Topic modeling for news articles

Yiding Weng

Introduction

We want to explore:

- How to extract useful information from massive news articles without going through all of them?
- What are the emphasized content of different news media ?
- What are the common topics of contemporary news articles ?

Approach

- Extract keywords from several news medias
- Use Latent Dirichlet Allocation (LDA) to group news articles into clusters of similar content.
- Define the topic of each news clusters based on its content and common keywords.

Data preparation

All the news: 143,000 articles from 15 American publications
Data Location: https://www.kaggle.com/snapcrack/all-the-news

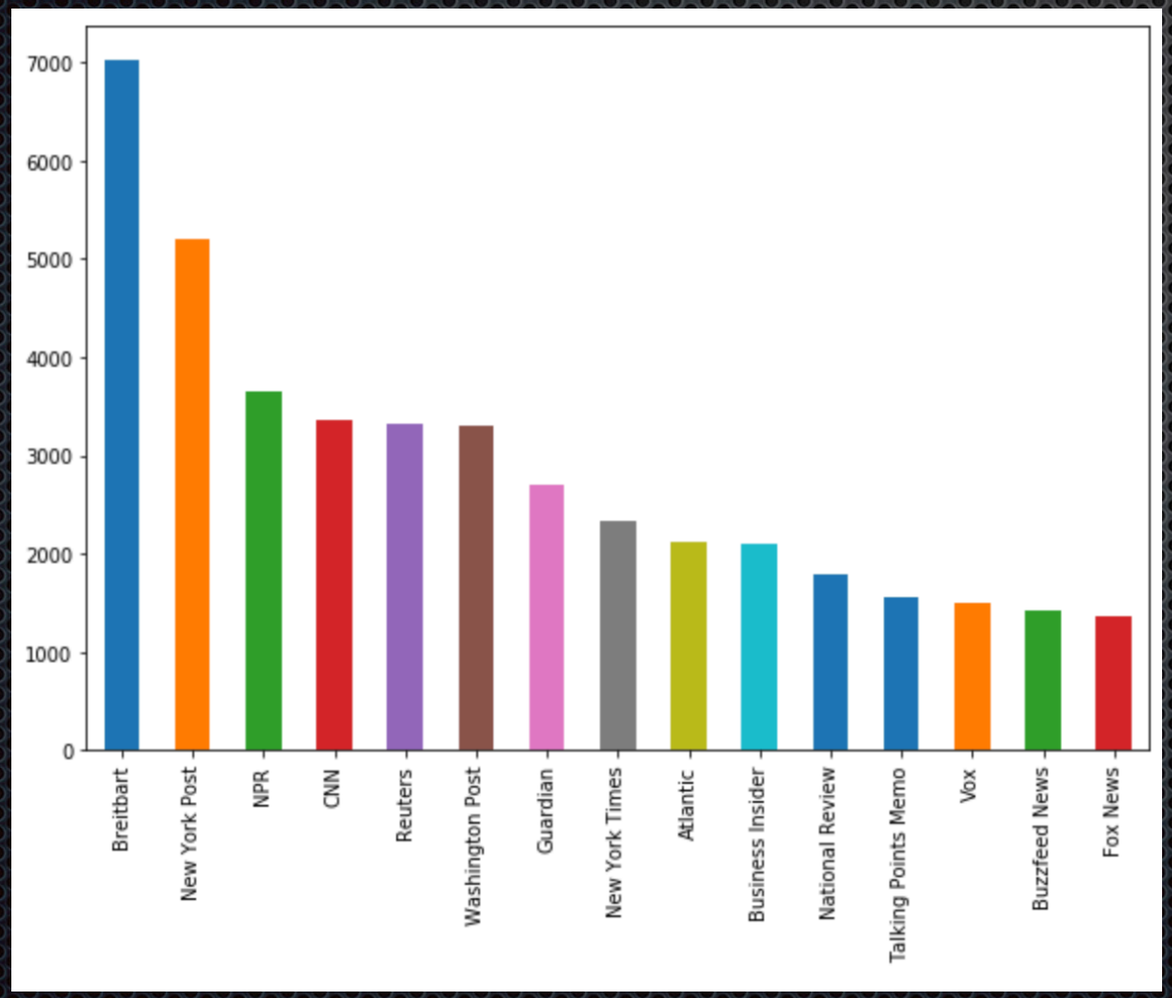
- **Data sampling ,** Due to the limited computing power, I took a sample of 30% of the total data
- **Data cleaning,** remove unwanted elements such as emails sign (@), newline and extra spaces with regular expressions.
- **Tokenization,** break down each sentence into a list of words
- **Creating Bigram and Trigram Models,** find out two or three words frequently occurring together in the document
-

Exploratory analysis

News article distribution

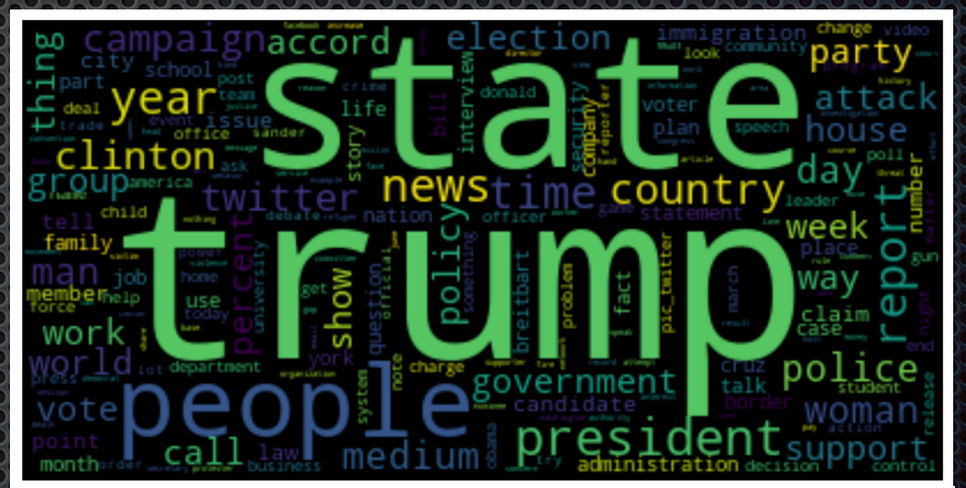
- Find the news medias that contribute the most number of articles within the dataset

.



Word cloud for keyword

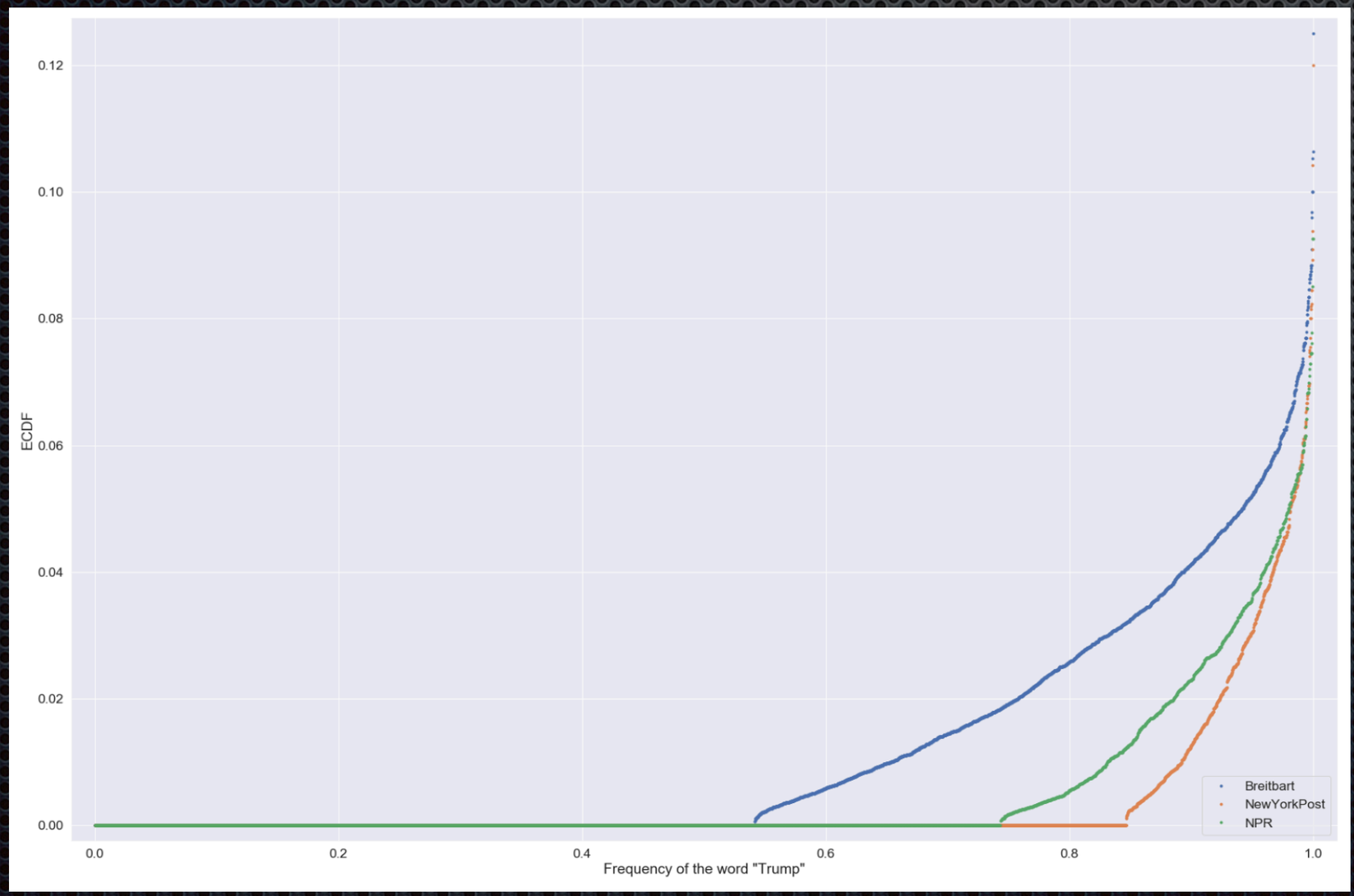
- The most frequently appeared word from different news media



Exploratory analysis

Word frequency of common keywords

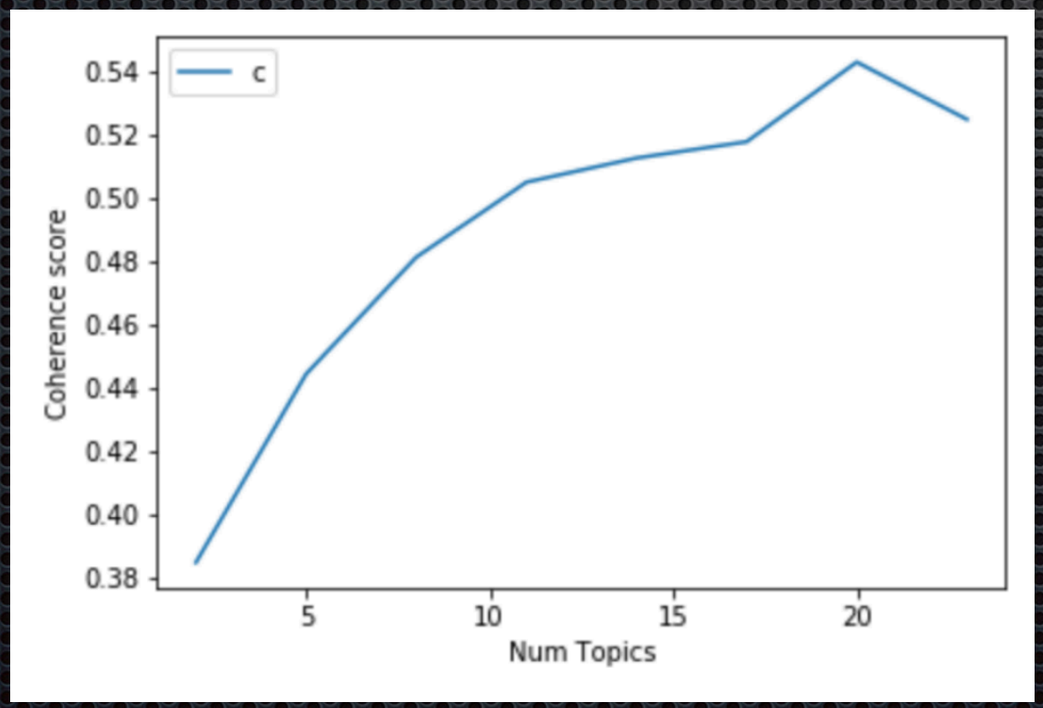
- The word ‘Trump’ has statistically different frequency across different new medias



Cluster modeling

- **Create the Dictionary and Corpus needed for Topic Modeling,** the two main inputs to the LDA topic model are the dictionary and the corpus, corpus shown is a mapping of (word_id, word_frequency)

- **Find the optimal number for news topics,** build many LDA models with different values of number of topics (k) and pick the one that gives the highest coherence value. Topic coherence value signifies the quality defined by % of conformance as per some predefined standards.

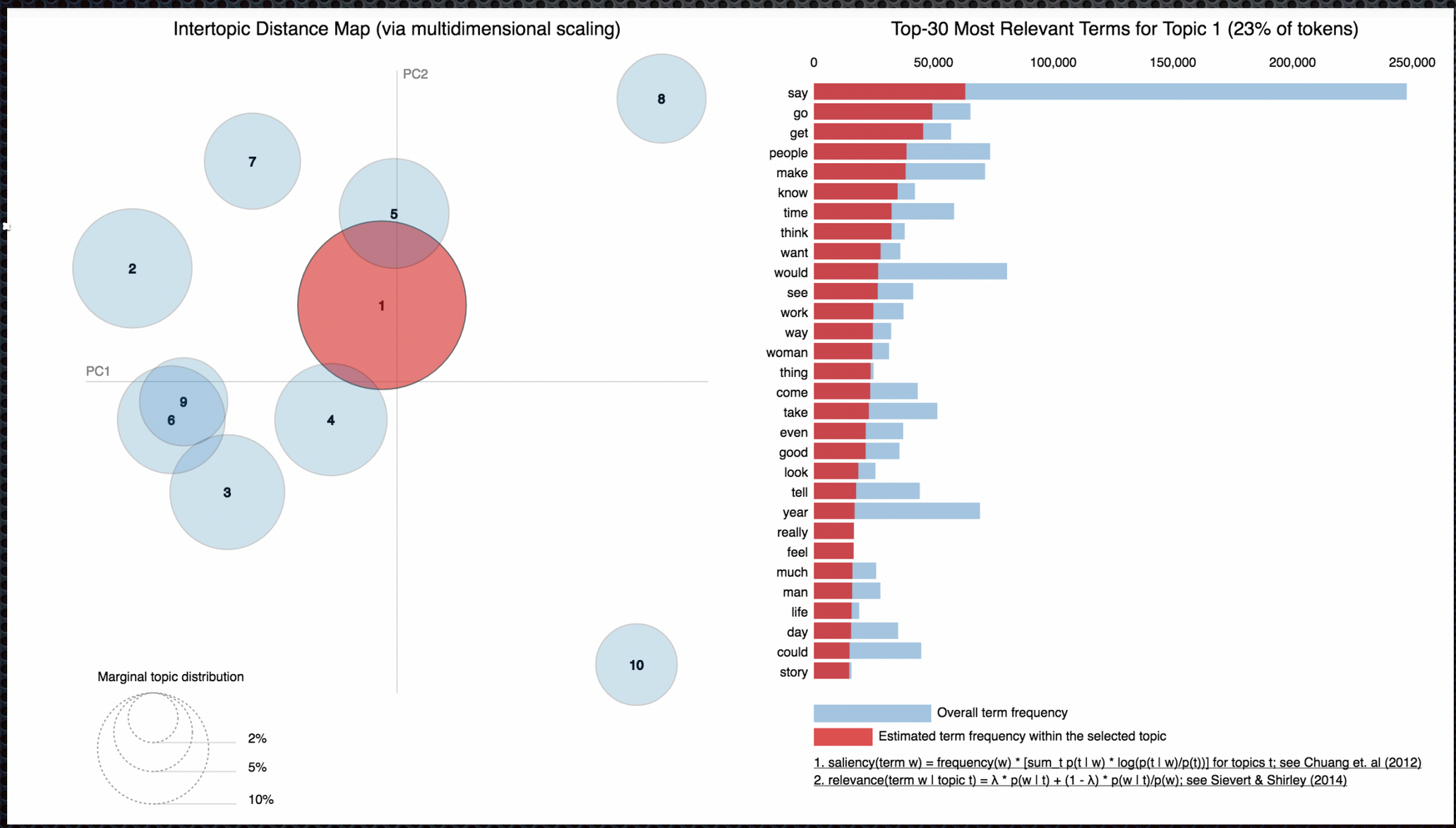


- **Choose the model with 10 topics**, pick the model that gave the highest CV before flattening out. Topic number going well above 10 have good coherence scores but may have repeated keywords in the topic

Cluster analysis

Visualize the topics-keywords

- pyLDAvis package's interactive chart helps to examine the produced topics and the associated keywords.



Cluster analysis

Go through each of the circle which represent different topics, from its own list of keywords we are able to estimate the theme of each topic

Topic 0 (circle 7): Crime
Topic 1 (circle 9): Foreign affairs
Topic 2 (circle 1): General News
Topic 3 (circle 5): Legal issues
Topic 4 (circle 3): Sports
Topic 5 (circle 10): Fashion & Entertainment
Topic 6 (circle 2): Election
Topic 7 (circle 8): Health & Environment
Topic 8 (circle 4): Education & Social work
Topic 9 (circle 6): Finance & Economy

Cluster analysis

Validating the result through human reading

- **From the top of the data frame of the notebook, we can choose the first five articles as a sample:**

0. Ivanka Trump just got booed in Germany for calling her father a champion of women

1. At Debate, Hillary Clinton Leaves Questions About Approach to Banks

2. Why Mars Is the Best Planet

3. Sketch To Impress: How An Oscar-Winning Designer Costumes The Stars

4. What to Make of the Saudi Shake-up

Cluster analysis

Validating the result through human reading

- **Though my reading, I extract the following keywords and give the topic to these articles**

0. Keywords: Ivanka Trump, women, presidency, equality, entrepreneurship **topic**: general news, election

1. Keywords: Hillary Clinton, Wall Street, banks, federal regulator, congress, economy, financial **topic**: Finance, economy

2. Keywords: Mars, water, Earth, life, planet, science, NASA, **topic**: Science, Environment

3. Keywords: Oscar, costume design, Cate Blanchett, dress, **topic**: Fashion, Entertainment

4. Keywords: Saudi Arabia, Salman, successor, ruler, monarchy, ambassador **topic**: Foreign news

- **Compare my defined topics with the ones defined by the model**

0.**my topic**: general news, election **model topic**: Fashion & Entertainment

1.**my topic**: Finance, economy **model topic**: Finance & Economy

2.**my topic**: Science, Environment **model topic**: Health & Environment

3.**my topic**: Fashion, Entertainment **model topic**: Fashion & Entertainment

4.**my topic**: Foreign news **model topic**: Foreign affair

Beside the first one, my conclusion is a bit different from the model conclusion, since the article itself span across multiple areas. But other than that the results are pretty much the same.

Conclusions

- The keywords and its related frequency suggest different emphasis within the news content from different media sources
- By comparing coherence score and the number of classifying groups, we decided to divide the news documents into ten groups given its corresponding optimal coherence score
- By comparing the modeling result with human interpretation of the article and judgment about the theme topic, we can conclude the model is reliable to divide news articles into 10 distinct groups
-

Future work

- **Better validation method,** validation can be improved in reducing bias by involving more people and reading larger amount of the articles sampled from different source of news media. A metric can be utilized to quantitively record the proportion of keywords identified by both human users and the model.

- **Text summarization,** to have a better understanding of the news article with more than keywords and topics, more work can be done in training the model to automatically summarize the new text.

Recommendations

In this project the LDA topic identifying model can identify main topics of large amount of news articles that's too much for human to process.

Since the modeling paradigm is only based on the word frequency, which means this model can be also used to identify topics and central information in other text sources such as legal, historical or financial documents, and even documents of other languages. So that more time can be saved from using man power to obtain the same amount of the information.

.

.