IST652 Final Project Report

What does NPR News focus on?

IST652 M002 - Jim Hwang / Woojin Park / Chiau Yin Yang

1. Objectives & Goals

This project is to explore news articles and use necessary python packages to quickly understand the context and topics without reading all of them. Furthermore, we would like to detect how articles clustered together if the articles were rendered into document-term matrices. And be able to calculate and aggregate word counts of articles by publishers to identify patterns and compare different characteristics. In this project, we mainly focus on one publisher - NPR, and try to understand the context based on the articles. We will analyze the topic from author and publisher perspectives.

The packages from advanced topic presentation we used in this project are **seaborn** for descriptive analysis, and **scikit-learn** for topic modeling.

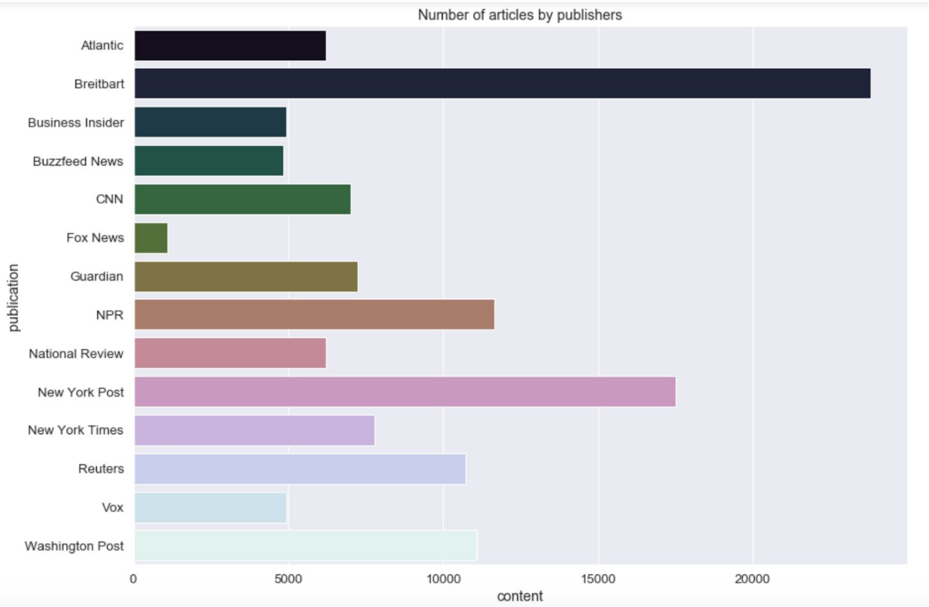
2. Dataset Description

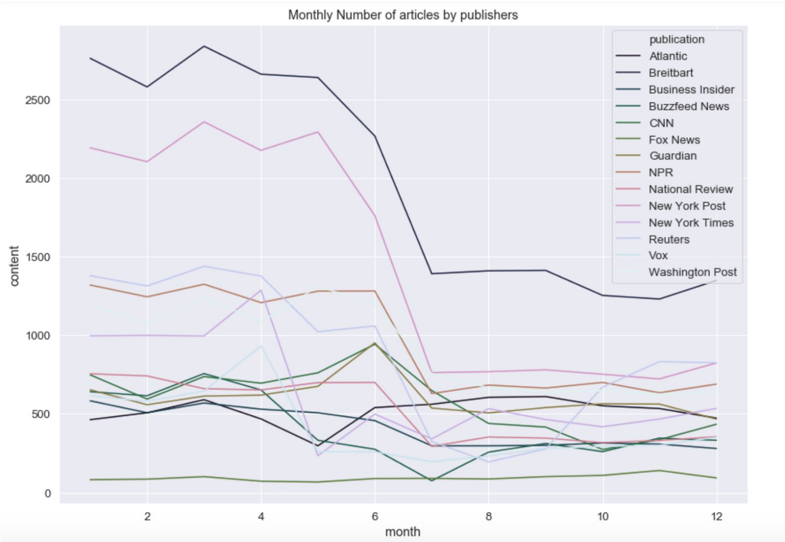
The source of data we used for this project is from Kaggle and there are 143,000 articles from 15 American publications such as the New York Times, Breitbart, CNN, NPR News, and so on. The original 3 datasets contain 10 columns, including the author of the articles, the publish date, publishers, article url, title and content itself. The articles primarily fall between the years of 2016 and July 2017 but there is a not-insignificant number of articles before then. In terms of data cleaning, we removed not important columns, such as URL, and drop null and duplicate values.

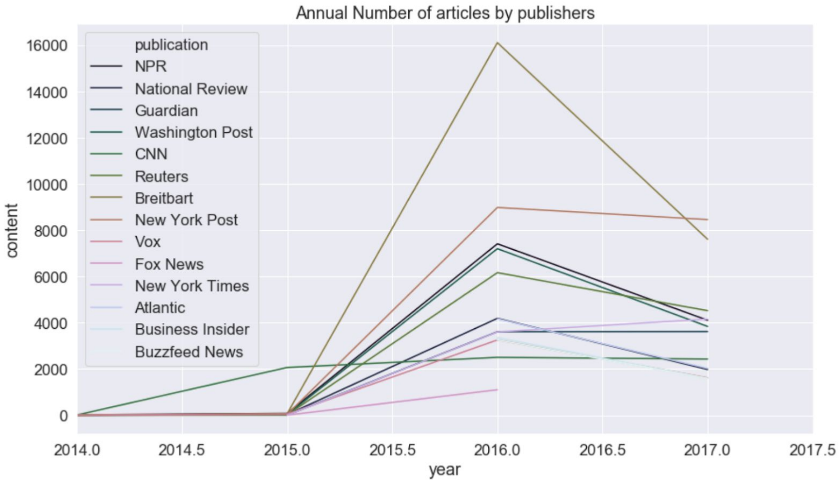
In addition, we want to concentrate on specific articles since, for example, some publishers only have a few authors or too many articles without authors, or authors in the publisher have less than 100 articles. Thus, we drilled down our scope to specific publishers or authors in the descriptive analysis step to build a more sophisticated and focused analysis. We first extracted authors that have at least 150 articles, and picked a publisher whose authors meet this requirements. Eventually, we chose National Public Radio (NPR) as our target publisher.

3. Descriptive Analysis

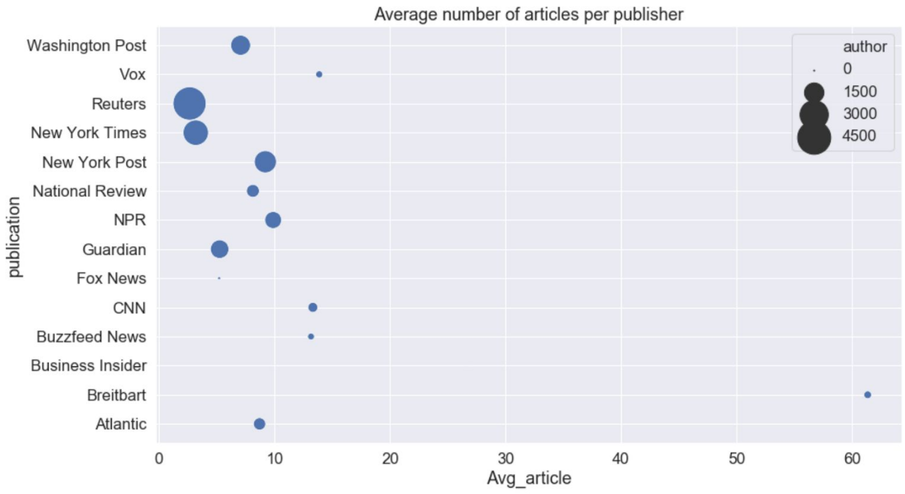
To quickly understand what the dataset is telling us, we aggregate different unit of analysis: (1) What is the number of articles by publisher, (2) by month (3) by year?


Number of articles by publishers
Atlantic
Breitbart
Business Insider
Buzzfeed News
CNN
Fox News
Guardian
NPR
National Review
New York Post
New York Times
Reuters
Vox
Washington Post
publication
0
5000
10000
15000
20000
content

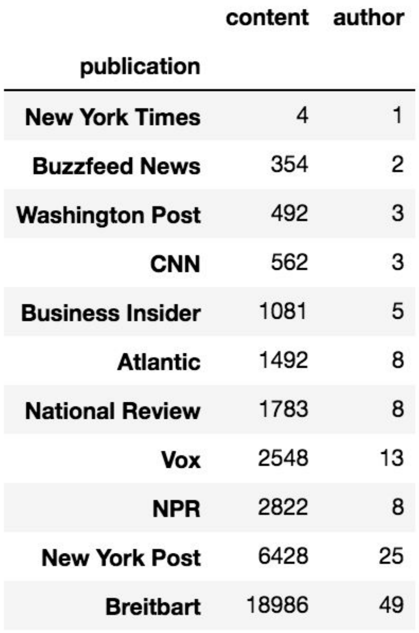

Monthly Number of articles by publishers
publication
Atlantic
Breitbart
Business Insider
Buzzfeed News
CNN
Fox News
Guardian
NPR
National Review
New York Post
New York Times
Reuters
Vox
Washington Post
0
500
1000
1500
2000
2500
content
2
4
6
8
10
12
month


Annual Number of articles by publishers
publication
NPR
National Review
Guardian
Washington Post
CNN
Reuters
Breitbart
New York Post
Vox
Fox News
New York Times
Atlantic
Business Insider
Buzzfeed News
0
2000
4000
6000
8000
10000
12000
14000
16000
content
2014.0
2014.5
2015.0
2015.5
2016.0
2016.5
2017.0
2017.5
year

(4) What is the average number of articles per publisher? (5) per author? -- The graph below shows the average number of article (x-axis) and the size of the bubble shows the number of author.

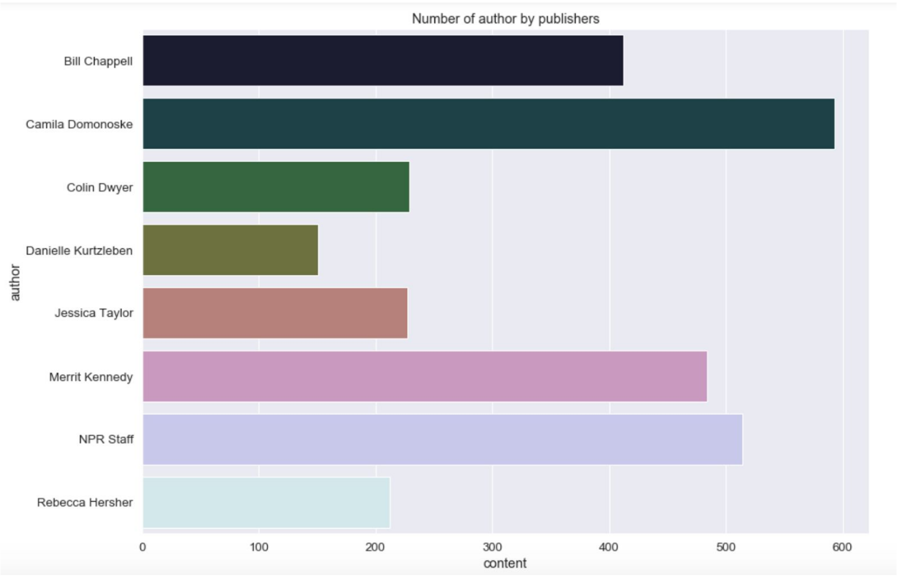


After cleaning the data, we think NPR has most even number of number of articles per author and total number of articles are appropriate.



	content	author
publication		
New York Times	4	1
Buzzfeed News	354	2
Washington Post	492	3
CNN	562	3
Business Insider	1081	5
Atlantic	1492	8
National Review	1783	8
Vox	2548	13
NPR	2822	8
New York Post	6428	25
Breitbart	18986	49

This is the total number of article per author in NPR.



4. Data Analysis results

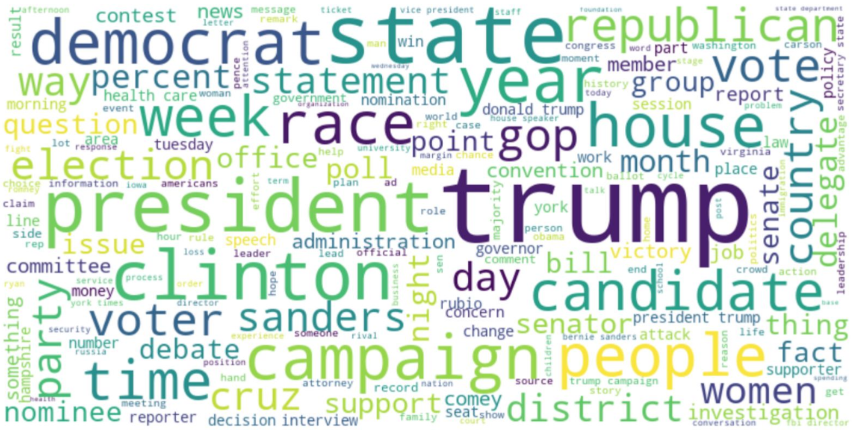
With a quick idea of how the dataset look like, we then moved on to advanced data analysis.

a. Corpus Analysis - Natural Language Processing (NLTK package)

We first wanted to see if each author will have different focused topics, so we aggregate all articles written by the same author and see if we can differentiate by the topics. In this section, with NLTK, we converted into lower-case, removed stopwords, punctuations, and make proper tokens. We performed tri-grams and 5-grams to understand the content a bit more than just single token.

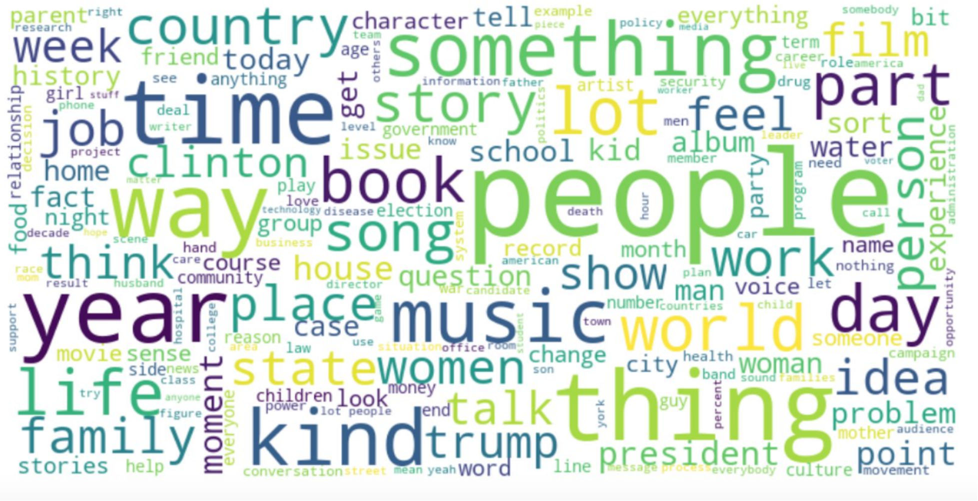
Since our goal is to figure out a few topics without reading those articles, we also use pos-tagging and only extracted nouns from those articles. Next, we applied word frequency and look at the first 20-50 words. To understand the topics faster, we created word cloud with around 200 words. Here are examples from some of the authors that, in my opinion, are most diverse.

1. This is from author Jessica Taylor, who obviously wrote many articles about 2016 presidential election and included nouns like Trump, democrat, clinton, voters etc. This also matched the major event in 2016.



2. This is from NPR staffs, where you can see the topics are a bit more diverse. For example, there are book, song, music, file, story, life etc. You can still see trump and clinton here but

they are not as big as you see in the previous word cloud. (meaning they did not happen as frequent)



Here are examples for the top 15 tri-gram and 5-gram, and you can see that they tell more information than just single words.

```
[('movies', 'trying', 'realistic', 'perhaps', 'way'),
 ('trying', 'realistic', 'perhaps', 'way', 'summon'),
 ('realistic', 'perhaps', 'way', 'summon', 'batman'),
 ('perhaps', 'way', 'summon', 'batman', 'bat'),
 ('way', 'summon', 'batman', 'bat', 'squeak'),
 ('summon', 'batman', 'bat', 'squeak', 'new'),
 ('batman', 'bat', 'squeak', 'new', 'research'),
 ('bat', 'squeak', 'new', 'research', 'bat'),
 ('squeak', 'new', 'research', 'bat', 'lab'),
 ('new', 'research', 'bat', 'lab', 'tel'),
 ('research', 'bat', 'lab', 'tel', 'aviv'),
 ('bat', 'lab', 'tel', 'aviv', 'university'),
 ('lab', 'tel', 'aviv', 'university', 'found'),
 ('tel', 'aviv', 'university', 'found', 'bats'),
 ('aviv', 'university', 'found', 'bats', 'vocalizing')]
```

Example for the top 15 tri-gram from a random author, you can see some of the 3 words that happened most frequent.

```
[(('black', 'lives', 'matter'), 6.590408319731462e-05),
 (('american', 'folklife', 'center'), 5.711687210433935e-05),
 (('new', 'york', 'city'), 4.393605546487642e-05),
 (('world', 'war', 'ii'), 3.514884437190114e-05),
 (('new', 'york', 'times'), 3.075523882541349e-05),
 (('affordable', 'care', 'act'), 2.196802773243821e-05),
 (('civil', 'rights', 'movement'), 2.196802773243821e-05),
 (('interview', 'highlights', 'contain'), 2.196802773243821e-05),
 (('lives', 'matter', 'movement'), 1.9771224959194387e-05),
 (('make', 'america', 'great'), 1.9771224959194387e-05),
 (('martin', 'luther', 'king'), 1.757442218595057e-05),
 (('mary', 'louise', 'kelly'), 1.757442218595057e-05),
 (('things', 'considered', 'host'), 1.5377619412706746e-05),
 (('u.', 's.', 'economy'), 1.5377619412706746e-05),
 (('u.', 's.', 'government'), 1.5377619412706746e-05)]
```

Based on the results from above (and more from the python notebook), we can tentatively conclude that tri-gram and 5-gram tell us more context from those articles, and they are ones that happen more frequent than any others. Additionally, we can use most frequent nouns to recommend hashtags, article category and so on.

b. Topic Modeling

From the analysis of NLP, we understand the better abstract and presume what are the issues or topics among articles. There are several ways to identify topics in the articles to see a bird-eye view of articles and understand what are the topics and how articles can be grouped by the topics. We decide NMF and LDA modeling to investigate and compare the foundings.

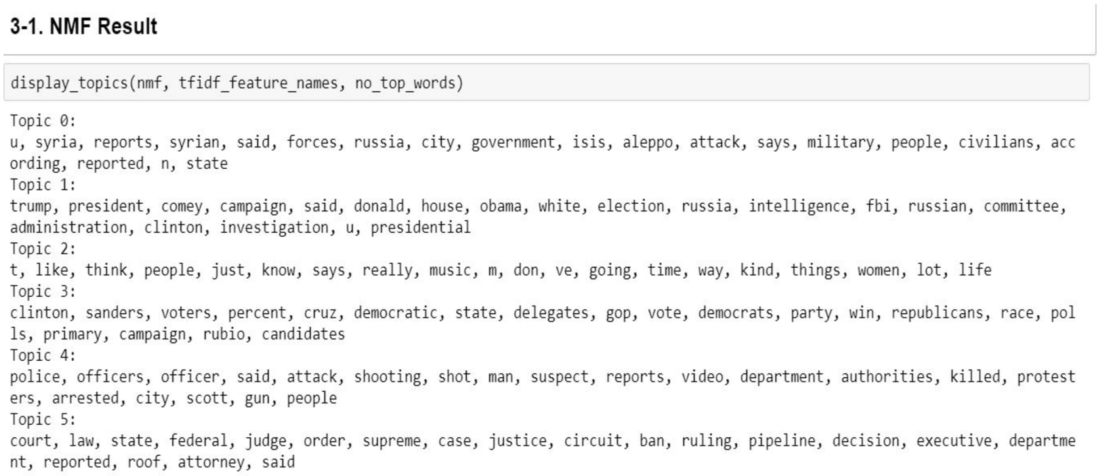
i. NMF

Non-negative Matrix Factorization (NMF) is one of the methods that can be used for topic modeling, which provides two matrices of topics that are factorized from Document-Term Matrix. One of two matrices is Document-Topic Matrix and the other is Topic-Term Matrix. What we focused on was to build models to have insights from a set of articles in several ways of analysis. Sci-kit learn (sklearn) allows NMF topic modeling which is more distance-based clustering method than Latent Dirichlet Allocation (LDA) topic modeling. Thus, NMF topic modeling is conducted as the first model using sklearn.

The data set used to conduct NMF topic modeling is the one from the previous analysis and the rows of NPR News. Since the data set was pre-processed from the previous steps, there was no further cleaning process needed.

The target data for the analysis was text data so that tokenization and vectorization should be run before building an actual modeling part. For tokenization, stop words and numbers were removed and max_df and min_df were set as 0.9 and 5 respectively due to not having too-frequent and less-frequent terms. 13,259 was the number of features from the vectorization process. Term-Frequency inverse Document Frequency (TFIDF) was applied to weight vectorized values to gain NMF in a more scaled way.

The top informative terms in topics help label each topic. From the words such as 'syria,' 'russia,' 'military,' and 'isis,' topic 0 can be labeled as something about 'International politics.' Topic 1 also has similar terms in topic 0, but more on 'Politics and government,' due to 'trump,' 'president,' 'house,' and 'administration.' Topic 2 was a bit vague than the two previous topics but can be labeled as 'Life style' from terms such as 'people,' 'music,' and 'life.' Topic 3 is also about politics but more on 'politics and election since this topic has 'voters,' 'percent,' 'delegates', and 'pol' as most informative terms. Other than the last four topics, topic 4 is about ‘Accidents and criminals’ labeled by ‘police,’ ‘shooting,’ and ‘protest.’ The last topic also has a different topic than others, ‘Law.’

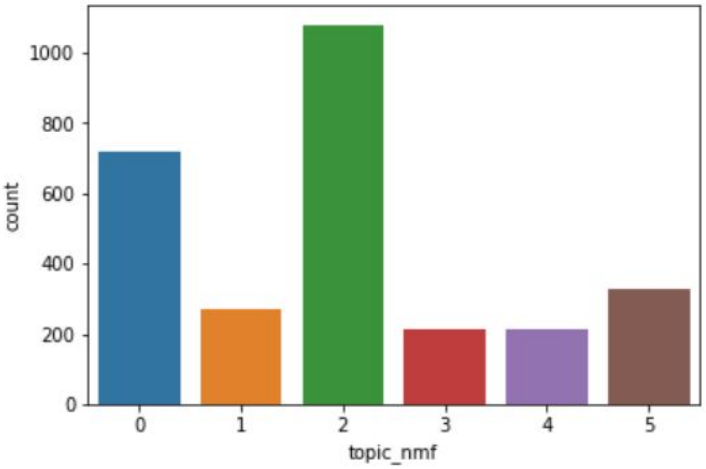


3-1. NMF Result

```
display_topics(nmf, tfidf_feature_names, no_top_words)
```

```
Topic 0:
u, syria, reports, syrian, said, forces, russia, city, government, isis, aleppo, attack, says, military, people, civilians, acc
ording, reported, n, state
Topic 1:
trump, president, comey, campaign, said, donald, house, obama, white, election, russia, intelligence, fbi, russian, committee,
administration, clinton, investigation, u, presidential
Topic 2:
t, like, think, people, just, know, says, really, music, m, don, ve, going, time, way, kind, things, women, lot, life
Topic 3:
clinton, sanders, voters, percent, cruz, democratic, state, delegates, gop, vote, democrats, party, win, republicans, race, pol
ls, primary, campaign, rubio, candidates
Topic 4:
police, officers, officer, said, attack, shooting, shot, man, suspect, reports, video, department, authorities, killed, protest
ers, arrested, city, scott, gun, people
Topic 5:
court, law, state, federal, judge, order, supreme, case, justice, circuit, ban, ruling, pipeline, decision, executive, departme
nt, reported, roof, attorney, said
```

The result below showed that topics are not perfectly evenly distributed but somehow they are distributed evenly through topic 1, 3, 4, and 5.



As a result, we can see that NPR News has contributed their articles about political issues such as international politics, elections, and government affairs followed by law and social issues.

ii. LDA

Alongside with NMF, Latent Dirichlet Allocation(LDA) topic modeling method is employed to investigate any meaningful and different insights from NMF. It is basically a ‘generative probabilistic model’ of a collection of composites made up of parts (composites are documents & the parts are words).

By applying this method, we can detect the number of the topics and the probabilities as the proportion of topic membership. Moreover, we apply this method to reduce the high dimensionality of our text dataset like the same logic of NMF which enables two matrices of topics to be factorized from Document-Term Matrix. Also, we use Term-Frequency inverse Document

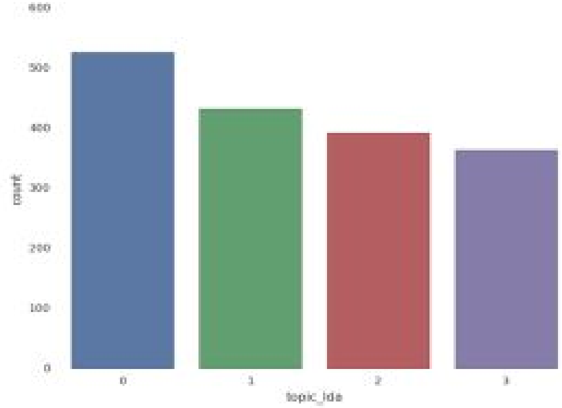
Frequency (TFIDF) to distinguish the different topics using the words in each topic and their corresponding weight.

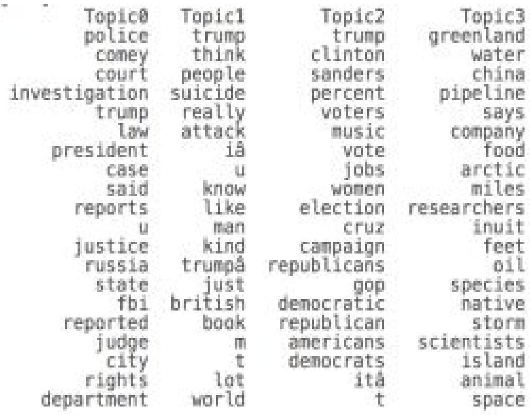
The data set used to conduct LDA topic modeling is identical of the one from the previous analysis applied to NMF, therefore, the data set was pre-processed from the previous steps, there was no further cleaning process needed.

In LDA modeling, we use Spark in databricks' customer cluster to build pipelines, validate and compare the model performance to excel the iterative computation, enabling MLlib to run fast and in-memory processing of relatively big text data. Likewise the NMF modeling, the target data for the analysis is text data so we build pipeline for each needed processing technique to prepare the input data by stages such as Stopwords removal, Tokenization, Vectorization, and TF-IDF.

After the final the pipeline is built, we fit and transform the data to have the best informative group of terms in each topic to make interpretation for LDA topic modeling output. Based on the abstract from corpus analysis we employed, we presume the optimal topic number K for our dataset would be 4 to 6. By changing the number K from 4 to 6, we can define the optimal K of LDA modeling with the most informative and distinguishable terms in each topics group as 4.

Basically, building the top 20 term matrix for each topic, we interpret and label each topic. From the terms such as 'investigation,' court,' law,' and justice' topic 0 can be labeled as Law and Investigations.' In Topic 1, the most frequent terms like 'suicide', 'attack', and 'people' explain mostly about 'Criminal and Accident'. Topic 2 is apparently about 'Politics and Election' since this topic has 'trump', 'clinton', 'voters', and 'campaign'. In LDA, other than the last 3 topics, topic 3 covers slightly different issues, the top terms in this topic are such as 'greenland', 'water', 'food', 'artic', 'native', and 'researchers'. Therefore, we label this one as 'Environment and Research' and it is interesting findings that is not appeared in 6 topics of NMF.





```
         Topic0    Topic1        Topic2       Topic3
         police     trump         trump    greenland
          comey     think       clinton        water
          court    people       sanders        china
  investigation   suicide       percent     pipeline
          trump    really        voters         says
            law    attack         music      company
      president        iâ          vote         food
           case         u          jobs       arctic
           said      know         women        miles
        reports      like      election  researchers
              u       man          cruz        inuit
        justice      kind      campaign         feet
         russia    trumpâ   republicans          oil
          state      just           gop      species
            fbi   british    democratic       native
       reported      book    republican        storm
          judge         m     americans   scientists
           city         t     democrats       island
         rights       lot           itâ       animal
     department     world             t        space
```

As a result from LDA, we are more convince that the NPR News has covered their articles mainly about political issues, government affairs followed by law and also environmental issues and research.

5. Conclusion & Inference

1) Corpus analysis for building Topic modeling

a) We apply the Corpus analysis for nouns to guess topics per author in NPR News

b) Use Tri-gram and 5-grams to better understand the abstract of text data

c) From the findings of our corpus analysis, not only we get abstract of major topics by authors but also detect major events of 2016 which are US presidential election, 2016 shooting of Dallas police officers, and Women in congress and so on.

d) These findings become key indicators for inference at the later analysis (Topic modeling).

2) Topic modeling

a) We Identify common/different topics between NMF and LDA which are about political issues such as international politics, elections, and government affairs followed by law and social issues in common, but LDA has clustered one apparently distinguishable topic about environment and research.

b) If we collect more data from NPR News, based on our NMF/LDA model's topic representing, we can suggest newly created 'hashtag' toward the NPR News.