Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

TEAM # 70

TEMPORAL EVOLUTION TOPIC MODELING VIA NON-NEGATIVE CP TENSOR DECOMPOSITION

PROBLEM STATEMENT:

Over the past few years, due to ease of internet access and the advent of various social media outlets like blogs, social network sites like twitter etc., we find ourselves amid an explosion of electronic document archives. This has led to the requirement of developing new tools using statistical and computational methods to turn this raw data into invaluable scientific and business insights. Dynamic topic modelling is one such method developed to automatically organize large document collections and determine/discover hidden pattern of words in the documents.
Dynamic topic modeling has been widely used in social and data sciences with the goal of learning latent topics that emerge, evolve, and fade over time. Currently there are several methods to tackle the problem of topic modeling; most of these methods ignored the temporal aspect in many document collections. The setback of this approach is that majority of these models are applied multiple times to extract topic evolution over a temporal domain. This iteration could lead to high variability in results as well as the interpretability of said results. Recently, analytical methods involving high order tensors and their decomposition like CP Tensor decomposition have been gaining in popularity. This is due to their ability to extract unique latent variables of a large dataset simultaneously in multiple dimensions. Due to constraints naturally imposed on the modeled data which results in the higher order tensor, the decomposed latent factors model important mixing information across all dimensions of the n_{th} order tensor.

OBJECTIVE:

Using dynamic topic modeling, we are going to identify latent topics using the dataset comprising news articles and essays and understand how different topics emerge, evolve and fade over time.
We will demonstrate not only the speed and efficiency of this method since we are working with a huge dataset, but its accuracy in determining the major topic from every month for each year.

DATA SOURCES AND DATA EXTRACTION:

2.7 million News articles and essays from 27 American publications ranging from 2016 to 2020. The source of our data is https://components.one/datasets/all-the-news-2-news-articles-dataset/ and following are the details of size of articles we are pulling each year with 138,547 Authors.

- 2016 – 604,511 articles
- 2017 – 640,493 articles
- 2018 – 553,588 articles
- 2019 – 655,456 articles
- 2020 – 234,830 articles

All the articles are taken from the diverse news sources like Vox, Business Insider, Reuters, TMZ, Vice, Vice News, Hyperallergic, TechCrunch, Axios, Refinery 29, The Verge, Mashable, People, Economist, CNN, Gizmodo, New Yorker, CNBC, Wired, New Republic, Fox News, The Hill, Politico, The New York Times, Buzzfeed News, Washington Post.

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

DATA PREPARATION:

We have documents of the size ~8 GB before any preprocessing. In order to do any further processing in a computationally efficient manner, we have employed Multiprocessing using various tools including AWS S3, DynamoDB, and SPARK etc. In order to ready, the data for analysis following steps were executed:

- The dataset (csv file) is initially stored in an AWS S3 bucket and it is then loaded into a Spark Cluster using AWS EMR services for further processing.
- Several NLP preprocessing procedures are applied to clean up the data. Following are the main Spark NLP functions we have used in the code to process text:
  - DocumentAssembler (): To get through the process in Spark NLP, we need to get raw data transformed into Document type at first. DocumentAssembler () is a special transformer that does this for us; it creates the first annotation of type Document which may be used by annotators down the road.
  - Tokenizer (): Tokenization is just the term used to describe the process of converting the normal text strings into a list of tokens (words that we want). It converts input text to streams of tokens, where each token is a separate word, punctuation sign, number/amount, date, e-mail, URL/URI, etc. The text output is saved to a new column called 'token'.
  - Normalizer (): Remove punctuation, numbers, and symbols from text from tokens and create the output to a new column called 'normalized'.
  - LemmatizerModel.pretrained (): To get the root of each word by performing lemmatization and create the output to a new column called 'lemma'. Words that have the same meaning, but different endings have been adjusted so that they all have the same form. For example, "include", "includes," and "included," would all be represented as "include." Lemmatization is closely related to stemming. The difference is that a stemmer operates on a single word without knowledge of the context, and therefore cannot discriminate between words which have different meanings depending on part of speech
  - StopWordsCleaner (): Used to remove all, stopwords from lemma column and create the output to a new column called 'clean_lemma'. A separate list created with all the 'stopwords' to be removed.
  - Finisher (): Finisher converts tokens to human-readable output.

- Once the data is cleaned, it is written back to AWS S3 as several partitions in the form of csv files.

- Working with the csv as is not computationally efficient and therefore we created a "Table" in AWS' NoSQL resource, DynamoDB. In DynamoDB, tables, items, and attributes are the core components that we worked with. A table is a collection of items, and each item is a collection of attributes. Furthermore, there are no primary keys, foreign keys, etc. here like there are for SQL type databases. Instead, it is a NoSQL database which uses partitions and sort keys. The partition acts as a type of a primary key in that it must be unique. However, more importantly, the data is "partitioned" into 10GB clusters based off the partition keys. From there, sort keys separate the data further.
- The data from the S3 bucket files are uploaded to the Dynamo dB as documents. For our schema, we have made it so that the years are the partition key, and then the sort keys are a combination of **[month].[day].[uid]** .For example, a random article from January 20th, 2019 would be encoded like:

Partition key	Sort Key
2019	1.20.342320

- The **[uid]** part in the sort key comes from a sequentially generated number from the original dataset. The first row in the data was given the value of 0, the second row the value of 1, the 380th row the value of 379, etc. There was about 2.5 million rows of data, so the [uid] number in the sort key will be some number between 0 and 2.5 million. All the queries in the backend are happening only dependent on the year and the first two values of the sort key.

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

- Read/Write in the DynamoDB is not free, therefore we created procedures to cache UIDs by Year, Month or Day and load them as needed when running the program.

Using above steps, we were able to process GBs of data within seconds on our own machines.

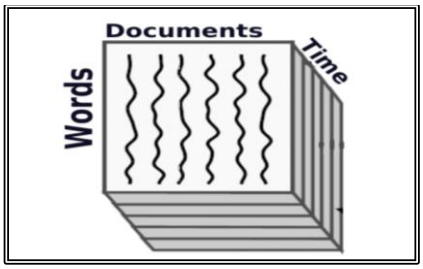
METHODOLOGY FOR DYNAMIC TOPIC MODELLING USING CP TENSOR DECOMPOSITION:

Topic modeling is a technique originally developed to discover the latent structure in large collections of text documents. A topic modeling algorithm processes a representation of each document in a corpus and produces a topic model that captures clusters of similar words as topics. One of the most commonly used is Latent Dirichlet Allocation (LDA) algorithm. In LDA, a document is modeled as a high-dimensional vector where each entry represents the count of a particular word in the document. Here, each document is assumed to be characterized by a particular set of topics. This is similar to the standard bag of words model assumption, and makes the individual words exchangeable. In LDA, both the order the words appear in a document and the order the documents appear in the corpus are oblivious to the model.

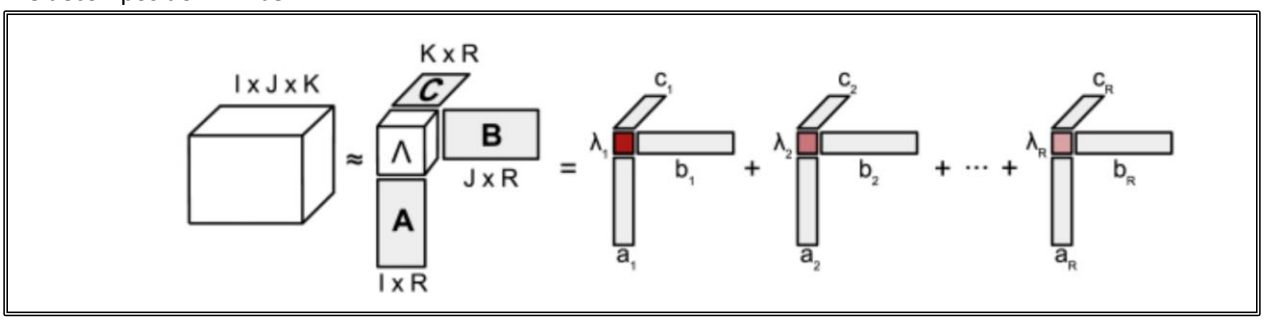
With this approach, therefore, information contained in the temporal dimension of the data is often neglected or underutilized. In a dynamic topic models the order of the documents plays a fundamental role. More precisely, the documents are grouped by time slice (e.g.: months or years) and it is assumed that the documents of each group come from a set of topics that evolved from the set of the previous slice.

We are using the method of nonnegative CANDECOMP/PARAPAC (CP) tensor decomposition (NNCPD), where the data tensor is directly decomposed into a minimal sum of outer products of nonnegative vectors, thereby preserving the temporal information.

Tensors (or multidimensional arrays) are a natural fit for representing data with multiple associated attributes. Tensor decompositions are a valuable, mathematically sound set of tools for exploratory analysis of multi-dimensional data and for capturing underlying multidimensional relationships. Tensor decompositions separate input data into patterns called components. Each component represents a latent behavior or correlation from within the dataset. For the dataset used in this project the tensor structure would be:



The decomposition will be:



Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

Here:

- A = Factor matrix of Word Importance vs Topics
- B = Factor matrix of Documents vs Topics
- C = Factor matrix of Months vs Topics

Tensor decomposition are a class of algorithm that are used for unsupervised pattern identification and, therefore this separation into components occurs without training.

The CP decomposition decomposes a tensor into a sum of a predefined number of component tensors. Each component consists of a vector of scores for each dimension of the original data, with one score in the vector for each element of the dimension. A scalar weight (lambda) term associated with each component captures the relative prominence of that component pattern in the original data. Entries within and across dimensions that score highly in a single component are correlated. At a high level, the method of decomposition involves a constrained model-fitting algorithm that proceeds by gradient descent.

The underlying mathematics of CP tensor decomposition using Alternating least squares can be explained as:

Algorithm 1 CP-ALS

function $[\boldsymbol{\lambda}, \{\mathbf{A}^{(n)}\}] = \text{CP-ALS}(\mathcal{X}, R)$ ▷ $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$
  Initialize factor matrices $\mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(N)}$
  repeat
    for $n = 1, \ldots, N$ **do**
      $\mathbf{V} \leftarrow \mathbf{A}^{(N)\mathsf{T}}\mathbf{A}^{(N)} \circledast \cdots \circledast \mathbf{A}^{(n+1)\mathsf{T}}\mathbf{A}^{(n+1)} \circledast \mathbf{A}^{(n-1)\mathsf{T}}\mathbf{A}^{(n-1)} \circledast \cdots \circledast \mathbf{A}^{(1)\mathsf{T}}\mathbf{A}^{(1)}$
      $\mathbf{Z}^{(n)} \leftarrow \mathbf{A}^{(N)} \odot \cdots \odot \mathbf{A}^{(n+1)} \odot \mathbf{A}^{(n-1)} \odot \cdots \odot \mathbf{A}^{(1)}$
      $\mathbf{W} \leftarrow \mathbf{X}_{(n)}\mathbf{Z}^{(n)}$
      Solve $\mathbf{A}^{(n)}\mathbf{V} = \mathbf{W}$ for $\mathbf{A}^{(n)}$
      Normalize columns of $\mathbf{A}^{(n)}$ and update $\boldsymbol{\lambda}$
    end for
  until termination criteria met
  return $\boldsymbol{\lambda}$, factor matrices $\{\mathbf{A}^{(n)}\}$
end function

Here, ʘ stands for outer product, λ is the weight vector , ⊛ stands for the "Khatri-Rao" product and 'R' is the rank(~ no. of topics).

A high-level overview of CP Tensor Decomposition Algorithm is as follows:

- For a given year sample n-documents from each month: Each document is viewed as a mixture of unobserved topics. Furthermore, each topic defines a multinomial distribution over a set of terms. Thus, for each word of each document, a topic is drawn from the mixture and a term is subsequently drawn from the multinomial distribution corresponding to that topic
- Vectorize all unique words from the sampled documents using TfidfVectorizer.
  - TfidfVectorizer: TF-IDF stands for **"Term Frequency — Inverse Document Frequency"**. This is a technique to quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus. This method is a widely used technique in Information Retrieval and Text Mining.
- Create a 3rd order tensor of dimensions (Words x Documents x Months) for the given year
- Perform CP Tensor Decomposition to find the latent features of the tensor
  - Latent Matrix (Words vs Topics)
  - Latent Matrix (Documents vs Topics)
  - Latent Matrix (Months vs Topics)

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

- Analyze how the different topics evolve over time using the Months vs Topics graph. Also get the list of words each topic consists of.

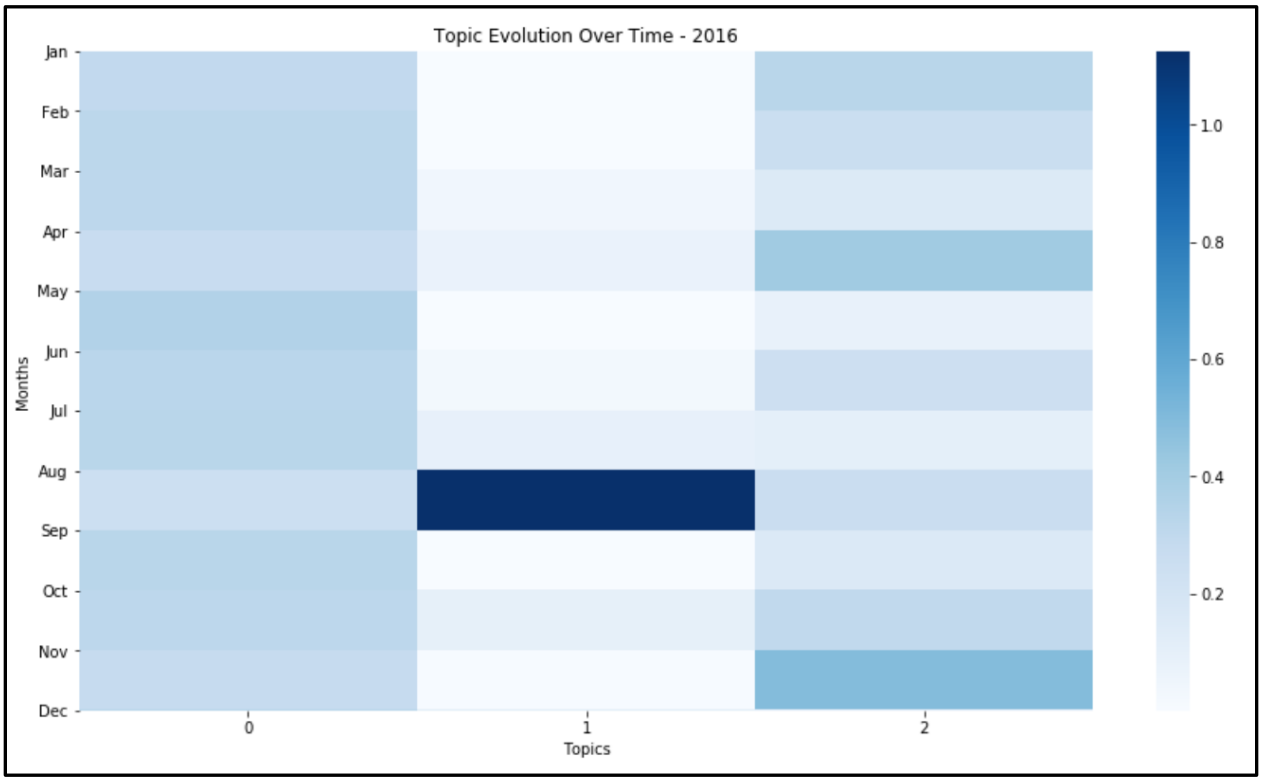
ANALYSIS AND RESULTS:

Topic Evolution Graphs

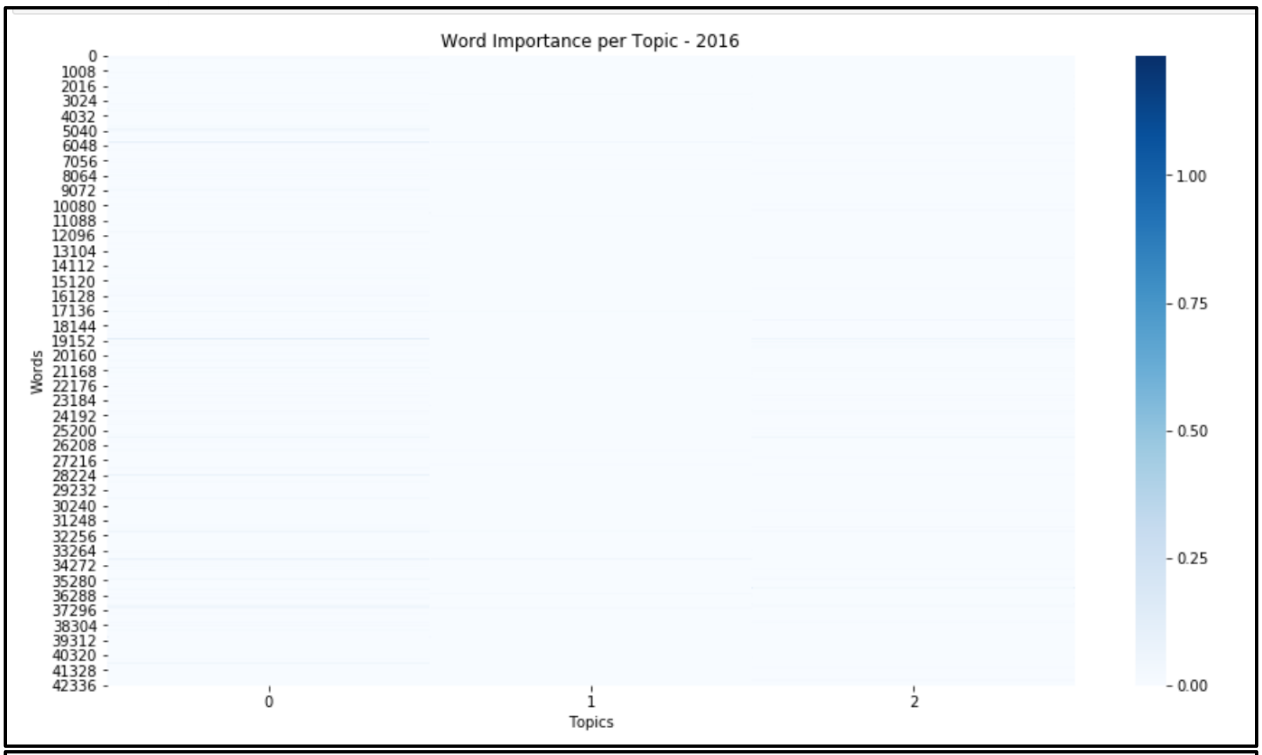
We created heatmaps to analyze the topic evolution throughout the year. The map is created using the topics generated on X-axis vs the months on Y-axis. The darker the gradient, the dominant is the topic in that month meaning the topic was an important news item and was present in several sampled articles. The change in the color gradient from light to dark signify how the topic is fading in and out of news.
Similarly, we have also used heatmaps to demonstrate which topic is present in what sampled document. Again, the gradient explains how dominant a particular topic is in that document. This can be further used to label documents depending on the topic composition, for ex. If the topic comprises words about the financial news like stocks, currencies etc. then we can group such document as such and then later use this for clustering the documents. We are not exploring this avenue of analysis as part of this project though.

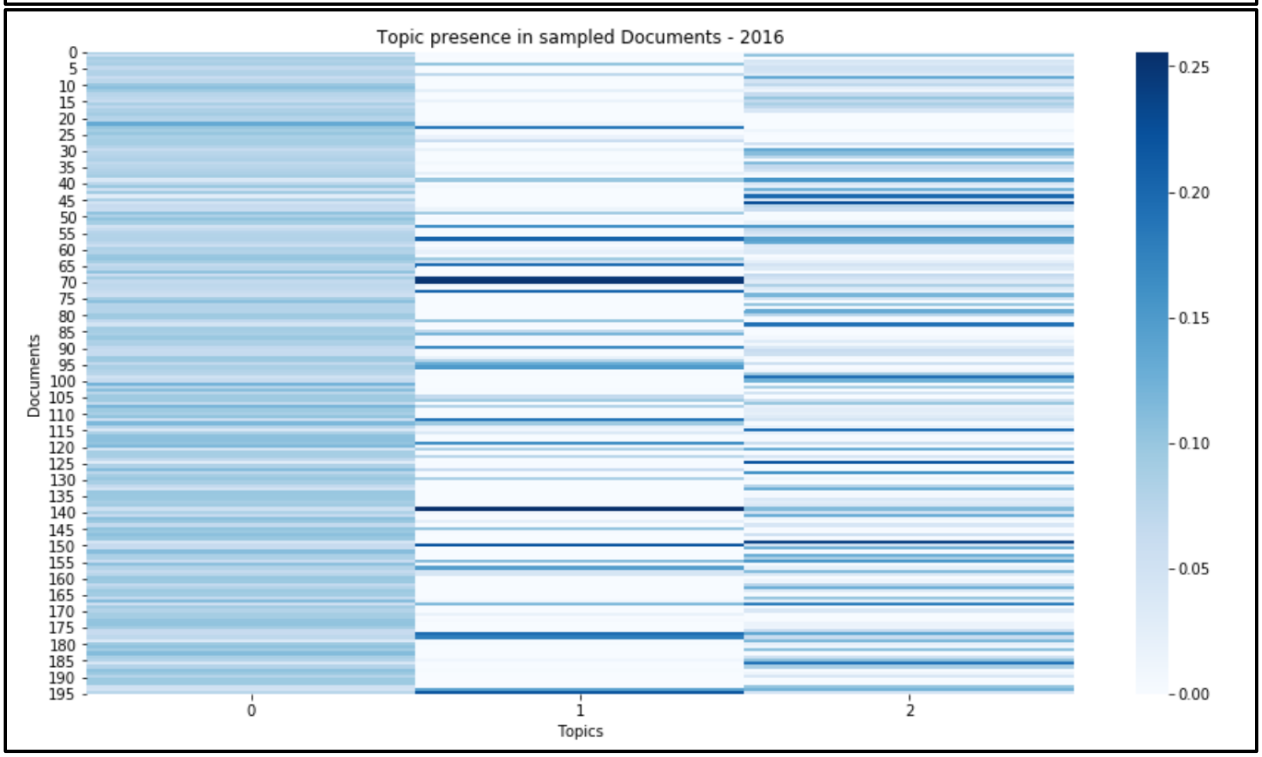
For the Year 2016:

The following graph shows the topic evolution throughout the year 2016. Topic 0 here is made up breaking news headlines each month and therefore is a medley of words.



Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)





Word per topic:

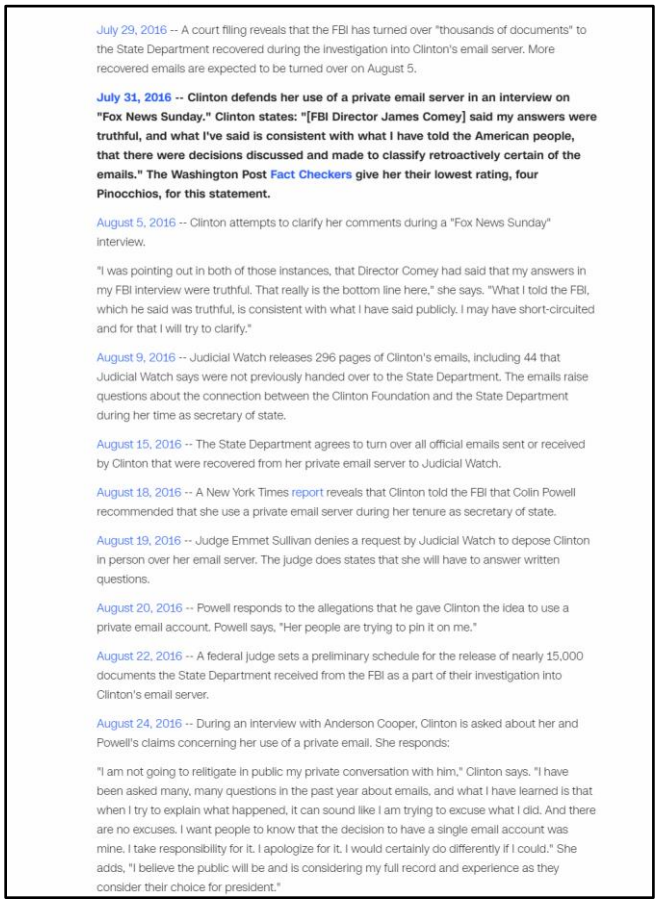
Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

```
Topic 0: ['york' 'year' 'write' 'world' 'work' 'woman' 'win' 'white' 'well' 'week'
 'wednesday' 'washington' 'vote' 'view' 'video' 'united' 'two' 'twitter'
 'turn' 'tuesday' 'trump' 'trade' 'time' 'thursday' 'three' 'text' 'tell'
 'technology' 'team' 'talk' 'take' 'system' 'support' 'subsidiary'
 'street' 'story' 'statement' 'state' 'start' 'star' 'something' 'small'
 'site' 'show' 'shoot' 'share' 'set' 'service' 'series' 'senate' 'sell'
 'see' 'second' 'season' 'sanders' 'run' 'rule' 'right' 'reuter' 'result'
 'republican' 'report' 'release' 'real' 'rate' 'raise' 'question'
 'publish' 'public' 'price' 'presidential' 'president' 'post' 'political'
 'policy' 'police' 'point' 'play' 'plan' 'place' 'photo' 'percent'
 'people' 'pay' 'party' 'part' 'open' 'one' 'official' 'offer' 'obama'
 'night' 'next' 'news' 'new' 'never' 'need' 'national' 'name' 'much'
 'move' 'month' 'money' 'monday' 'million' 'might' 'member' 'meet' 'media'
 'mean' 'may' 'market' 'many' 'man' 'make' 'low' 'love' 'lot' 'look'
 'long' 'live' 'line' 'life' 'let' 'leave' 'lead' 'law' 'late' 'last'
 'kill' 'keep' 'june' 'july' 'job' 'issue' 'interest' 'increase' 'include'
 'house' 'home' 'hold' 'hit' 'hill' 'high' 'help' 'health' 'happen' 'gun'
 'group' 'great' 'government' 'good' 'global' 'give' 'game' 'fund'
 'friend' 'friday' 'four' 'former' 'follow' 'first' 'find' 'film' 'feel'
 'federal' 'family' 'fall' 'face' 'expect' 'every' 'event' 'even' 'euro'
 'end' 'election' 'edit' 'economy' 'early' 'donald' 'discussion'
 'democratic' 'democrat' 'deal' 'dc' 'day' 'create' 'court' 'country'
 'corp' 'control' 'continue' 'content' 'company' 'communication' 'comment'
 'come' 'close' 'clinton' 'city' 'child' 'change' 'case' 'car' 'capitol'
 'candidate' 'campaign' 'call' 'buy' 'business' 'build' 'bring' 'billion'
 'bill' 'big' 'believe' 'begin' 'bank' 'bad' 'back' 'attack' 'ask'
 'around' 'apple' 'another' 'announce' 'american' 'allow' 'add' 'accord']

Topic 1: ['woman' 'voter' 'vote' 'trump' 'tax' 'state' 'republican' 'presidential'
 'poll' 'nominee' 'hillary' 'hill' 'gop' 'emails' 'election' 'donald'
 'democrat' 'debate' 'clinton' 'campaign']

Topic 2: ['year' 'text' 'stock' 'share' 'sales' 'rise' 'revenue' 'reuter' 'report'
 'rate' 'quarter' 'price' 'percent' 'per' 'oil' 'newsroom' 'million' 'may'
 'market' 'index' 'ibes' 'high' 'group' 'first' 'financial' 'fall' 'euro'
 'earnings' 'company' 'chinese' 'china' 'bengaluru' 'bank' 'april']
```

For the year 2016, we got a lot of overlapping topics therefore, we reduced the number of topics to get rid of the overlapping words. Now we have 3 topics each relating to particular set of news.

Topic 1 has words like ‘Hilary’,’gop’,’emails’ and is in and out of news throughout the year. This is expected since its clear that it is about the US election which was a major news item in 2016.

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)



July 29, 2016 -- A court filing reveals that the FBI has turned over "thousands of documents" to the State Department recovered during the investigation into Clinton's email server. More recovered emails are expected to be turned over on August 5.

July 31, 2016 -- Clinton defends her use of a private email server in an interview on "Fox News Sunday." Clinton states: "[FBI Director James Comey] said my answers were truthful, and what I've said is consistent with what I have told the American people, that there were decisions discussed and made to classify retroactively certain of the emails." The Washington Post Fact Checkers give her their lowest rating, four Pinocchios, for this statement.

August 5, 2016 -- Clinton attempts to clarify her comments during a "Fox News Sunday" interview.

"I was pointing out in both of those instances, that Director Comey had said that my answers in my FBI interview were truthful. That really is the bottom line here," she says. "What I told the FBI, which he said was truthful, is consistent with what I have said publicly. I may have short-circuited and for that I will try to clarify."

August 9, 2016 -- Judicial Watch releases 296 pages of Clinton's emails, including 44 that Judicial Watch says were not previously handed over to the State Department. The emails raise questions about the connection between the Clinton Foundation and the State Department during her time as secretary of state.

August 15, 2016 -- The State Department agrees to turn over all official emails sent or received by Clinton that were recovered from her private email server to Judicial Watch.

August 18, 2016 -- A New York Times report reveals that Clinton told the FBI that Colin Powell recommended that she use a private email server during her tenure as secretary of state.

August 19, 2016 -- Judge Emmet Sullivan denies a request by Judicial Watch to depose Clinton in person over her email server. The judge does states that she will have to answer written questions.

August 20, 2016 -- Powell responds to the allegations that he gave Clinton the idea to use a private email account. Powell says, "Her people are trying to pin it on me."

August 22, 2016 -- A federal judge sets a preliminary schedule for the release of nearly 15,000 documents the State Department received from the FBI as a part of their investigation into Clinton's email server.

August 24, 2016 -- During an interview with Anderson Cooper, Clinton is asked about her and Powell's claims concerning her use of a private email. She responds:

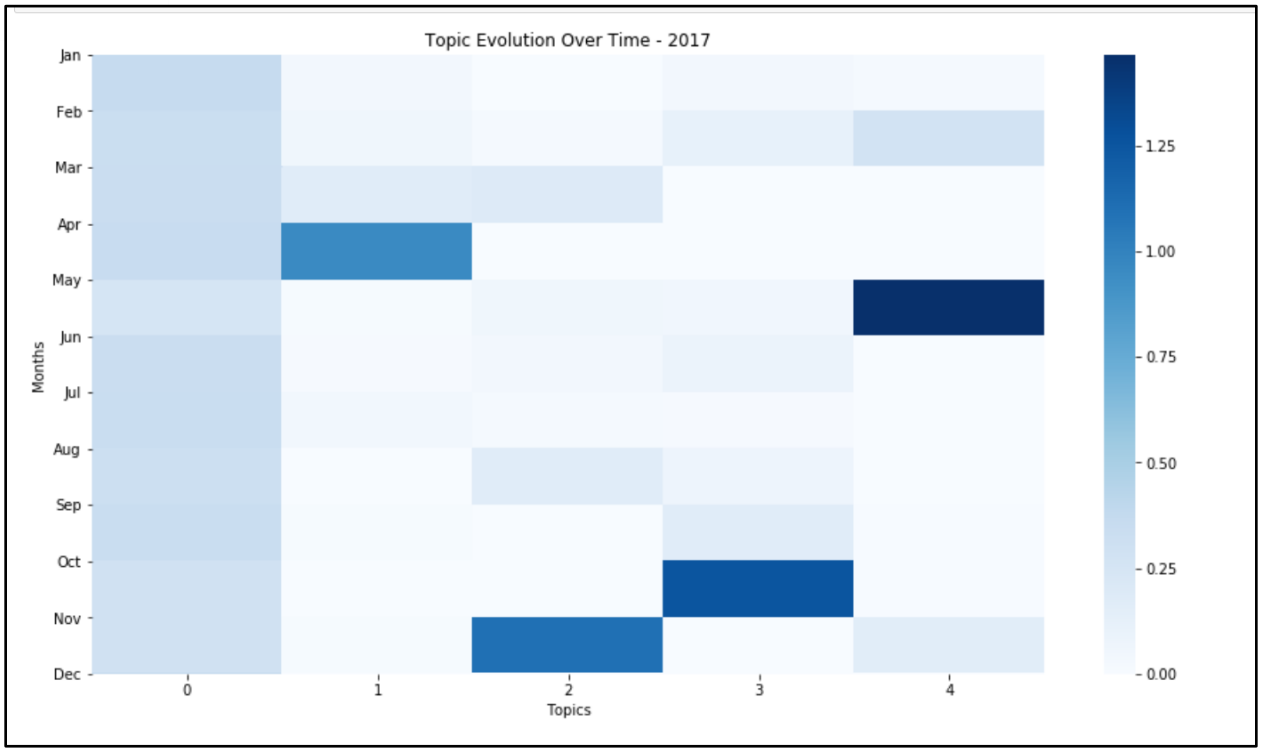
"I am not going to relitigate in public my private conversation with him," Clinton says. "I have been asked many, many questions in the past year about emails, and what I have learned is that when I try to explain what happened, it can sound like I am trying to excuse what I did. And there are no excuses. I want people to know that the decision to have a single email account was mine. I take responsibility for it. I apologize for it. I would certainly do differently if I could." She adds, "I believe the public will be and is considering my full record and experience as they consider their choice for president."

Topic 3 in turn comprises all the financial news and is dominant throughout the year.

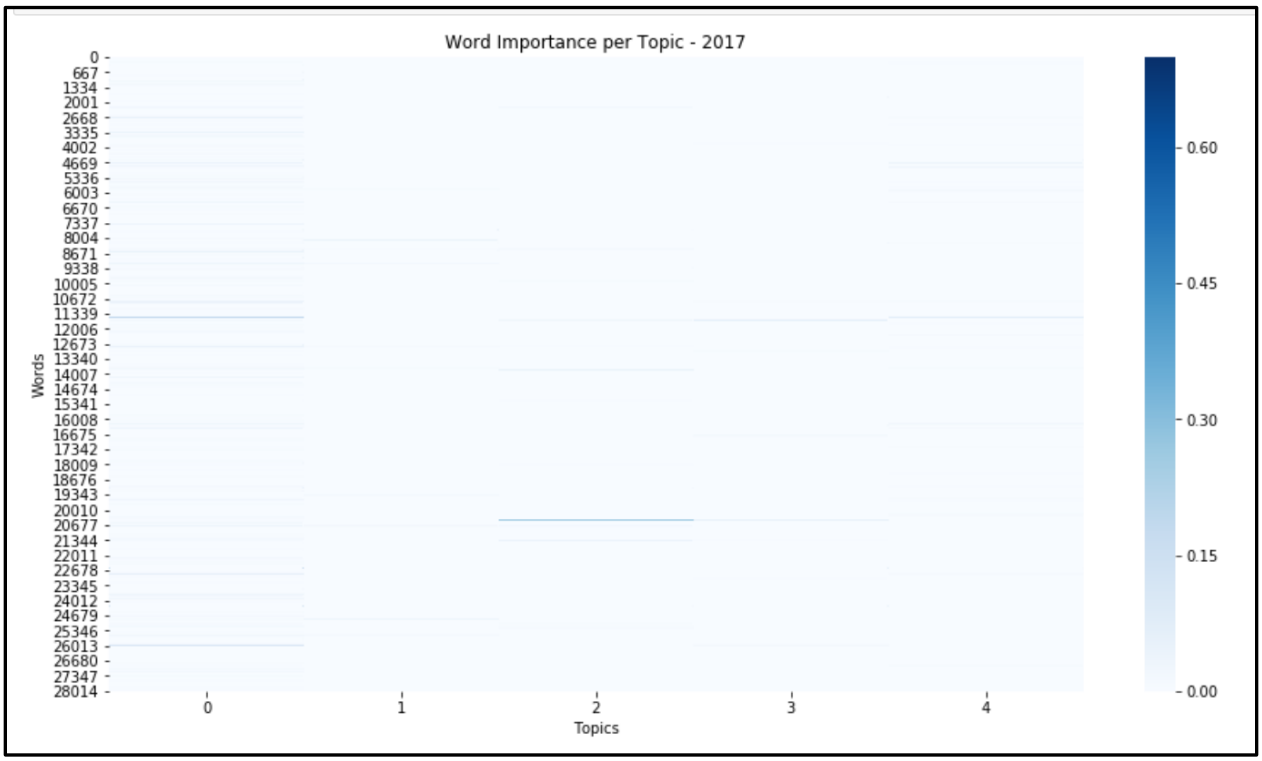
For Year 2017:

Topic evolution for this year is visualized as below:

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

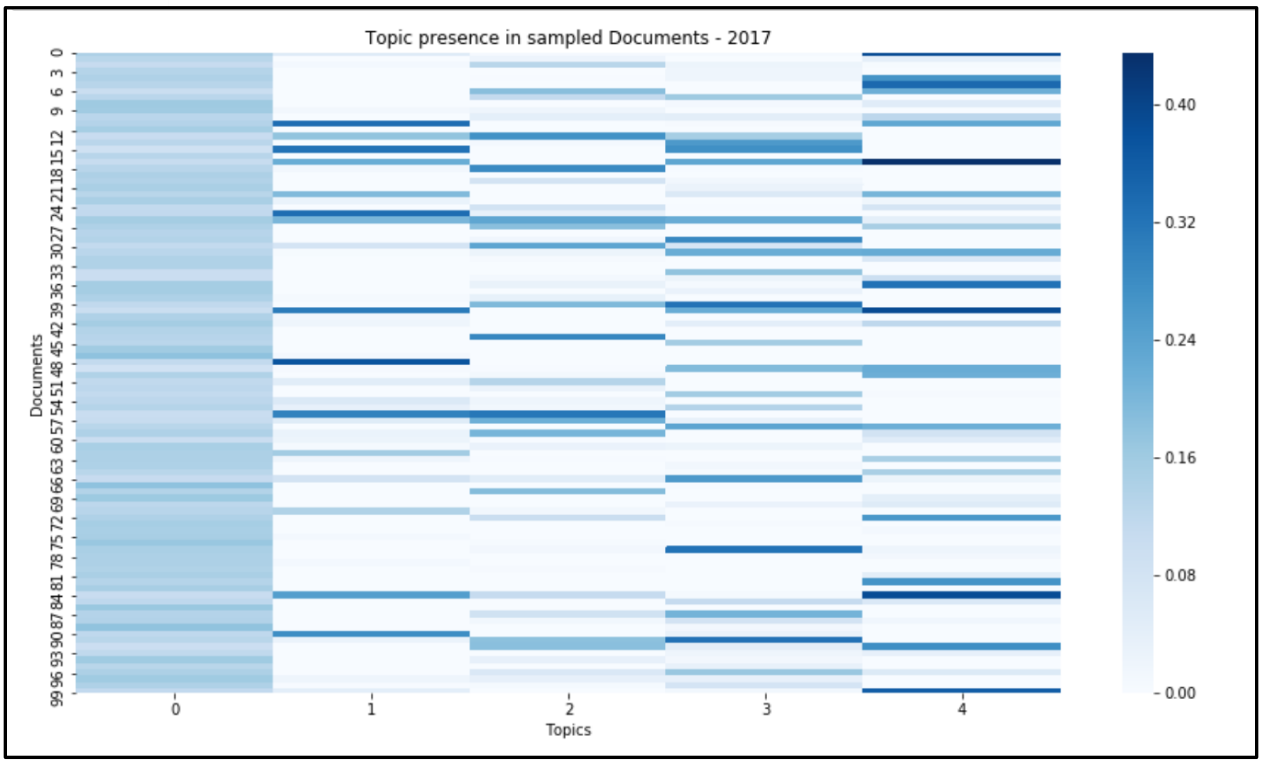


Words importance per topic:



Graph to visualize the Topic presence in the sampled document:

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)



Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

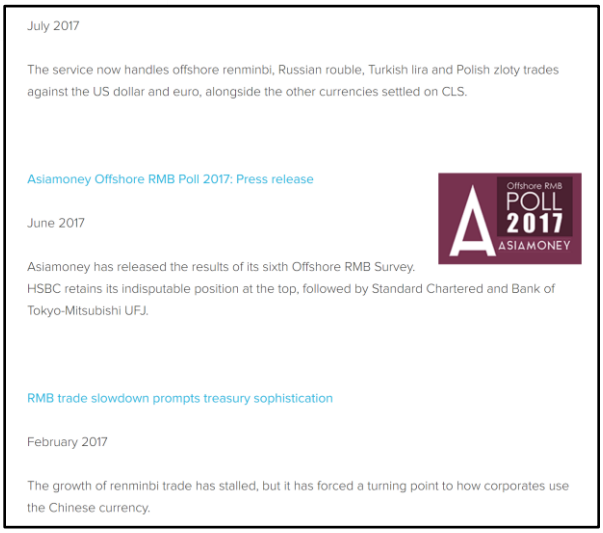
Words per topic:

```
Topic 0: ['york' 'year' 'write' 'world' 'work' 'woman' 'win' 'white' 'well' 'week'
 'wednesday' 'watch' 'washington' 'vote' 'view' 'video' 'united' 'two'
 'tuesday' 'trump' 'trade' 'today' 'time' 'thursday' 'three' 'text' 'tell'
 'team' 'tax' 'talk' 'take' 'support' 'street' 'story' 'stock' 'statement'
 'state' 'start' 'star' 'speak' 'site' 'sign' 'show' 'share' 'set'
 'service' 'senate' 'see' 'security' 'second' 'russian' 'russia' 'run'
 'rule' 'rise' 'right' 'reuter' 'republican' 'report' 'release' 'record'
 'rate' 'question' 'public' 'price' 'president' 'power' 'post' 'policy'
 'police' 'point' 'play' 'plan' 'place' 'percent' 'people' 'pay' 'party'
 'part' 'order' 'open' 'one' 'official' 'office' 'number' 'north' 'next'
 'news' 'new' 'need' 'national' 'name' 'much' 'move' 'month' 'money'
 'monday' 'minister' 'million' 'member' 'meet' 'media' 'may' 'market'
 'many' 'man' 'make' 'love' 'lot' 'look' 'long' 'life' 'leave' 'leader'
 'lead' 'law' 'late' 'last' 'keep' 'june' 'july' 'job' 'issue'
 'investigation' 'international' 'interest' 'increase' 'include' 'house'
 'home' 'hold' 'hill' 'high' 'help' 'health' 'happen' 'group' 'government'
 'good' 'give' 'game' 'fund' 'friday' 'former' 'force' 'follow' 'first'
 'find' 'financial' 'file' 'fight' 'feel' 'federal' 'family' 'face'
 'expect' 'executive' 'every' 'even' 'end' 'election' 'edit' 'early'
 'department' 'democrat' 'deal' 'day' 'court' 'country' 'cost' 'corp'
 'continue' 'company' 'comment' 'come' 'close' 'city' 'chinese' 'child'
 'chief' 'change' 'case' 'care' 'capital' 'call' 'business' 'build'
 'black' 'billion' 'bill' 'big' 'begin' 'bank' 'back' 'attack' 'ask'
 'around' 'april' 'app' 'another' 'announce' 'american' 'agreement'
 'administration' 'add' 'accord']

Topic 1: ['zloty' 'yuan' 'yoc' 'versus' 'text' 'swedish' 'stockholm' 'sek' 'sec'
 'sa' 'revenue' 'reuter' 'presumably' 'prelim' 'placement' 'newsroom'
 'net' 'natera' 'million' 'march' 'increase' 'hubstyle' 'headline'
 'gsmatt' 'googlrgzthn' 'euro' 'dgc' 'company' 'chinese' 'beijing'
 'bambora' 'ag']

Topic 2: ['yogaworks' 'thomson' 'text' 'shekel' 'share' 'second' 'sales' 'rgt'
 'revenue' 'reuter' 'quarter' 'qtrly' 'profit' 'percent' 'per' 'oibda'
 'net' 'million' 'loss' 'ibes' 'hy' 'healthcare' 'earnings' 'dinar'
 'company' 'cent' 'biosystems' 'bhd' 'aug' 'attributable' 'approximately'
 'adjust']

Topic 3: ['zall' 'yen' 'versus' 'total' 'text' 'share' 'september' 'revenue'
 'reuter' 'repurchase' 'quarter' 'per' 'offer' 'nov' 'newsroom' 'million'
 'ibes' 'eur' 'earnings' 'company' 'bmtc' 'billion' 'ag']

Topic 4: ['trump' 'senate' 'president' 'house' 'health' 'care' 'bill']
```

Looking at Topic 1, we see it relates to financial news spanning different entities. It is dominant during the month of March, April which is the end of fiscal year when lot of companies release the fiscal year results.

If we do search we will see there were news articles relating to 'Yuan', 'Euro','Zolty' which are different currencies:

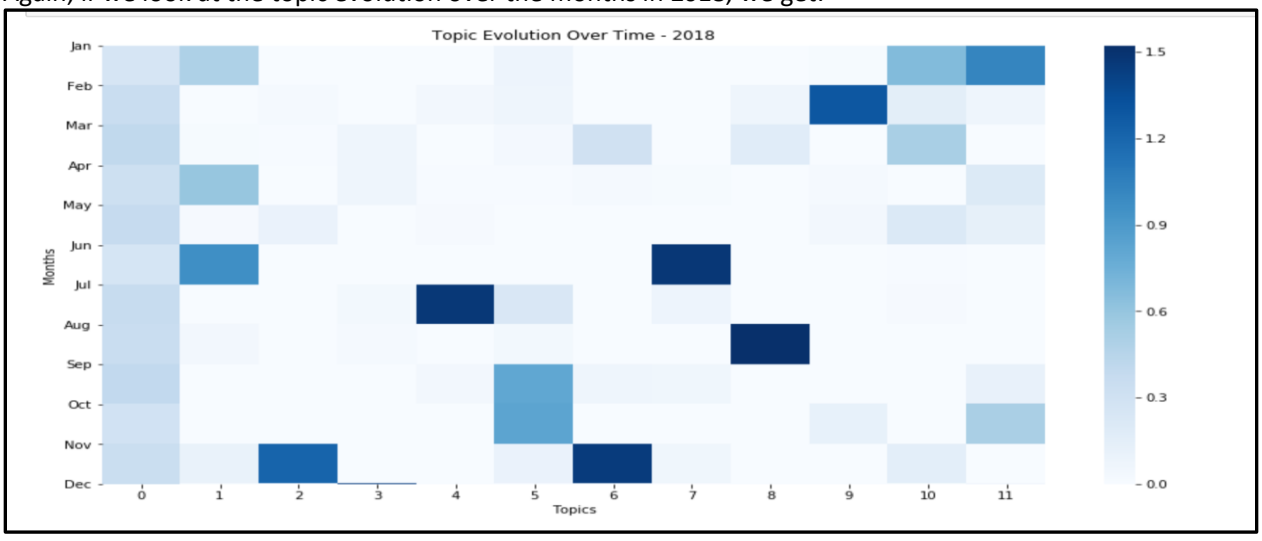
Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)



Similarly, Topic 4 is related to 'Health care' bill and was dominant is the month of May 2017:

For Year 2018:

Again, if we look at the topic evolution over the months in 2018, we get:



Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

The topic 0 is a medley of words and is dominant throughout the year. This is mainly because it comprises of the major “breaking news” headlines. For other topics we can see the months where each topic was dominant and how it fades in and out.

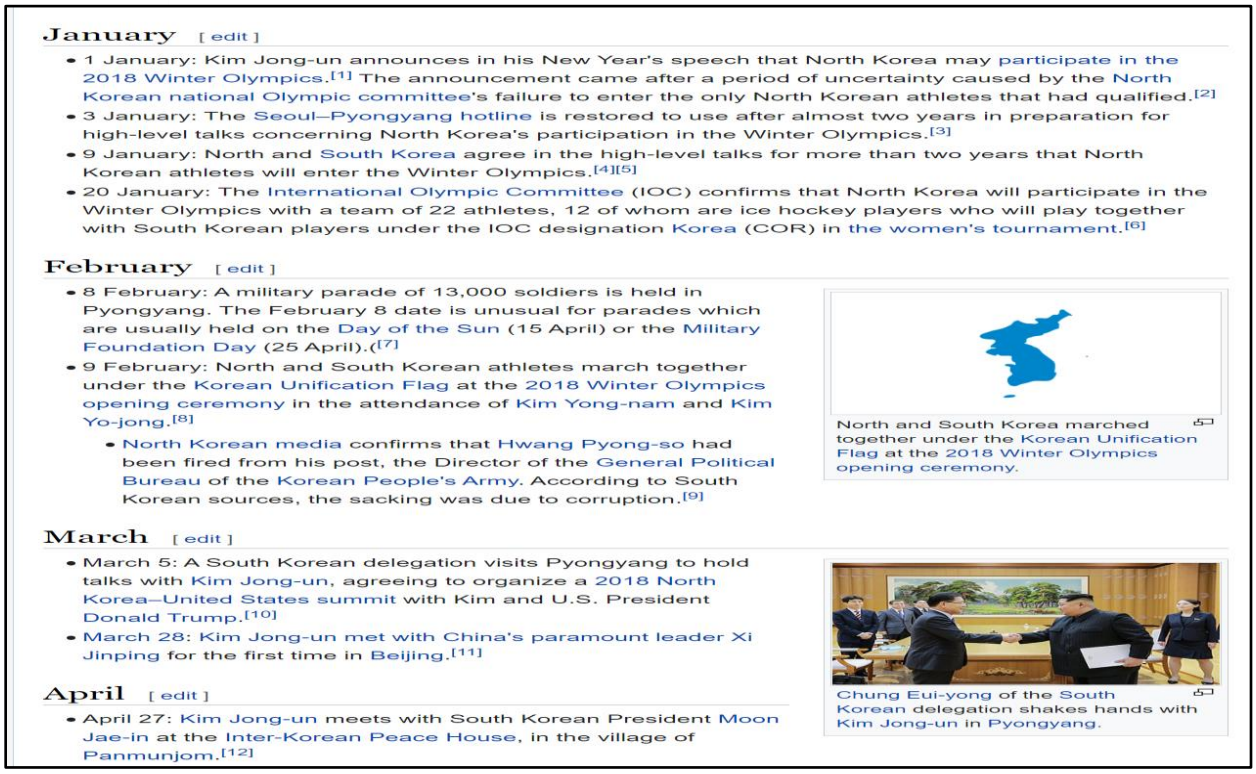
Words in each topic:

```
Topic 0: ['york' 'year' 'write' 'world' 'work' 'woman' 'win' 'white' 'well' 'week'
 'wednesday' 'washington' 'vote' 'video' 'united' 'two' 'tuesday' 'trump'
 'trade' 'time' 'thursday' 'three' 'tell' 'team' 'talk' 'take' 'system'
 'support' 'story' 'statement' 'state' 'start' 'star' 'spend' 'show'
 'shoot' 'set' 'service' 'series' 'see' 'security' 'second' 'season'
 'school' 'run' 'rule' 'rise' 'right' 'reuter' 'republican' 'report'
 'release' 'read' 'public' 'price' 'president' 'post' 'political' 'policy'
 'police' 'point' 'play' 'plan' 'place' 'percent' 'people' 'party' 'part'
 'open' 'one' 'oil' 'official' 'office' 'offer' 'next' 'news' 'new' 'need'
 'national' 'much' 'move' 'month' 'money' 'monday' 'minister' 'million'
 'might' 'member' 'meet' 'media' 'mean' 'may' 'market' 'many' 'man' 'make'
 'love' 'lot' 'look' 'long' 'life' 'leave' 'lead' 'law' 'late' 'last'
 'keep' 'job' 'issue' 'include' 'house' 'home' 'hold' 'hill' 'high' 'help'
 'health' 'happen' 'group' 'great' 'government' 'good' 'give' 'game'
 'friday' 'former' 'force' 'follow' 'first' 'find' 'feel' 'federal'
 'family' 'expect' 'every' 'even' 'end' 'election' 'edit' 'early'
 'democrat' 'deal' 'day' 'data' 'court' 'country' 'continue' 'company'
 'comment' 'come' 'city' 'china' 'child' 'change' 'case' 'car' 'call'
 'business' 'build' 'billion' 'big' 'begin' 'bank' 'back' 'ask' 'around'
 'another' 'announce' 'american' 'add' 'accord']

Topic 1: ['yuan' 'trade' 'text' 'steel' 'shareholder' 'share' 'reuter' 'profit'
 'percent' 'per' 'pay' 'news' 'net' 'million' 'march' 'loss' 'increase'
 'holdings' 'headline' 'eur' 'dividend' 'company' 'chinese' 'cash'
 'billion' 'beijing' 'april']

Topic 2: ['trump' 'senator' 'senate' 'republican' 'nominee' 'nomination'
 'kavanaugh' 'judiciary' 'judge' 'hill' 'hear' 'hawley' 'grassley' 'ford'
 'dr' 'confirmation' 'committee' 'blasey' 'allegation']

Topic 3: ['versus' 'text' 'share' 'revenue' 'reuter' 'quarter' 'profit' 'per'
 'nongaap' 'net' 'million' 'loss' 'feb' 'euro' 'earnings' 'company'
 'billion']

Topic 4: ['woman' 'vote' 'trump' 'sexual' 'senate' 'republican' 'kavanaugh'
 'assault']
```

```
Topic 5: ['trade' 'stock' 'sp' 'share' 'quarter' 'profit' 'price' 'percent' 'high'
 'forecast' 'company' 'billion' 'bank' 'analyst']

Topic 6: ['theresa' 'proposal' 'minister' 'may' 'liverpool' 'labour' 'eu' 'deal'
 'british' 'britain' 'brexit']

Topic 7: ['white' 'trump' 'subpoena' 'sanders' 'president' 'nunberg' 'house'
 'campaign']

Topic 8: ['white' 'trump' 'president' 'omarosa' 'mcgahn' 'manigault' 'house' 'hill']

Topic 9: ['view' 'thomson' 'text' 'share' 'revenue' 'reuter' 'rayonier' 'quarterly'
 'per' 'nongaap' 'million' 'may' 'ipg' 'intermolecular' 'ibes' 'earnings'
 'dividend' 'company']

Topic 10: ['united' 'south' 'seoul' 'pyeongchang' 'putin' 'nuclear' 'north' 'korean'
 'korea' 'kim' 'jong' 'delegation']

Topic 11: ['trump' 'president' 'immigration' 'epa' 'democrat']
```

If we look at the Topic 10, it is about South Korea and North Korea. On doing a quick search we find that indeed these two countries were in news during the months of Jan, Feb, and March 2018 when the topic 10 is dominant. The two countries were in talks about the PyeongChang Winter Olympics.

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)



January [edit]

- 1 January: Kim Jong-un announces in his New Year's speech that North Korea may participate in the 2018 Winter Olympics.[1] The announcement came after a period of uncertainty caused by the North Korean national Olympic committee's failure to enter the only North Korean athletes that had qualified.[2]
- 3 January: The Seoul–Pyongyang hotline is restored to use after almost two years in preparation for high-level talks concerning North Korea's participation in the Winter Olympics.[3]
- 9 January: North and South Korea agree in the high-level talks for more than two years that North Korean athletes will enter the Winter Olympics.[4][5]
- 20 January: The International Olympic Committee (IOC) confirms that North Korea will participate in the Winter Olympics with a team of 22 athletes, 12 of whom are ice hockey players who will play together with South Korean players under the IOC designation Korea (COR) in the women's tournament.[6]

February [edit]

- 8 February: A military parade of 13,000 soldiers is held in Pyongyang. The February 8 date is unusual for parades which are usually held on the Day of the Sun (15 April) or the Military Foundation Day (25 April).[7]
- 9 February: North and South Korean athletes march together under the Korean Unification Flag at the 2018 Winter Olympics opening ceremony in the attendance of Kim Yong-nam and Kim Yo-jong.[8]
  - North Korean media confirms that Hwang Pyong-so had been fired from his post, the Director of the General Political Bureau of the Korean People's Army. According to South Korean sources, the sacking was due to corruption.[9]

North and South Korea marched together under the Korean Unification Flag at the 2018 Winter Olympics opening ceremony.

March [edit]

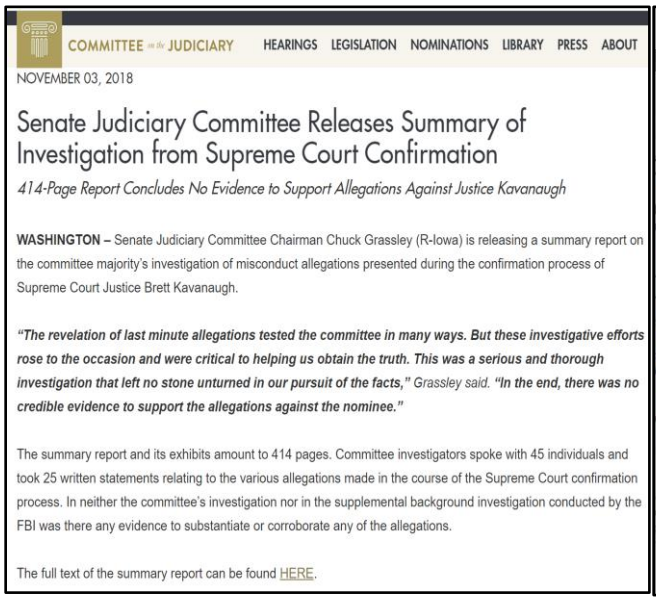
- March 5: A South Korean delegation visits Pyongyang to hold talks with Kim Jong-un, agreeing to organize a 2018 North Korea–United States summit with Kim and U.S. President Donald Trump.[10]
- March 28: Kim Jong-un met with China's paramount leader Xi Jinping for the first time in Beijing.[11]

Chung Eui-yong of the South Korean delegation shakes hands with Kim Jong-un in Pyongyang.

April [edit]

- April 27: Kim Jong-un meets with South Korean President Moon Jae-in at the Inter-Korean Peace House, in the village of Panmunjom.[12]

Looking at the words in Topic 2, we can see that it is about the allegations against Supreme court Justice Brett Kavanaugh. The topic is most prominent in the month of Nov/Dec which was when the Senate Judiciary committee released its Summary of investigation report.



COMMITTEE on the JUDICIARY HEARINGS LEGISLATION NOMINATIONS LIBRARY PRESS ABOUT

NOVEMBER 03, 2018

Senate Judiciary Committee Releases Summary of Investigation from Supreme Court Confirmation

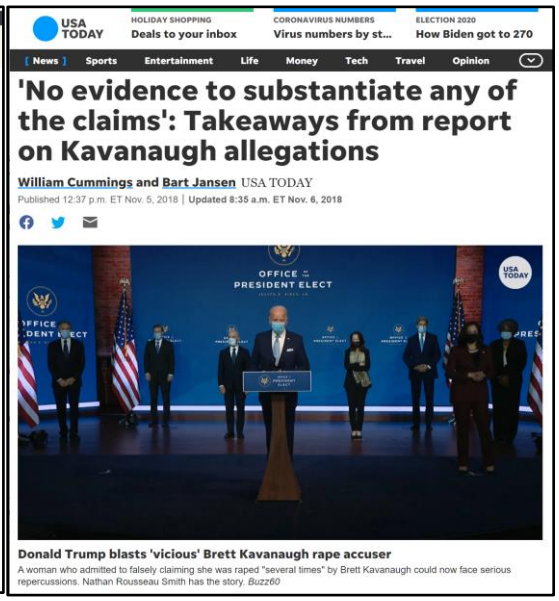
414-Page Report Concludes No Evidence to Support Allegations Against Justice Kavanaugh

WASHINGTON – Senate Judiciary Committee Chairman Chuck Grassley (R-Iowa) is releasing a summary report on the committee majority's investigation of misconduct allegations presented during the confirmation process of Supreme Court Justice Brett Kavanaugh.

"The revelation of last minute allegations tested the committee in many ways. But these investigative efforts rose to the occasion and were critical to helping us obtain the truth. This was a serious and thorough investigation that left no stone unturned in our pursuit of the facts," *Grassley said.* ***"In the end, there was no credible evidence to support the allegations against the nominee."***

The summary report and its exhibits amount to 414 pages. Committee investigators spoke with 45 individuals and took 25 written statements relating to the various allegations made in the course of the Supreme Court confirmation process. In neither the committee's investigation nor in the supplemental background investigation conducted by the FBI was there any evidence to substantiate or corroborate any of the allegations.

The full text of the summary report can be found HERE.



USA TODAY | HOLIDAY SHOPPING Deals to your inbox | CORONAVIRUS NUMBERS Virus numbers by st... | ELECTION 2020 How Biden got to 270

[News] Sports Entertainment Life Money Tech Travel Opinion

'No evidence to substantiate any of the claims': Takeaways from report on Kavanaugh allegations

William Cummings and Bart Jansen USA TODAY

Published 12:37 p.m. ET Nov. 5, 2018 | Updated 8:35 a.m. ET Nov. 6, 2018

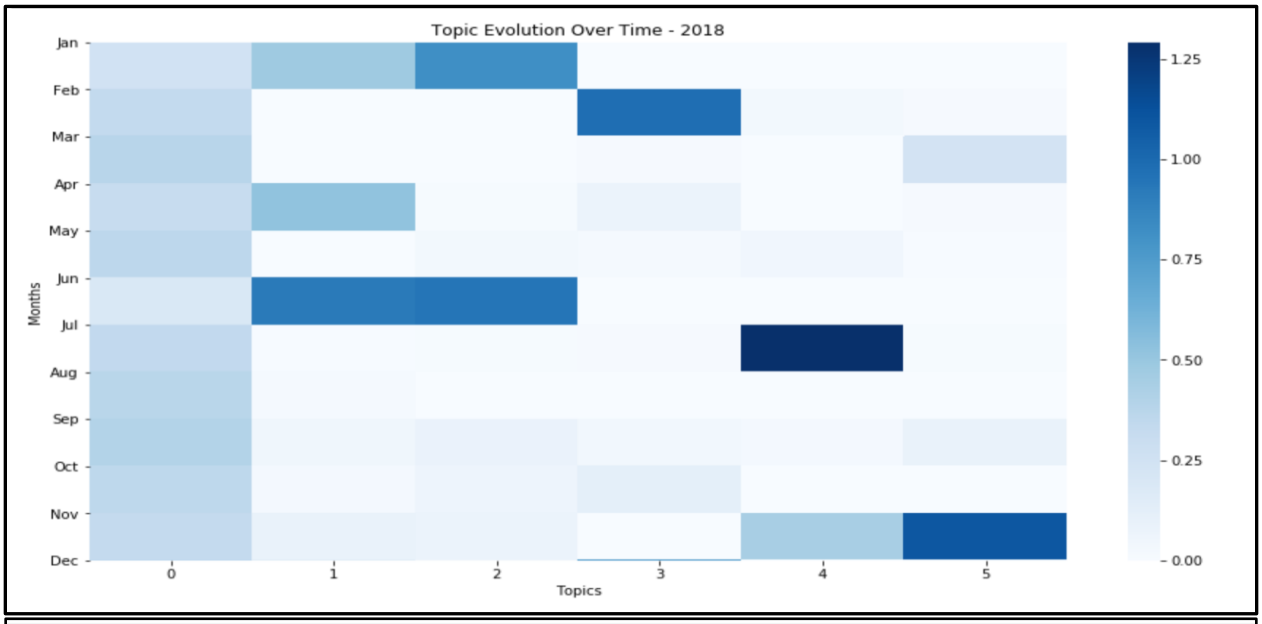
Donald Trump blasts 'vicious' Brett Kavanaugh rape accuser

A woman who admitted to falsely claiming she was raped "several times" by Brett Kavanaugh could now face serious repercussions. Nathan Rousseau Smith has the story. *Buzz60*

On looking at all the 12 topics for 2018 year, we see a lot of overlap like topic 2 and topic 4 have overlapping words. We can see that several topics have “trump”, “presidential” etc. In order to get more robust topics without such

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

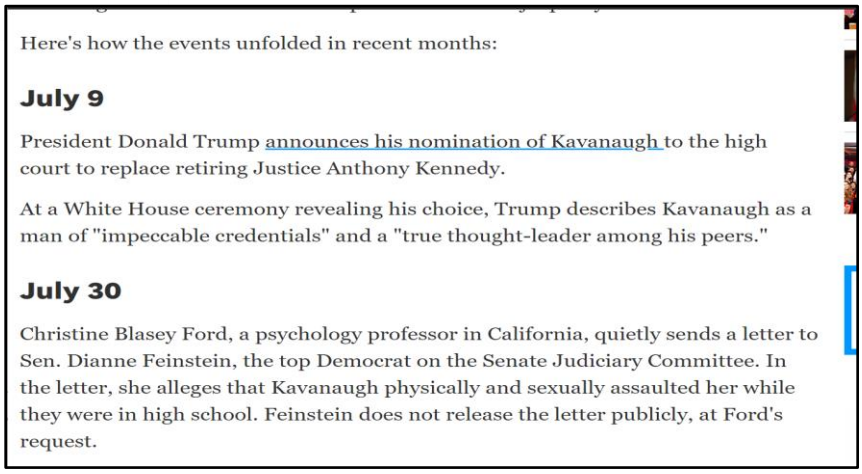
overlaps, we decreased the number of topics in order to put all such "overlapping" words into one topic. The number of topics was set to 6 and this gave us below results:




```
Topic 0: ['york' 'year' 'write' 'world' 'work' 'woman' 'win' 'white' 'whether'
 'well' 'week' 'wednesday' 'washington' 'vote' 'view' 'video' 'united'
 'two' 'tuesday' 'trump' 'trade' 'today' 'time' 'thursday' 'three' 'tell'
 'team' 'talk' 'take' 'system' 'support' 'street' 'story' 'stock'
 'statement' 'state' 'start' 'star' 'spend' 'speak' 'something' 'site'
 'show' 'shoot' 'share' 'set' 'service' 'series' 'see' 'security' 'second'
 'season' 'school' 'run' 'rule' 'rise' 'right' 'reuter' 'republican'
 'report' 'release' 'read' 'rate' 'question' 'publish' 'public' 'price'
 'president' 'post' 'political' 'policy' 'police' 'point' 'play' 'plan'
 'place' 'percent' 'people' 'pay' 'party' 'part' 'open' 'one' 'oil'
 'official' 'office' 'offer' 'number' 'north' 'next' 'news' 'new' 'need'
 'national' 'name' 'much' 'move' 'month' 'money' 'monday' 'minister'
 'million' 'might' 'member' 'meet' 'media' 'mean' 'may' 'market' 'many'
 'man' 'make' 'love' 'lot' 'look' 'long' 'list' 'life' 'leave' 'leader'
 'lead' 'law' 'late' 'last' 'large' 'keep' 'july' 'john' 'job' 'issue'
 'investment' 'interest' 'include' 'house' 'hour' 'home' 'hold' 'hill'
 'high' 'help' 'health' 'happen' 'group' 'great' 'government' 'good'
 'give' 'game' 'friday' 'former' 'force' 'follow' 'first' 'fire' 'find'
 'feel' 'federal' 'family' 'fall' 'face' 'expect' 'every' 'even' 'end'
 'election' 'edit' 'early' 'department' 'democrat' 'deal' 'day' 'data'
 'court' 'country' 'corp' 'continue' 'content' 'company' 'comment' 'come'
 'close' 'claim' 'city' 'china' 'child' 'chief' 'change' 'case' 'car'
 'call' 'business' 'build' 'border' 'billion' 'big' 'begin' 'bank' 'back'
 'authority' 'ask' 'around' 'another' 'announce' 'american' 'amazon'
 'administration' 'add' 'accord']

Topic 1: ['yuan' 'trade' 'text' 'steel' 'shareholder' 'share' 'reuter' 'profit'
 'percent' 'per' 'pay' 'paidin' 'news' 'net' 'million' 'march' 'loss'
 'increase' 'holdings' 'headline' 'eur' 'dollar' 'dividend' 'company'
 'chinese' 'cash' 'billion' 'beijing' 'april']

Topic 2: ['woman' 'white' 'trump' 'president' 'mueller' 'immigration' 'house' 'epa'
 'democrat' 'campaign' 'administration']

Topic 3: ['view' 'thomson' 'text' 'share' 'revenue' 'reuter' 'range' 'per'
 'oncolytics' 'nongaap' 'million' 'may' 'loss' 'ibes' 'gaap' 'feb'
 'earnings' 'dividend' 'company' 'biotech']

Topic 4: ['woman' 'voter' 'vote' 'trump' 'sexual' 'senate' 'republican' 'kavanaugh'
 'ford' 'assault']

Topic 5: ['union' 'theresa' 'proposal' 'prime' 'party' 'opposition' 'northern'
 'minister' 'may' 'liverpool' 'labour' 'european' 'eu' 'england' 'deal'
 'corbyn' 'british' 'britain' 'brexit' 'bloc']
```

Now, we see that the previous topic 2 (about Supreme court Justice Brett Kavanaugh allegations) is topic 4 but has the more clear topic evolution following the progression of the news about this event.

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)



Here's how the events unfolded in recent months:

July 9

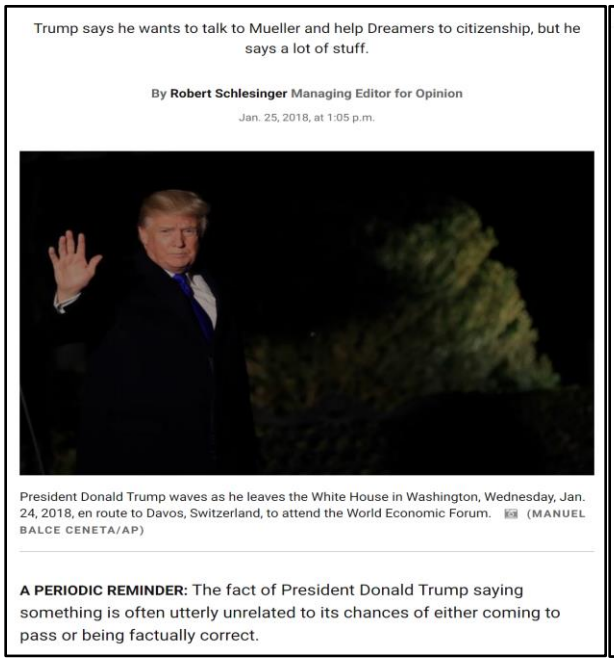
President Donald Trump announces his nomination of Kavanaugh to the high court to replace retiring Justice Anthony Kennedy.

At a White House ceremony revealing his choice, Trump describes Kavanaugh as a man of "impeccable credentials" and a "true thought-leader among his peers."

July 30

Christine Blasey Ford, a psychology professor in California, quietly sends a letter to Sen. Dianne Feinstein, the top Democrat on the Senate Judiciary Committee. In the letter, she alleges that Kavanaugh physically and sexually assaulted her while they were in high school. Feinstein does not release the letter publicly, at Ford's request.

Topic 2 is about two the US immigration as well as the Mueller Investigation, which were the biggest news items in the months of Jan and Feb 2018 and the topic 2 evolution, aligns clearly with that timeline.



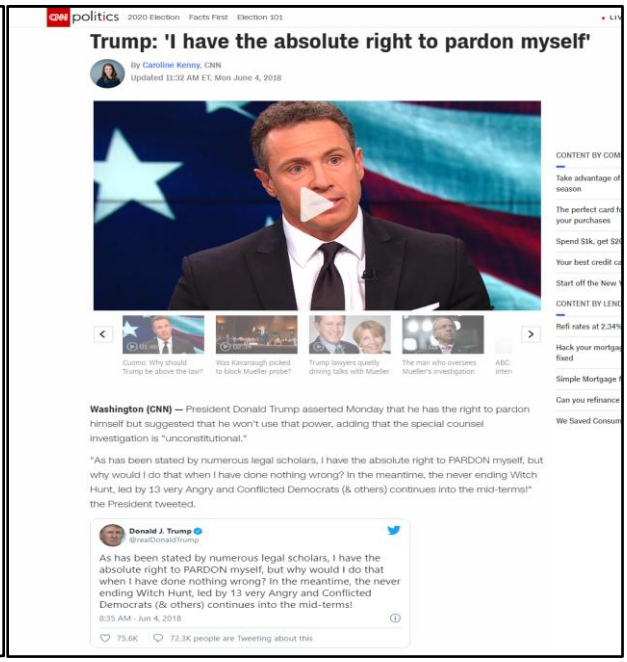
Trump says he wants to talk to Mueller and help Dreamers to citizenship, but he says a lot of stuff.

By **Robert Schlesinger** Managing Editor for Opinion

Jan. 25, 2018, at 1:05 p.m.

President Donald Trump waves as he leaves the White House in Washington, Wednesday, Jan. 24, 2018, en route to Davos, Switzerland, to attend the World Economic Forum. (MANUEL BALCE CENETA/AP)

A PERIODIC REMINDER: The fact of President Donald Trump saying something is often utterly unrelated to its chances of either coming to pass or being factually correct.



politics 2020 Election Facts First Election 101

Trump: 'I have the absolute right to pardon myself'

By Caroline Kenny, CNN
Updated 11:32 AM ET, Mon June 4, 2018

Washington (CNN) — President Donald Trump asserted Monday that he has the right to pardon himself but suggested that he won't use that power, adding that the special counsel investigation is "unconstitutional."

"As has been stated by numerous legal scholars, I have the absolute right to PARDON myself, but why would I do that when I have done nothing wrong? In the meantime, the never ending Witch Hunt, led by 13 very Angry and Conflicted Democrats (& others) continues into the mid-terms!" the President tweeted.

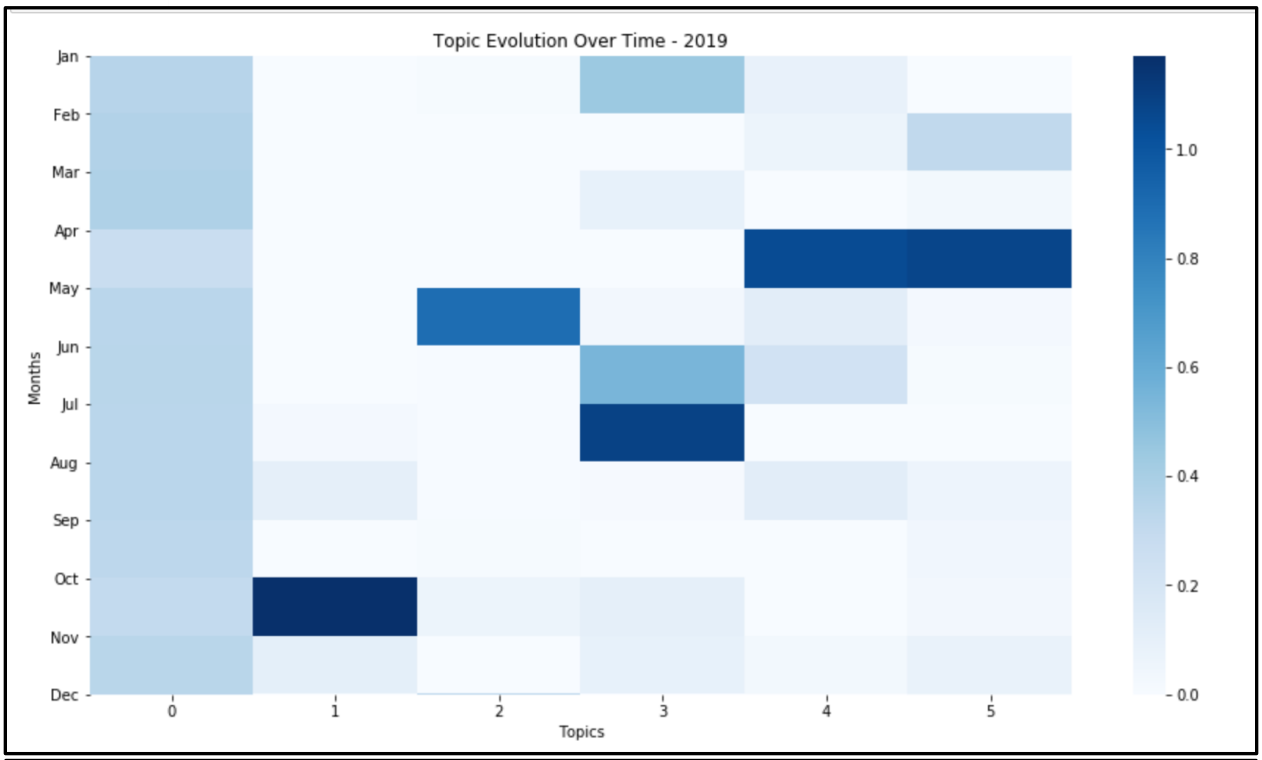
Donald J. Trump @realDonaldTrump

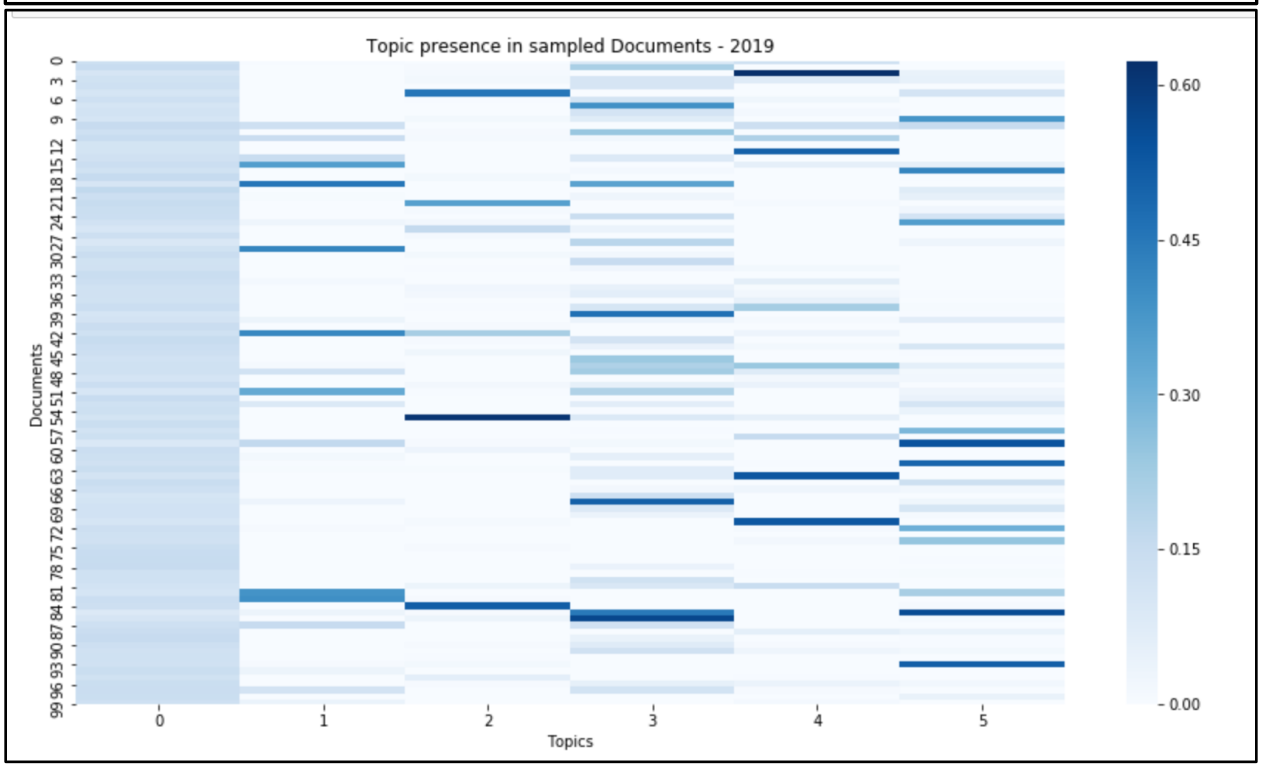
As has been stated by numerous legal scholars, I have the absolute right to PARDON myself, but why would I do that when I have done nothing wrong? In the meantime, the never ending Witch Hunt, led by 13 very Angry and Conflicted Democrats (& others) continues into the mid-terms!

8:35 AM - Jun 4, 2018

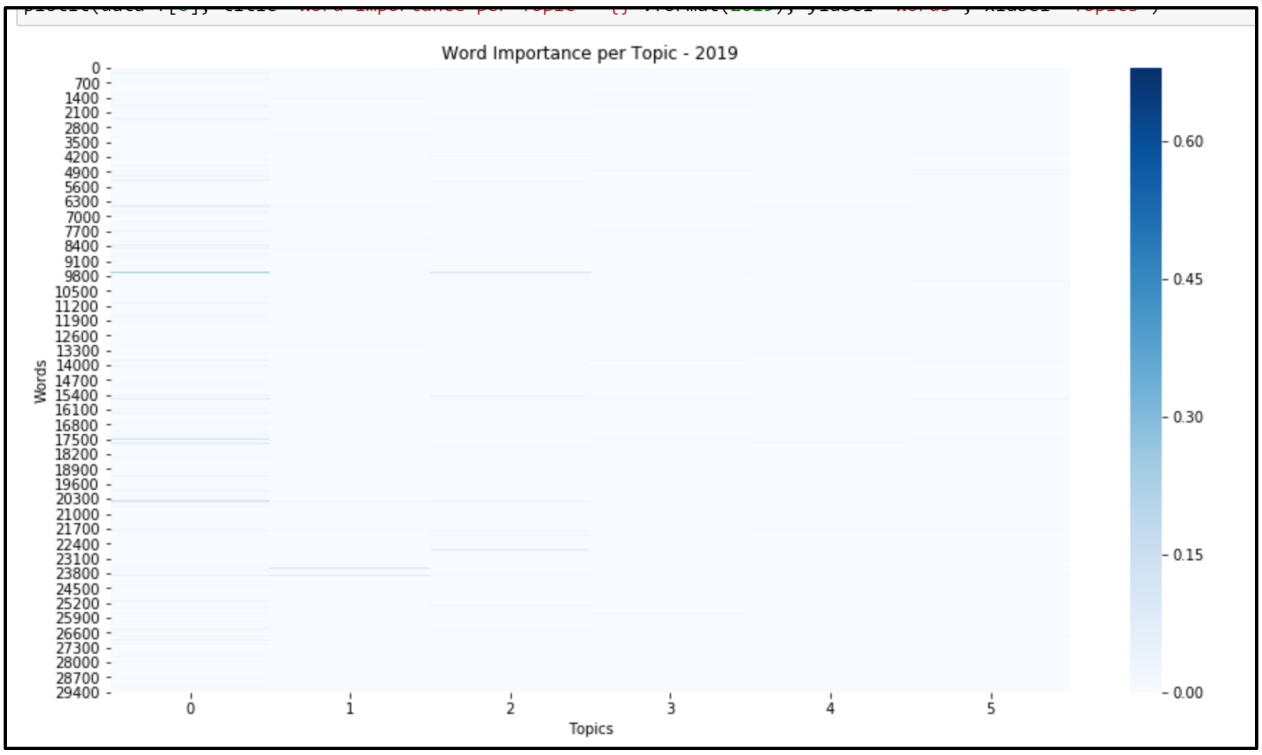
For the Year 2019: Let us look at the topic evolution for this year:

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)





Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

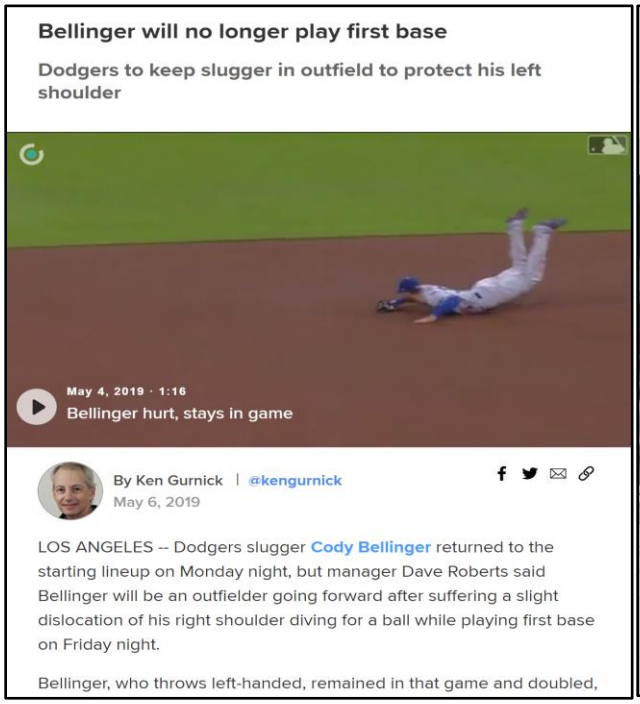


Word in each topic:

```
Topic 0: ['york' 'year' 'write' 'world' 'work' 'woman' 'win' 'white' 'whether'
 'well' 'week' 'wednesday' 'washington' 'war' 'vote' 'view' 'video'
 'united' 'union' 'two' 'turn' 'tuesday' 'trump' 'trade' 'today' 'time'
 'thursday' 'three' 'tell' 'technology' 'team' 'target' 'talk' 'take'
 'system' 'support' 'sunday' 'street' 'story' 'stop' 'statement' 'state'
 'start' 'star' 'spend' 'speak' 'south' 'something' 'social' 'small'
 'site' 'sign' 'show' 'share' 'set' 'service' 'series' 'see' 'security'
 'second' 'season' 'school' 'sales' 'run' 'rule' 'rise' 'right' 'reuter'
 'result' 'require' 'republican' 'report' 'remain' 'release' 'recent'
 'read' 'rate' 'raise' 'question' 'quarter' 'publish' 'public' 'product'
 'process' 'prime' 'price' 'presidential' 'president' 'power' 'post'
 'political' 'policy' 'police' 'point' 'play' 'plan' 'place' 'phone' 'per'
 'people' 'pay' 'party' 'part' 'order' 'open' 'one' 'oil' 'official'
 'office' 'offer' 'number' 'next' 'news' 'new' 'need' 'national' 'name'
 'much' 'move' 'month' 'money' 'monday' 'minister' 'million' 'member'
 'meet' 'media' 'mean' 'may' 'market' 'many' 'man' 'make' 'major' 'low'
 'love' 'lot' 'look' 'long' 'live' 'little' 'list' 'line' 'life' 'level'
 'leave' 'least' 'leader' 'lead' 'law' 'launch' 'late' 'last' 'large'
 'keep' 'june' 'july' 'job' 'issue' 'iran' 'investment' 'investigation'
 'interview' 'international' 'interest' 'increase' 'include' 'house'
 'home' 'hold' 'hit' 'hill' 'high' 'help' 'health' 'head' 'happen'
 'growth' 'grow' 'group' 'great' 'government' 'google' 'good' 'global'
 'give' 'game' 'future' 'fund' 'full' 'friday' 'former' 'foreign' 'force'
 'follow' 'first' 'firm' 'find' 'financial' 'film' 'feel' 'federal'
 'family' 'fall' 'facebook' 'face' 'expect' 'executive' 'every' 'even'
 'european' 'end' 'election' 'edit' 'economy' 'economic' 'early'
 'democratic' 'democrat' 'decision' 'deal' 'day' 'data' 'cut' 'create'
 'court' 'country' 'cost' 'continue' 'content' 'company' 'comment' 'come'
 'close' 'climate' 'claim' 'city' 'chinese' 'china' 'child' 'chief'
 'charge' 'change' 'central' 'case' 'car' 'capital' 'candidate' 'campaign'
 'call' 'buy' 'business' 'build' 'border' 'billion' 'bill' 'big' 'biden'
 'believe' 'begin' 'bank' 'back' 'attack' 'ask' 'around' 'apple' 'appear'
 'app' 'another' 'announce' 'american' 'amazon' 'allow' 'agreement'
 'administration' 'add' 'action' 'across' 'account' 'accord']

Topic 1: ['witness' 'vote' 'trump' 'trial' 'senate' 'schumer' 'schiff' 'republican'
 'president' 'political' 'party' 'mcconnell' 'impeachment' 'impeach'
 'house' 'election' 'democratic' 'democrat' 'committee' 'article']

Topic 2: ['win' 'walk' 'victory' 'tworun' 'two' 'threegame' 'three' 'th'
 'strikeouts' 'strike' 'stella' 'starter' 'sixth' 'six' 'single' 'seven'
 'second' 'season' 'scoreless' 'score' 'ryu' 'run' 'red' 'pitch' 'ninth'
 'lead' 'jose' 'jansen' 'innings' 'inning' 'homered' 'homer' 'home' 'hit'
 'gonsolin' 'give' 'game' 'four' 'five' 'first' 'fielder' 'field' 'eighth'
 'eight' 'earn' 'double' 'dodger' 'cub' 'cincinnati' 'bellinger' 'base'
 'ball' 'angel']

Topic 3: ['wall' 'trade' 'tax' 'stock' 'shutdown' 'rise' 'percent' 'market' 'index'
 'data' 'bank']

Topic 4: ['trump' 'tell' 'special' 'sentence' 'russia' 'president' 'pardon'
 'muellers' 'mueller' 'manaforts' 'manafort' 'lawyer' 'investigation'
 'giuliani' 'fraud' 'counsel' 'conway' 'congress' 'committee' 'cohens'
 'cohen' 'campaign' 'barr' 'attorney']

Topic 5: ['vote' 'trade' 'percent' 'nodeal' 'may' 'market' 'march' 'london'
 'european' 'eu' 'deal' 'british' 'britain' 'brexit']
```

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

If we look at the Topic 2 here, it comprises all the 'Baseball' related words – Players (Bellinger, Gonsolin, and Jansen), Teams (Cincinnati red, Dodger's) and other terms (Homer, Homered, Pitch, Strike etc.).This topic was dominant during the month of May 2019. On doing a search for news articles in May 2019, there are several news articles about the mentioned players and the teams at that time:



Bellinger will no longer play first base

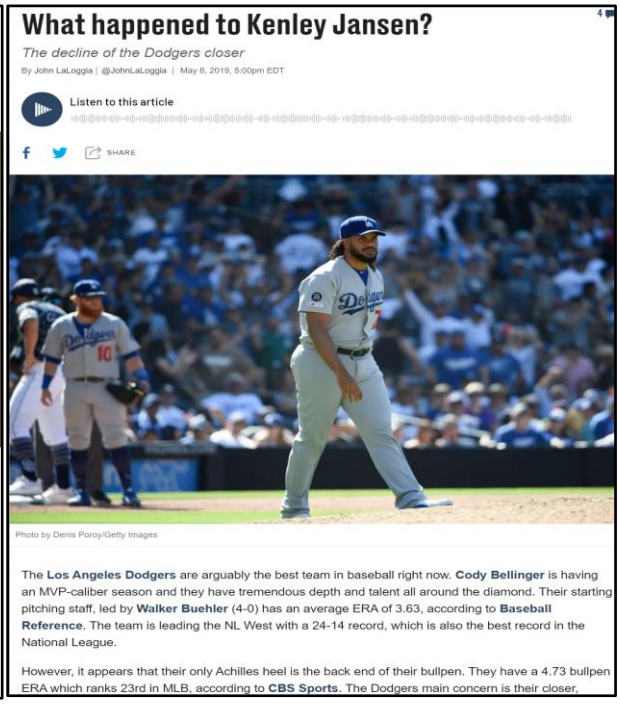
Dodgers to keep slugger in outfield to protect his left shoulder

May 4, 2019 · 1:16
Bellinger hurt, stays in game

By Ken Gurnick | @kengurnick
May 6, 2019

LOS ANGELES -- Dodgers slugger **Cody Bellinger** returned to the starting lineup on Monday night, but manager Dave Roberts said Bellinger will be an outfielder going forward after suffering a slight dislocation of his right shoulder diving for a ball while playing first base on Friday night.

Bellinger, who throws left-handed, remained in that game and doubled,



What happened to Kenley Jansen?

The decline of the Dodgers closer

By John LaLoggia | @JohnLaLoggia | May 8, 2019, 5:00pm EDT

Listen to this article

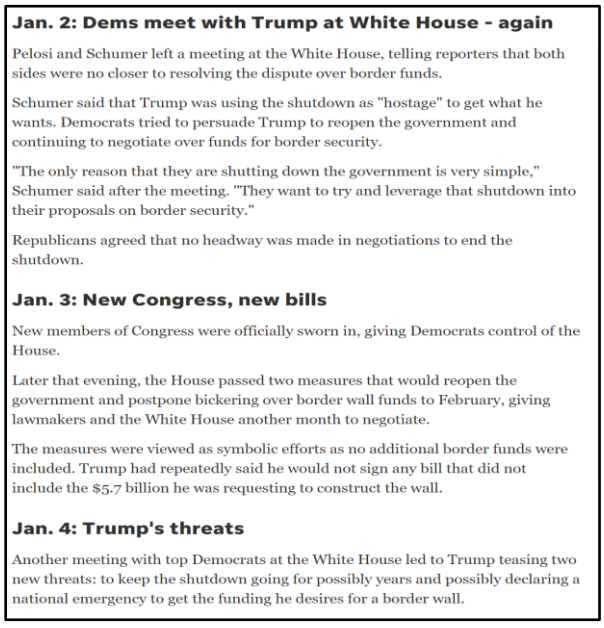
SHARE

Photo by Denis Poroy/Getty Images

The **Los Angeles Dodgers** are arguably the best team in baseball right now. **Cody Bellinger** is having an MVP-caliber season and they have tremendous depth and talent all around the diamond. Their starting pitching staff, led by **Walker Buehler** (4-0) has an average ERA of 3.63, according to **Baseball Reference**. The team is leading the NL West with a 24-14 record, which is also the best record in the National League.

However, it appears that their only Achilles heel is the back end of their bullpen. They have a 4.73 bullpen ERA which ranks 23rd in MLB, according to **CBS Sports**. The Dodgers main concern is their closer,

Topic 3 has words 'wall', 'shutdown' and was dominant in the month of Jan, March and then again in Jun, Jul of the year. This is one topic which has faded in - out throughout the year. This topic deals with the US government shutdown done in the month of Jan 2019 for the funding of 'We can verify this is we do simple search of the timeline. There were few court judgements in the month of Jul/Aug that made this topic dominant again.



Jan. 2: Dems meet with Trump at White House - again

Pelosi and Schumer left a meeting at the White House, telling reporters that both sides were no closer to resolving the dispute over border funds.

Schumer said that Trump was using the shutdown as "hostage" to get what he wants. Democrats tried to persuade Trump to reopen the government and continuing to negotiate over funds for border security.

"The only reason that they are shutting down the government is very simple," Schumer said after the meeting. "They want to try and leverage that shutdown into their proposals on border security."

Republicans agreed that no headway was made in negotiations to end the shutdown.

Jan. 3: New Congress, new bills

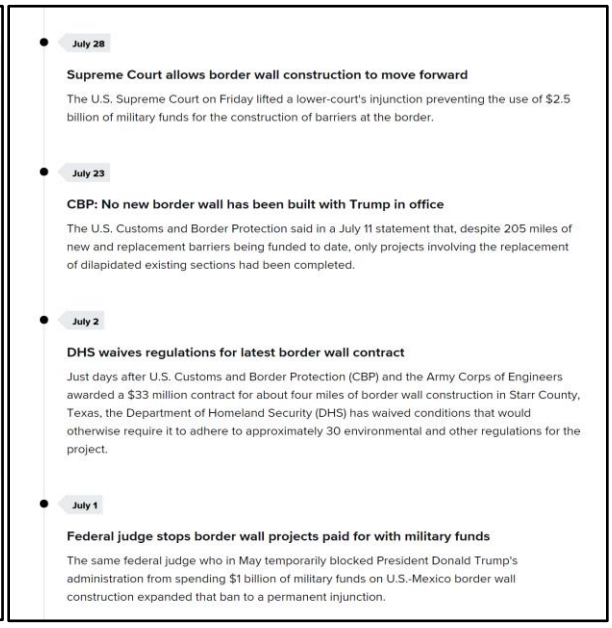
New members of Congress were officially sworn in, giving Democrats control of the House.

Later that evening, the House passed two measures that would reopen the government and postpone bickering over border wall funds to February, giving lawmakers and the White House another month to negotiate.

The measures were viewed as symbolic efforts as no additional border funds were included. Trump had repeatedly said he would not sign any bill that did not include the $5.7 billion he was requesting to construct the wall.

Jan. 4: Trump's threats

Another meeting with top Democrats at the White House led to Trump teasing two new threats: to keep the shutdown going for possibly years and possibly declaring a national emergency to get the funding he desires for a border wall.



July 28

Supreme Court allows border wall construction to move forward

The U.S. Supreme Court on Friday lifted a lower-court's injunction preventing the use of $2.5 billion of military funds for the construction of barriers at the border.

July 23

CBP: No new border wall has been built with Trump in office

The U.S. Customs and Border Protection said in a July 11 statement that, despite 205 miles of new and replacement barriers being funded to date, only projects involving the replacement of dilapidated existing sections had been completed.

July 2

DHS waives regulations for latest border wall contract

Just days after U.S. Customs and Border Protection (CBP) and the Army Corps of Engineers awarded a $33 million contract for about four miles of border wall construction in Starr County, Texas, the Department of Homeland Security (DHS) has waived conditions that would otherwise require it to adhere to approximately 30 environmental and other regulations for the project.

July 1

Federal judge stops border wall projects paid for with military funds

The same federal judge who in May temporarily blocked President Donald Trump's administration from spending $1 billion of military funds on U.S.-Mexico border wall construction expanded that ban to a permanent injunction.

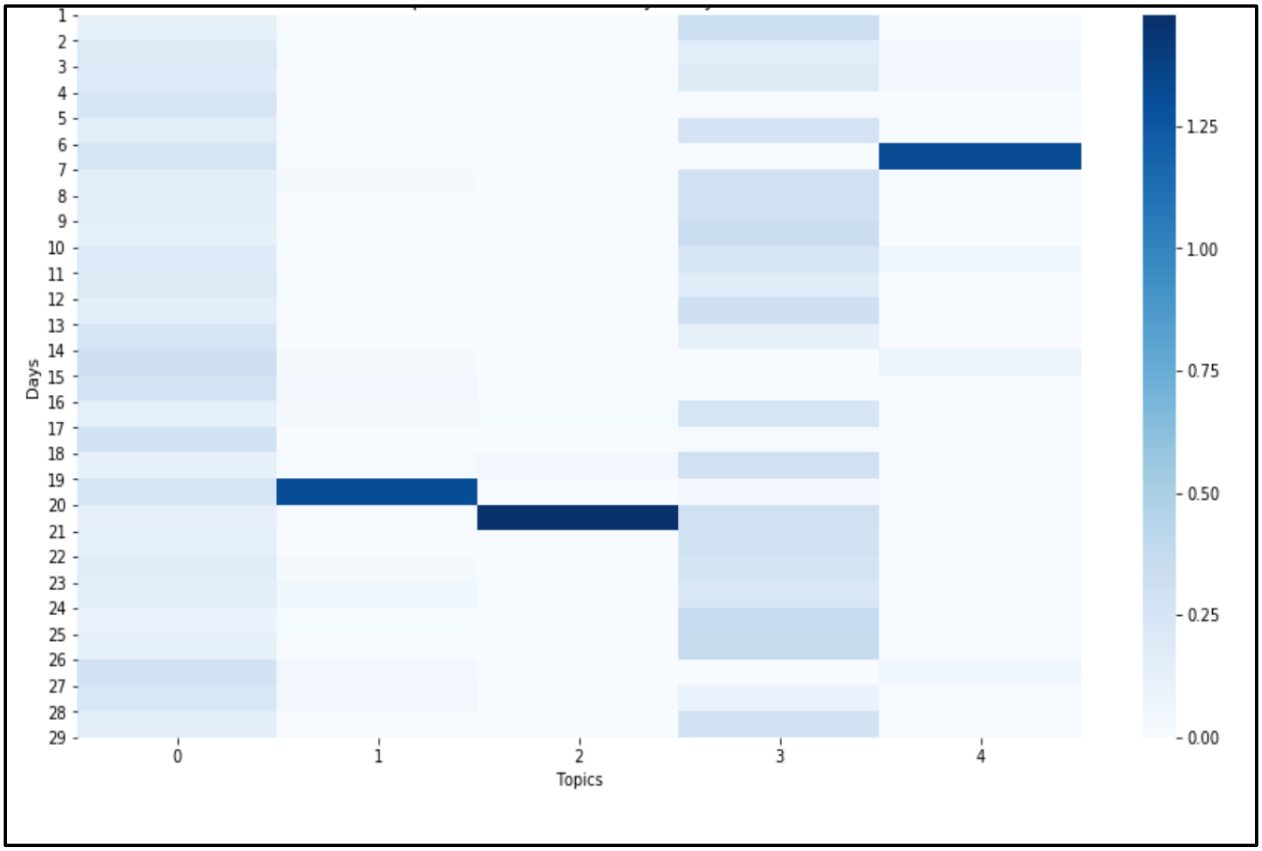
Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

Topic 4 is about the "Muellar Investigation" and likewise the topic evolution follows the timeline of the underlying events:

For Year 2020:

Year 2020 has data only for four months, that is, from Jan -Apr and of which only the month of Feb and March have major chunk of data. Jan month has only 5 document and April only 11. Therefore, we will be running monthly analysis instead of yearly to see the topic evolution for the months of Feb and March.

Topic Evolution for the month of Feb:



Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

Words per topic:

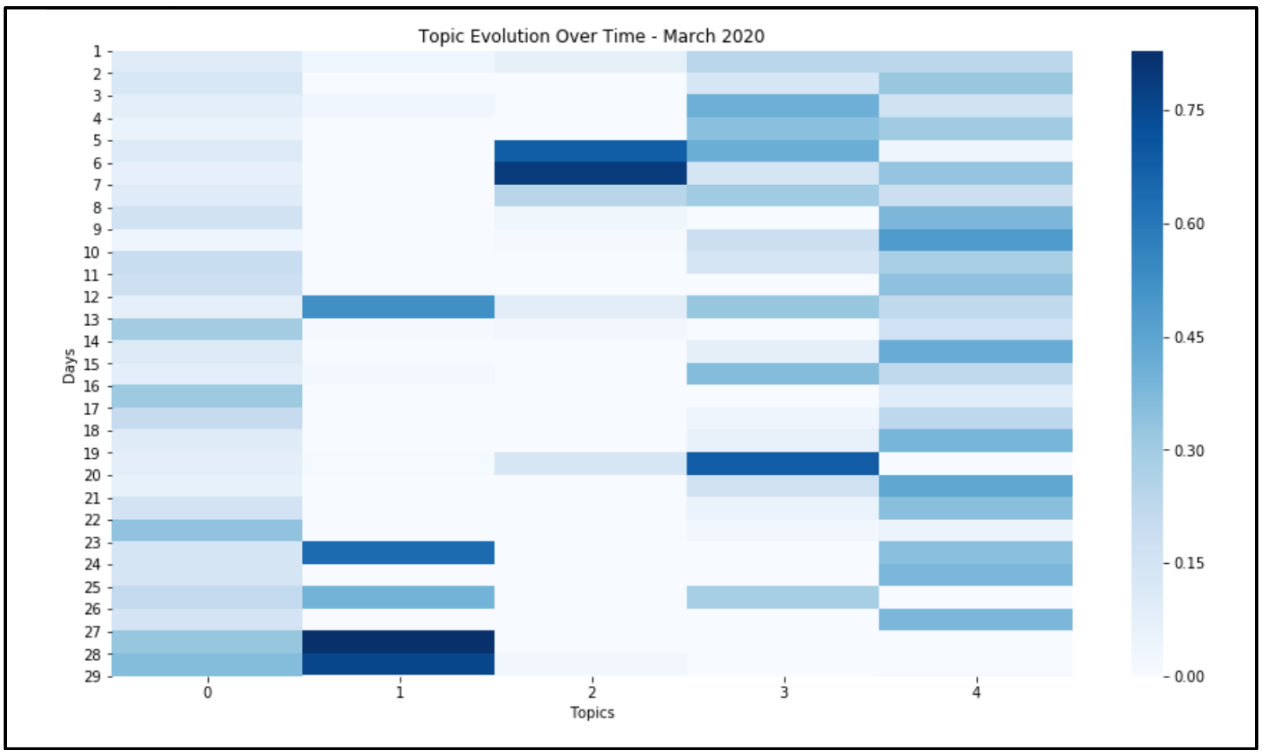
```
Topic 0: ['year' 'work' 'woman' 'win' 'week' 'vote' 'virus' 'two' 'trump' 'time'
 'test' 'tell' 'take' 'sunday' 'story' 'state' 'south' 'show' 'see'
 'saturday' 'sanders' 'reuter' 'report' 'president' 'point' 'people'
 'party' 'outbreak' 'one' 'official' 'news' 'new' 'national' 'many' 'make'
 'lead' 'last' 'include' 'house' 'hill' 'health' 'government' 'good'
 'game' 'friday' 'former' 'first' 'feb' 'edit' 'democratic' 'death' 'day'
 'country' 'coronavirus' 'company' 'come' 'city' 'chinese' 'china' 'case'
 'candidate' 'campaign' 'bloomberg' 'biden' 'back' 'american' 'accord']

Topic 1: ['win' 'voter' 'south' 'sanders' 'oddson' 'nevada' 'mccain' 'favorite'
 'democratic' 'clyburn' 'clinton' 'carolina']

Topic 2: ['story' 'require' 'read' 'prime' 'membership' 'full' 'article']

Topic 3: ['year' 'text' 'reuter' 'report' 'million' 'market' 'feb' 'coronavirus'
 'company' 'billion' 'bank']

Topic 4: ['win' 'warren' 'sanders' 'recall' 'iowa' 'hampshire' 'buttigieg']
```

The topic 4 has words ‘Buttigieg’, ‘Iowa’ which is the news relating to Presidential candidate Pete Buttigieg from Iowa and when searched aligns correctly with the topic evolution timeline as can be seen below:

Topic 3 correctly identifies Coronavirus Pandemic as the major news in the month of Feb 2020. It can be seen that was in news almost daily.

Topic Evolution for the month of March:

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)




```
Topic 0: ['york' 'year' 'write' 'world' 'worker' 'work' 'week' 'virus' 'two'
 'trump' 'time' 'test' 'tell' 'take' 'sunday' 'story' 'stay' 'state'
 'spread' 'social' 'show' 'see' 'saturday' 'reuter' 'report' 'quarantine'
 'public' 'president' 'positive' 'people' 'patient' 'pandemic' 'outbreak'
 'order' 'one' 'official' 'number' 'news' 'new' 'need' 'national'
 'ministry' 'medical' 'measure' 'may' 'mask' 'march' 'many' 'make' 'life'
 'keep' 'insider' 'include' 'house' 'hospital' 'homepage' 'home' 'help'
 'health' 'government' 'good' 'friday' 'first' 'even' 'emergency' 'edit'
 'death' 'day' 'crisis' 'covid' 'country' 'coronavirus' 'confirm'
 'company' 'come' 'close' 'city' 'case' 'care' 'call' 'business'
 'american']

Topic 1: ['story' 'prime' 'insider' 'exclusively' 'business' 'available']

Topic 2: ['win' 'voter' 'vote' 'super' 'state' 'south' 'sanders' 'primary' 'poll'
 'democratic' 'delegate' 'carolina' 'candidate' 'campaign' 'biden']

Topic 3: ['biden']

Topic 4: ['reuter' 'march' 'coronavirus' 'company' 'bank']
```

Topic 2 above is about the democrat primaries in ‘Carolina’ state. This news was prevalent in the first week of March:

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

March 2020

- **March 1**: Buttigieg dropped out of the race.[362]
- **March 2**:
  - Klobuchar dropped out of the race.[363]
  - Buttigieg, Klobuchar, and O'Rourke endorsed Biden during an evening rally in Texas.[364]
- **March 3**:
  - Super Tuesday was held.[365]
  - Biden won Alabama, Arkansas, Massachusetts, Maine, Minnesota, North Carolina, Oklahoma, Tennessee, Texas, and Virginia.
  - Bloomberg won American Samoa.
  - Sanders won California, Colorado, Utah, and Vermont.
  - Voting in the Democrats Abroad primary began.[366]
- **March 4**: Bloomberg dropped out of the race and endorsed Biden.[367]
- **March 5**: Warren dropped out of the race.[368]
- **March 6**: Delaney endorsed Biden.[369]
- **March 8**: Harris endorsed Biden.[370]
- **March 9**: Booker endorsed Biden.[371]
- **March 10**:
  - Biden won Idaho, Michigan, Mississippi, Missouri, and Washington.[372][373]
  - Sanders won North Dakota.[374]
  - Yang endorsed Biden.[375]
- **March 13**:
  - A national emergency was declared due to the coronavirus pandemic.[376] Several presidential primaries were soon rescheduled
  - The Louisiana primary was rescheduled from April 4 to June 20 due to coronavirus concerns.[377]
- **March 14**:

Above we saw how the topic is evolving/Fading every month for each year or every day for each month. It must be noted here that there are some topic like “Brexit”, “Mueller Investigation” that are a recurrent topic and are present for most of the analyzed years. The time evolution for these also follow the timeline of actual news articles relating to these events.

Computational efficiency

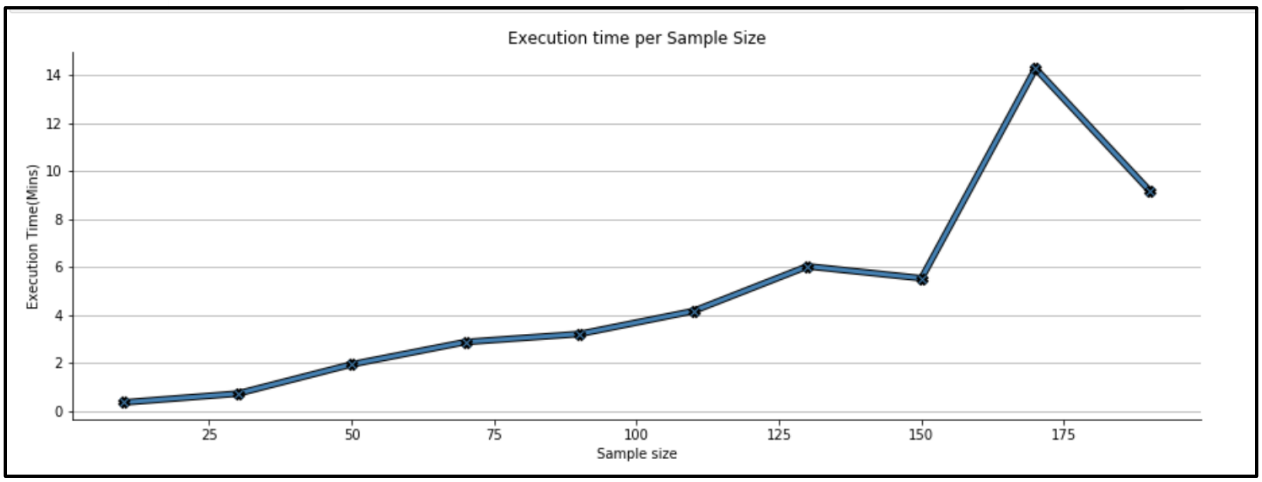
We also analyzed the computational efficiency of the program by experimenting with different sample size and analyzing the execution time each time. Given that we are dealing with a large dataset, the tensor size can grow exponentially as the sample size increases:

```
10
(120, 7455)
Tensor shape : (7455, 10, 12)
--- 20.538015842437744 seconds ---
30
(360, 15447)
Tensor shape : (15447, 30, 12)
--- 42.64621663093567 seconds ---
50
(600, 20235)
Tensor shape : (20235, 50, 12)
--- 115.95428085327148 seconds ---
70
(840, 23987)
Tensor shape : (23987, 70, 12)
--- 171.93645572662354 seconds ---
90
(1080, 29748)
Tensor shape : (29748, 90, 12)
--- 192.01537227630615 seconds ---
```

```
110
(1320, 32396)
Tensor shape : (32396, 110, 12)
--- 249.50250792503357 seconds ---
130
(1560, 35974)
Tensor shape : (35974, 130, 12)
--- 361.47305560112 seconds ---
150
(1800, 38372)
Tensor shape : (38372, 150, 12)
--- 331.57483887672424 seconds ---
170
(2040, 43194)
Tensor shape : (43194, 170, 12)
--- 856.4454109668732 seconds ---
190
(2280, 45853)
Tensor shape : (45853, 190, 12)
--- 550.8820140361786 seconds ---
```

Above we see that the tensor size is increasing as the sample size is increasing, for ex. for sample size 150, the tensor is of size 38372 x 150 x 12(~69 million parameters are parsed). One thing to note here is that we are not restricting the features here. The graph of execution time vs the sample size:

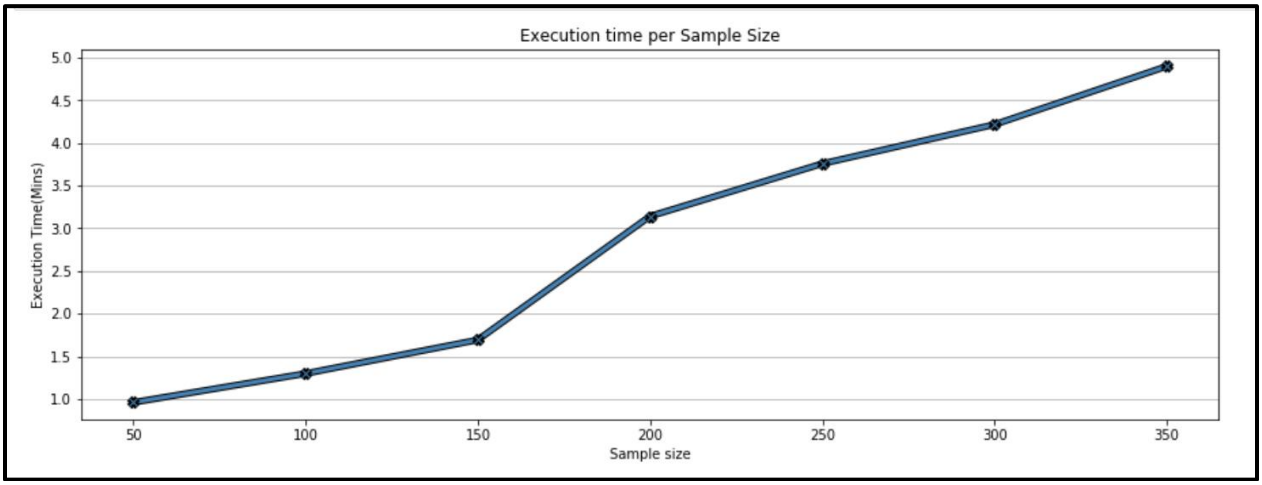
Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)



Now if we restrict the number of features vectorized to 10000 and then observe the execution time for each sample size, we obtain below results:

```
50
(600, 10000)
Tensor shape : (10000, 50, 12)
--- 57.59834575653076 seconds ---
100
(1200, 10000)
Tensor shape : (10000, 100, 12)
--- 77.89110660552979 seconds ---
150
(1800, 10000)
Tensor shape : (10000, 150, 12)
--- 101.55580759048462 seconds ---
200
(2400, 10000)
Tensor shape : (10000, 200, 12)
--- 188.49518990516663 seconds ---
```

```
250
(3000, 10000)
Tensor shape : (10000, 250, 12)
--- 225.47543835639954 seconds ---
300
(3600, 10000)
Tensor shape : (10000, 300, 12)
--- 253.0161817073822 seconds ---
350
(4200, 10000)
Tensor shape : (10000, 350, 12)
--- 293.9427137374878 seconds ---
```



As seen above that, even for a sample size of 350, the tensor decomposition is done less than 5 mins and tensor size here is 10000 x 350 x 12 (~42 mil parameters are parsed).

Gangwanea_Akanksha (agangwanea3@gatech.edu)(GT ID:903548537)
Scott_Kelly (ksims35@gatech.edu) (GT ID:903469283)
Waghere_Shubhangi (swaghere3@gatech.edu) (GT ID:903415553)

CONCLUSION:

Dynamic topic modelling using CP tensor decomposition did give quite accurate results even though we sampled only very small subset from the huge datasets. As part of this project, we covered one aspect of analysis but DTM has vast applications and can be further extended to achieve document clustering, Document correlation etc. It can also be used to automatically organizing, searching, browsing documents.

Above we also visualized what topic is present in what sampled document. This can be further used to determine the document similarity or topic similarity.

APPENDIX

1) On Large-Scale Dynamic Topic Modeling with Nonnegative CP Tensor Decomposition Miju Ahn, Nicole Eikmeier, Jamie Haddock, Lara Kassab, Alona Kryshchenko, Kathryn Leonard, Deanna Needell, R. W. M. A. Madushani, Elena Sizikova, Chuntian Wang
2) Topic modeling for analysis of big data tensor decompositions Thomas S. Henretty, M. Harper Langston, Muthu Baskaran, James Ezick, and Richard Lethin Reservoir Labs, 632 Broadway Suite 803, New York, NY, USA, 10012
3) Dynamic Topic Models David M. Blei BLEI@CS.PRINCETON.EDU Computer Science Department, Princeton University, Princeton, NJ 08544, USA John D. Lafferty LAFFERTY@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh PA 15213, USA

4) Learning Evolving and Emerging Topics in Social Media: A Dynamic NMF approach with Temporal Regularization Ankan Saha Department of Computer Science University of Chicago, Chicago IL 60637 ankans@cs.uchicago.edu Vikas Sindhwani IBM T.J. Watson Research Center Yorktown Heights, NY 10598 vsindhw@us.ibm.com
5) Chong Wang and David M Blei. “Collaborative topic modeling for recommending scientific articles”. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011, pp. 448–456.

6) A Practical Randomized CP Tensor Decomposition: Casey Battaglino, Grey Ballard, Tamara G. Kolda