In [1]:
```
import pandas as pd
```

In [2]:
```
# Read article from National Public Radio. -- filename 'npr.csv'
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.columns
```

Out[3]:
```
Index(['Article'], dtype='object')
```

In [4]:
```
len(npr['Article'])
```

Out[4]:
```
11992
```

Note: There are no labels for the articles

In [5]:
```
# To look at the first article

npr['Article'][0]
```

Out[5]:
```
'In the Washington of 2016, even when the policy can be bipartisan, th
e politics cannot. And in that sense, this year shows little sign of e
nding on Dec. 31. When President Obama moved to sanction Russia over i
ts alleged interference in the U. S. election just concluded, some Rep
ublicans who had long called for similar or more severe measures could
scarcely bring themselves to approve. House Speaker Paul Ryan called t
he Obama measures "appropriate" but also "overdue" and "a prime exampl
e of this administration's ineffective foreign policy that has left Am
erica weaker in the eyes of the world." Other GOP leaders sounded much
the same theme. "[We have] been urging President Obama for years to ta
ke strong action to deter Russia's worldwide aggression, including its
operations," wrote Rep. Devin Nunes,  . chairman of the House Intellig
ence Committee. "Now with just a few weeks left in office, the preside
nt has suddenly decided that some stronger measures are indeed warrant
```

ed." Appearing on CNN, frequent Obama critic Trent Franks, . called f or "much tougher" actions and said three times that Obama had "finally found his tongue." Meanwhile, at and on Fox News, various spokesmen for Trump said Obama's real target was not the Russians at all but t he man poised to take over the White House in less than three weeks. T hey spoke of Obama trying to "tie Trump's hands" or "box him in," mean ing the would be forced either to keep the sanctions or be at odds w ith Republicans who want to be tougher still on Moscow. Throughout 201 6, Trump has repeatedly called not for sanctions but for closer ties w ith Russia, including cooperation in the fight against ISIS. Russia ha s battled ISIS in Syria on behalf of that country's embattled dictator , Bashar Assad, bombing the besieged city of Aleppo that fell to Ass ad's forces this week. During the campaign, Trump even urged Russia to "find" missing emails from the private server of his opponent, Hillary Clinton. He has exchanged public encomiums with Russian President Vlad imir Putin on several occasions and added his doubts about the current U. S. levels of support for NATO — Putin's longtime nemesis. There have also been suggestions that Trump's extensive business dealings wi th various Russians are the reason he refuses to release his tax retur ns. All those issues have been disquieting to some Republicans for man y months. Sens. John McCain, . and Lindsay Graham, . C. prominent se nior members of the Armed Services Committee, have accepted the assess ment of 17 U. S. intelligence agencies regarding the role of Russia in the hacking of various Democratic committees last year. That includes the FBI and CIA consensus that the Russian goal was not just to discre dit American democracy but to defeat Clinton and elect Trump. They say the great majority of their Senate colleagues agree with them, and McC ain has slated an Armed Services hearing on cyberthreats for Jan. 5. B ut the politicizing of the Russian actions — the idea that they hel ped Trump win — has also made the issue difficult for Republican le aders. It has allowed Trump supporters to push back on the intelligenc e agencies and say the entire issue is designed to undermine Trump's l egitimacy. Senate Majority Leader Mitch McConnell has so far resisted calls for a select committee to look into the Russian interference in the 2016 campaign. He has said it is enough for Sen. Richard Burr, . C. to look into it as chairman of the Senate Intelligence Committee. T ypically, Republican leaders and spokesmen say there is no evidence th at the actual voting or tallying on Nov. 8 was compromised, and that i s true. But it is also a red herring, as interference in those functio ns has not been alleged and is not the focus of the U. S. intelligence agencies' concern. For his part, Trump has shown little interest in de lving into what happened. He has cast doubt on the U. S. intelligence reports to date and suggested "no one really knows what happened." He also has suggested that computers make it very difficult to know who i s using them. This week, Trump said it was time to "get on with our li ves and do more important things." However, at week's end he did agree to have an intelligence briefing on the subject next week. The has n ot wanted the daily intelligence briefings available to him in recent weeks, preferring that they be given to the men he has chosen as his v ice president (Mike Pence) and national security adviser (Mike Flynn)

with Trump taking them only occasionally. The irony of this controversy arising at the eleventh hour of the Obama presidency can scarcely be overstated, and it defines the dilemma facing both the outgoing president and the incoming party in control. Obama appears to have been reluctant to retaliate against the Russian hacking before the election for fear of seeming to interfere with the election himself. The Republicans, meanwhile, have for years called for greater confrontation with the Russians, with Obama usually resisting. Obama did join with NATO in punishing the Russians with economic sanctions over the annexation of Crimea. Those sanctions may have been painful, coming as they did alongside falling prices for oil — the commodity that keeps the Russian economy afloat. On other occasions, despite Russian provocations through surrogates in Syria and elsewhere, Obama did not make overt moves to force Russia's hand. That includes occasions when Russia was believed to be hacking critical computer systems in neighboring Ukraine, Estonia and Poland. But this week, following a chorus of confirmation from the U. S. intelligence community regarding the Russian role in computer hacking in the political campaign, Obama acted. He imposed a set of mostly diplomatic actions such as sanctioning some Russian officials, closing two diplomatic compounds and expelling 35 Russian diplomats. There may have been more damaging measures taken covertly, and some Russophobes in Washington held out hope for that. But the visible portion of the program scarcely amounted to major retribution. And Putin saw fit to diminish the Obama sanctions further by declining to respond. Although his government has steadfastly denied any interference in the U. S. election, Putin rejected his own foreign minister's recommended package of responses. (He even sent an invitation for U. S. diplomats to send their children to a holiday party in Moscow.) That allowed Putin to appear for the moment to be "the bigger man," even as he spurned Obama and kept up what has looked like a public bromance with Trump, who tweeted: "Great move on delay (by V. Putin) I always knew he was very smart!" At the moment it may seem that the overall Russia question amounts to the first crisis facing the Trump presidency. Whether forced by this campaign interference issue or not, Trump must grasp the nettle of a relationship Mitt Romney once called the greatest threat to U. S. security in the world. To be sure, Trump needs to dispel doubts about his ability to stand up to Putin, who has bullied and cajoled his way to center stage in recent world affairs. But Trump also seems determined to turn the page on past U. S. commitments, from free trade philosophy to funding of NATO and the United Nations. And if his Twitter account is any guide, Trump shows little concern about the conundrum others perceive to be facing him. Above all, Trump has shown himself determined to play by his own rules. A year ago, many were confident that would not work for him in the world of presidential politics. We are about to find out whether it works for him in the Oval Office.'

In [6]:
```
npr['Article'][4000]
```

Out[6]: 'The headline shocked the world of the surface Navy: Seven sailors aboard the destroyer USS Fitzgerald were killed, and other crew members

injured, when the warship collided with a cargo vessel off Japan. As t he Navy family grieves, both it and the wider world are asking the sam e question: How did this happen? The short answer is that no one knows — yet. Official inquiries into what led up to the encounter could ta ke months or more. The Navy and the U. S. Coast Guard both likely will eventually issue reports that describe what happened and could make re commendations for preventing another such accident. ”I will not specul ate on how long these investigations will last,” said Vice Adm. Joseph Aucoin, commander of the Navy’s 7th Fleet. The Fitzgerald and the othe r ships of Destroyer Squadron 15, based outside Tokyo, fall under his authority. There are clues, however, that explain how something like t he Fitzgerald’s collision could happen, including photographs of the s hips involved, navigation data about the container ship ACX Crystal an d the experience the Navy has had with past mishaps. The $1. 8 billion Fitzgerald is one of the most modern and technologically advanced wars hips afloat, capable of using its powerful sensors to look up into spa ce, if necessary, and reach up to hit targets there with its battery o f missiles. The destroyer still has a human crew, however, most of whi ch was likely asleep around 2:30 a. m. local time when it collided wit h the Crystal. There was no moon over the waters south of Tokyo Bay, a ccording to local accounts, and the channel there is frequently crowde d with ships on their way into and out of the Japanese capital. Vessel s of all sizes sail to other ports in Asia or head east into the vast Pacific. Sailors in the Fitzgerald’s combat information center and on its bridge are responsible for using the ship’s sensors to plot the lo cation of each one, as well as the directions they’re headed and the s peed at which they’re sailing. Officers and sailors must at all times keep what the Navy calls good ”situational awareness” about not only w hat their own ship is doing, but about what might be ahead in the next patch of ocean where the Fitzgerald wants to sail. In 2012 a sibling o f the Fitzgerald, the destroyer USS Porter, was in a congested, seaw ay called the Strait of Hormuz — the ribbon of water that connects the Persian Gulf with the Arabian Sea — when it collided with an oi l tanker. The Navy’s investigation later found that as sailors tried t o keep track of the traffic all around them, including those ships hea ded the other direction, they lost focus on their own immediate course ahead. When the tanker Otowasan suddenly loomed ahead, Cmdr. Martin Ar riola ordered the Porter to turn left to cross ahead of the huge other ship to avoid a crash. But he hadn’t done so with enough time, and n ot even ordering full speed at the last minute could get the destroyer safely clear. The Otowasan hit the Porter along its right — or star board — side, in a location on the ship very near where the ACX Cry stal hit the Fitzgerald early Saturday. But when the sun came up and p hotos appeared of both ships, they revealed the Crystal had damage on the left or port side of its bow — suggesting it might have been tr aveling in the same direction as the Fitzgerald. It may have been trai ling the smaller destroyer at a perpendicular angle that stayed relati vely the same even as the distance between the ships closed: ”constant bearing, decreasing range.” If the crew of the Fitzgerald was watching what was ahead of them and got used to the presence of the container s

hip on their starboard quarter because it didn't appear to be moving i
n either direction relative to the destroyer — even though it was g
etting closer all the while — the sailors might not have realized w
hat was happening until they were in extremis. Another similar possibi
lity: the Fitzgerald wanted to sail east, say, and its course crossed
over that of the Crystal, heading north. The destroyer might have been
like someone trying to get across a busy street, thinking it could get
out of the way of the oncoming cars in time — in this case, a misca
lculation. Investigators will focus closely on what the crews on both
ships were doing. When the fast attack submarine USS Hartford collided
with the amphibious transport USS New Orleans in 2009, discipline on t
he sub was lax, the Navy later found. The Hartford's captain never cam
e into the control room during the transit through the crowded Strait
of Hormuz. The navigator was in the wardroom listening to his iPod. It
's possible that no one was on the bridge of the Crystal — even hug
e container ships are comparatively lightly crewed, compared with Navy
ships, and unlike warships, often use an autopilot. In the wide open P
acific, mariners sometimes let "Iron Mike" take the helm. After a seri
es of accidents, the U. S. Coast Guard warned mariners last year about
the dangers involved with relying too heavily on autopilot. The Fitzge
rald's bridge almost certainly was crewed, by sailors and officers on
the overnight "midwatch" and those are the watchstanders who may have
made the critical decisions about what to do or not do before the coll
ision. Were they managing a whole screen full of contacts and too dist
racted to notice the one bearing down on them? Or was it a quiet night
with so little to do that the crew became bored and complacent? Invest
igators will conduct interviews, review navigational data and could ev
en listen to recordings of what happened on the bridge, like the one e
ventually released from the Porter. One detail already is known: The F
itzgerald's commanding officer, Cmdr. Bryce Benson, was in his cabin a
t the time of the accident, 7th Fleet's Aucoin said. The captain's com
partment is located on the starboard side of the ship that was crushed
by the Crystal, and Benson was hurt — the Japanese Coast Guard took
him to shore by helicopter. Other sailors were berthed in compartments
farther below decks, which were flooded by the Crystal's bulbous bow.
In all, two berthing compartments and one machinery space, which house
s one of the gas turbines for making the ship's electrical power, quic
kly filled with seawater. "Heroic efforts prevented the flooding from
catastrophically spreading, which could have caused the ship to founde
r or sink," Aucoin said. "It could have been much worse." The Fitzgera
ld limped into Tokyo under its own power the crew of the Aegis comba
tant used a magnetic compass and their backup instruments to get home
with only one of the ship's two propellers. The destroyer now needs mi
llions of dollars' worth of repairs, including a visit to a dry dock,
before it could be ready to take another mission. The Navy identified
the seven sailors who died in the accident on Sunday evening. Acting N
avy Secretary Sean Stackley vowed that service officials would answer
the question everyone is now asking: how it could have happened. "In d
ue time, the United States Navy will fully investigate the cause of th
is tragedy," he said, "and I ask all of you to keep the Fitzgerald fam

```
ilies in your thoughts and prayers as we begin the task of answering t
he many questions before us.”'
```

In [7]:
```
from sklearn.feature_extraction.text import CountVectorizer

cv = CountVectorizer(max_df=0.9, min_df = 2, stop_words = 'english')
```

In [8]:
```
dtm = cv.fit_transform(npr['Article'])
```

In [9]:
```
dtm
```

Out[9]:
```
<11992x54777 sparse matrix of type '<class 'numpy.int64'>'
        with 3033388 stored elements in Compressed Sparse Row format>
```

LDA

In [10]:
```
from sklearn.decomposition import LatentDirichletAllocation
```

In [11]:
```
LDA = LatentDirichletAllocation(n_components=7, random_state=42)
```

In [12]:
```
LDA.fit(dtm)
```

Out[12]:
```
LatentDirichletAllocation(batch_size=128, doc_topic_prior=None,
             evaluate_every=-1, learning_decay=0.7,
             learning_method='batch', learning_offset=10.0,
             max_doc_update_iter=100, max_iter=10, mean_change_tol=0.0
01,
             n_components=7, n_jobs=None, n_topics=None, perp_tol=0.1,
             random_state=42, topic_word_prior=None,
             total_samples=1000000.0, verbose=0)
```

Grab the vocabulary of words

In [13]:
```
#Grab the vocabulary of words
len(cv.get_feature_names())
```

Out[13]:
```
54777
```

In [14]:
```
type(cv.get_feature_names())
```

Out[14]:
```
list
```

In [15]:
```
cv.get_feature_names()[4000]
```

Out[15]:
```
'atizado'
```

In [16]:
```
cv.get_feature_names()[50000]
```

Out[16]:
```
'transcribe'
```

In [18]:
```
# Generate 10 random feature names
import random

for i in range(10):
    random_word_id = random.randint(0, 54777)
    print("index = ", i, "random_word_id = ", random_word_id,
          "random feature name = ", cv.get_feature_names()[random_word_i
```

```
index =  0 random_word_id =  1425 random feature name =  accentuated
index =  1 random_word_id =  15925 random feature name =  edification
index =  2 random_word_id =  3049 random feature name =  anodyne
index =  3 random_word_id =  39666 random feature name =  rashida
index =  4 random_word_id =  50943 random feature name =  unchaperoned
index =  5 random_word_id =  25570 random feature name =  interplaneta
ry
index =  6 random_word_id =  33655 random feature name =  nobleman
index =  7 random_word_id =  49051 random feature name =  thaad
index =  8 random_word_id =  25153 random feature name =  inimitable
index =  9 random_word_id =  30350 random feature name =  masseur
```

In [19]:
```
for i in range(10):
    random_word_id = random.randint(0, 54777)
    print("index = ", i, "random_word_id = ", random_word_id,
          "random feature name = ", cv.get_feature_names()[random_word_i
```

```
index =  0 random_word_id =  34604 random feature name =  openness
index =  1 random_word_id =  50374 random feature name =  troubadours
index =  2 random_word_id =  8710 random feature name =  cervantes
index =  3 random_word_id =  44177 random feature name =  sheldrick
index =  4 random_word_id =  21919 random feature name =  gutted
index =  5 random_word_id =  21335 random feature name =  grandmaster
index =  6 random_word_id =  6881 random feature name =  brazilians
index =  7 random_word_id =  46146 random feature name =  spoilers
index =  8 random_word_id =  32421 random feature name =  mountaineers
index =  9 random_word_id =  24694 random feature name =  incidence
```

Grab the Topics

In [21]:
```
len(LDA.components_)
```

Out[21]:
```
7
```

In [22]:
```
type(LDA.components_)
```

Out[22]:
```
numpy.ndarray
```

In [23]:
```
LDA.components_.shape
```

Out[23]:
```
(7, 54777)
```

In [24]:
```
LDA.components_
```

Out[24]:
```
array([[8.64332806e+00, 2.38014333e+03, 1.42900522e-01, ...,
        1.43006821e-01, 1.42902042e-01, 1.42861626e-01],
       [2.76191749e+01, 5.36394437e+02, 1.42857148e-01, ...,
        1.42861973e-01, 1.42857147e-01, 1.42906875e-01],
       [7.22783888e+00, 8.24033986e+02, 1.42857148e-01, ...,
        6.14236247e+00, 2.14061364e+00, 1.42923753e-01],
       ...,
       [3.11488651e+00, 3.50409655e+02, 1.42857147e-01, ...,
        1.42859912e-01, 1.42857146e-01, 1.42866614e-01],
       [4.61486388e+01, 5.14408600e+01, 3.14281373e+00, ...,
        1.43107628e-01, 1.43902481e-01, 2.14271779e+00],
       [4.93991422e-01, 4.18841042e+02, 1.42857151e-01, ...,
        1.42857146e-01, 1.43760101e-01, 1.42866201e-01]])
```

Grab the highest probability words per topic

In [26]:
```
single_topic = LDA.components_[0]
```

In [27]:
```
single_topic.argsort()
```

Out[27]:
```
array([ 2475, 18302, 35285, ..., 22673, 42561, 42993])
```

In [28]:
```
# Aside -- To understand what argsort does
```

In [29]:
```
import numpy as np
```

In [30]:
```
arr = np.array([10, 200, 1])
```

In [31]:
```
arr
```

Out[31]:
```
array([ 10, 200,   1])
```

In [32]:
```
arr.argsort()
```

Out[32]:
```
array([2, 0, 1])
```

In [33]:
```
# We want the index of the top 10 valiues i.e. greatest values in single
single_topic.argsort()[-10:]
```

Out[33]:
```
array([33390, 36310, 21228, 10425, 31464,  8149, 36283, 22673, 42561,
       42993])
```

In [34]:
```
top_ten_words = single_topic.argsort()[-10:]
for index in top_ten_words:
    print("index = ", index, cv.get_feature_names()[index])
```

```
index =  33390 new
index =  36310 percent
index =  21228 government
index =  10425 company
index =  31464 million
index =  8149 care
index =  36283 people
index =  22673 health
index =  42561 said
index =  42993 says
```

In [36]:
```
top_twenty_words = single_topic.argsort()[-20:]
for index in top_twenty_words:
    print("index = ", index, cv.get_feature_names()[index])
```

```
index =  38079 president
index =  46581 state
index =  48643 tax
index =  25406 insurance
index =  50426 trump
index =  10421 companies
index =  32089 money
index =  54403 year
index =  18349 federal
index =  1 000
index =  33390 new
index =  36310 percent
index =  21228 government
index =  10425 company
index =  31464 million
index =  8149 care
index =  36283 people
index =  22673 health
index =  42561 said
index =  42993 says
```

In [39]:
```
list(LDA.components_)
```

Out[39]:
```
[array([8.64332806e+00, 2.38014333e+03, 1.42900522e-01, ...,
        1.43006821e-01, 1.42902042e-01, 1.42861626e-01]),
 array([2.76191749e+01, 5.36394437e+02, 1.42857148e-01, ...,
        1.42861973e-01, 1.42857147e-01, 1.42906875e-01]),
 array([7.22783888e+00, 8.24033986e+02, 1.42857148e-01, ...,
        6.14236247e+00, 2.14061364e+00, 1.42923753e-01]),
 array([1.75214142e+00, 9.00736692e+02, 1.42857148e-01, ...,
        1.42944048e-01, 1.43107445e-01, 1.42857144e-01]),
 array([3.11488651e+00, 3.50409655e+02, 1.42857147e-01, ...,
        1.42859912e-01, 1.42857146e-01, 1.42866614e-01]),
 array([46.14863883, 51.44085996,  3.14281373, ...,  0.14310763,
         0.14390248,  2.14271779]),
 array([4.93991422e-01, 4.18841042e+02, 1.42857151e-01, ...,
        1.42857146e-01, 1.43760101e-01, 1.42866201e-01])]
```

In [43]:
```
for i, topic in enumerate(LDA.components_):
    print(f"The TOP 15 words for TOPIC # {i}")
    print([cv.get_feature_names()[index]  for index in topic.argsort()[-
    print('\n\n')
```

```
The TOP 15 words for TOPIC # 0
['companies', 'money', 'year', 'federal', '000', 'new', 'percent', 'go
```

```
vernment', 'company', 'million', 'care', 'people', 'health', 'said', '
says']
```

```
The TOP 15 words for TOPIC # 1
['military', 'house', 'security', 'russia', 'government', 'npr', 'repo
rts', 'says', 'news', 'people', 'told', 'police', 'president', 'trump'
, 'said']
```

```
The TOP 15 words for TOPIC # 2
['way', 'world', 'family', 'home', 'day', 'time', 'water', 'city', 'ne
w', 'years', 'food', 'just', 'people', 'like', 'says']
```

```
The TOP 15 words for TOPIC # 3
['time', 'new', 'don', 'years', 'medical', 'disease', 'patients', 'jus
t', 'children', 'study', 'like', 'women', 'health', 'people', 'says']
```

```
The TOP 15 words for TOPIC # 4
['voters', 'vote', 'election', 'party', 'new', 'obama', 'court', 'repu
blican', 'campaign', 'people', 'state', 'president', 'clinton', 'said'
, 'trump']
```

```
The TOP 15 words for TOPIC # 5
['years', 'going', 've', 'life', 'don', 'new', 'way', 'music', 'really
', 'time', 'know', 'think', 'people', 'just', 'like']
```

```
The TOP 15 words for TOPIC # 6
['student', 'years', 'data', 'science', 'university', 'people', 'time'
, 'schools', 'just', 'education', 'new', 'like', 'students', 'school',
'says']
```

Attaching Discovered Topic Labels to Original Articles

In [45]:
```
dtm
```

Out[45]:
```
<11992x54777 sparse matrix of type '<class 'numpy.int64'>'
        with 3033388 stored elements in Compressed Sparse Row format>
```

In [46]:
```
dtm.shape
```

Out[46]:
```
(11992, 54777)
```

In [47]:
```
len(npr)
```

Out[47]:
```
11992
```

In [49]:
```
topic_results = LDA.transform(dtm)
```

In [50]:
```
topic_results.shape
```

Out[50]:
```
(11992, 7)
```

In [51]:
```
topic_results[0]
```

Out[51]:
```
array([1.61040465e-02, 6.83341493e-01, 2.25376318e-04, 2.25369288e-04,
       2.99652737e-01, 2.25479379e-04, 2.25497980e-04])
```

In [54]:
```
# Round to 2 decimal places
topic_results[0].round(2)
```

Out[54]:
```
array([0.02, 0.68, 0.  , 0.  , 0.3 , 0.  , 0.  ])
```

In [56]:
```
topic_results[0].argmax()
```

Out[56]:
```
1
```

In [57]:
```
npr.head()
```

Out[57]:

	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 [58]:
```
topic_results.argmax(axis = 1)
```

Out[58]:
```
array([1, 1, 1, ..., 3, 4, 0])
```

In [59]:
```
npr['topics'] = topic_results.argmax(axis = 1)
```

In [61]:
```
npr.head(10)
```

Out[61]:

	Article	topics
0	In the Washington of 2016, even when the polic...	1
1	Donald Trump has used Twitter — his prefe...	1
2	Donald Trump is unabashedly praising Russian...	1
3	Updated at 2:50 p. m. ET, Russian President Vl...	1
4	From photography, illustration and video, to d...	2
5	I did not want to join yoga class. I hated tho...	3
6	With a who has publicly supported the debunk...	3
7	I was standing by the airport exit, debating w...	2
8	If movies were trying to be more realistic, pe...	3
9	Eighteen years ago, on New Year's Eve, David F...	2

In [ ]:
```

```