In [2]:
```
#analyze a collection of tweets
import pandas as pd
from nltk import word_tokenize, ngrams
from collections import Counter
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /home/adel/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

Out[2]:
```
True
```

In [3]:
```
#Import data
data = pd.read_csv('files/tweets.csv')
data.head()
```

Out[3]:

	date	content
0	2009-05-04 13:54:25	Be sure to tune in and watch John Doe on Late ...
1	2009-05-04 20:00:10	John Doe will be appearing on The View tomorro...
2	2009-05-08 08:38:08	John Doe reads Top Ten Financial Tips on Late ...
3	2009-05-08 15:40:15	New Blog Post: Celebrity Apprentice Finale and...
4	2009-05-12 09:07:28	"My persona will never be that of a wallflower...

In [4]:
```
#Convert content to a list of content
content = list(data['content'])
```

In [5]:
```
len(content)
```

Out[5]:
```
43352
```

In [6]:
```
#Create a corpus
corpus = []
for item in content:
  corpus.extend([word.lower() for word in word_tokenize(item) if any(c.isalpha() for c in word)])
```

In [7]:
```
# Check corpus
len(corpus)
```

Out[7]:
```
850502
```

In [8]:
```
corpus[:10]
```

Out[8]:
```
['be', 'sure', 'to', 'tune', 'in', 'and', 'watch', 'john', 'doe', 'on']
```

In [9]:
```
#Display all 3-grams
ngram = Counter(ngrams(corpus, 3))
```

In [10]:
```
ngram.most_common(10)
```

Out[10]:
```
[(('america', 'great', 'again'), 537),
 (('the', 'united', 'states'), 529),
 (('i', 'will', 'be'), 522),
 (('make', 'america', 'great'), 501),
 (('run', 'for', 'president'), 397),
 (('one', 'of', 'the'), 353),
 (('the', 'fake', 'news'), 347),
 (('the', 'white', 'house'), 288),
 (('all', 'of', 'the'), 280),
 (('thank', 'you', 'to'), 275)]
```

In [11]:
```
#Pretty print
for gram, freq in ngram.most_common(10):
  print(f'Frequency: {freq} -> {gram}')
```

```
Frequency: 537 -> ('america', 'great', 'again')
Frequency: 529 -> ('the', 'united', 'states')
Frequency: 522 -> ('i', 'will', 'be')
Frequency: 501 -> ('make', 'america', 'great')
Frequency: 397 -> ('run', 'for', 'president')
Frequency: 353 -> ('one', 'of', 'the')
Frequency: 347 -> ('the', 'fake', 'news')
Frequency: 288 -> ('the', 'white', 'house')
Frequency: 280 -> ('all', 'of', 'the')
Frequency: 275 -> ('thank', 'you', 'to')
```

In [12]:
```
#with 4-grams
ngram = Counter(ngrams(corpus, 4))
```

In [13]:
```
ngram.most_common(10)
```

Out[13]:
```
[(('make', 'america', 'great', 'again'), 489),
 (('the', 'great', 'state', 'of'), 173),
 (('the', 'fake', 'news', 'media'), 167),
 (('art', 'of', 'the', 'deal'), 160),
 (('of', 'the', 'united', 'states'), 141),
 (('the', 'art', 'of', 'the'), 137),
 (('in', 'the', 'history', 'of'), 131),
 (('my', 'complete', 'and', 'total'), 116),
 (('complete', 'and', 'total', 'endorsement'), 116),
 (('i', 'will', 'be', 'interviewed'), 113)]
```

In [14]:
```
ngram = Counter(ngrams(corpus, 5))
ngram.most_common(10)
```

Out[14]:
```
[(('the', 'art', 'of', 'the', 'deal'), 134),
 (('my', 'complete', 'and', 'total', 'endorsement'), 115),
 (('has', 'my', 'complete', 'and', 'total'), 106),
 (('it', 'was', 'my', 'great', 'honor'), 90),
 (('was', 'my', 'great', 'honor', 'to'), 87),
 (('to', 'make', 'america', 'great', 'again'), 82),
 (('i', 'will', 'be', 'interviewed', 'on'), 65),
 (('president', 'of', 'the', 'united', 'states'), 61),
 (('in', 'the', 'history', 'of', 'our'), 57),
 (('in', 'the', 'great', 'state', 'of'), 56)]
```