

June 6, 2021

Quinn E Knudsen

IST 736

Syracuse University

Hot Topic

Using Topic Modeling to Determine Themes in Popular News Articles

**Table of Contents**

**Introduction**……………………………………………………………………………......pg 2

Areas of Focus …………………………………………………………….……......pg 2

BBC News..............………………………………………………………................pg 2

ABC News………. ………………………………………………………….…...... pg 3

**Analysis**……………………………………………………………………………….….…pg 4

About the Data ...……………………………………………………………….…...pg 4

Analysis 1: BBC News Topic Model……………………………………………..... pg 4

Data Cleaning & Prep..……………………………………………….….…..pg 4

Analysis 2: ABC News Topic Model……………………………………………..... pg 7

Data Cleaning & Prep..……………………………………………….….…..pg 7

**Results**…....………………………………………………………………………..………..pg 9

Analysis 1: BBC News Topic Model ……………………………………….…...….pg 9

*Research Question 1………*……………………………………………….....pg 9

Analysis 2: ABC News Topic Model ……………………………………….…...….pg 12

*Research Question 2………*………………………………………………....pg 12

**Conclusions**…………………………………………………………………………….…...pg 14

**Introduction**

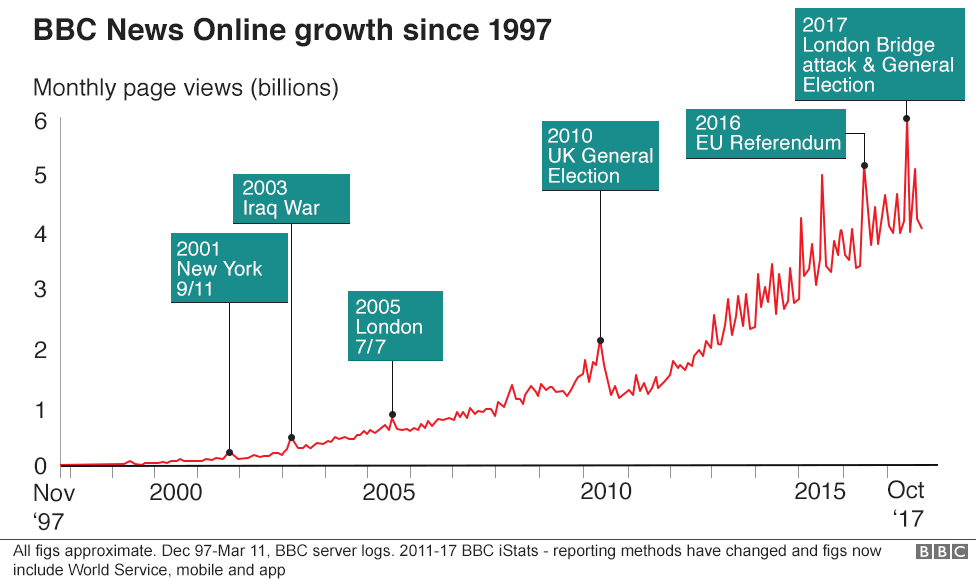
**Areas of Focus**

The scope of this paper is to explore Latent Dirichlet Allocation (LDA) utilizing data scraped from BBC-News using newsapi.org and a series of ABC news headlines collected and stored in a csv. LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. The overarching goal of LDA is to create a series of unique topics with embedded words that are related to one another. This process is a combination of art and science and requires some degree of human interpretation to derive meaning and define the optimal number of topics for the model.

This paper will utilize Python packages such as sklearn and CountVectorizer to convert text corpuses into sparse matrixes for supervised machine learning classification. Implications and future considerations will be discussed.

**BBC News**

BBC News is a British news network that offers up to the minute news on local, global and US & Canada breaking news. It’s online platform, launched in 1997, is one of the most popular websites in the UK with 14 million global viewers each month. BBC reaches ¼ of it’s audience via the website. As exemplified in Figure 1 below, BBC has grown exponentially with boosts when major global news occurs. This makes intuitive sense as people turn to the web when hot current events transpire. It is expected that the popularity of web-based news will continue to grow in the coming years.



*Figure 1.*Growth of BBC’s online web utilization from launch to October 2017.

*News API*

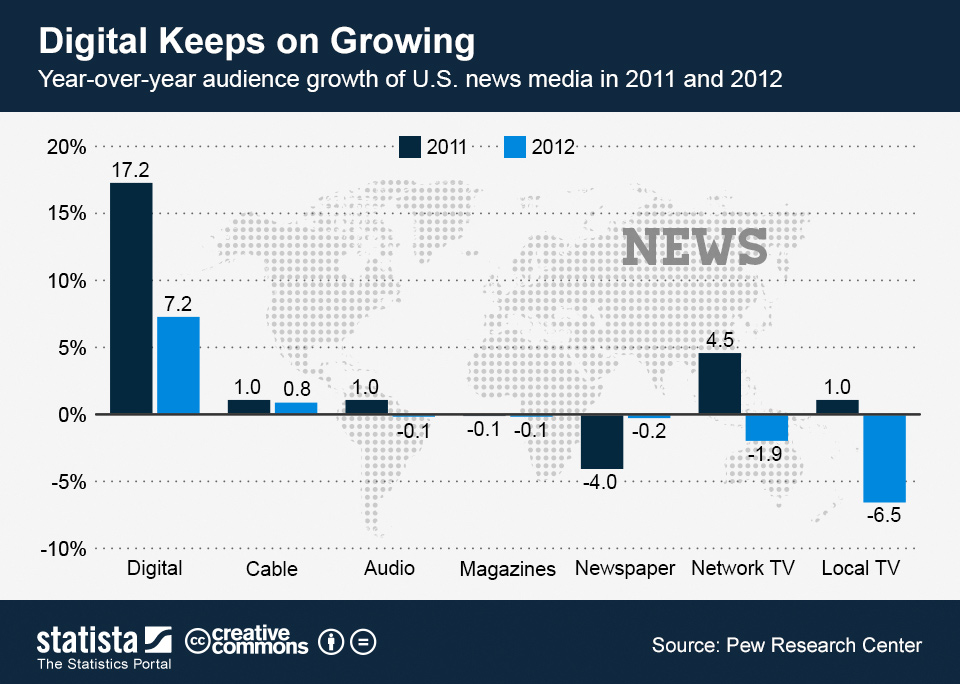
News API is a simple rest API that returns JSON search results from current and historical news articles published by over 75,000 worldwide sources. It is free for developers to use after signing up and creating a unique API key. Post can be scraped by entering the API key and defining the parameters desired. For example, the BBC news was scraped by defining ‘bbc-news’ as the source and sorting by the top 75 articles.

**ABC News**

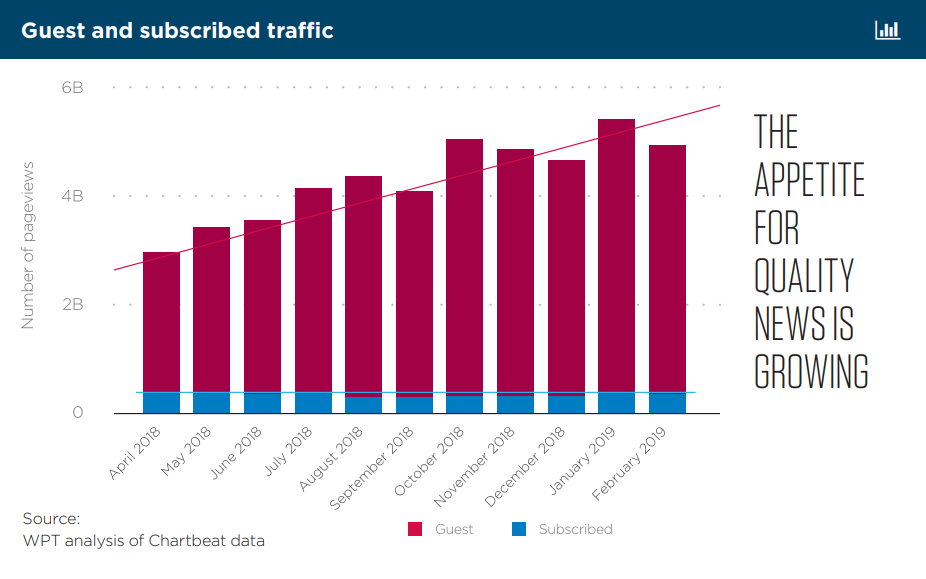
ABC News is the news division of Walt Disney Television's ABC broadcast network based out of New York, New York. Similar to BBC, ABC covers global and local news with a prioritization on news happening in the United States.

**Growth of Online News**

Digital news subscriber numbers worldwide have increased 208% over five years to 2018, and are expected to grow by a further 13% in 2019. While the nostalgia of the paper and print news still captures a demographic, the rate at which humans consume information has rapidly shifted. Waiting to get news from a paper and print article has the flair of waiting to watch a favorite TV show on tape recording when everyone around has already seen it and want to talk about it. Figures 2 and 3 demonstrate the advancement of news consumption with a clear prioritization on digital methods.



*Figure 2.*Growth of online web news consumption.



*Figure 3.*Growth of online web news consumption.

With the market demand and technological advancements being made, the abundance of news on the web continues to grow each day. This enhances the value of web scarping this data and wrangling it into key topic areas. It becomes logistically impossible to keep track of all the news by organized topics even with a devoted team trying to track it manually. As a result, the practical value topic modeling provides in this domain is clearly evident.

**Analysis**

**About the Data**

*Analysis 1: BBC News Data*

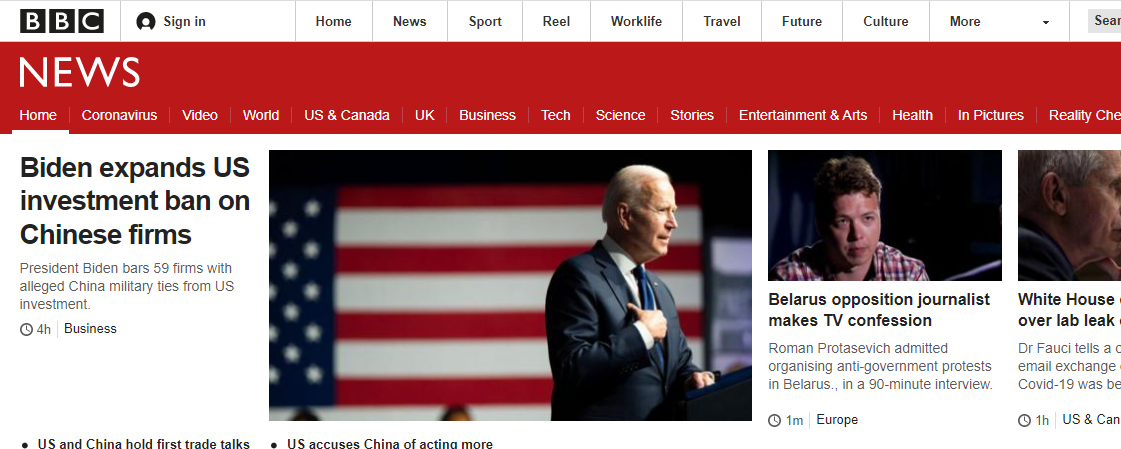
**Data Cleaning & Prep**

This section will detail the steps taken to clean and prepare the datasets for further modeling using pictures to walk through the before during and aftereffects of the data preparation process.

*Before Prep*

Prior to any preprocessing, as described, the dataset existed as a webpage with underlying text. Before conversion to a document term matrix it was read in as a json file. The dataset appears Figure 4.

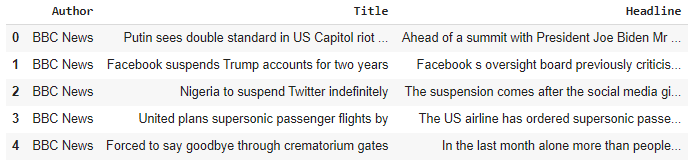
*Reading in the Data and Initial Preparation*



*Figure 4.*BBC headline news snippet.

*Reading in the Data and Initial Preparation*

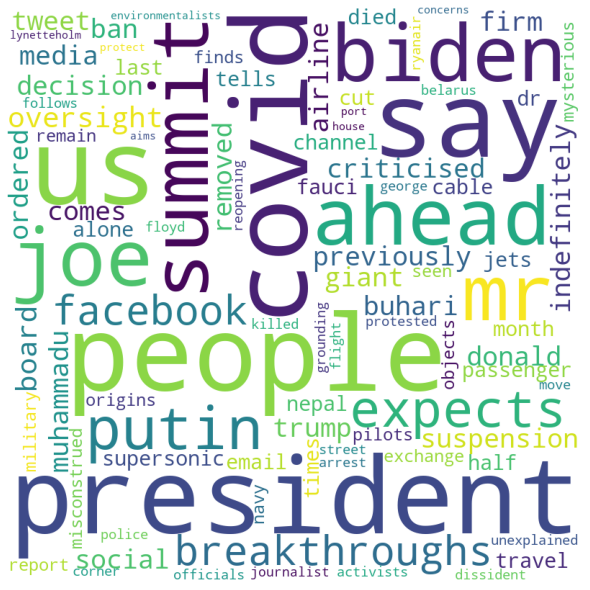
Data was imported into Python as a json file and converted to a data frame with columns as depicted in the Table 1 below.



*Table 1.* Simple structured data frame of BBC articles.

*Development of Word Clouds for Frequency*

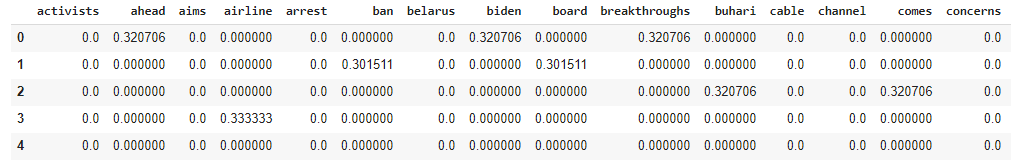
Once again, leveraging the nltk stopwords and custom words “go” and “going”, the text was stripped of common English words and tokenized. Each tokenized word was converted to lowercase for consistency. These words were then visualized in a word cloud to show frequency as demonstrated in Figure 5 on the next page.

****

*Figure 5.* The most commonly used words from BBC.com after removing stop words.

*Tf-idf Matrix*

The data was then passed into an empty list before tf-idf was used to transform the list based upon normalized word frequency. The array was then converted to a data frame as in analysis 1 and resulted in the following (see Table 2 for details).



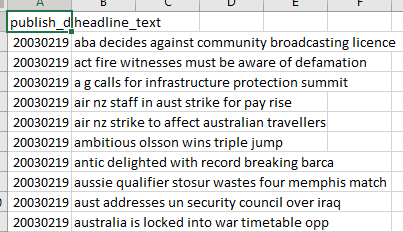
*Table 2.* Tf-idf data frame of BBC article headline words.

*Analysis 2: ABC News Data*

**Data Cleaning & Prep**

*Before Prep*

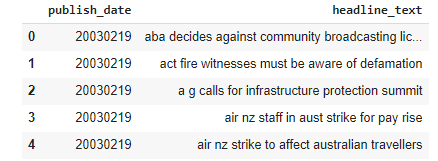
Prior to any preprocessing, as described, the dataset appears as below in Figure 6.



*Figure 6.* Example extract of the raw data file.

*Reading in the Data and Initial Preparation*

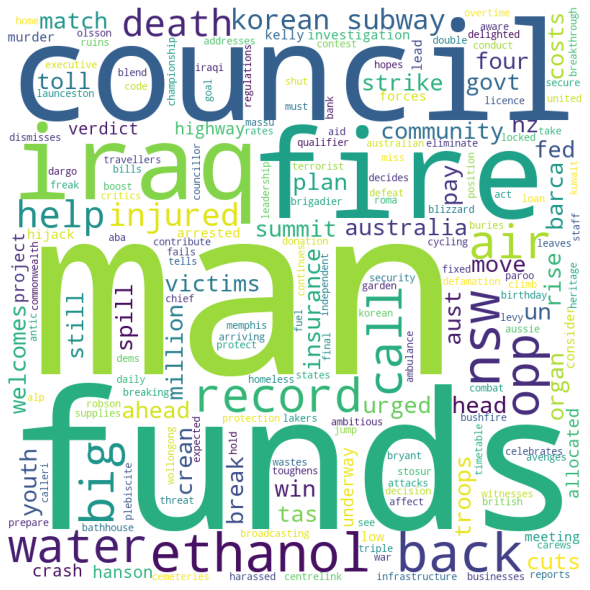
Data was imported into Python as a csv file and a data frame was created for the final clean version with columns as depicted in the Table 3 below.



*Table 3.* Simple structured data frame.

*Development of Word Clouds for Frequency*

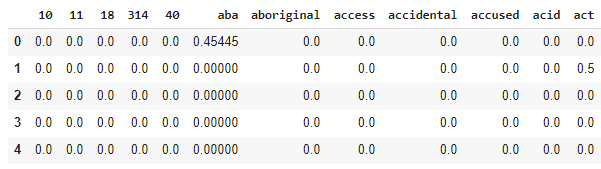
Once again, leveraging the nltk stopwords and custom words “go” and “going”, the text was stripped of common English words and tokenized. Each tokenized word was converted to lowercase for consistency. These words were then visualized in a word cloud to show frequency as demonstrated in Figure 7 on the next page.



*Figure 7.* The most commonly used words in the ABC news headlines after removing stop words.

*Tf-idf Matrix*

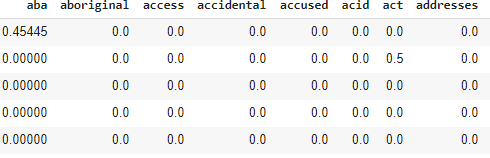
The unlabeled data was then passed into an empty list before CountVectorizer was used to transform the list based upon word occurrence. The array was then converted to a data frame as in analysis 1 and resulted in the following (see Table 4 for details).



*Table 4.* TF-idf demonstrates the normalization and stop word removal

*Tf-idf Featuring Numeric Removal*

An additional step was taken to convert the document term matrix into a normalized document term matrix using tf-idf transformation without numbers. This transformation takes the normalized value for which the word occurs in the text matrix (proportionality). Common English stop words were removed from this sparse data frame along with numbers. This conversion can be found in Table 5 below.



*Table 5.* TF-idf demonstrates the normalization and stop word removal.

**Results**

**LDA of BBC Headlines**

*Research Question 1:* What are the most popular topics on BBC news as defined by the headlines?

*Coherence Based Topic Modeling*

One of the most arduous tasks in LDA is defining the number of topics that make the most conceptual sense. Research indicates some science can be added to the art by using Jaccard similarity and coherence for each number of topics modeled. Coherence in this case measures a single topic by the degree of semantic similarity between high scoring words in the topic (do these words co-occur across the text corpus). This topic will be questioned as a best practice (and successful Python implementation) when meeting with the research team on June 6th, 2021.

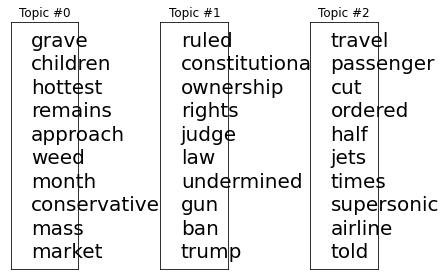
After the June 6th meeting, it was determined to move forward with model coherence as a determining factor for the number of topics to generate as a purely exploratory task. An initial LDA model coherence score of .53594 was achieved using ten topics. Coherence was improved to .629323 using a four-topic approach. Code was then utilized to determine the optimal coherence score adjusting alpha, beta and the number of topics. A learning here is that the execution time (`10 minutes) might not scale well with very large datasets and should be considered when computational constraints exist. A list of the best 5 configurations can be seen in Table 6 below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Validation\_Set** | **Topics** | **Alpha** | **Beta** | **Coherence** |
| 75% Corpus | 6 | 0.91 | 0.91 | 0.756872943 |
| 75% Corpus | 8 | 0.91 | 0.61 | 0.748411418 |
| 75% Corpus | 7 | 0.91 | 0.91 | 0.747848099 |
| 75% Corpus | 10 | 0.61 | 0.91 | 0.746711805 |
| 100% Corpus | 2 | 0.31 | 0.01 | 0.746441616 |

*Table 6.* Best coherence scores when adjusting alpha, beta and topic number.

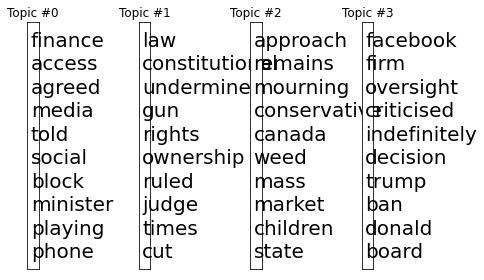
*Guess and Check Topic Modeling*

The first attempt looked at a collection of three topics as displayed in Figure 8. The 3-topic model demonstrated themes around the growing cannabis market and it’s relation to conservative ideologies. Embedded in this topic may be an unrelated issue of mass graves for children. This indicates adding more topics might help. The second topic created was around gun laws and changing policies with an administration shift from the Trump regime. Finally, the third topic is about the new jets being released which will cut travel times down significantly.



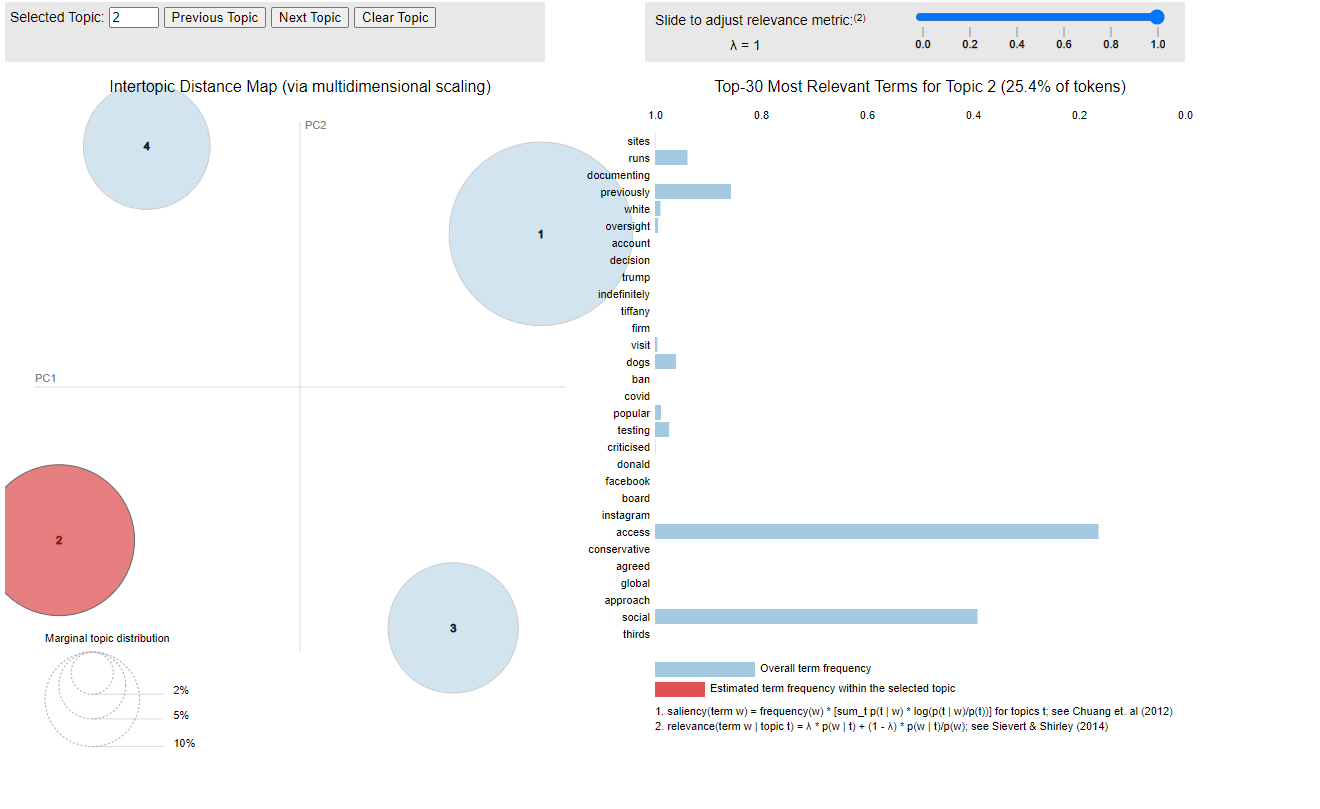
*Figure 8.* Three topic model example displaying key words.

Using six topics proved to be too many and it was decided that 4 topics provides author defined coherence in understanding the latest trends. Expanding upon the previous themes, it is clear that there are other themes such as blocking social media and Facebook’s controversial ban of Donald Trump.



*Figure 9.* Four topic model example displaying key words.

Additional code was leveraged in order to build app-like dashboard to view the top terms by model topic. This dashboard, as shown in Figure 10 provides an interactive way to look at the number of topics and the words that are most associated with them.



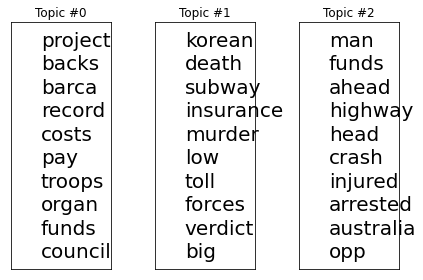
*Figure 10.* Four topic model dashboard example displaying key words.

**LDA of BBC Headlines**

*Research Question 2:* What are the most popular topics on ABC’s online news headlines?

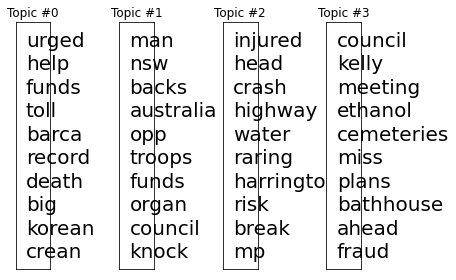
Using the methodology for topic modeling outlined in the analysis of research question 1, results defined in the subsequent section follow a similar thought process.

The first three topic approach demonstrates a theme around funds and costs related to troops, but the overall theme is not entirely clear. The second topic appears to be related to some misfortune in Korea involving a subway murder and the verdict behind it involving some level of insurance. Finally, the third topic appears to highlight a highway crash that occurred in Australia in which a man was injured and arrested.



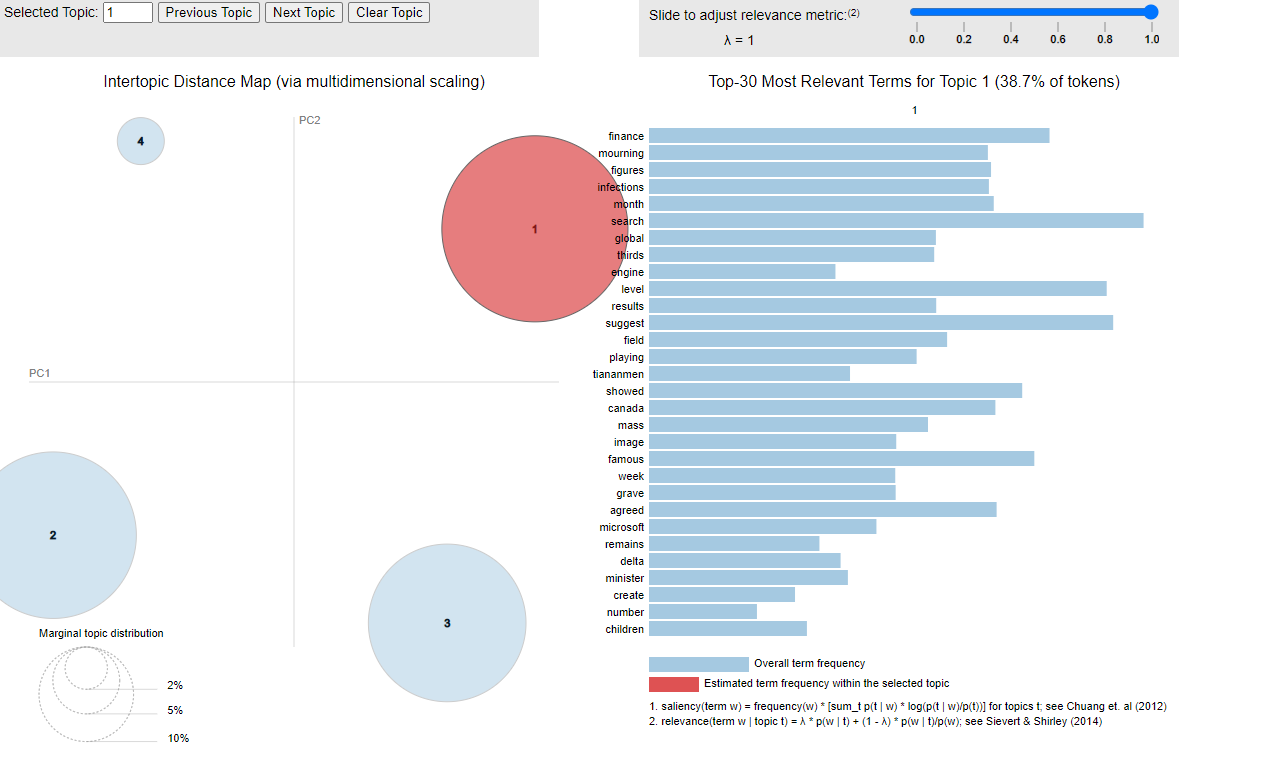
*Figure 11.* Three topic model example displaying key words.

A four-topic approach added some level of confusion with a noteworthy finding that the word “council appear at the top of topic 4, but also occurred within topic number 2. A number of other model scenarios were tested-however as the n-number of topics increases, it is not always a guarantee that the interpretability/coherence increase.



*Figure 11.* Four topic model example displaying key words.

Finally, the same code was leveraged to create a user dashboard interface displaying n-number of topics (here defined as 4).



*Figure 12.* Four topic model example displaying key words.

**Conclusion**

Major news channels publish anywhere from 250-1,200 articles per day online. As established, this trend has continued to increase in volume as the shift from paper news to digital news becomes increasingly evident. In fact, the number of daily news articles published online has risen 30% rise over the past decade ([The Atlantic](https://www.theatlantic.com/technology/archive/2016/05/how-many-stories-do-newspapers-publish-per-day/483845/)).

Sifting through hundreds of articles for even one major publisher to establish themes could be an overwhelming task even if someone devoted their entire existence to the cause. Multiply this time commitment across multiple news platforms, blog posts and the scale soon exceeds possibility or reasonability for any person of team or persons to pursue. When considering the volume of *historical* posts available, another means is clearly required to establish core topics over any length of time.

Topic modeling provides that solution for researchers and practitioners by creating thematic representations of large quantities of text. This research provided evidence for the utilization of topic modeling for news article themes. The application of topic modeling extends to areas such as company/customer reviews, employee survey feedback and stock market trading blogs. Choosing the right number of topics can be an iterative process whether using a computation to determine the right number or purely using domain knowledge to guess and check into the most coherent depiction of the text.

While topic modeling requires a mixture of art and science it can provide tremendous insight translating a potentially untapped resource of rich data for businesses and researchers alike.