AI&ML_ASSESSMENT 2

POONAM PRAFUL SHRISHRIMAL_s8075211

2023-09-06

Objective of the Assignment

The goal of the assignment is to apply topic modeling using LDA to a set of State of the Union speeches. We aim to gain insights into how various topics are distributed within these speeches, filter speeches based on specific topics by determining the topic's percentage in each speech, rank topics, and conduct a comprehensive analysis of how topics have evolved over different time periods within the dataset.

1. Import required libraries and Data

```
# set options
options(stringsAsFactors = F)         # no automatic data transformation
options("scipen" = 100, "digits" = 4) # suppress math annotation
# load packages
library(knitr)
library(DT)
library(tm)
library(topicmodels)
library(reshape2)
library(ggplot2)
library(wordcloud)
library(pals)
library(SnowballC)
library(lda)
library(ldatuning)
library(flextable)
```

```
# load data
df=read.csv("E:/VU SYDNEY/AI&ML/Speechdata.csv")
```

2. Review the Dataset

Reviewing the data is a critical step in any data analysis or machine learning project. It helps ensure the quality, integrity, and suitability of the data for the intended analysis or modeling

task. It allows us to understand the data structure, type and also help us to identify errors and inconsistencies.

From below we can see that the data is spread across 1790 to 1903 i.e., the speeches during these many years at the State of Union by respective president.

```
head(df)  # View the first few rows
```

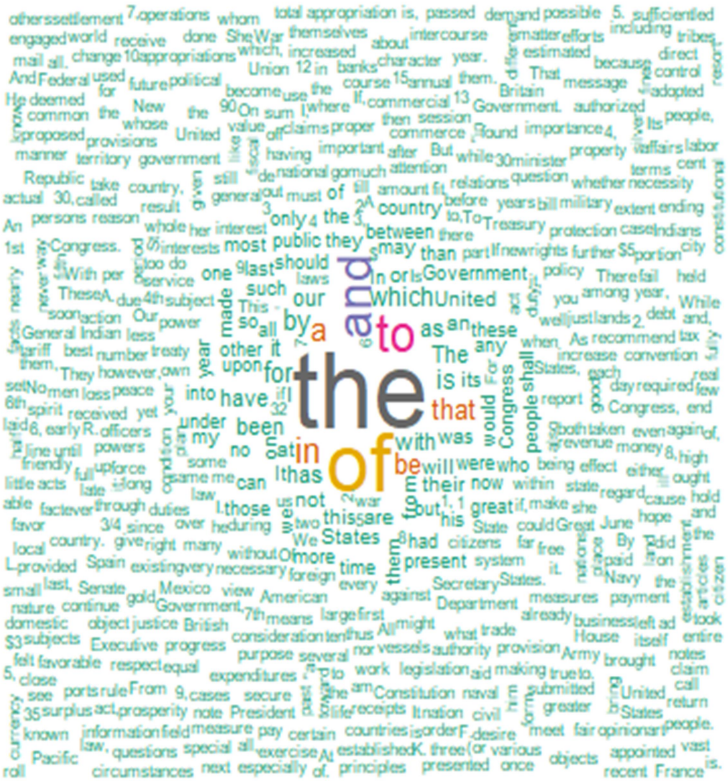
```
##   doc_id speech_doc_id                speech_type         president
date
## 1      1             1 State of the Union Address George Washington 1790-
01-08
## 2      2             1 State of the Union Address George Washington 1790-
01-08
## 3      3             1 State of the Union Address George Washington 1790-
01-08
## 4      4             1 State of the Union Address George Washington 1790-
01-08
## 5      5             1 State of the Union Address George Washington 1790-
01-08
## 6      6             1 State of the Union Address George Washington 1790-
01-08
##
text
## 1

Fellow-Citizens of the Senate and House of Representatives:
## 2 I embrace with great satisfaction the opportunity which now presents
itself\nof congratulating you on the present favorable prospects of our
public\naffairs. The recent accession of the important state of North
Carolina to\nthe Constitution of the United States (of which official
information has\nbeen received), the rising credit and respectability of our
country, the\ngeneral and increasing good will toward the government of the
Union, and\nthe concord, peace, and plenty with which we are blessed are
circumstances\nauspicious in an eminent degree to our national prosperity.
## 3                                                    In resuming your
consultations for the general good you can not but derive\nencouragement from
the reflection that the measures of the last session\nhave been as
satisfactory to your constituents as the novelty and\ndifficulty of the work
allowed you to hope. Still further to realize their\nexpectations and to
secure the blessings which a gracious Providence has\nplaced within our reach
will in the course of the present important session\ncall for the cool and
deliberate exertion of your patriotism, firmness, and\nwisdom.
## 4
Among the many interesting objects which will engage your attention that
of\nproviding for the common defense will merit particular regard. To
be\nprepared for war is one of the most effectual means of preserving peace.
## 5
A free people ought not only to be armed, but disciplined; to which end
```

```
a\nuniform and well-digested plan is requisite; and their safety and
interest\nrequire that they should promote such manufactories as tend to
render them\nindependent of others for essential, particularly military,
supplies.
## 6
The proper establishment of the troops which may be deemed
indispensable\nwill be entitled to mature consideration. In the arrangements
which may be\nmade respecting it it will be of importance to conciliate the
comfortable\nsupport of the officers and soldiers with a due regard to
economy.

tail(df)  # View the last few rows

##      doc_id speech_doc_id                speech_type          president
## 8828   8828          115 State of the Union Address Theodore Roosevelt
## 8829   8829          115 State of the Union Address Theodore Roosevelt
## 8830   8830          115 State of the Union Address Theodore Roosevelt
## 8831   8831          115 State of the Union Address Theodore Roosevelt
## 8832   8832          115 State of the Union Address Theodore Roosevelt
## 8833   8833          115 State of the Union Address Theodore Roosevelt
##            date
## 8828 1903-12-07
## 8829 1903-12-07
## 8830 1903-12-07
## 8831 1903-12-07
## 8832 1903-12-07
## 8833 1903-12-07
##
text
## 8828
The Indian agents should not be dependent for their appointment or\ntenure of
office upon considerations of partisan politics; the practice\nof appointing,
when possible, ex-army officers or bonded\nsuperintendents to the vacancies
that occur is working well. Attention\nis invited to the widespread
illiteracy due to lack of public schools\nin the Indian Territory. Prompt
heed should be paid to the need of\neducation for the children in this
Territory.
## 8829
In my last annual Message the attention of the Congress was called to\nthe
necessity of enlarging the safety-appliance law, and it is\ngratifying to
note that this law was amended in important respects.\nWith the increasing
railway mileage of the country, the greater number\nof men employed, and the
use of larger and heavier equipment, the\nurgency for renewed effort to
prevent the loss of life and limb upon\nthe railroads of the country,
particularly to employees, is apparent.\nFor the inspection of water craft
and the Life-Saving Service upon the\nwater the Congress has built up an
elaborate body of protective\nlegislation and a thorough method of inspection
and is annually\nspending large sums of money. It is encouraging to observe
that the\nCongress is alive to the interests of those who are employed upon
our\nwonderful arteries of commerce--the railroads--who so safely
```

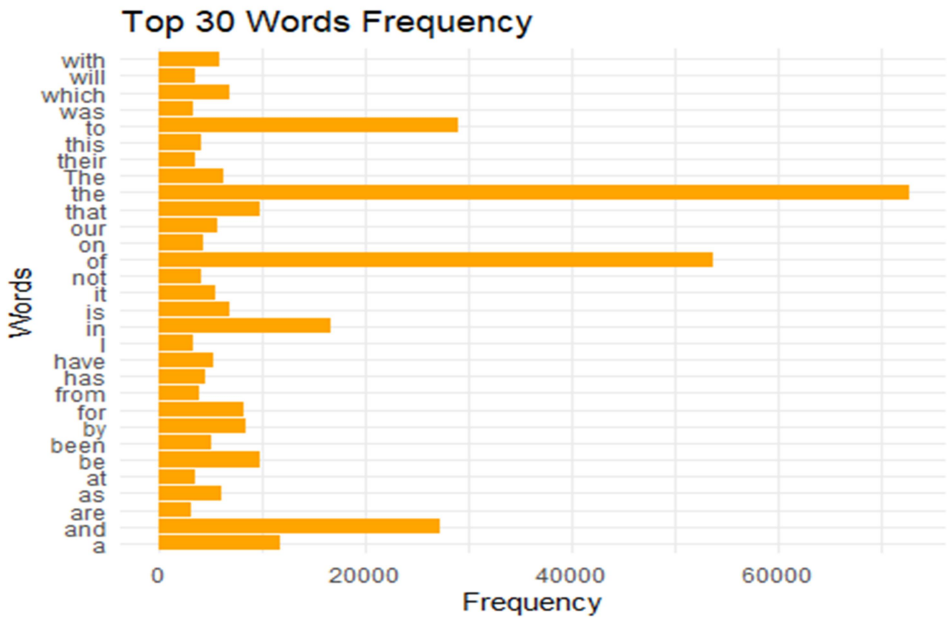
transport\nmillions of passengers and billions of tons of freight. The Federal\ninspection, of safety appliances, for which the Congress is now making\nappropriations, is a service analogous to that which the Government has\nupheld for generations in regard to vessels, and it is believed will\nprove of great practical benefit, both to railroad employees and the\ntraveling public. As the greater part of commerce is interstate and\nexclusively under the control of the Congress the needed safety and\nuniformity must be secured by national legislation.
8830
No other class of our citizens deserves so well of the Nation as those\nto whom the Nation owes its very being, the veterans of the civil war.\nSpecial attention is asked to the excellent work of the Pension Bureau\nin expediting and disposing of pension claims. During the fiscal year\nending July 1, 1903, the Bureau settled 251,982 claims, an average of\n825 claims for each working day of the year. The number of settlements\nsince July 1, 1903, has been in excess of last year's average,\napproaching 1,000 claims for each working day, and it is believed that\nthe work of the Bureau will be current at the close of the present\nfiscal year.
8831 During the year ended June 30 last 25,566 persons were appointed\nthrough competitive examinations under the civil-service rules. This\nwas 12,672 more than during the preceding year, and 40 per cent of\nthose who passed the examinations. This abnormal growth was largely\noccasioned by the extension of classification to the rural\nfree-delivery service and the appointment last year of over 9,000 rural\ncarriers. A revision of the civil-service rules took effect on April 15\nlast, which has greatly improved their operation. The completion of the\nreform of the civil service is recognized by good citizens everywhere\nas a matter of the highest public importance, and the success of the\nmerit system largely depends upon the effectiveness of the rules and\nthe machinery provided for their enforcement. A very gratifying spirit\nof friendly co-operation exists in all the Departments of the\nGovernment in the enforcement and uniform observance of both the letter\nand spirit of the civil-service act. Executive orders of July 3, 1902;\nMarch 26, 1903, and July 8, 1903, require that appointments of all\nunclassified laborers, both in the Departments at Washington and in the\nfield service, shall be made with the assistance of the United States\nCivil Service Commission, under a system of registration to test the\nrelative fitness of applicants for appointment or employment. This\nsystem is competitive, and is open to all citizens of the United States\nqualified in respect to age, physical ability, moral character,\nindustry, and adaptability for manual labor; except that in case of\nveterans of the Civil War the element of age is omitted. This system of\nappointment is distinct from the classified service and does not\nclassify positions of mere laborer under the civil-service act and\nrules. Regulations in aid thereof have been put in operation in several\nof the Departments and are being gradually extended in other parts of\nthe service. The results have been very satisfactory, as extravagance\nhas been checked by decreasing the number of unnecessary positions and\nby increasing the efficiency of the employees remaining.
8832
The Congress, as the result of a thorough investigation of the\ncharities and

reformatory institutions in the District of Columbia, by\na joint select committee of the two Houses which made its report in\nMarch, 1898, created in the act approved June 6, 1900, a board of\ncharities for the District of Columbia, to consist of five residents of\nthe District, appointed by the President of the United States, by and\nwith the advice and consent of the Senate, each for a term of three\nyears, to serve without compensation. President McKinley appointed five\nmen who had been active and prominent in the public charities in\nWashington, all of whom upon taking office July 1, 1900, resigned from\nthe different charities with which they had been connected. The members\nof the board have been reappointed in successive years. The board\nserves under the Commissioners of the District of Columbia. The board\ngave its first year to a careful and impartial study of the special\nproblems before it, and has continued that study every year in the\nlight of the best practice in public charities elsewhere. Its\nrecommendations in its annual reports to the Congress through the\nCommissioners of the District of Columbia "for the economical and\nefficient administration of the charities and reformatories of the\nDistrict of Columbia," as required by the act creating it, have been\nbased upon the principles commended by the joint select committee of\nthe Congress in its report of March, 1898, and approved by the best\nadministrators of public charities, and make for the desired\nsystematization and improvement of the affairs under its supervision.\nThey are worthy of favorable consideration by the Congress.
8833
The effect of the laws providing a General Staff for the Army and for\nthe more effective use of the National Guard has been excellent. Great\nimprovement has been made in the efficiency of our Army in recent\nyears. Such schools as those erected at Fort Leavenworth and Fort Riley\nand the institution of fall maneuver work accomplish satisfactory\nresults. The good effect of these maneuvers upon the National Guard is\nmarked, and ample appropriation should be made to enable the guardsmen\nof the several States to share in the benefit. The Government should as\nsoon as possible secure suitable permanent camp sites for military\nmaneuvers in the various sections of the country. The service thereby\nrendered not only to the Regular Army, but to the National Guard of the\nseveral States, will be so great as to repay many times over the\nrelatively small expense. We should not rest satisfied with what has\nbeen done, however. The only people who are contented with a system of\npromotion by mere seniority are those who are contented with the\ntriumph of mediocrity over excellence. On the other hand, a system\nwhich encouraged the exercise of social or political favoritism in\npromotions would be even worse. But it would surely be easy to devise a\nmethod of promotion from grade to grade in which the opinion of the\nhigher officers of the service upon the candidates should be decisive\nupon the standing and promotion of the latter. Just such a system now\nobtains at West Point. The quality of each year's work determines the\nstanding of that year's class, the man being dropped or graduated into\nthe next class in the relative position which his military superiors\ndecide to be warranted by his merit. In other words, ability, energy,\nfidelity, and all other similar qualities determine the rank of a man\nyear after year in West Point, and his standing in the Army

```
when he\ngraduates from West Point; but from that time on, all effort to
find\nwhich man is best or worst, and reward or punish him accordingly,
is\nabandoned; no bri

#Let’s plot a world cloud

# Load the required library
library(wordcloud)

# Tokenize the "text" column into words
words <- unlist(strsplit(df$text, " "))

# Calculate word frequencies
word_freq <- table(words)

# Sort the word frequencies in descending order
sorted_word_freq <- sort(word_freq, decreasing = TRUE)

# Create a word cloud with all words
wordcloud(words = names(sorted_word_freq),
          freq = sorted_word_freq,
          scale = c(3, 0.5), random.order = FALSE,
          colors = brewer.pal(8, "Dark2"))
```

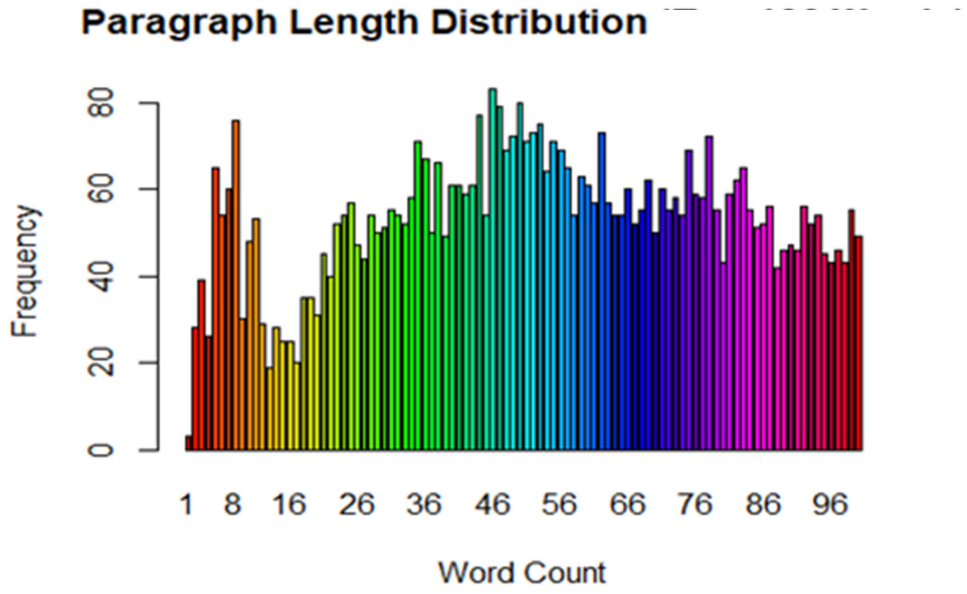


```r
#Let’s plot a frequency plot of words


# Tokenize the "text" column into words
words <- unlist(strsplit(df$text, " "))

# Calculate word frequencies
word_freq <- table(words)

# Sort the word frequencies in descending order
sorted_word_freq <- sort(word_freq, decreasing = TRUE)

# Select the top 30 words
top_30_words <- head(sorted_word_freq, 30)

# Create a data frame with appropriate column names
plot_data <- data.frame(Words = names(top_30_words), Frequency =
as.numeric(top_30_words))

# Create a frequency plot
ggplot(plot_data, aes(x = Words, y = Frequency)) +
  geom_bar(stat = "identity", fill = "orange") +
  labs(title = "Top 30 Words Frequency", x = "Words", y = "Frequency") +
  theme_minimal() +
  coord_flip()
```



After reviewing the data, it is clear that we need to clean it. There are lot of symbols and stopwords used which will not serve meaningful in our analysis and will give unexpected results so data cleaning and pre-processing is must here.

```r
# Calculate paragraph lengths
paragraph_lengths <- sapply(df$text, function(x) length(unlist(strsplit(x, " "))))

# Create a bar plot with colors and limit to 100 words
top_100_lengths <- paragraph_lengths[paragraph_lengths <= 100]  # Filter to 100 words or fewer

# Create a color palette (adjust the number of colors as needed)
colors <- rainbow(length(unique(top_100_lengths)))

# Create a bar plot with colors
barplot(table(top_100_lengths), main = "Paragraph Length Distribution",
        xlab = "Word Count", ylab = "Frequency", col = colors)
```



The above graph shows the paragraph length distribution.

3. Data Cleaning and Pre-processing

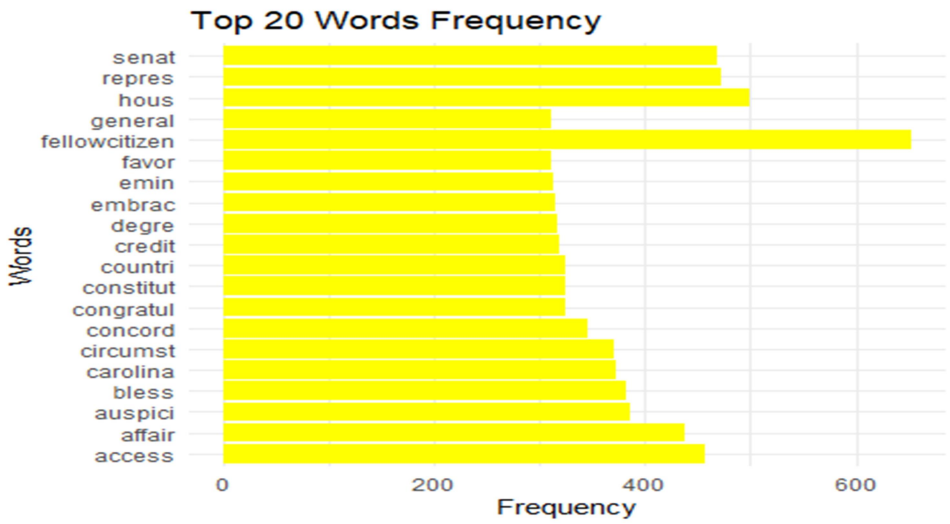
Cleaning data is crucial because it ensures the accuracy, reliability, and effectiveness of your analysis and decision-making processes.

```r
#Lets create a word cloud

# Create a corpus from the text data
corpus <- Corpus(DataframeSource(df))
```

```r
# Preprocessing chain
processedCorpus <- tm_map(corpus, content_transformer(tolower))  # Convert to
lowercase
processedCorpus <- tm_map(processedCorpus, removePunctuation)  # Remove
punctuation
processedCorpus <- tm_map(processedCorpus, removeNumbers)  # Remove numbers

# Remove inbuilt English stopwords
processedCorpus <- tm_map(processedCorpus, removeWords, stopwords("en"))

# Stemming (if desired, you can skip this step if it doesn't improve results)
processedCorpus <- tm_map(processedCorpus, stemDocument, language = "en")

# Strip extra whitespace
processedCorpus <- tm_map(processedCorpus, stripWhitespace)

# compute document term matrix with terms >= minimumFrequency
minimumFrequency <- 5
DTM <- DocumentTermMatrix(processedCorpus, control = list(bounds =
list(global = c(minimumFrequency, Inf))))

# due to vocabulary pruning, we have empty rows in our DTM
# LDA does not like this. So we remove those docs from the
# DTM and the metadata
sel_idx <- slam::row_sums(DTM) > 0
DTM <- DTM[sel_idx, ]
df<- df[sel_idx, ]

# Get the terms (words) from the DTM
terms <- Terms(DTM)

# Create a named vector of word frequencies
word_freq <- as.vector(slam::row_sums(DTM))

# Sort the word frequencies in descending order
sorted_word_freq <- sort(word_freq, decreasing = TRUE)

# Filter the word frequencies to include only non-zero values
non_zero_word_freq <- sorted_word_freq[sorted_word_freq > 0]

# Create a word cloud with all words
wordcloud(words = terms,
          freq = non_zero_word_freq,
          scale = c(2, 0.5), random.order = FALSE,
          colors = brewer.pal(8, "Dark2"))
```


```
#Lets plot a frequency plot of words

# Create a corpus from the text data
corpus <- Corpus(DataframeSource(df))

# Preprocessing chain
processedCorpus <- tm_map(corpus, content_transformer(tolower))  # Convert to lowercase
processedCorpus <- tm_map(processedCorpus, removePunctuation)  # Remove punctuation
processedCorpus <- tm_map(processedCorpus, removeNumbers)  # Remove numbers

# Remove inbuilt English stopwords
processedCorpus <- tm_map(processedCorpus, removeWords, stopwords("en"))

# Stemming (if desired, you can skip this step if it doesn't improve results)
processedCorpus <- tm_map(processedCorpus, stemDocument, language = "en")

# Strip extra whitespace
processedCorpus <- tm_map(processedCorpus, stripWhitespace)
# compute document term matrix with terms >= minimumFrequency
minimumFrequency <- 5
DTM <- DocumentTermMatrix(processedCorpus, control = list(bounds =
list(global = c(minimumFrequency, Inf))))
# due to vocabulary pruning, we have empty rows in our DTM
# LDA does not like this. So we remove those docs from the
# DTM and the metadata
sel_idx <- slam::row_sums(DTM) > 0
```

```r
DTM <- DTM[sel_idx, ]
df<- df[sel_idx, ]
# Get the terms (words) from the DTM
terms <- Terms(DTM)

# Create a named vector of word frequencies
word_freq <- as.vector(slam::row_sums(DTM))

# Sort the word frequencies in descending order
sorted_word_freq <- sort(word_freq, decreasing = TRUE)

# Filter the word frequencies to include only non-zero values
non_zero_word_freq <- sorted_word_freq[sorted_word_freq > 0]

# Select the top N words (e.g., top 20 words)
top_N <- 20
top_terms <- terms[1:top_N]
top_freq <- non_zero_word_freq[1:top_N]

# Create a data frame for the bar plot
plot_data <- data.frame(Words = top_terms, Frequency = top_freq)

# Create a bar plot with flipped coordinates and minimal theme
library(ggplot2)
ggplot(plot_data, aes(x = Frequency, y = Words)) +
geom_bar(stat = "identity", fill = "yellow") +
labs(title = "Top 20 Words Frequency", x = "Frequency", y = "Words") +
theme_minimal() +
theme(legend.position = "none") # Remove the legend for fill
```



After cleaning the data this world cloud now make sense and we can tell about the content of the speech. These are just the top 30 words but still

4. Model Building

4.1 Finding Value of K

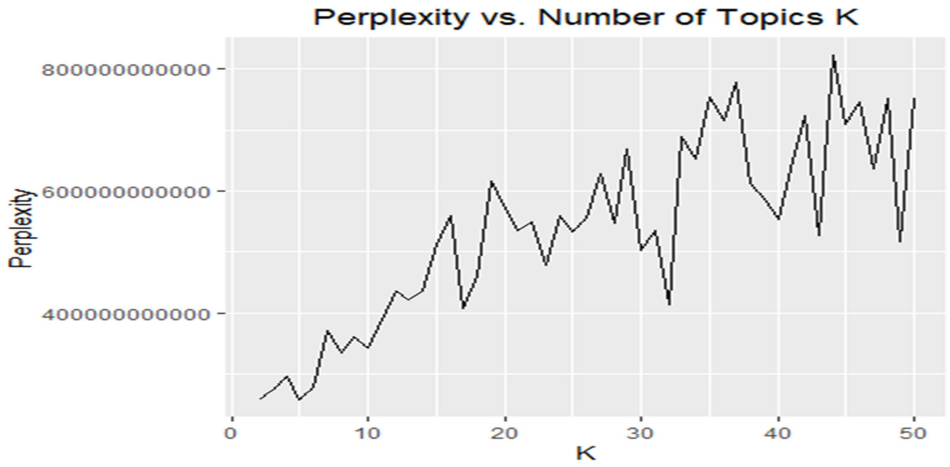
For models with parameters like Latent Dirichlet Allocation (LDA), selecting the number of topics i.e., K, is a crucial step. The choice of the optimal K depends on various factors. If K is set too low, the collection is segmented into only a few broad semantic contexts. Conversely, if K is excessively high, the collection is split into numerous topics, some of which might overlap, while others become challenging to interpret.

4.1.1 Perplexity

We will use a perplexity approach first and try to get the number of topics through minimum perplexity.

```
set.seed(87460945)

# Calculate folds
idxs <- sample(seq_len(9))
folds <- split(idxs, rep(1:3, each = 3, length.out = 9))

# Define number of topics
topics <- seq(2, 50, 1)

# Create data frame for storing results
results <- data.frame()

# Perform cross-validation
for (k in topics) {
  scores <- c()
  for (i in 1:3) {
    test_idx <- folds[[i]]
    train_idx <- setdiff(unlist(folds, use.names = FALSE), test_idx)

    test <- DTM[test_idx, ]
    train <- DTM[train_idx, ]

    LDA.out <- LDA(train, k, method = "Gibbs")

    # Calculate perplexity for the test data
    log_likelihood <- logLik(LDA.out, newdata = test)
    n_tokens_test <- sum(test)
    perplexity <- exp(-sum(log_likelihood) / n_tokens_test)

    scores <- c(scores, perplexity)
  }
  temp <- data.frame("K" = k, "Perplexity" = mean(scores))
  results <- rbind(results, temp)
}
```

```r
# Plot Perplexity vs. K
library(ggplot2)
ggplot(results, aes(x = K, y = Perplexity)) +
  geom_line() +
  ggtitle("Perplexity vs. Number of Topics K") +
  theme(plot.title = element_text(hjust = 0.5))
```



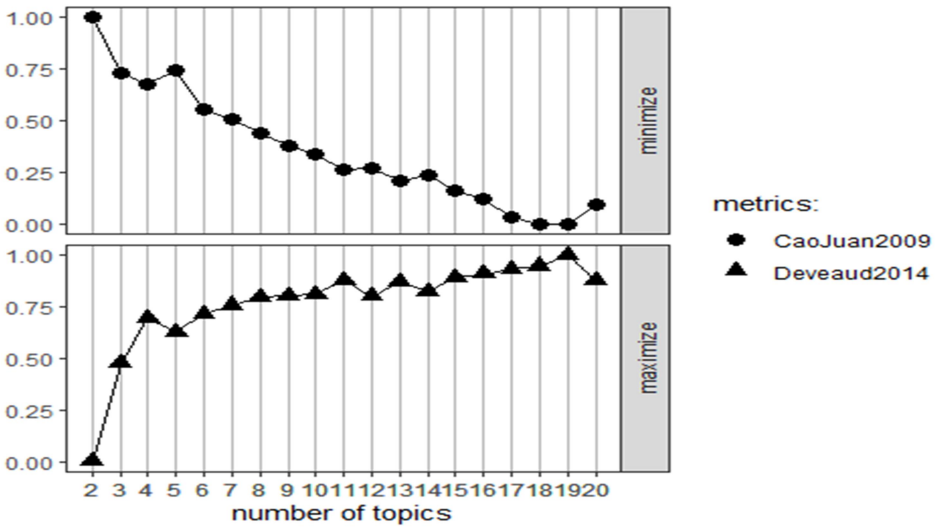

```r
# Find the row with the minimum perplexity
min_perplexity_row <- results[which.min(results$Perplexity), ]

# Extract the minimum perplexity value and corresponding K value
min_perplexity <- min_perplexity_row$Perplexity
corresponding_K <- min_perplexity_row$K

# Print the results
cat("Minimum Perplexity:", min_perplexity, "\n")

## Minimum Perplexity: 258255023132

cat("Corresponding K (Number of Topics):", corresponding_K, "\n")

## Corresponding K (Number of Topics): 5
```

From the above graph of perplexity though the minimum perplexity is at 5 number of topics but still we can see that there are multiple elbow points so we need to consider other approaches or a combination of methods to make an informed decision

4.1.2 FindTopicNumber

We will use a FindTopicNumber approach to determining the number of topics which explores a range of values rather than fixating on a specific number. We will use only two metrics (CaoJuan2009 and Deveaud2014)

```
# create models with different number of topics
result <- ldatuning::FindTopicsNumber(
  DTM,
  topics = seq(from = 2, to = 20, by = 1),
  metrics = c("CaoJuan2009",  "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 77),
  verbose = TRUE
)

## fit models... done.
## calculate metrics:
##   CaoJuan2009... done.
##   Deveaud2014... done.

#Plot the results
FindTopicsNumber_plot(result)

## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use
"none" instead as
## of ggplot2 3.3.4.
## ℹ The deprecated feature was likely used in the ldatuning package.
##   Please report the issue at <https://github.com/nikita-
moor/ldatuning/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



For our minima we see CaoJuan2009 suggests 20 and our maxima we find Deveaud2014 also suggests around 18 and 20. For our first analysis, however, we choose a thematic "resolution" of K = 20 topics.

4.2 Model Fitting

```
# number of topics
K <- 20

# set random number generator seed
set.seed(9161)

# compute the LDA model, inference via 1000 iterations of Gibbs sampling
topicModel <- LDA(DTM, K, method="Gibbs", control=list(iter = 500, verbose =
25))
```

```
## K = 20; V = 4479; M = 8811
## Sampling 500 iterations!
## Iteration 25 ...
## Iteration 50 ...
## Iteration 75 ...
## Iteration 100 ...
## Iteration 125 ...
## Iteration 150 ...
## Iteration 175 ...
## Iteration 200 ...
## Iteration 225 ...
## Iteration 250 ...
## Iteration 275 ...
## Iteration 300 ...
## Iteration 325 ...
## Iteration 350 ...
## Iteration 375 ...
## Iteration 400 ...
## Iteration 425 ...
## Iteration 450 ...
## Iteration 475 ...
## Iteration 500 ...
## Gibbs sampling completed!
```

The topic model inference results in two (approximate) posterior probability distributions: a distribution theta over K topics within each document and a distribution beta over V terms within each topic, where V represents the length of the vocabulary of the collection (V = 4278). Let's take a closer look at these results:

```
# have a look a some of the results (posterior distributions)
tmResult <- posterior(topicModel)
# format of the resulting object
attributes(tmResult)
```

```
## $names
## [1] "terms"  "topics"
```

```
nTerms(DTM)              # lengthOfVocab
```

```
## [1] 4479

# topics are probability distributions over the entire vocabulary
beta <- tmResult$terms   # get beta from results
dim(beta)                # K distributions over nTerms(DTM) terms

## [1]   20 4479

rowSums(beta)            # rows in beta sum to 1

##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1

nDocs(DTM)               # size of collection

## [1] 8811

# for every document we have a probability distribution of its contained
topics
theta <- tmResult$topics
dim(theta)               # nDocs(DTM) distributions over K topics

## [1] 8811   20

rowSums(theta)[1:10]     # rows in theta sum to 1

##  1  2  3  4  5  6  7  8  9 10
##  1  1  1  1  1  1  1  1  1  1
```

Let's take a look at the 10 most likely terms within the term probabilities beta of the inferred topics (only the first 8 are shown below).

For the next steps, we want to give the topics more descriptive names than just numbers. Therefore, we simply concatenate the five most likely terms of each topic to a string that represents a pseudo-name for each topic

Let us now look more closely at the distribution of topics within individual documents. To this end, we visualize the distribution in 3 sample documents. Let us first take a look at the contents of three sample documents:

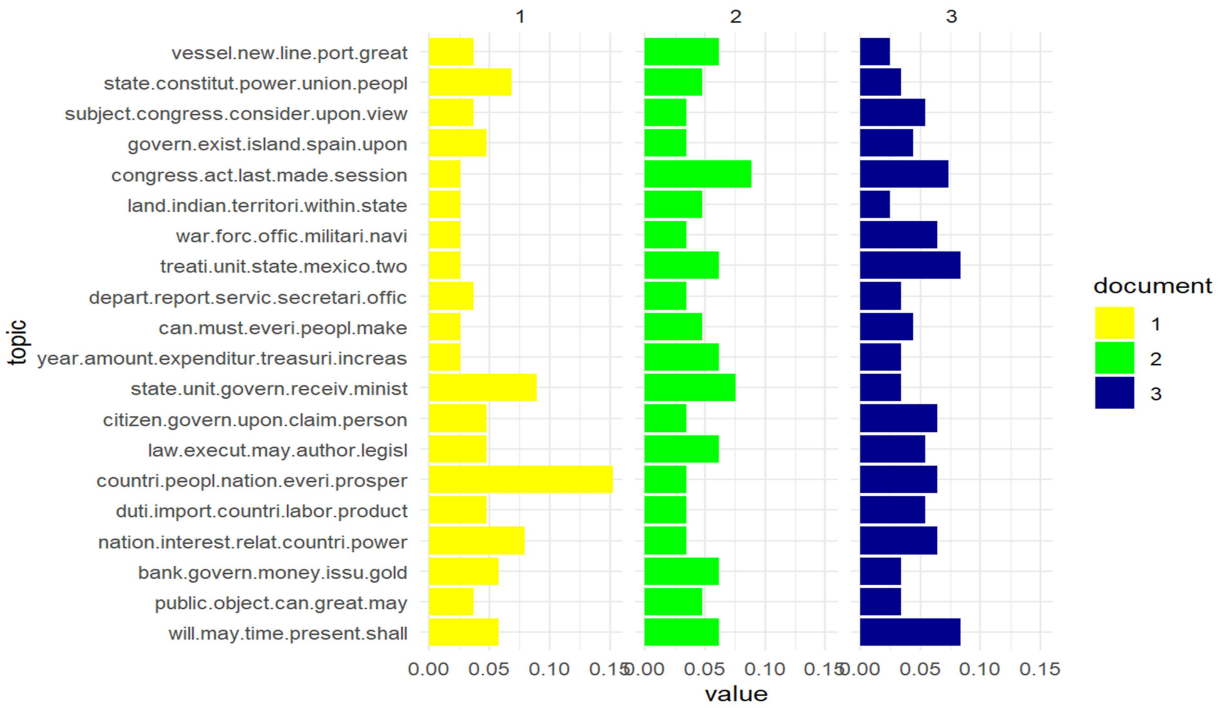
```
exampleIds <- c(2, 100, 200)
lapply(corpus[exampleIds], as.character)

## $`2`
## [1] "I embrace with great satisfaction the opportunity which now presents
itself\nof congratulating you on the present favorable prospects of our
public\naffairs. The recent accession of the important state of North
Carolina to\nthe Constitution of the United States (of which official
information has\nbeen received), the rising credit and respectability of our
country, the\ngeneral and increasing good will toward the government of the
```

```
Union, and\nthe concord, peace, and plenty with which we are blessed are
circumstances\nauspicious in an eminent degree to our national prosperity."
##
## $`100`
## [1] "Provision is likewise requisite for the reimbursement of the loan
which has\nbeen made of the Bank of the United States, pursuant to the
eleventh\nsection of the act by which it is incorporated. In fulfilling the
public\nstipulations in this particular it is expected a valuable saving will
be\nmade."
##
## $`200`
## [1] "After many delays and disappointments arising out of the European
war, the\nfinal arrangements for fulfilling the engagements made to the Dey
and\nRegency of Algiers will in all present appearance be crowned with
success,\nbut under great, though inevitable, disadvantages in the
pecuniary\ntransactions occasioned by that war, which will render further
provision\nnecessary. The actual liberation of all our citizens who were
prisoners in\nAlgiers, while it gratifies every feeling of heart, is itself
an earnest of\na satisfactory termination of the whole negotiation. Measures
are in\noperation for effecting treaties with the Regencies of Tunis and
Tripoli."

exampleIds <- c(2, 100, 200)
print(paste0(exampleIds[1], ": ", substr(content(corpus[[exampleIds[1]]]), 0,
400), '...'))

## [1] "2: I embrace with great satisfaction the opportunity which now
presents itself\nof congratulating you on the present favorable prospects of
our public\naffairs. The recent accession of the important state of North
Carolina to\nthe Constitution of the United States (of which official
information has\nbeen received), the rising credit and respectability of our
country, the\ngeneral and increasing good will ..."

print(paste0(exampleIds[2], ": ", substr(content(corpus[[exampleIds[2]]]), 0,
400), '...'))

## [1] "100: Provision is likewise requisite for the reimbursement of the
loan which has\nbeen made of the Bank of the United States, pursuant to the
eleventh\nsection of the act by which it is incorporated. In fulfilling the
public\nstipulations in this particular it is expected a valuable saving will
be\nmade...."

print(paste0(exampleIds[3], ": ", substr(content(corpus[[exampleIds[3]]]), 0,
400), '...'))

## [1] "200: After many delays and disappointments arising out of the
European war, the\nfinal arrangements for fulfilling the engagements made to
the Dey and\nRegency of Algiers will in all present appearance be crowned
with success,\nbut under great, though inevitable, disadvantages in the
pecuniary\ntransactions occasioned by that war, which will render further
provision\nnecessary. The actual liberation of all ..."
```

5. MODEL EVALUATION AND RESULTS

Let's visualize the topic distributions within the documents

```
N <- length(exampleIds)
# get topic proportions form example documents
topicProportionExamples <- theta[exampleIds,]
colnames(topicProportionExamples) <- topicNames
vizDataFrame <- melt(cbind(data.frame(topicProportionExamples), document =
factor(1:N)), variable.name = "topic", id.vars = "document")
ggplot(data = vizDataFrame, aes(topic, value, fill = document), ylab =
"proportion") +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  coord_flip() +
  facet_wrap(~ document, ncol = N) +
  scale_fill_manual(values = c("yellow", "green",
"darkblue"))+theme_minimal()
```



The diagram above illustrates the distribution of topics within a document based on the model. In the current model, all three documents exhibit some degree of each topic, but a few topics dominate each document

The allocation of topics within a document can be adjusted using the Alpha parameter of the model. Higher Alpha priors for topics lead to a more uniform distribution of topics within a

document, while lower Alpha priors ensure that the inference process concentrates the probability on a select few topics for each document

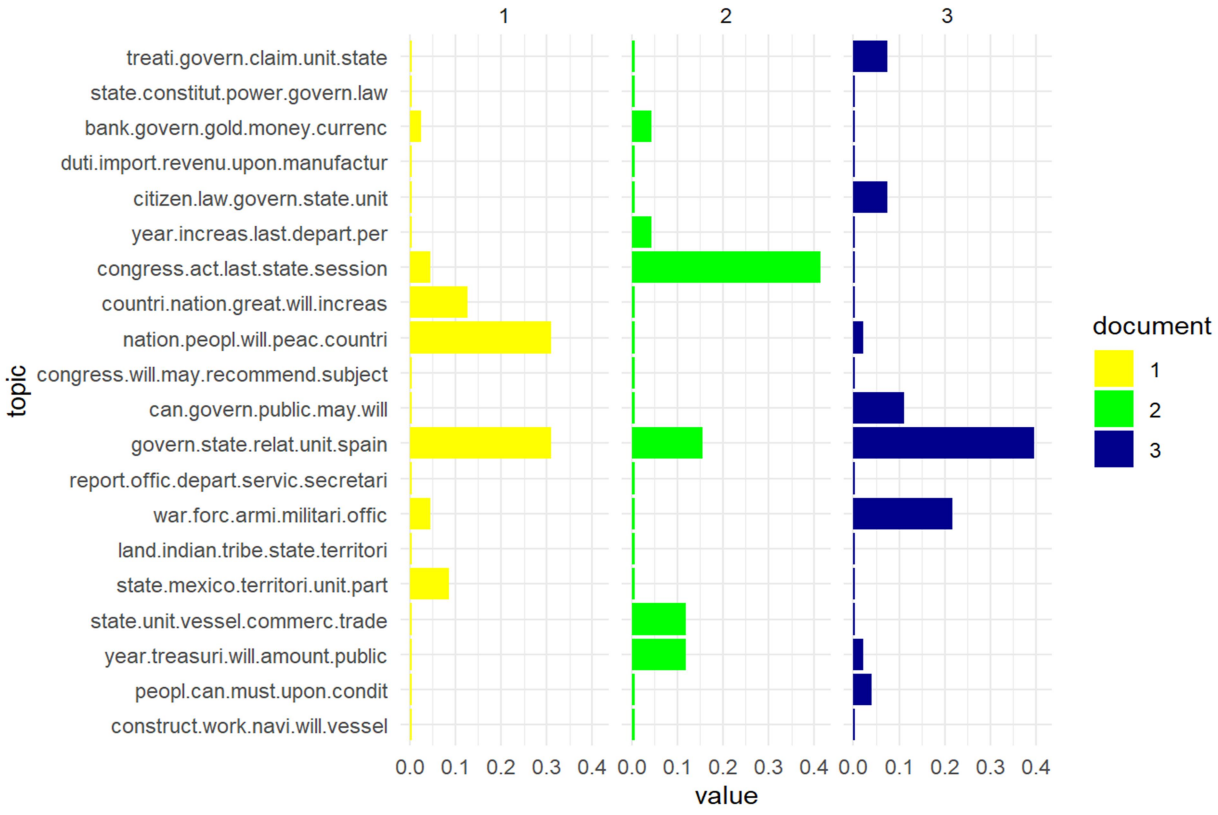
In the previous model calculation, the Alpha prior was automatically estimated to best fit the data (maximizing the overall model probability). However, this automatic estimate may not align with the preferences of an analyst. Depending on our analytical goals, we may desire a distribution of topics in the model that is either more concentrated or more evenly spread

Now, let's modify the Alpha prior to a lower value to observe its impact on the topic distributions in the model.

```
# see alpha from previous model
attr(topicModel, "alpha")

## [1] 2.5

topicModel2 <- LDA(DTM, K, method="Gibbs", control=list(iter = 500, verbose =
25, alpha = 0.2))

## K = 20; V = 4479; M = 8811
## Sampling 500 iterations!
## Iteration 25 ...
## Iteration 50 ...
## Iteration 75 ...
## Iteration 100 ...
## Iteration 125 ...
## Iteration 150 ...
## Iteration 175 ...
## Iteration 200 ...
## Iteration 225 ...
## Iteration 250 ...
## Iteration 275 ...
## Iteration 300 ...
## Iteration 325 ...
## Iteration 350 ...
## Iteration 375 ...
## Iteration 400 ...
## Iteration 425 ...
## Iteration 450 ...
## Iteration 475 ...
## Iteration 500 ...
## Gibbs sampling completed!

tmResult <- posterior(topicModel2)
theta <- tmResult$topics
beta <- tmResult$terms
topicNames <- apply(terms(topicModel2, 5), 2, paste, collapse = " ")  # reset
topicnames
```

Now visualize the topic distributions in the three documents again and see the difference.

```
# get topic proportions form example documents
topicProportionExamples <- theta[exampleIds,]
colnames(topicProportionExamples) <- topicNames
vizDataFrame <- melt(cbind(data.frame(topicProportionExamples), document =
factor(1:N)), variable.name = "topic", id.vars = "document")
ggplot(data = vizDataFrame, aes(topic, value, fill = document), ylab =
"proportion") +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  coord_flip() +
  facet_wrap(~ document, ncol = N) +
  scale_fill_manual(values = c("yellow", "green",
"darkblue"))+theme_minimal()
```



Now we can see that the inference process distributes the probability mass on a few topics for each document.

5.1 Topic Ranking

Initially, we aim to establish a more significant arrangement of the most important terms for each topic by assigning them a particular score. This concept of reordering terms shares similarities with the TF-IDF (Term Frequency-Inverse Document Frequency) approach. Essentially, if a term frequently appears at higher levels relative to its probability, it becomes

less valuable in terms of describing the topic. Consequently, this scoring method gives preference to terms that effectively represent a topic.

```
# re-rank top topic terms for topic names
topicNames <- apply(lda::top.topic.words(beta, 5, by.score = T), 2, paste,
collapse = " ")
```

What are the defining topics within a collection? There are different approaches to find out which can be used to bring the topics into a certain order.

5.1.1 Approach 1

We sort topics according to their probability within the entire collection

```
# What are the most probable topics in the entire collection?
topicProportions <- colSums(theta) / nDocs(DTM)  # mean probabilities over
all paragraphs
names(topicProportions) <- topicNames     # assign the topic names we created
before
sort(topicProportions, decreasing = TRUE) # show summed proportions in
decreased order
```

```
##               public can may object power
##                                    0.06427
##           relat spain govern state minist
##                                    0.06358
##    congress recommend attent will subject
##                                    0.05979
##         nation peac peopl prosper countri
##                                    0.05968
##      report depart offic servic secretari
##                                    0.05924
##           treati claim convent senat unit
##                                    0.05602
##           state constitut power law union
##                                    0.05321
##      year treasuri amount expenditur debt
##                                    0.05197
##           congress act last session state
##                                    0.05124
##             citizen law claim case person
##                                    0.04882
##                  peopl can must labor men
##                                    0.04805
## countri nation increas product agricultur
##                                    0.04803
##              war armi forc militari offic
##                                    0.04695
##              state unit vessel port trade
##                                    0.04636
##           construct navi vessel work ship
```

```
##                                     0.04414
##                 year increas per last cent
##                                     0.04195
##             bank gold currenc silver money
##                                     0.04122
##     duti revenu tariff manufactur product
##                                     0.04088
##           land indian tribe territori acr
##                                     0.03839
##          mexico state territori texa unit
##                                     0.03622

# What are the most probable topics in the entire collection?
topicProportions <- colSums(theta) / nDocs(DTM)  # mean probabilities over
all paragraphs
names(topicProportions) <- topicNames     # assign the topic names we created
before
sort(topicProportions, decreasing = TRUE) # show summed proportions in
decreased order

##                public can may object power
##                                     0.06427
##           relat spain govern state minist
##                                     0.06358
##    congress recommend attent will subject
##                                     0.05979
##         nation peac peopl prosper countri
##                                     0.05968
##      report depart offic servic secretari
##                                     0.05924
##           treati claim convent senat unit
##                                     0.05602
##           state constitut power law union
##                                     0.05321
##      year treasuri amount expenditur debt
##                                     0.05197
##           congress act last session state
##                                     0.05124
##             citizen law claim case person
##                                     0.04882
##                  peopl can must labor men
##                                     0.04805
## countri nation increas product agricultur
##                                     0.04803
##              war armi forc militari offic
##                                     0.04695
##              state unit vessel port trade
##                                     0.04636
##           construct navi vessel work ship
##                                     0.04414
```

```
##                year increas per last cent
##                                   0.04195
##            bank gold currenc silver money
##                                   0.04122
##     duti revenu tariff manufactur product
##                                   0.04088
##           land indian tribe territori acr
##                                   0.03839
##          mexico state territori texa unit
##                                   0.03622
```

```
soP <- sort(topicProportions, decreasing = TRUE)
paste(round(soP, 5), ":", names(soP))
```

```
##  [1] "0.06427 : public can may object power"
##  [2] "0.06358 : relat spain govern state minist"
##  [3] "0.05979 : congress recommend attent will subject"
##  [4] "0.05968 : nation peac peopl prosper countri"
##  [5] "0.05924 : report depart offic servic secretari"
##  [6] "0.05602 : treati claim convent senat unit"
##  [7] "0.05321 : state constitut power law union"
##  [8] "0.05197 : year treasuri amount expenditur debt"
##  [9] "0.05124 : congress act last session state"
## [10] "0.04882 : citizen law claim case person"
## [11] "0.04805 : peopl can must labor men"
## [12] "0.04803 : countri nation increas product agricultur"
## [13] "0.04695 : war armi forc militari offic"
## [14] "0.04636 : state unit vessel port trade"
## [15] "0.04414 : construct navi vessel work ship"
## [16] "0.04195 : year increas per last cent"
## [17] "0.04122 : bank gold currenc silver money"
## [18] "0.04088 : duti revenu tariff manufactur product"
## [19] "0.03839 : land indian tribe territori acr"
## [20] "0.03622 : mexico state territori texa unit"
```

We recognize some topics that are way more likely to occur in the corpus than others. These describe rather general thematic coherence. Other topics correspond more to specific contents.

5.1.2 Approach 2

We count how often a topic appears as a primary topic within a paragraph This method is also called Rank-1

```
countsOfPrimaryTopics <- rep(0, K)
names(countsOfPrimaryTopics) <- topicNames
for (i in 1:nDocs(DTM)) {
  topicsPerDoc <- theta[i, ] # select topic distribution for document i
  # get first element position from ordered list
  primaryTopic <- order(topicsPerDoc, decreasing = TRUE)[1]
  countsOfPrimaryTopics[primaryTopic] <- countsOfPrimaryTopics[primaryTopic]
```

```
+ 1
}
sort(countsOfPrimaryTopics, decreasing = TRUE)
```

```
##           relat spain govern state minist
##                                        685
##           treati claim convent senat unit
##                                        594
##      report depart offic servic secretari
##                                        588
##               public can may object power
##                                        557
##         nation peac peopl prosper countri
##                                        550
##      year treasuri amount expenditur debt
##                                        484
##    congress recommend attent will subject
##                                        468
##           construct navi vessel work ship
##                                        440
##           state constitut power law union
##                                        433
##              state unit vessel port trade
##                                        429
##              war armi forc militari offic
##                                        418
##             citizen law claim case person
##                                        401
##            bank gold currenc silver money
##                                        395
##           land indian tribe territori acr
##                                        375
##                  peopl can must labor men
##                                        364
##           congress act last session state
##                                        360
##                year increas per last cent
##                                        355
## countri nation increas product agricultur
##                                        353
##     duti revenu tariff manufactur product
##                                        297
##          mexico state territori texa unit
##                                        265
```

```
so <- sort(countsOfPrimaryTopics, decreasing = TRUE)
paste(so, ":", names(so))
```

```
##  [1] "685 : relat spain govern state minist"
##  [2] "594 : treati claim convent senat unit"
```

```
##  [3] "588 : report depart offic servic secretari"
##  [4] "557 : public can may object power"
##  [5] "550 : nation peac peopl prosper countri"
##  [6] "484 : year treasuri amount expenditur debt"
##  [7] "468 : congress recommend attent will subject"
##  [8] "440 : construct navi vessel work ship"
##  [9] "433 : state constitut power law union"
## [10] "429 : state unit vessel port trade"
## [11] "418 : war armi forc militari offic"
## [12] "401 : citizen law claim case person"
## [13] "395 : bank gold currenc silver money"
## [14] "375 : land indian tribe territori acr"
## [15] "364 : peopl can must labor men"
## [16] "360 : congress act last session state"
## [17] "355 : year increas per last cent"
## [18] "353 : countri nation increas product agricultur"
## [19] "297 : duti revenu tariff manufactur product"
## [20] "265 : mexico state territori texa unit"
```

We observe that when we arrange topics using the Rank-1 method, topics characterized by fairly distinct thematic cohesiveness are positioned towards the top of the list.

This organization of topics can be applied to subsequent analysis procedures, including interpreting the semantic content of topics within the collection, examining time series data related to the most significant topics, or filtering the original collection based on particular sub-topics.

5.2 Filtering the Documents

The availability of topic probabilities for each document, or in our case, each paragraph, within a topic model enables us to employ it for thematic filtration of a collection. As part of the filtering process, we choose to retain only those documents that surpass a specific threshold in terms of their probability value for particular topics. For instance, we may opt to retain every document that contains more than 20 percent of topic X.

In the subsequent steps, we will filter documents based on their topic content and illustrate how this impacts the overall number of documents over time.

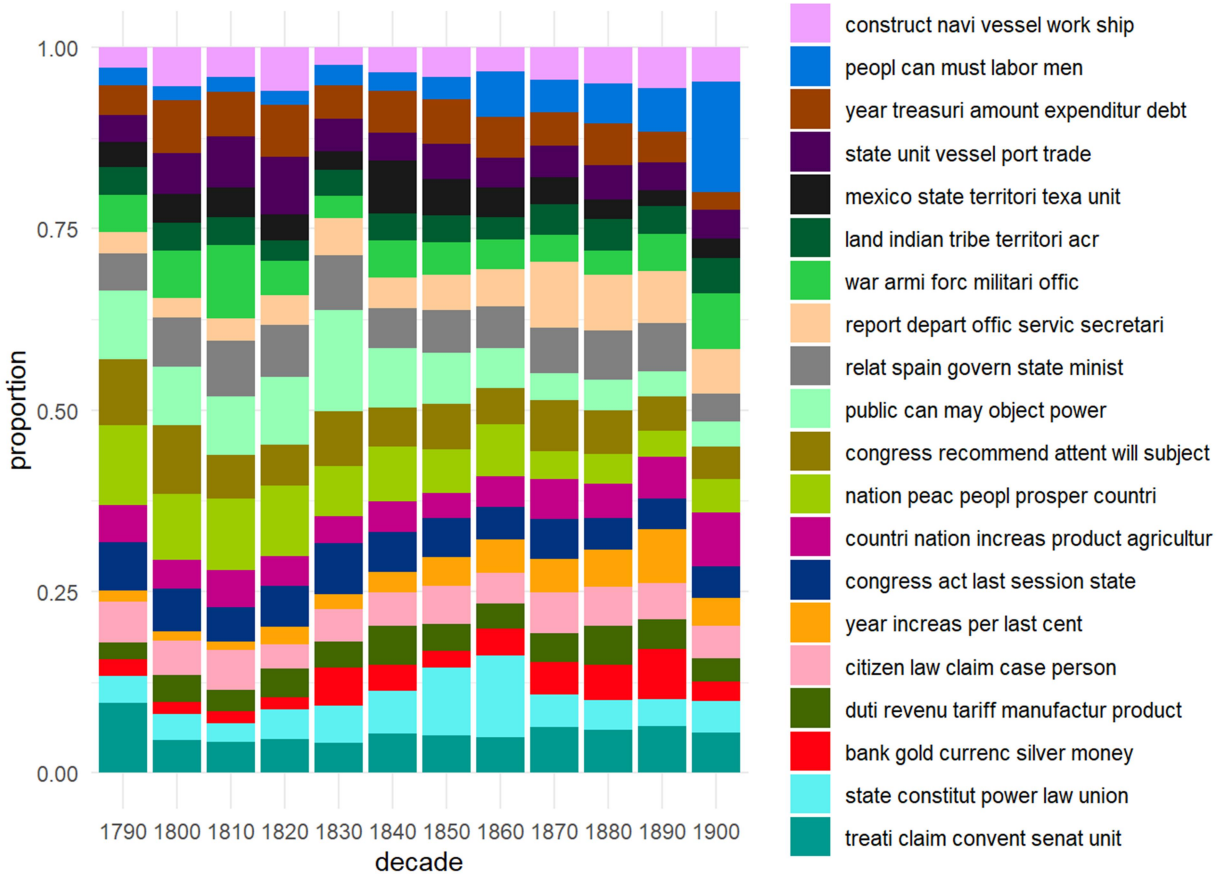
```
topicToFilter <- 6  # you can set this manually ...
# ... or have it selected by a term in the topic name (e.g. 'children')
topicToFilter <- grep('children', topicNames)[1]
topicThreshold <- 0.2
selectedDocumentIndexes <- which(theta[, topicToFilter] >= topicThreshold)
filteredCorpus <- corpus[selectedDocumentIndexes]
# show length of filtered corpus
filteredCorpus

## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 4
## Content:  documents: 0
```

Our filtered corpus contains 0 documents related to the topic NA to at least 20 %

5.3 Topic Proportions over Time

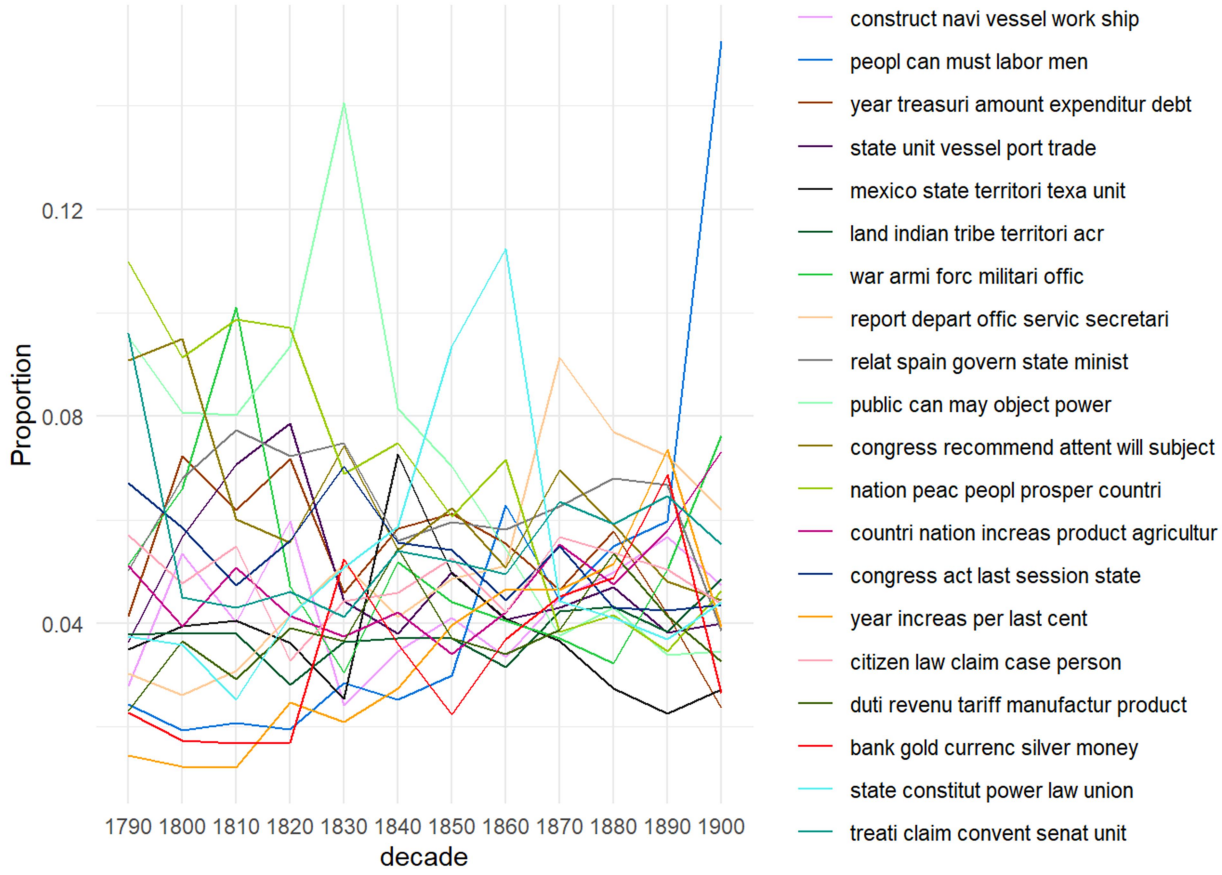
Finally, we take a broader look at the evolution of topics within the dataset across different time periods. To achieve this, we calculate the average topic proportions for each decade, considering all State of the Union speeches. These consolidated topic proportions can then be represented visually, for example, in the form of a bar plot.

```
# append decade information for aggregation
df$decade <- paste0(substr(df$date, 0, 3), "0")
# get mean topic proportions per decade
topic_proportion_per_decade <- aggregate(theta, by = list(decade =
df$decade), mean)
# set topic names to aggregated columns
colnames(topic_proportion_per_decade)[2:(K+1)] <- topicNames
# reshape data frame
vizDataFrame <- melt(topic_proportion_per_decade, id.vars = "decade")
# plot topic proportions per decade as bar plot
ggplot(vizDataFrame, aes(x=decade, y=value, fill=variable)) +
geom_bar(stat = "identity") + ylab("proportion") +
scale_fill_manual(values = paste0(alphabet(20), "FF"), name = "decade") +
theme(axis.text.x = element_text(angle = 90, hjust = 1))+theme_minimal()
```



#Plotting a line chart

```
# Load required libraries
library(ggplot2)

# Your data preparation code (append decade information and aggregate)
remains the same

# Reshape data frame for a line plot
vizDataFrame <- melt(topic_proportion_per_decade, id.vars = "decade")

# Create a line plot for topic proportions per decade
ggplot(vizDataFrame, aes(x = decade, y = value, color = variable, group =
variable)) +
  geom_line() +
  ylab("Proportion") +
  scale_color_manual(values = paste0(alphabet(20), "FF"), name = "Topic") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+theme_minimal()
```



The visualization illustrates that in the initial decades, topics related to the interaction between the federal government and individual states, along with internal conflicts, clearly take precedence. In contrast, contemporary State of the Union (SOTU) addresses primarily revolve around security matters and economic concerns, signifying their heightened importance.

Conclusion:

In essence, this assignment has not only equipped us with a deeper understanding of the content and evolution of State of the Union speeches but also demonstrated the power of topic modeling as a valuable tool for uncovering hidden patterns and trends in large textual datasets. By accomplishing these objectives, we have enhanced our ability to analyze and interpret historical political discourse, providing valuable insights for researchers and policymakers alike.