia 996033.402886769.
Converged at iteration 12: center shift 0.00080799952796410551 within tolerance 0.0011096
20681239686.
Initialization complete
Iteration 0, inertia 1550450.6284790493.
Iteration 1, inertia 1151210.4361219571.
Iteration 2, inertia 1137194.7642577698.
Iteration 3, inertia 1128946.310314767.
Iteration 4, inertia 1121102.1018958369.
Iteration 5, inertia 1110640.8213093085.
Iteration 6, inertia 1093940.1503232163.
Iteration 7, inertia 1070401.1041904916.
Iteration 8, inertia 1045187.1790854204.
Iteration 9, inertia 1024310.1757900064.
Iteration 10, inertia 1010792.4042434094.
Iteration 11, inertia 1003907.397925228.
Iteration 12, inertia 1000420.1224045076.
Iteration 13, inertia 998421.7395614588.
Iteration 14, inertia 997338.9648979363.
Iteration 15, inertia 996797.261590145.
Iteration 16, inertia 996471.2072900049.
Iteration 17, inertia 996228.9460894582.
Iteration 18, inertia 996075.8947273863.
Iteration 19, inertia 996003.8898725911.
Converged at iteration 19: center shift 0.000748731953406195 within tolerance 0.00110962
0681239686.
Initialization complete
Iteration 0, inertia 1550406.2929406762.
Iteration 1, inertia 1155753.7180632993.
Iteration 2, inertia 1082360.558992974.
Iteration 3, inertia 1041229.8452670004.
Iteration 4, inertia 1020318.2737591977.
Iteration 5, inertia 1008664.1004240076.
Iteration 6, inertia 1003093.8971177947.
Iteration 7, inertia 1000000.7323716764.
Iteration 8, inertia 998238.5938196504.
Iteration 9, inertia 997254.277455633.
Iteration 10, inertia 996743.2250198094.
Iteration 11, inertia 996428.8155552761.
Iteration 12, inertia 996200.2303082383.
Iteration 13, inertia 996059.8465460294.
Converged at iteration 13: center shift 0.000992771175273876 within tolerance 0.00110962
0681239686.
Initialization complete
Iteration 0, inertia 1344250.558173859.
Iteration 1, inertia 1020776.5829565604.
Iteration 2, inertia 1006628.2046340295.
Iteration 3, inertia 1001991.924081785.
Iteration 4, inertia 999985.9759780444.
Iteration 5, inertia 998638.4544078903.
Iteration 6, inertia 997843.4702906207.
Iteration 7, inertia 997277.175441651.
Iteration 8, inertia 996794.6327437846.
Iteration 9, inertia 996475.6956098749.
Iteration 10, inertia 996302.9075408303.
Iteration 11, inertia 996178.5407919907.
Iteration 12, inertia 996079.1143993122.
Iteration 13, inertia 995999.8721795405.
```

```
Converged at iteration 13: center shift 0.0010488352825846797 within tolerance 0.0011096
20681239686.
Initialization complete
Iteration 0, inertia 1384354.7884290419.
Iteration 1, inertia 1062701.4599220217.
Iteration 2, inertia 1039642.1221731947.
Iteration 3, inertia 1025608.3227785777.
Iteration 4, inertia 1017092.739315414.
Iteration 5, inertia 1010457.8312873873.
Iteration 6, inertia 1006453.5171653291.
Iteration 7, inertia 1003856.3178039716.
Iteration 8, inertia 1002062.3625862853.
Iteration 9, inertia 1000683.6352019911.
Iteration 10, inertia 999543.8964452269.
Iteration 11, inertia 998574.7615419964.
Iteration 12, inertia 997858.2390487683.
Iteration 13, inertia 997293.1796318426.
Iteration 14, inertia 996803.7996371195.
Iteration 15, inertia 996485.8134823443.
Iteration 16, inertia 996308.1049978202.
Iteration 17, inertia 996182.8489349551.
Iteration 18, inertia 996082.9797624865.
Iteration 19, inertia 996004.0659039896.
Iteration 20, inertia 995936.9311892035.
Converged at iteration 20: center shift 0.0006082368225012965 within tolerance 0.0011096
20681239686.
Initialization complete
Iteration 0, inertia 2651628.8696994907.
Iteration 1, inertia 1130177.9889352107.
Iteration 2, inertia 1129197.1105727423.
Iteration 3, inertia 1128293.2354355294.
Iteration 4, inertia 1127201.0856140177.
Iteration 5, inertia 1125664.7473575547.
Iteration 6, inertia 1123525.2718074308.
Iteration 7, inertia 1120578.8339077819.
Iteration 8, inertia 1116487.3296536258.
Iteration 9, inertia 1109829.6682026484.
Iteration 10, inertia 1098266.5873149328.
Iteration 11, inertia 1080087.1686797047.
Iteration 12, inertia 1056865.4031234416.
Iteration 13, inertia 1037118.2428230159.
Iteration 14, inertia 1024457.305859624.
Iteration 15, inertia 1016053.9161001501.
Iteration 16, inertia 1009787.0026763418.
Iteration 17, inertia 1006011.6229861734.
Iteration 18, inertia 1003567.0149228588.
Iteration 19, inertia 1001864.7464708453.
Iteration 20, inertia 1000513.3837498587.
Iteration 21, inertia 999385.0925825813.
Iteration 22, inertia 998463.2079426572.
Iteration 23, inertia 997765.9841677428.
Iteration 24, inertia 997206.5424910712.
Iteration 25, inertia 996747.9524624799.
Iteration 26, inertia 996452.146743401.
Iteration 27, inertia 996289.4080454275.
Iteration 28, inertia 996168.0793255643.
Iteration 29, inertia 996074.8967690866.
Iteration 30, inertia 995996.4508457335.
Converged at iteration 30: center shift 0.0010192454927505155 within tolerance 0.0011096
20681239686.
```

In [26]:
```python
dataset_predict['prediction_PCA_two'] = clustered_PCA_two
dataset_predict['prediction_PCA_two'] = dataset_predict.apply(
    lambda x: (0 if x['prediction_PCA_two']==1 else 1), axis=1)
dataset_predict.head(10)
```
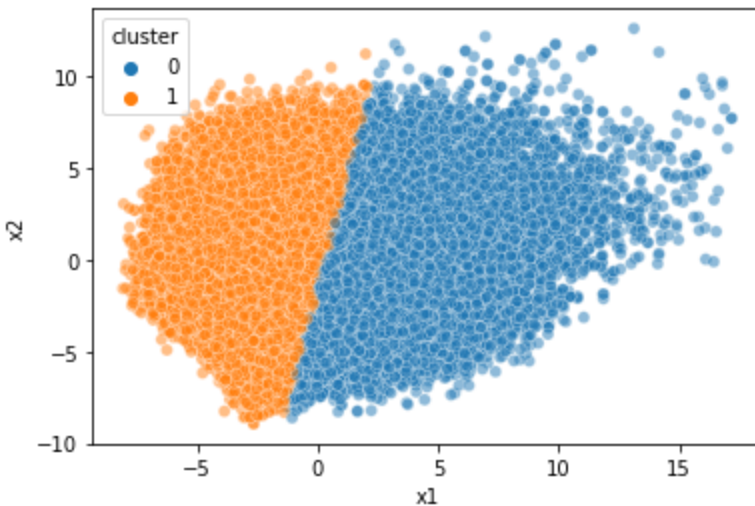
	text	label	prediction_two	prediction_PCA_two
0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	1	1	1
2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	1	0	0
3	Bobby Jindal, raised Hindu, uses story of Chri...	0	1	1
4	SATAN 2: Russia unvelis an image of its terrif...	1	0	0
5	About Time! Christian Group Sues Amazon and SP...	1	0	0
6	DR BEN CARSON TARGETED BY THE IRS: "I never ha...	1	0	0
7	HOUSE INTEL CHAIR On Trump-Russia Fake Story: ...	1	0	0
8	Sports Bar Owner Bans NFL Games...Will Show Only...	1	1	1
9	Latest Pipeline Leak Underscores Dangers Of Da...	1	0	0
10	GOP Senator Just Smacked Down The Most Puncha...	1	1	1

In [27]:
```python
PCA_df = pd.DataFrame(pca_result)
PCA_df['cluster'] = clustered_PCA_two
PCA_df.columns = ['x1','x2','cluster']
k_means_figure = sns.scatterplot(data=PCA_df,x='x1',y='x2',hue='cluster',legend="full",a
```


In [28]:
```python
print(str(accuracy(dataset_predict, 'label', 'prediction_PCA_two')) + '% Articles Correc
```

69.80863049890267% Articles Correctly Clustered

In [29]:
```python
dataset_predict['corpus'] = corpus
cluster_0 = []
cluster_1 = []
for i in range(0,len(dataset_predict)):
    corpora = dataset_predict['corpus'].iloc[i].split()
    if dataset_predict['prediction_PCA_two'].iloc[i] == 0:
        for x in corpora:
            cluster_0.append(x)
    elif dataset_predict['prediction_PCA_two'].iloc[i] == 1:
        for x in corpora:
            cluster_1.append(x)
```

In [30]:
```python
clusters_all = [cluster_0, cluster_1]
title = ['Category0','Category1']
index = 0
for cluster_inst in clusters_all:
    most_freq_start, most_freq_mid = most_frequent(cluster_inst)
    print(title[index] + ' Articles: ')
```

```
    index =+ 1
    print('Top 1% common words: ' + str(most_freq_start))
    print('Top 10% common words: ' + str(most_freq_mid))
    print()
```

Category0 Articles:
Top 1% common words: ['trump', 'state', 'would', 'president', 'government', 'republica
n', 'house', 'people', 'official', 'clinton']
Top 10% common words: ['hearing', 'expert', 'individual', 'allow', 'agent', 'dollar', 'i
raqi', 'climate', 'terrorism', 'spending', 'proposal']

Category1 Articles:
Top 1% common words: ['trump', 'people', 'would', 'clinton', 'state', 'president', 'amer
ican', 'republican', 'hillary', 'obama']
Top 10% common words: ['belief', 'outside', 'breitbart', 'nearly', 'sunday', 'victim',
'worker', 'thousand', 'understand', 'perhaps', 'economy']

In [31]:
```python
wcss = []
for i in range(1,20):
 kmeans = cluster.KMeans(n_clusters=i,init='k-means++',max_iter=300,n_init=10,random_sta
 kmeans.fit(pca_result)
 wcss.append(kmeans.inertia_)
 #print("Cluster", i, "Intertia", kmeans.inertia_)
plt.plot(range(1,20),wcss)
plt.title('The Elbow Curve')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```
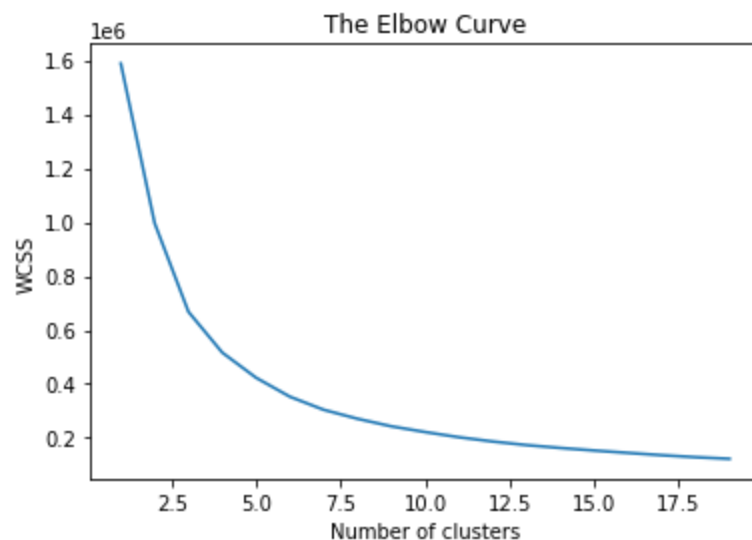

In [32]:
```python
Kmeans_PCA_ten = cluster.KMeans(n_clusters=10,init='k-means++',max_iter=500,verbose=True
clustered_PCA_ten = Kmeans_PCA_ten.fit_predict(pca_result)
```

Initialization complete
Iteration 0, inertia 268317.50930172007.
Iteration 1, inertia 239861.554601915.
Iteration 2, inertia 232855.0611244911.
Iteration 3, inertia 229458.89647359838.
Iteration 4, inertia 227632.66578054454.
Iteration 5, inertia 226528.66640123402.
Iteration 6, inertia 225782.32749865053.
Iteration 7, inertia 225211.0565361768.
Iteration 8, inertia 224781.60601742336.
Iteration 9, inertia 224469.26439259958.
Iteration 10, inertia 224212.88323543168.
Iteration 11, inertia 223973.473386796.
Iteration 12, inertia 223757.35519612837.
Iteration 13, inertia 223535.08543604912.

```
Iteration 14, inertia 223325.1149468134.
Iteration 15, inertia 223165.37489242863.
Iteration 16, inertia 223016.1908353007.
Iteration 17, inertia 222876.48018861478.
Iteration 18, inertia 222763.73934339915.
Iteration 19, inertia 222659.83277610366.
Iteration 20, inertia 222572.12704482125.
Iteration 21, inertia 222495.2059916145.
Iteration 22, inertia 222432.91606504237.
Iteration 23, inertia 222384.8627184083.
Iteration 24, inertia 222347.4809224637.
Iteration 25, inertia 222315.3157327197.
Iteration 26, inertia 222289.7081726862.
Iteration 27, inertia 222271.60689289472.
Converged at iteration 27: center shift 0.0006482516250330816 within tolerance 0.0011096
20681239686.
Initialization complete
Iteration 0, inertia 288226.594566519.
Iteration 1, inertia 251459.593886216.
Iteration 2, inertia 246215.38450981205.
Iteration 3, inertia 242889.51349077193.
Iteration 4, inertia 240365.28938898485.
Iteration 5, inertia 238308.03465943216.
Iteration 6, inertia 236643.68715475727.
Iteration 7, inertia 235347.22306927043.
Iteration 8, inertia 234416.22803748096.
Iteration 9, inertia 233752.50133881392.
Iteration 10, inertia 233312.0901913289.
Iteration 11, inertia 232968.21786859565.
Iteration 12, inertia 232705.496365962.
Iteration 13, inertia 232504.9767834391.
Iteration 14, inertia 232351.87643892213.
Iteration 15, inertia 232229.19955585204.
Iteration 16, inertia 232125.97585650266.
Iteration 17, inertia 232035.67395338768.
Iteration 18, inertia 231918.07072481638.
Iteration 19, inertia 231798.95987610944.
Iteration 20, inertia 231675.24452194947.
Iteration 21, inertia 231553.3760153856.
Iteration 22, inertia 231440.27352263755.
Iteration 23, inertia 231329.8621921694.
Iteration 24, inertia 231233.96925918842.
Iteration 25, inertia 231149.11058980422.
Iteration 26, inertia 231060.80438237533.
Iteration 27, inertia 230957.5725262181.
Iteration 28, inertia 230843.24053270102.
Iteration 29, inertia 230706.6610019068.
Iteration 30, inertia 230557.3282361793.
Iteration 31, inertia 230386.4699355548.
Iteration 32, inertia 230182.99298105316.
Iteration 33, inertia 229965.6397958066.
Iteration 34, inertia 229717.7644488147.
Iteration 35, inertia 229400.83245536772.
Iteration 36, inertia 229061.4753523522.
Iteration 37, inertia 228742.72318585851.
Iteration 38, inertia 228496.4078533277.
Iteration 39, inertia 228285.13621035338.
Iteration 40, inertia 228073.44640057112.
Iteration 41, inertia 227903.44303359278.
Iteration 42, inertia 227740.937704852.
Iteration 43, inertia 227588.0130386447.
Iteration 44, inertia 227464.39357230708.
Iteration 45, inertia 227347.4021994123.
Iteration 46, inertia 227246.52237139564.
Iteration 47, inertia 227166.11098499535.
Iteration 48, inertia 227103.92385995964.
```

```
Iteration 49, inertia 227062.3169324577.
Iteration 50, inertia 227031.87469314015.
Iteration 51, inertia 227008.79527701612.
Iteration 52, inertia 226986.8853374675.
Iteration 53, inertia 226966.59276210854.
Iteration 54, inertia 226949.4363309555.
Iteration 55, inertia 226930.48005906685.
Iteration 56, inertia 226910.97957598077.
Iteration 57, inertia 226889.20163921162.
Iteration 58, inertia 226872.68961624667.
Iteration 59, inertia 226860.0480183947.
Converged at iteration 59: center shift 0.0007602361645065378 within tolerance 0.0011096
20681239686.
Initialization complete
Iteration 0, inertia 275353.4008525042.
Iteration 1, inertia 242700.72420518534.
Iteration 2, inertia 235821.64759556836.
Iteration 3, inertia 233129.1431338867.
Iteration 4, inertia 231518.6029347621.
Iteration 5, inertia 230190.76899879036.
Iteration 6, inertia 228995.3435161024.
Iteration 7, inertia 227935.43304024122.
Iteration 8, inertia 227028.16203124533.
Iteration 9, inertia 226342.57585596805.
Iteration 10, inertia 225794.41670205616.
Iteration 11, inertia 225365.12145711828.
Iteration 12, inertia 225046.42541157448.
Iteration 13, inertia 224772.99512659424.
Iteration 14, inertia 224482.69155699055.
Iteration 15, inertia 224198.7315074647.
Iteration 16, inertia 223946.35498308088.
Iteration 17, inertia 223733.1437970465.
Iteration 18, inertia 223535.97225444953.
Iteration 19, inertia 223327.62640575858.
Iteration 20, inertia 223125.92893672062.
Iteration 21, inertia 222934.48627785733.
Iteration 22, inertia 222744.76502520792.
Iteration 23, inertia 222599.373553692.
Iteration 24, inertia 222485.63281256263.
Iteration 25, inertia 222386.5856065099.
Iteration 26, inertia 222307.88682128763.
Iteration 27, inertia 222247.74570223977.
Iteration 28, inertia 222194.7123441275.
Iteration 29, inertia 222147.0013650438.
Iteration 30, inertia 222108.45192726774.
Iteration 31, inertia 222077.4551027421.
Iteration 32, inertia 222051.11589710816.
Iteration 33, inertia 222030.62096412838.
Iteration 34, inertia 222012.8568634145.
Iteration 35, inertia 221997.0795867805.
Converged at iteration 35: center shift 0.0010744906990255633 within tolerance 0.0011096
20681239686.
Initialization complete
Iteration 0, inertia 282032.04174362.
Iteration 1, inertia 246755.50583586228.
Iteration 2, inertia 238249.68899174978.
Iteration 3, inertia 234276.06787668235.
Iteration 4, inertia 231934.20777476823.
Iteration 5, inertia 230279.2219288173.
Iteration 6, inertia 228913.0060103702.
Iteration 7, inertia 227669.3648332685.
Iteration 8, inertia 226510.1579791927.
Iteration 9, inertia 225429.91338017664.
Iteration 10, inertia 224540.80507634315.
Iteration 11, inertia 223875.17425735272.
Iteration 12, inertia 223366.27441663164.
```

```
Iteration 13, inertia 222995.94141853374.
Iteration 14, inertia 222733.98058332008.
Iteration 15, inertia 222553.39202183124.
Iteration 16, inertia 222424.78325584886.
Iteration 17, inertia 222324.61040466087.
Iteration 18, inertia 222256.63656499234.
Iteration 19, inertia 222195.09877276162.
Iteration 20, inertia 222144.09439447834.
Iteration 21, inertia 222106.96610077884.
Iteration 22, inertia 222075.30575895094.
Iteration 23, inertia 222042.35589345003.
Iteration 24, inertia 222013.3615902018.
Iteration 25, inertia 221985.55682718655.
Iteration 26, inertia 221962.14553186027.
Iteration 27, inertia 221939.25871887352.
Iteration 28, inertia 221920.93231725134.
Converged at iteration 28: center shift 0.001074246625003622 within tolerance 0.00110962
0681239686.
Initialization complete
Iteration 0, inertia 285314.88062957756.
Iteration 1, inertia 261244.62000140088.
Iteration 2, inertia 253613.28265372053.
Iteration 3, inertia 247959.0561428228.
Iteration 4, inertia 243256.91354812402.
Iteration 5, inertia 239056.503856059.
Iteration 6, inertia 234903.97984247303.
Iteration 7, inertia 231123.9203089105.
Iteration 8, inertia 228086.11083369036.
Iteration 9, inertia 226079.3320763181.
Iteration 10, inertia 224745.47222731763.
Iteration 11, inertia 223882.59423927046.
Iteration 12, inertia 223364.52916650442.
Iteration 13, inertia 223039.8774473265.
Iteration 14, inertia 222840.4643050297.
Iteration 15, inertia 222700.33149818762.
Iteration 16, inertia 222598.33208288418.
Iteration 17, inertia 222520.49163897143.
Iteration 18, inertia 222453.57398403017.
Iteration 19, inertia 222399.75050490032.
Iteration 20, inertia 222356.77400651283.
Iteration 21, inertia 222327.22452248968.
Iteration 22, inertia 222303.87486724154.
Iteration 23, inertia 222284.76966071065.
Iteration 24, inertia 222267.7113520088.
Iteration 25, inertia 222252.61066138756.
Converged at iteration 25: center shift 0.0009541826425857803 within tolerance 0.0011096
20681239686.
Initialization complete
Iteration 0, inertia 276464.4965376644.
Iteration 1, inertia 242133.83195864822.
Iteration 2, inertia 233414.7445754503.
Iteration 3, inertia 229578.43169491517.
Iteration 4, inertia 227424.84687501652.
Iteration 5, inertia 226117.5706155045.
Iteration 6, inertia 225160.0712985771.
Iteration 7, inertia 224514.5031292184.
Iteration 8, inertia 224023.3737581545.
Iteration 9, inertia 223637.88862299547.
Iteration 10, inertia 223324.96837425022.
Iteration 11, inertia 223084.42456931097.
Iteration 12, inertia 222918.13933677252.
Iteration 13, inertia 222787.2763237582.
Iteration 14, inertia 222669.84988968074.
Iteration 15, inertia 222583.07463050418.
Iteration 16, inertia 222502.62291266778.
Iteration 17, inertia 222439.9119272577.
```

```
Iteration 18, inertia 222395.56835222163.
Iteration 19, inertia 222359.1925206547.
Iteration 20, inertia 222329.75860406033.
Iteration 21, inertia 222304.83181376837.
Iteration 22, inertia 222283.47420562422.
Iteration 23, inertia 222266.6483583643.
Iteration 24, inertia 222251.72714144803.
Converged at iteration 24: center shift 0.0008590755035894423 within tolerance 0.0011096
20681239686.
Initialization complete
Iteration 0, inertia 275499.89810016274.
Iteration 1, inertia 247273.13349837184.
Iteration 2, inertia 239813.35684193793.
Iteration 3, inertia 235721.74235532485.
Iteration 4, inertia 233378.90349635677.
Iteration 5, inertia 231909.36055333653.
Iteration 6, inertia 230803.6825684673.
Iteration 7, inertia 229933.68986272524.
Iteration 8, inertia 229274.11009522507.
Iteration 9, inertia 228765.9987899828.
Iteration 10, inertia 228370.14588590089.
Iteration 11, inertia 228055.3245388146.
Iteration 12, inertia 227817.52155724814.
Iteration 13, inertia 227623.35554358948.
Iteration 14, inertia 227485.56063206308.
Iteration 15, inertia 227367.9577462893.
Iteration 16, inertia 227259.2527475321.
Iteration 17, inertia 227140.47013697645.
Iteration 18, inertia 227032.98070218274.
Iteration 19, inertia 226941.22926890102.
Iteration 20, inertia 226873.76505996924.
Iteration 21, inertia 226822.87414229647.
Iteration 22, inertia 226771.85301878638.
Iteration 23, inertia 226724.8665524839.
Iteration 24, inertia 226682.22356497266.
Iteration 25, inertia 226638.82958589462.
Iteration 26, inertia 226586.50605465882.
Iteration 27, inertia 226540.2017920015.
Iteration 28, inertia 226493.90236979854.
Iteration 29, inertia 226457.28467426545.
Iteration 30, inertia 226421.07325643385.
Iteration 31, inertia 226388.55571504743.
Iteration 32, inertia 226359.33331828707.
Iteration 33, inertia 226334.770921559.
Iteration 34, inertia 226313.58800332053.
Iteration 35, inertia 226294.649903986.
Iteration 36, inertia 226277.3097069387.
Iteration 37, inertia 226260.74608462217.
Converged at iteration 37: center shift 0.00075926756493659 within tolerance 0.001109620
681239686.
Initialization complete
Iteration 0, inertia 291262.4351182924.
Iteration 1, inertia 240878.9360684432.
Iteration 2, inertia 236595.94605363254.
Iteration 3, inertia 235736.95492493556.
Iteration 4, inertia 235168.22918780235.
Iteration 5, inertia 234672.45396134024.
Iteration 6, inertia 234275.12112918665.
Iteration 7, inertia 233915.2889582006.
Iteration 8, inertia 233566.32589594834.
Iteration 9, inertia 233189.4447842192.
Iteration 10, inertia 232691.26393393663.
Iteration 11, inertia 232048.512387882.
Iteration 12, inertia 231190.21620664463.
Iteration 13, inertia 230198.67607438946.
Iteration 14, inertia 229249.11694558573.
```

```
Iteration 15, inertia 228414.67675617227.
Iteration 16, inertia 227702.78605393207.
Iteration 17, inertia 227139.8608983202.
Iteration 18, inertia 226705.66050587117.
Iteration 19, inertia 226347.0619036242.
Iteration 20, inertia 226039.0748813324.
Iteration 21, inertia 225803.87047118566.
Iteration 22, inertia 225605.88287951364.
Iteration 23, inertia 225438.00352063242.
Iteration 24, inertia 225293.08232170445.
Iteration 25, inertia 225160.15707345452.
Iteration 26, inertia 225060.64175863544.
Iteration 27, inertia 224957.67920252838.
Iteration 28, inertia 224862.6260536416.
Iteration 29, inertia 224781.77627378883.
Iteration 30, inertia 224706.6112080909.
Iteration 31, inertia 224646.34072261304.
Iteration 32, inertia 224592.0487578676.
Iteration 33, inertia 224541.65306074853.
Iteration 34, inertia 224492.63243758102.
Iteration 35, inertia 224449.02047252405.
Iteration 36, inertia 224404.66874362028.
Iteration 37, inertia 224371.6628979779.
Iteration 38, inertia 224343.69339810897.
Iteration 39, inertia 224313.10221110316.
Iteration 40, inertia 224287.5374437714.
Iteration 41, inertia 224266.27064433065.
Iteration 42, inertia 224246.65315357046.
Iteration 43, inertia 224230.0959026395.
Converged at iteration 43: center shift 0.0010239215693658836 within tolerance 0.0011096
20681239686.
Initialization complete
Iteration 0, inertia 273489.5608450799.
Iteration 1, inertia 246040.44618487346.
Iteration 2, inertia 242641.33387317767.
Iteration 3, inertia 240556.86161579227.
Iteration 4, inertia 238924.47071337822.
Iteration 5, inertia 237499.77062027936.
Iteration 6, inertia 236023.86637174076.
Iteration 7, inertia 234492.28750777844.
Iteration 8, inertia 232931.3763144013.
Iteration 9, inertia 231644.68567540206.
Iteration 10, inertia 230619.1733758723.
Iteration 11, inertia 229918.46499955817.
Iteration 12, inertia 229419.04151484848.
Iteration 13, inertia 229054.57470396516.
Iteration 14, inertia 228787.99349974593.
Iteration 15, inertia 228594.41392149398.
Iteration 16, inertia 228424.22750973102.
Iteration 17, inertia 228270.95588301748.
Iteration 18, inertia 228120.12476482638.
Iteration 19, inertia 227974.73344257107.
Iteration 20, inertia 227822.36211298223.
Iteration 21, inertia 227677.3723217461.
Iteration 22, inertia 227534.4357186797.
Iteration 23, inertia 227396.8137377745.
Iteration 24, inertia 227266.38413869226.
Iteration 25, inertia 227143.47312942005.
Iteration 26, inertia 227029.51385928434.
Iteration 27, inertia 226935.14572583017.
Iteration 28, inertia 226856.22923093097.
Iteration 29, inertia 226778.78418920806.
Iteration 30, inertia 226699.62999485323.
Iteration 31, inertia 226607.1378305605.
Iteration 32, inertia 226469.69024028757.
Iteration 33, inertia 226308.9009748737.
```

```
Iteration 34, inertia 226149.15471132414.
Iteration 35, inertia 226014.651650182.
Iteration 36, inertia 225884.11363736683.
Iteration 37, inertia 225771.6811703417.
Iteration 38, inertia 225661.72333367108.
Iteration 39, inertia 225562.84623562702.
Iteration 40, inertia 225462.27683053107.
Iteration 41, inertia 225376.0071252339.
Iteration 42, inertia 225293.2053358277.
Iteration 43, inertia 225217.41199123664.
Iteration 44, inertia 225154.30659375555.
Iteration 45, inertia 225086.5618215329.
Iteration 46, inertia 225009.52516488015.
Iteration 47, inertia 224927.01403183545.
Iteration 48, inertia 224845.80925958985.
Iteration 49, inertia 224780.41827855533.
Iteration 50, inertia 224719.354450859.
Iteration 51, inertia 224661.04863590194.
Iteration 52, inertia 224599.97689242722.
Iteration 53, inertia 224538.3924157433.
Iteration 54, inertia 224479.49321292152.
Iteration 55, inertia 224413.04785481296.
Iteration 56, inertia 224350.11342037408.
Iteration 57, inertia 224283.31455084233.
Iteration 58, inertia 224205.9880681001.
Iteration 59, inertia 224126.30649070107.
Iteration 60, inertia 224032.952305192.
Iteration 61, inertia 223940.76063750847.
Iteration 62, inertia 223837.82445831492.
Iteration 63, inertia 223703.37589012377.
Iteration 64, inertia 223544.46436950203.
Iteration 65, inertia 223360.98506217636.
Iteration 66, inertia 223169.55667599285.
Iteration 67, inertia 222957.78090443817.
Iteration 68, inertia 222762.3927018984.
Iteration 69, inertia 222589.2430206479.
Iteration 70, inertia 222452.33166032992.
Iteration 71, inertia 222342.91999695468.
Iteration 72, inertia 222251.2588338449.
Iteration 73, inertia 222171.48514979766.
Iteration 74, inertia 222112.96154471298.
Iteration 75, inertia 222079.70061635264.
Iteration 76, inertia 222049.95402041043.
Iteration 77, inertia 222027.6882640351.
Iteration 78, inertia 222012.36929361487.
Converged at iteration 78: center shift 0.0009943101763602489 within tolerance 0.0011096
20681239686.
Initialization complete
Iteration 0, inertia 271024.36810805224.
Iteration 1, inertia 242695.48925077048.
Iteration 2, inertia 236022.3002764128.
Iteration 3, inertia 232281.44288314966.
Iteration 4, inertia 229910.81621154814.
Iteration 5, inertia 228287.06980002593.
Iteration 6, inertia 227065.67657225285.
Iteration 7, inertia 226062.98491058985.
Iteration 8, inertia 225363.6099280888.
Iteration 9, inertia 224858.0244551415.
Iteration 10, inertia 224468.85374012685.
Iteration 11, inertia 224147.2880140246.
Iteration 12, inertia 223883.12839703856.
Iteration 13, inertia 223693.39110404789.
Iteration 14, inertia 223517.5616063996.
Iteration 15, inertia 223359.96344214323.
Iteration 16, inertia 223207.22067752312.
Iteration 17, inertia 223073.83339688357.
```

```
Iteration 18, inertia 222941.60753246702.
Iteration 19, inertia 222799.76419409047.
Iteration 20, inertia 222658.03529729604.
Iteration 21, inertia 222530.79729065587.
Iteration 22, inertia 222411.38615694625.
Iteration 23, inertia 222302.07748015196.
Iteration 24, inertia 222216.5695180272.
Iteration 25, inertia 222154.93990478478.
Iteration 26, inertia 222107.00123285994.
Iteration 27, inertia 222076.1241175065.
Iteration 28, inertia 222049.5182923278.
Iteration 29, inertia 222024.156248965.
Iteration 30, inertia 222005.93245046263.
Converged at iteration 30: center shift 0.0010381627528728023 within tolerance 0.0011096
20681239686.
```

In [33]:
```
dataset_predict['prediction_PCA_ten'] = clustered_PCA_ten
dataset_predict.head(10)
```

Out[33]:

	text	label	prediction_two	prediction_PCA_two	corpus	prediction_PCA_ten
0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	1	1	1	law enforcement high alert following threat co...	7
2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	1	0	0	unbelievable obama attorney general say charlo...	8
3	Bobby Jindal, raised Hindu, uses story of Chri...	0	1	1	bobby jindal raised hindu us story christian c...	2
4	SATAN 2: Russia unvelis an image of its terrif...	1	0	0	satan 2 russia unvelis image terrifying new su...	8
5	About Time! Christian Group Sues Amazon and SP...	1	0	0	time christian group sue amazon splc designati...	1
6	DR BEN CARSON TARGETED BY THE IRS: "I never ha...	1	0	0	dr ben carson targeted irs never audit spoke n...	4
7	HOUSE INTEL CHAIR On Trump-Russia Fake Story: ...	1	0	0	house intel chair trump russia fake story evid...	4
8	Sports Bar Owner Bans NFL Games...Will Show Only...	1	1	1	sport bar owner ban nfl game show true america...	7
9	Latest Pipeline Leak Underscores Dangers Of Da...	1	0	0	latest pipeline leak underscore danger dakota ...	6
10	GOP Senator Just Smacked Down The Most Puncha...	1	1	1	gop senator smacked punchable alt right nazi i...	9

```
In [34]:  PCA_df = pd.DataFrame(pca_result)
          PCA_df['cluster'] = clustered_PCA_ten
          PCA_df.columns = ['x1','x2','cluster']
          k_means_figure = sns.scatterplot(data=PCA_df,x='x1',y='x2',hue='cluster',legend="full",a
```
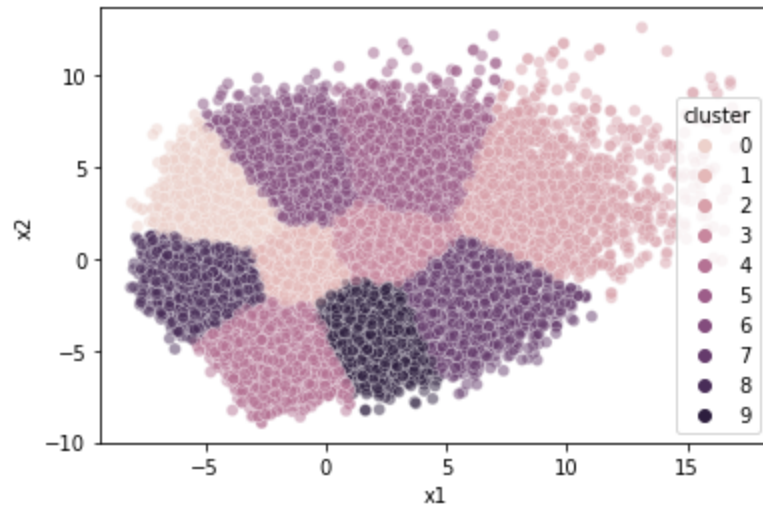


```
In [35]:  zero_sum = [0] * 10

          for index, row in dataset_predict.loc[dataset_predict['label'] == 0].iterrows():
              if row['prediction_PCA_ten'] == 0:
                  zero_sum[0] += 1
              elif row['prediction_PCA_ten'] == 1:
                  zero_sum[1] += 1
              elif row['prediction_PCA_ten'] == 2:
                  zero_sum[2] += 1
              elif row['prediction_PCA_ten'] == 3:
                  zero_sum[3] += 1
              elif row['prediction_PCA_ten'] == 4:
                  zero_sum[4] += 1
              elif row['prediction_PCA_ten'] == 5:
                  zero_sum[5] += 1
              elif row['prediction_PCA_ten'] == 6:
                  zero_sum[6] += 1
              elif row['prediction_PCA_ten'] == 7:
                  zero_sum[7] += 1
              elif row['prediction_PCA_ten'] == 8:
                  zero_sum[8] += 1
              elif row['prediction_PCA_ten'] == 9:
                  zero_sum[9] += 1

          print(zero_sum)
```

[7901, 5122, 1340, 3014, 1412, 2161, 5795, 649, 6476, 1158]

```
In [36]:  one_sum = [0] * 10

          for index, row in dataset_predict.loc[dataset_predict['label'] == 1].iterrows():
              if row['prediction_PCA_ten'] == 0:
                  one_sum[0] += 1
              elif row['prediction_PCA_ten'] == 1:
                  one_sum[1] += 1
              elif row['prediction_PCA_ten'] == 2:
                  one_sum[2] += 1
              elif row['prediction_PCA_ten'] == 3:
                  one_sum[3] += 1
              elif row['prediction_PCA_ten'] == 4:
                  one_sum[4] += 1
              elif row['prediction_PCA_ten'] == 5:
                  one_sum[5] += 1
```

```python
        elif row['prediction_PCA_ten'] == 6:
            one_sum[6] += 1
        elif row['prediction_PCA_ten'] == 7:
            one_sum[7] += 1
        elif row['prediction_PCA_ten'] == 8:
            one_sum[8] += 1
        elif row['prediction_PCA_ten'] == 9:
            one_sum[9] += 1

print(one_sum)
```

```
[928, 5562, 1410, 5347, 5745, 1672, 1569, 5025, 907, 8344]
```

In [37]:
```python
total_sum = [sum(value) for value in zip(zero_sum, one_sum)]
percent_fake = []
for i in range(10):
    percent_fake.append(str(zero_sum[i] / total_sum[i] * 100) + "%")
column_name = ['Fake(0)', 'Real(1)']
data = {'Fake(0)': zero_sum, 'Real(1)': one_sum, 'Total': total_sum, 'Percent Fake': per
df = pd.DataFrame(data)
df.head(10)
```

Out[37]:

	Fake(0)	Real(1)	Total	Percent Fake
0	7901	928	8829	89.48918337297542%
1	5122	5562	10684	47.9408461250468%
2	1340	1410	2750	48.72727272727273%
3	3014	5347	8361	36.04831957899773%
4	1412	5745	7157	19.72893670532346%
5	2161	1672	3833	56.378815549178185%
6	5795	1569	7364	78.69364475828354%
7	649	5025	5674	11.438138879097638%
8	6476	907	7383	87.7150209941758%
9	1158	8344	9502	12.186908019364344%

In [38]:
```python
dataset_predict['corpus'] = corpus
cluster_0 = []
cluster_1 = []
cluster_2 = []
cluster_3 = []
cluster_4 = []
cluster_5 = []
cluster_6 = []
cluster_7 = []
cluster_8 = []
cluster_9 = []
for i in range(0,len(dataset_predict)):
    corpora = dataset_predict['corpus'].iloc[i].split()
    if dataset_predict['prediction_PCA_ten'].iloc[i] == 0:
        for x in corpora:
            cluster_0.append(x)
    elif dataset_predict['prediction_PCA_ten'].iloc[i] == 1:
        for x in corpora:
            cluster_1.append(x)
    elif dataset_predict['prediction_PCA_ten'].iloc[i] == 2:
        for x in corpora:
            cluster_2.append(x)
    elif dataset_predict['prediction_PCA_ten'].iloc[i] == 3:
```

```
            for x in corpora:
                cluster_3.append(x)
        elif dataset_predict['prediction_PCA_ten'].iloc[i] == 4:
            for x in corpora:
                cluster_4.append(x)
        elif dataset_predict['prediction_PCA_ten'].iloc[i] == 5:
            for x in corpora:
                cluster_5.append(x)
        elif dataset_predict['prediction_PCA_ten'].iloc[i] == 6:
            for x in corpora:
                cluster_6.append(x)
        elif dataset_predict['prediction_PCA_ten'].iloc[i] == 7:
            for x in corpora:
                cluster_7.append(x)
        elif dataset_predict['prediction_PCA_ten'].iloc[i] == 8:
            for x in corpora:
                cluster_8.append(x)
        elif dataset_predict['prediction_PCA_ten'].iloc[i] == 9:
            for x in corpora:
                cluster_9.append(x)
```

```
In [39]: clusters_all = [cluster_0, cluster_1, cluster_2, cluster_3, cluster_4, cluster_5, cluste
         titles = ['Category0','Category1','Category2','Category3','Category4','Category5','Categ
         index = 0
         for cluster_inst in clusters_all:
             most_freq_start, most_freq_mid = most_frequent(cluster_inst)
             print(titles[index] + ' Articles: ')
             index += 1
             print('Top 1% Common Words: ' + str(most_freq_start))
             print('Top 10% Common Words: ' + str(most_freq_mid))
             print()
```

```
Category0 Articles:
Top 1% Common Words: ['state', 'trump', 'would', 'president', 'government', 'reuters',
'united', 'official', 'country', 'security']
Top 10% Common Words: ['adding', 'proposed', 'using', 'passed', 'failed', 'ahead', 'pari
s', 'activity', 'brexit', 'potential', 'lebanon']

Category1 Articles:
Top 1% Common Words: ['trump', 'clinton', 'state', 'president', 'republican', 'would',
'obama', 'campaign', 'people', 'house', 'election']
Top 10% Common Words: ['stand', 'church', 'civil', 'calling', 'thousand', 'front', 'doll
ar', 'nearly', 'current', 'college']

Category2 Articles:
Top 1% Common Words: ['people', 'trump', 'would', 'world', 'american', 'state', 'thing',
'could', 'think', 'right']
Top 10% Common Words: ['feeling', 'california', 'figure', 'heard', 'space', 'current',
'front', 'church', 'couple', 'financial', 'following']

Category3 Articles:
Top 1% Common Words: ['trump', 'clinton', 'would', 'state', 'people', 'president', 'repu
blican', 'american', 'obama', 'campaign']
Top 10% Common Words: ['carolina', 'rubio', 'agency', 'middle', 'recently', 'individua
l', 'especially', 'looking', 'probably', 'wanted', 'attorney']

Category4 Articles:
Top 1% Common Words: ['trump', 'clinton', 'twitter', 'hillary', 'president', 'video', 'd
onald', 'email', 'campaign', 'obama']
Top 10% Common Words: ['msnbc', 'without', 'crime', 'governor', 'effort', 'fraud', '21wi
re', 'racist', 'stand', 'website', 'fight']

Category5 Articles:
Top 1% Common Words: ['state', 'people', 'would', 'trump', 'government', 'american', 'pr
esident', 'obama', 'country', 'could']
```

```
Top 10% Common Words: ['access', 'freedom', 'civil', 'resident', 'although', 'region',
'following', 'value', 'similar', 'george', 'peace']

Category6 Articles:
Top 1% Common Words: ['state', 'would', 'trump', 'government', 'president', 'people', 'c
ountry', 'united', 'official', 'could']
Top 10% Common Words: ['island', 'terrorism', 'third', 'account', 'afghanistan', 'analys
t', 'situation', 'protection', 'response', 'barack']

Category7 Articles:
Top 1% Common Words: ['trump', 'people', 'clinton', 'would', 'donald', 'president', 'hil
lary', 'republican', 'right', 'think']
Top 10% Common Words: ['attention', 'whatever', 'bring', 'breitbart', 'remark', 'followi
ng', 'piece', 'voice', 'heart', 'company', 'nomination']

Category8 Articles:
Top 1% Common Words: ['trump', 'president', 'reuters', 'state', 'house', 'north', 'kore
a', 'would', 'republican', 'russia', 'russian']
Top 10% Common Words: ['issued', 'bureau', 'however', 'confirmation', 'florida', 'associ
ate', 'refugee', 'ukraine', 'family', 'hariri']

Category9 Articles:
Top 1% Common Words: ['trump', 'clinton', 'president', 'hillary', 'donald', 'people', 'r
epublican', 'twitter', 'would', 'video', 'campaign']
Top 10% Common Words: ['speaking', 'journalist', 'major', 'among', 'abortion', 'others',
'continue', 'telling', 'reality', 'across']
```
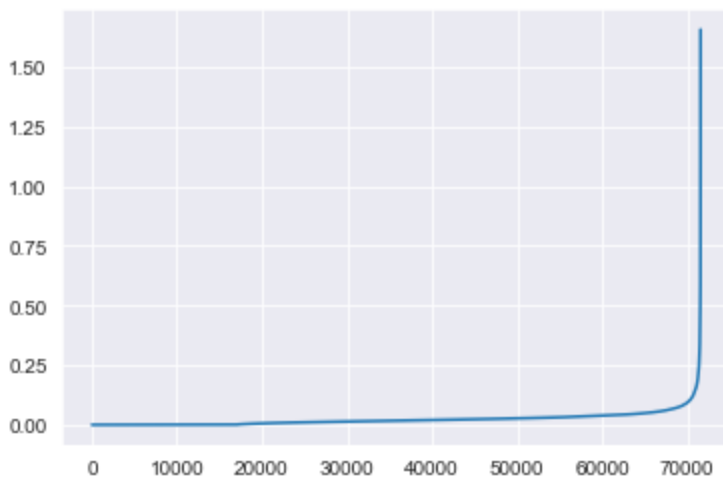
In [311…
```python
neigh = NearestNeighbors(n_neighbors=2)
nbrs = neigh.fit(pca_result)
distances, indices = nbrs.kneighbors(pca_result)
```

In [312…
```python
distances = np.sort(distances, axis=0)
distances = distances[:,1]
plt.plot(distances)
```

Out[312]: `[<matplotlib.lines.Line2D at 0x18330928d90>]`

In [332…
```python
dbscan_PCA = cluster.DBSCAN(eps=0.2, min_samples=5, )
dbscan_PCA.fit(pca_result)
clusters = dbscan_PCA.labels_
print(len(clusters))
```

```
71537
```

In [333…
```python
colors = ['royalblue', 'maroon', 'forestgreen', 'mediumorchid', 'tan', 'deeppink', 'oliv
vectorizer = np.vectorize(lambda x: colors[x % len(colors)])
```

In [337…
```python
dataset_predict['prediction_dbscan_pca'] = clusters
```

```
dataset_predict.head(100)
```

Out[337]:

	text	label	prediction_two	prediction_PCA_two	prediction_PCA_ten	corpus	prediction_dbsca
0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	1	1	1	9	law enforcement high alert following threat co...	
2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	1	0	0	6	unbelievable obama attorney general say charlo...	
3	Bobby Jindal, raised Hindu, uses story of Chri...	0	1	1	5	bobby jindal raised hindu us story christian c...	
4	SATAN 2: Russia unvelis an image of its terrif...	1	0	0	6	satan 2 russia unvelis image terrifying new su...	
5	About Time! Christian Group Sues Amazon and SP...	1	0	0	8	time christian group sue amazon splc designati...	
...	...	...	...	...	...	...	
97	Boiler Room EP #124 – Weather Warfare & CNN Go...	1	0	1	2	boiler room ep 124 weather warfare cnn goblin ...	
98	COLLEGE REPUBLICANS PRESIDENT Attacked by Anti...	1	1	1	2	college republican president attacked antifa l...	
99	BEYONCE DOUBLES DOWN... Debuts #LEMONADE, Another...	1	1	1	7	beyonce double debut lemonade another race bai...	
100	One person shot in Portland as anti-Trump prot...	0	0	0	0	one person shot portland anti trump protester ...	
101	For Helping Immigrants, Chobani's Founder Draw...	0	0	0	3	helping immigrant chobani founder draw threat ...	

100 rows × 7 columns

```python
def dbscan_grid_search(X_data, lst, clst_count, eps_space = 0.5,
                       min_samples_space = 5, min_clust = 0, max_clust = 10):

    # Importing counter to count the amount of data in each cluster
    from collections import Counter


    # Starting a tally of total iterations
    n_iterations = 0


    # Looping over each combination of hyperparameters
    for eps_val in eps_space:
        for samples_val in min_samples_space:

            dbscan_grid = cluster.DBSCAN(eps = eps_val,
                                         min_samples = samples_val)


            # fit_transform
            clusters = dbscan_grid.fit_predict(X = X_data)


            # Counting the amount of data in each cluster
            cluster_count = Counter(clusters)


            # Saving the number of clusters
            n_clusters = sum(abs(pd.np.unique(clusters))) - 1


            # Increasing the iteration tally with each run of the loop
            n_iterations += 1


            # Appending the lst each time n_clusters criteria is reached
            if n_clusters >= min_clust and n_clusters <= max_clust:

                lst.append([eps_val,
                            samples_val,
                            n_clusters])


            clst_count.append(cluster_count)

    # Printing grid search summary information
    print(f"""Search Complete. \nYour list is now of length {len(lst)}. """)
    print(f"""Hyperparameter combinations checked: {n_iterations}. \n""")
```

```python
dbscan_clusters = []
cluster_count = []

dbscan_grid_search(X_data = pca_result,
                   lst = dbscan_clusters,
                   clst_count = cluster_count,
                   eps_space = pd.np.arange(0.01, 10, 1),
                   min_samples_space = pd.np.arange(1, 20, 5),
                   min_clust = 0,
                   max_clust = 10)
```

```
MemoryError                               Traceback (most recent call last)
c:\Users\Liam's Computer\Documents\projects\IBM_Course_Projects\unsupervised ml\proj3.ip
ynb Cell 44' in <cell line: 4>()
      <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/I
BM_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000065?line=0'>1</a> dbscan_clusters
 = []
      <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/I
BM_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000065?line=1'>2</a> cluster_count =
 []
----> <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/I
BM_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000065?line=3'>4</a> dbscan_grid_sea
rch(X_data = pca_result,
      <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/I
BM_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000065?line=4'>5</a>
    lst = dbscan_clusters,
      <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/I
BM_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000065?line=5'>6</a>
    clst_count = cluster_count,
      <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/I
BM_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000065?line=6'>7</a>
    eps_space = pd.np.arange(0.01, 10, 1),
      <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/I
BM_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000065?line=7'>8</a>
    min_samples_space = pd.np.arange(1, 20, 5),
      <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/I
BM_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000065?line=8'>9</a>
    min_clust = 0,
     <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/IB
M_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000065?line=9'>10</a>
    max_clust = 10)

c:\Users\Liam's Computer\Documents\projects\IBM_Course_Projects\unsupervised ml\proj3.ip
ynb Cell 43' in dbscan_grid_search(X_data, lst, clst_count, eps_space, min_samples_spac
e, min_clust, max_clust)
     <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/IB
M_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000066?line=16'>17</a> dbscan_grid =
 cluster.DBSCAN(eps = eps_val,
     <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/IB
M_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000066?line=17'>18</a>
      min_samples = samples_val)
     <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/IB
M_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000066?line=20'>21</a> # fit_transfor
m
---> <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/IB
M_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000066?line=21'>22</a> clusters = dbs
can_grid.fit_predict(X = X_data)
     <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/IB
M_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000066?line=24'>25</a> # Counting the
 amount of data in each cluster
     <a href='vscode-notebook-cell:/c%3A/Users/Liam%27s%20Computer/Documents/projects/IB
M_Course_Projects/unsupervised%20ml/proj3.ipynb#ch0000066?line=25'>26</a> cluster_count
 = Counter(clusters)

File c:\Users\Liam's Computer\scoop\apps\python\current\lib\site-packages\sklearn\cluste
r\_dbscan.py:458, in DBSCAN.fit_predict(self, X, y, sample_weight)
    433 def fit_predict(self, X, y=None, sample_weight=None):
    434     """Compute clusters from a data or distance matrix and predict labels.
    435
    436     Parameters
  (...)
    456         Cluster labels. Noisy samples are given the label -1.
    457     """
--> 458     self.fit(X, sample_weight=sample_weight)
    459     return self.labels_

File c:\Users\Liam's Computer\scoop\apps\python\current\lib\site-packages\sklearn\cluste
```

```
r\_dbscan.py:420, in DBSCAN.fit(self, X, y, sample_weight)
    418 # A list of all core samples found.
    419 core_samples = np.asarray(n_neighbors >= self.min_samples, dtype=np.uint8)
--> 420 dbscan_inner(core_samples, neighborhoods, labels)
    422 self.core_sample_indices_ = np.where(core_samples)[0]
    423 self.labels_ = labels

File sklearn\cluster\_dbscan_inner.pyx:36, in sklearn.cluster._dbscan_inner.dbscan_inner
()

MemoryError: bad allocation
```

In [ ]: