What does NPR News focus on?

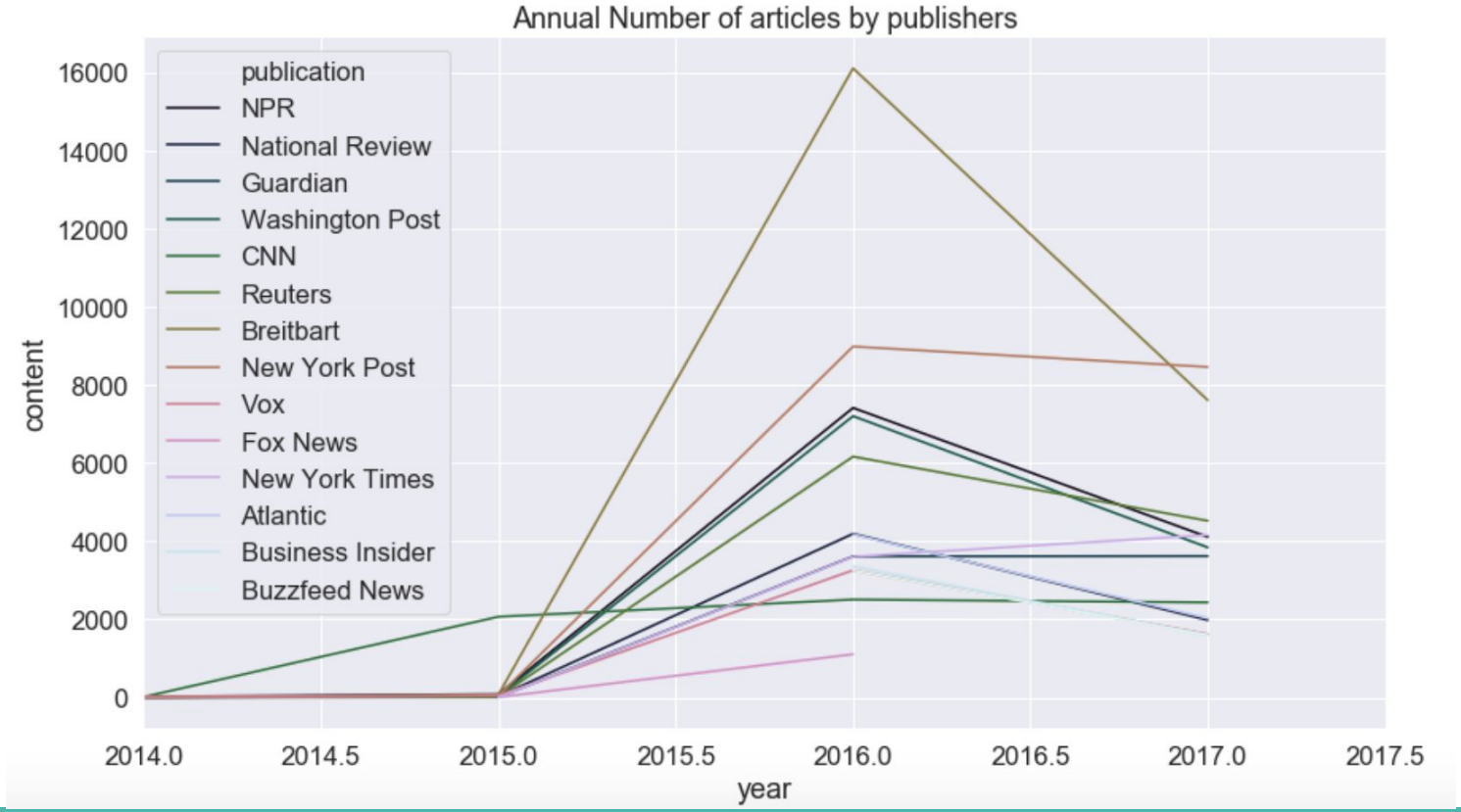
IST652_M002_Final Project

Jim Hwang
Woojin Park
Chiau Yin Yang

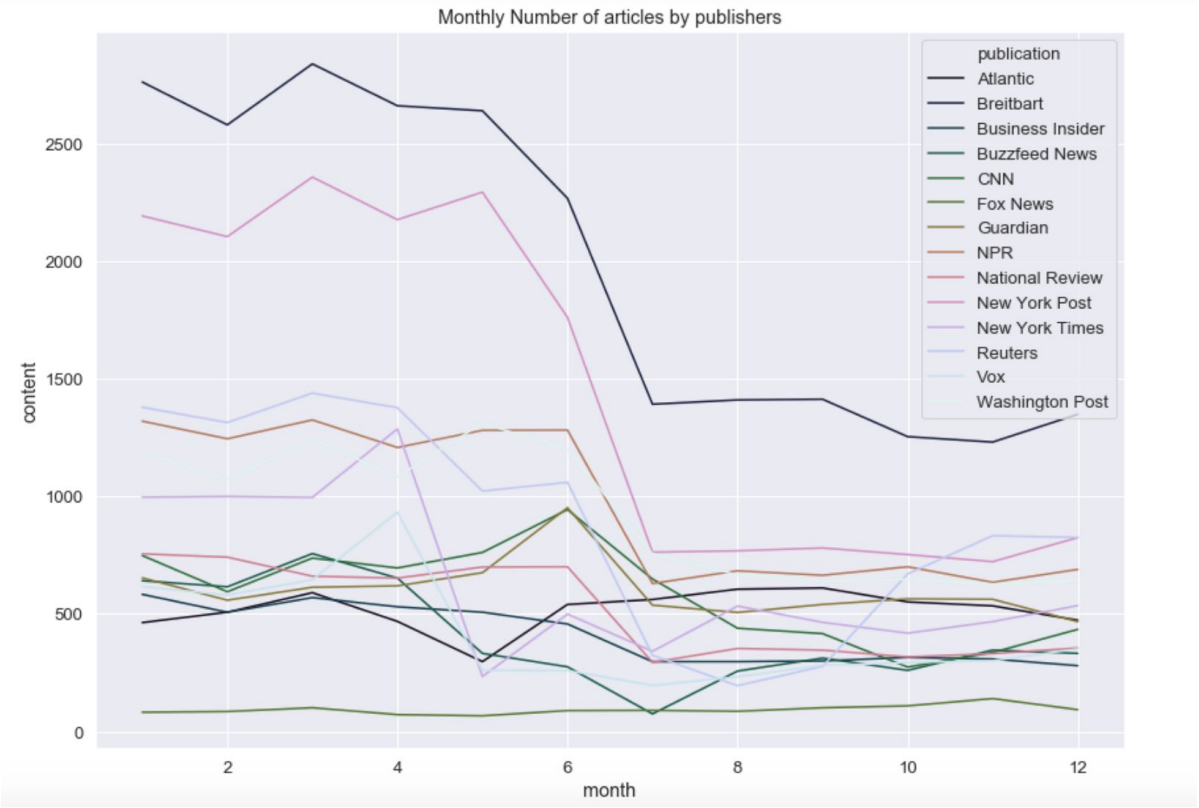
Briefing

- **Data source**: 3 dataset from Kaggle, and there are 143,000 articles from 15 American publications -
  - https://www.kaggle.com/snapcrack/all-the-news#articles1.csv
  - Structured dataset - csv
- **What questions do we answer? - Unit of analysis**
  - How many articles are there by publishers, authors, monthly, and yearly?
  - What is the average number of article per author by publisher?
  - What does a author write about? Are there specific and obvious topics they focus?
  - What topics/content does NPR focus on?

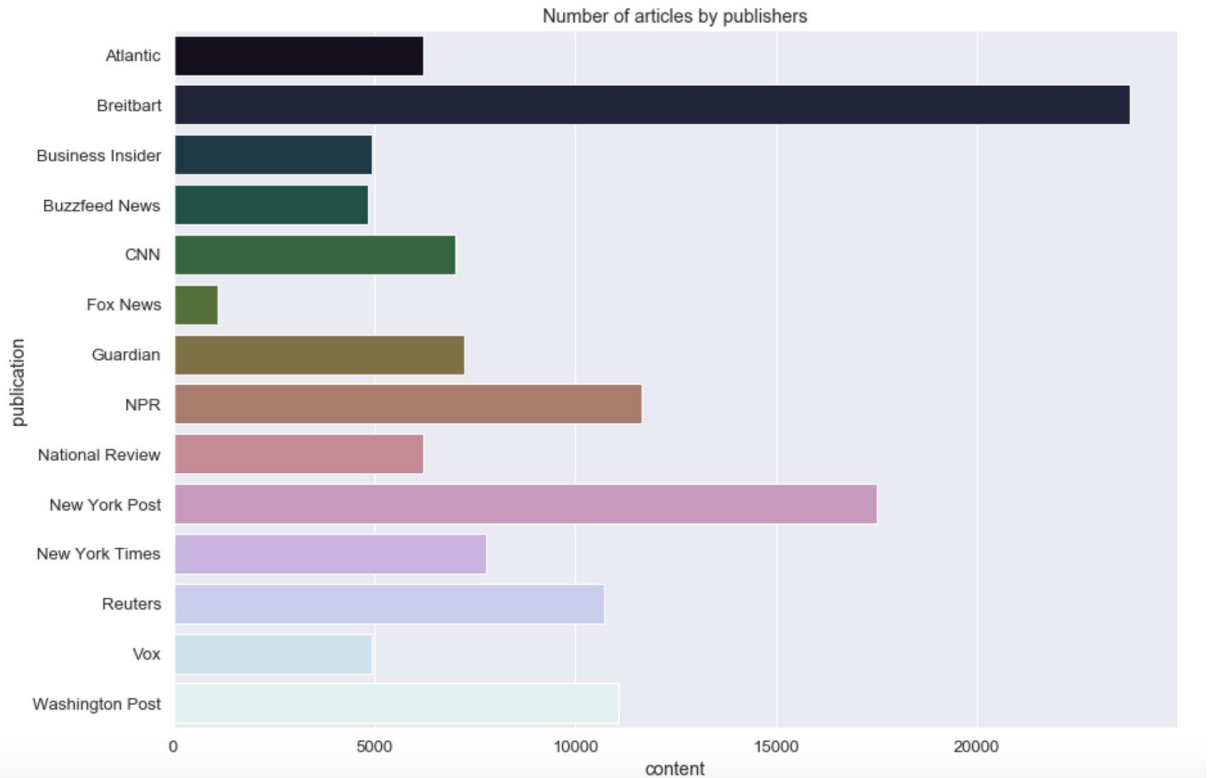
How many articles were there each year?



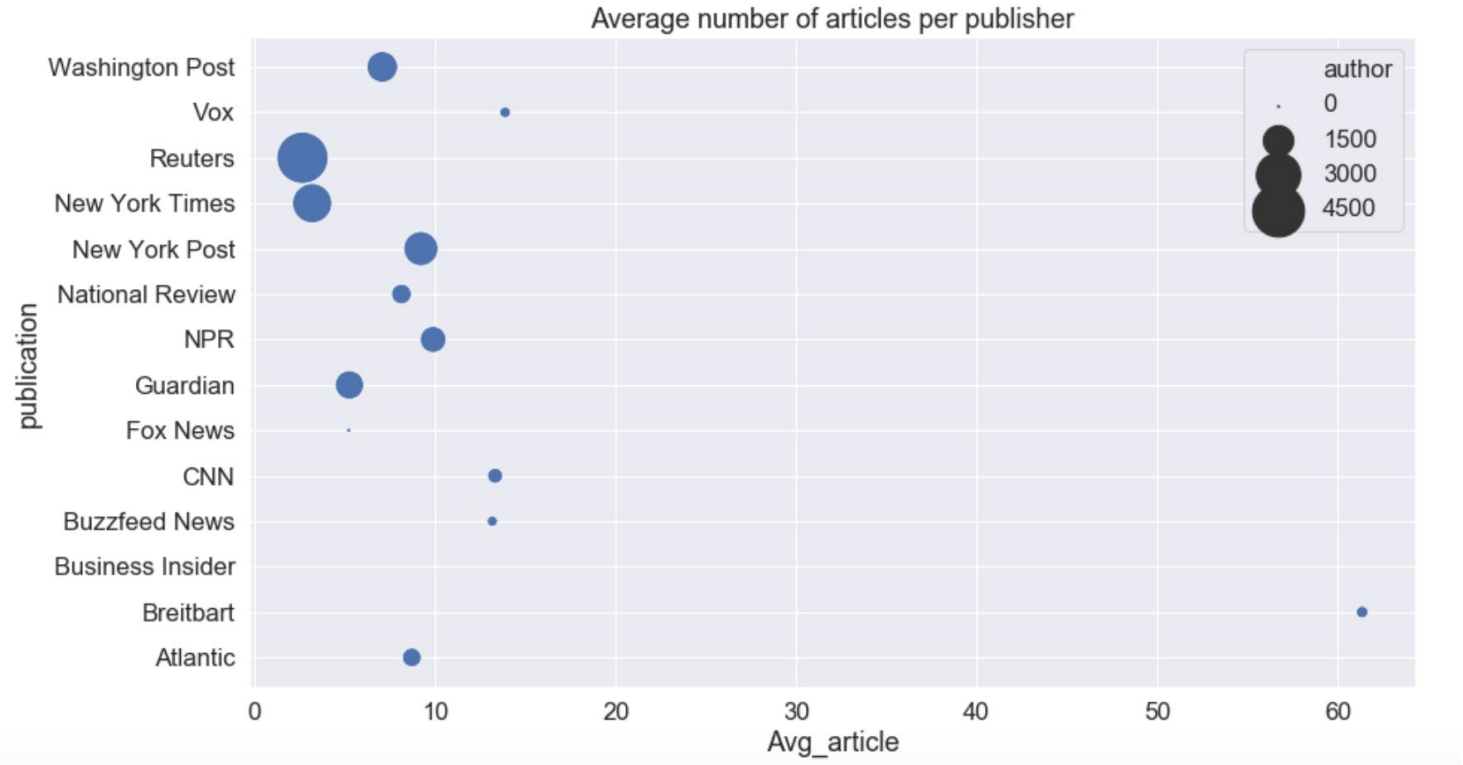
How many articles were there each month?



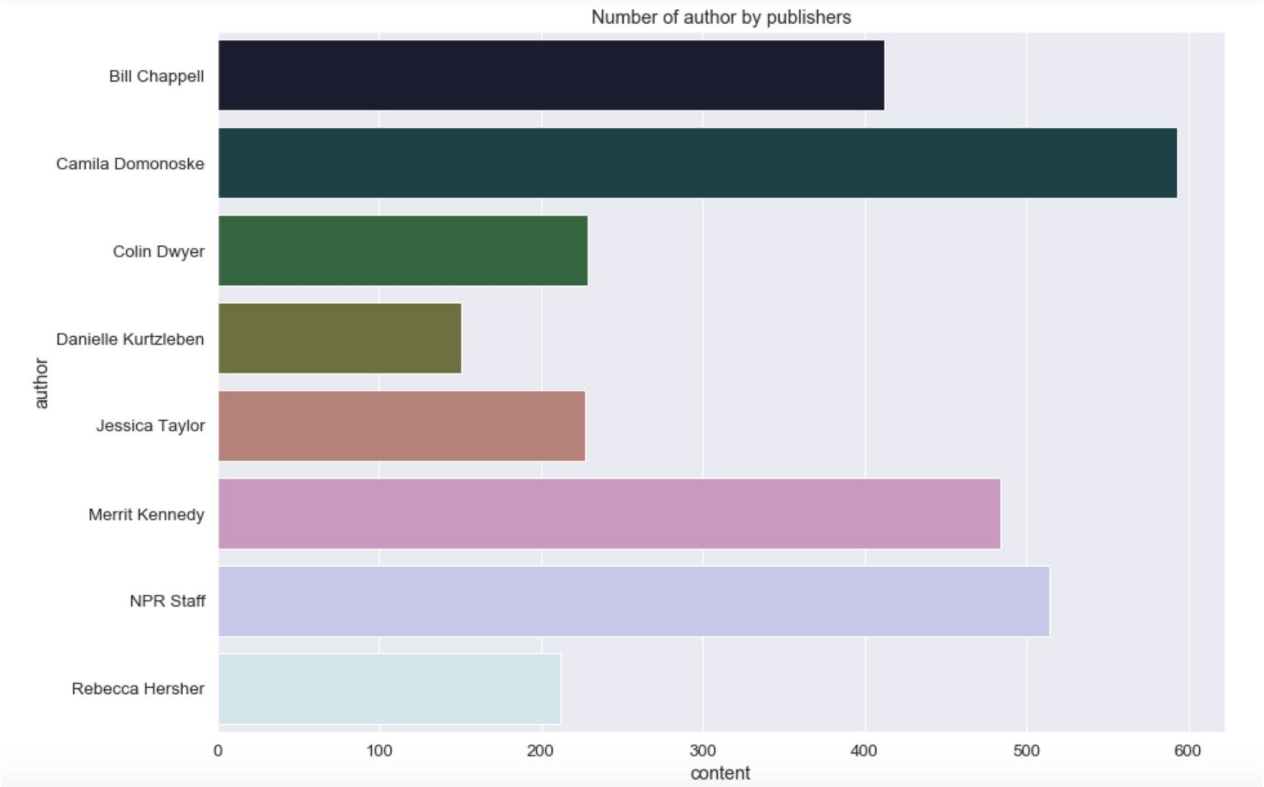
How many articles per publisher?



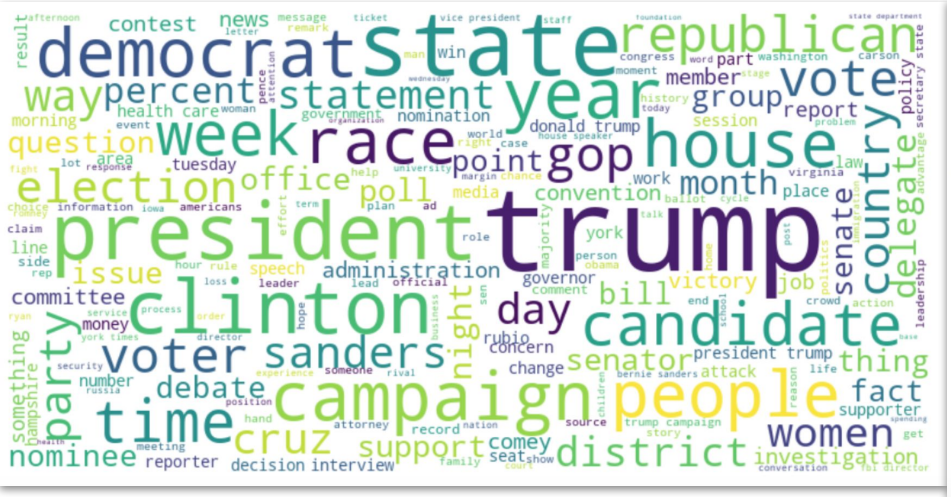
What is average number of articles per publisher?



Among all NPR articles, how many articles per author?



WordCloud - Jessica Taylor / NPR Staff



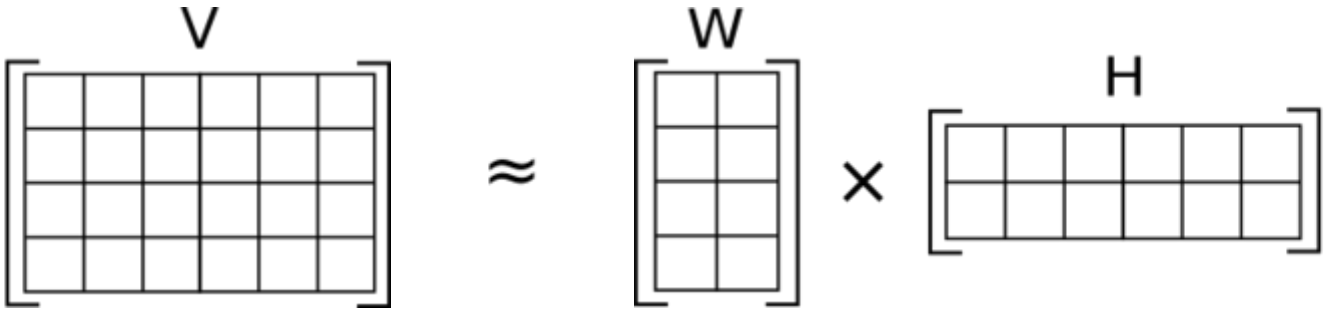
What we learned from the context?

Major events

- 2016 US presidential election / Russian president inference
- 2016 shooting of Dallas police officers
- The 2016 Summer Olympics in Rio de Janeiro, Brazil
- The Syrian refugee crisis
- Standing Rock and the Dakota Access Pipeline Protests
- Women in congress

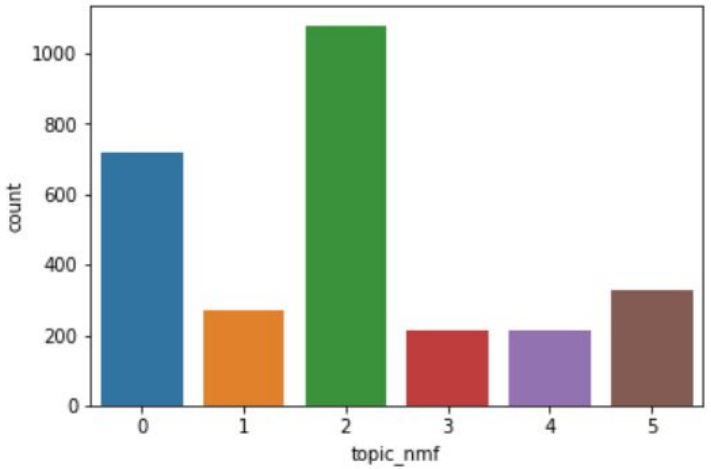
Topic Modeling - NMF

- **Non-negative Matrix Factorization (NMF)**
  - Decomposition Technique
  - V = WH
  - Sparsity (Text mining)
  - Distance measure (TF-IDF)



Topic Modeling - NMF

- Topic 0: Int'l politics
- Topic 1: Politics and government
- Topic 2: Life style
- Topic 3: Politics and election
- Topic 4: Accident and criminals
- Topic 5: 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
```

Topic Modeling - LDA

- **LDA?**

  Latent Dirichlet allocation(LDA) is a “generative probabilistic model” of a collection of composites made up of parts (Composites are documents & the parts are words)

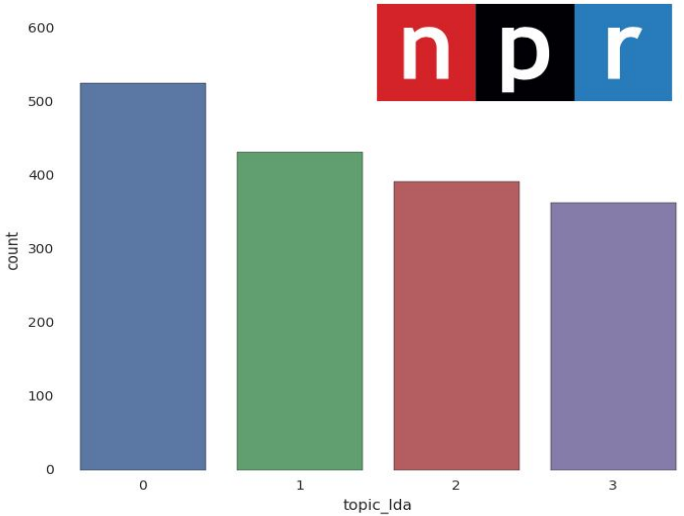
➢ Detect Number of topics and the probabilities as the proportion of topics membership
(Vs. k-means-only belong to cluster)

➢ Way of reducing the dimensionality

➢ Distinguish different topics using the words in each topic and their corresponding weights (TF-IDF)

Topic Modeling - LDA

- Topic 0 : Law & Investigations
- Topic 1 : Criminal & Accident
- Topic 2 : Politics & Election
- Topic 3 : Environment & Research



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

Conclusion

- **Corpus analysis for building Topic modeling**
  - Use nouns to guess topics
  - Tri-gram and 5-grams to better understand the abstract
- **Topic modeling**
  - Identify common/different topics between NMF and LDA
  - Use topics to recommend hashtag
  - NPR focuses on topics, like politics, society issues, science and so on.

Thank you :)