In [1]:

```
import pandas as pd
```

In [2]:

```
npr = pd.read_csv('npr.csv')
npr.head()
```

Out[2]:

	Article
0	In the Washington of 2016, even when the polic...
1	Donald Trump has used Twitter — his prefe...
2	Donald Trump is unabashedly praising Russian...
3	Updated at 2:50 p. m. ET, Russian President Vl...
4	From photography, illustration and video, to d...

In [3]:

```
npr.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 1 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   Article  200 non-null    object
dtypes: object(1)
memory usage: 1.7+ KB
```

Text Preprocessing

Text Cleaning

- **Remove Punctuation**
- **Remove Stopwords**
- **Stemming/Lemmatization**

In [4]:

```
import nltk
import re
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
wnl = WordNetLemmatizer()

corpus=[]
for i in range(len(npr)):
    rp = re.sub('[^a-zA-Z]'," ",npr['Article'][i])
    rp = rp.lower()
    rp = rp.split()
    rp = [wnl.lemmatize(word) for word in rp if not word in set(stopwords.words('english'))]
    rp = " ".join(rp)
    corpus.append(rp)

print(corpus)
```

```
['washington even policy bipartisan politics cannot sense year show little sign ending dec president obama move
d sanction russia alleged interference u election concluded republican long called similar severe measure could
scarcely bring approve house speaker paul ryan called obama measure appropriate also overdue prime example admi
nistration ineffective foreign policy left america weaker eye world gop leader sounded much theme urging presid
ent obama year take strong action deter russia worldwide aggression including operation wrote rep devin nunes c
hairman house intelligence committee week left office president suddenly decided stronger measure indeed warran
ted appearing cnn frequent obama critic trent frank called much tougher action said three time obama finally fo
und tongue meanwhile fox news various spokesman trump said obama real target russian man poised take white hous
e le three week spoke obama trying tie trump hand box meaning would forced either keep sanction odds republican
want tougher still moscow throughout trump repeatedly called sanction closer tie russia including cooperation f
ight isi russia battled isi syria behalf country embattled dictator bashar assad bombing besieged city aleppo f
ell assad force week campaign trump even urged russia find missing email private server opponent hillary clinto
n exchanged public encomium russian president vladimir putin several occasion added doubt current u level suppo
rt nato putin longtime nemesis also suggestion trump extensive business dealing various russian reason refuse r
elease tax return issue disquieting republican many month sen john mccain lindsay graham c prominent senior mem
ber armed service committee accepted assessment u intelligence agency regarding role russia hacking various dem
ocratic committee last year includes fbi cia consensus russian goal discredit american democracy defeat clinton
elect trump say great majority senate colleague agree mccain slated armed service hearing cyberthreats jan poli
ticizing russian action idea helped trump win also made issue difficult republican leader allowed trump support
```

Vectorization

In [5]:

```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
X = cv.fit_transform(corpus)
```

Modelling using LDA

In [6]:

```
from sklearn.decomposition import LatentDirichletAllocation

model = LatentDirichletAllocation(n_components=4)

model.fit(X)
```

Out[6]:

```
LatentDirichletAllocation(n_components=4)
```

In [7]:

```
topic_results = model.transform(X)
```

In [8]:

```
topic_results[0]
```

Out[8]:

```
array([3.80782888e-04, 3.56898170e-01, 3.84134965e-04, 6.42336912e-01])
```

In [9]:

```
topic_results[0].argmax()
```

Out[9]:

```
3
```

This means that our model thinks that the first article belongs to topic #2.

Combining with Original Data

In [10]:

```
npr['group'] = topic_results.argmax(axis=1)
```

In [11]:

```
npr.head()
```

Out[11]:

	Article	group
0	In the Washington of 2016, even when the polic...	3
1	Donald Trump has used Twitter — his prefe...	3
2	Donald Trump is unabashedly praising Russian...	3
3	Updated at 2:50 p. m. ET, Russian President Vl...	3
4	From photography, illustration and video, to d...	0

Showing Top Words Per Topic

In [12]:

```
for index,topic in enumerate(model.components_):
    print(f'THE TOP 10 WORDS FOR TOPIC #{index}')
    print([cv.get_feature_names()[i] for i in topic.argsort()[-10:]])
    print('\n')
```

```
THE TOP 10 WORDS FOR TOPIC #0

C:\Users\admin\anaconda3\lib\site-packages\sklearn\utils\deprecation.py:87: FutureWarning: Function get_feature_nam
es is deprecated; get_feature_names is deprecated in 1.0 and will be removed in 1.2. Please use get_feature_names_o
ut instead.
  warnings.warn(msg, category=FutureWarning)

['think', 'student', 'new', 'school', 'time', 'people', 'one', 'year', 'like', 'say']


THE TOP 10 WORDS FOR TOPIC #1
['say', 'obama', 'house', 'time', 'year', 'people', 'trump', 'one', 'republican', 'said']


THE TOP 10 WORDS FOR TOPIC #2
['could', 'new', 'time', 'said', 'would', 'one', 'people', 'like', 'year', 'say']


THE TOP 10 WORDS FOR TOPIC #3
['new', 'people', 'russian', 'intelligence', 'also', 'president', 'russia', 'said', 'say', 'trump']
```