Changepoint Analysis of Linguistics in R

1 Loading All Packages

```
library(stringr)
library(tidytext)
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(plotly)
```

2 The Dataset

Similar to the work (https://subroy13.github.io/NLPinR/Reports/chapter2.html) before, we shall be using US Presidential Speeches Dataset (https://github.com/kfogel/presidential-speeches) by Kary Fogel (https://github.com/kfogel) in his github respository. I have gone ahead and downloaded the repository locally, and then unzipped it.

All the speeches till today 2020-05-17 11:01:42 is available in his repository, within the data folder, and we are going to analyze this rich dataset.

```
speeches <- list.files('../datasets/presidential-speeches/')
head(speeches)
```

```
[1] "1789-04-30-first-inaugural-address.txt"
[2] "1789-10-03-thanksgiving-proclamation.txt"
[3] "1790-01-08-first-annual-message-congress.txt"
[4] "1790-12-08-second-annual-message-congress.txt"
[5] "1790-12-29-talk-chiefs-and-counselors-seneca-nation.txt"
[6] "1791-10-25-third-annual-message-congress.txt"
```

Now, as we have seen before, the first inaugural address by George Washington in 1789, contains a lot of old english of Victorian age, which almost no people uses Nowadays. Therefore, if we wish to analyze the whole span of presidency at once, we shall be in a lot of trouble (as you have seen the classifiers perform bad) as there will be different words with very similar meanings, one possibly being an evoluted version of another.

Therefore, it is better to split up the whole timespan into different time period, over which the linguistic characteristics changes. To achieve this goal of detecting this points of change in linguistic characteristics, we use Changepoint Detection Analysis.

3 Term Frequency as a time series

First, we shall load our cleaned dataset containing the speeches of each president as we obtained in our previous post.

```
speech_df <- readRDS('../datasets/US-president-speech.Rds')
speech_df[1:100, ]
```

text <chr>
fellow citizens of the senate and the house of representatives

text
<chr>
among the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order and received on the fourteenth day of the present month on the one hand i was summoned by my country whose voice i can never hear but with veneration and love from a retreat which i had chosen with the fondest predilection and in my flattering hopes with an immutable decision as the asylum of my declining years a retreat which was rendered every day more necessary as well as more dear to me by the addition of habit to inclination and of frequent interruptions in my health to the gradual waste committed on it by time on the other hand the magnitude and difficulty of the trust to which the voice of my country called me being sufficient to awaken in the wisest and most experienced of her citizens a distrustful scrutiny into his qualification could not but overwhelm with dispondence one who inheriting inferior endowments from nature and unpractised in the duties of civil administration ought to be peculiarly conscious of his own deficencies in this conflict of emotions all i dare aver is that it has been my faithful study to collect my duty from a just appreciation of every circuance by which it might be affected all i dare hope is that if in executing this task i have been too much swayed by a grateful remembrance of former instances or by an affectionate sensibility to this transcendent proof of the confidence of my fellow citizens and have thence too little consulted my incapacity as well as disinclination for the weighty and untried cares before me my error will be palliated by the motives which misled me and its consequences be judged by my country with some share of the partiality in which they originated
such being the impressions under which i have in obedience to the public summons repaired to the present station it would be peculiarly improper to omit in this first official act my fervent supplications to that almighty being who rules over the universe who presides in the councils of nations and whose providential aids can supply every human defect that his benediction may consecrate to the liberties and happiness of the people of the united states a government instituted by thelves for these essential purposes and may enable every instrument employed in its administration to execute with success the functions allotted to his charge in tendering this homage to the great author of every public and private good i assure myself that it expresses your sentiments not less than my own nor those of my fellow citizens at large less than either no people can be bound to acknowledge and adore the invisible hand which conducts the affairs of men more than the people of the united states every step by which they have advanced to the character of an independent nation seeto have been distinguished by some token of providential agency and in the important revolution just accomplished in the system of their united government the tranquil deliberations and voluntary consent of so many distinct communities from which the event has resulted cannot be compared with the means by which most governments have been established without some return of pious gratitude along with an humble anticipation of the future blessings which the past seem to presage these reflections arising out of the present crisis have forced thelves too strongly on my mind to be suppressed you will join with me i trust in thinking that there are none under the influence of which the proceedings of a new and free government can more auspiciously commence
by the article establishing the executive department it is made the duty of the president to recommend to your consideration such measures as he shall judge necessary and expedient the circuances under which i now meet you will acquit me from entering into that subject farther than to refer to the great constitutional charter under which you are assembled and which in defining your powers designates the objects to which your attention is to be given it will be more consistent with those circuances and far more congenial with the feelings which actuate me to substitute in place of a recommendation of particular measures the tribute that is due to the talents the rectitude and the patriotism which adorn the characters selected to devise and adopt them in these honorable qualifications i behold the surest pledges that as on one side no local prejudices or attachments no seperate views nor party animosities will misdirect the comprehensive and equal eye which ought to watch over this great assemblage of communities and interests so on another that the foundations of our national policy will be laid in the pure and immutable principles of private morality and the pre eminence of a free government be exemplified by all the attributes which can win the affections of its citizens and command the respect of the world
i dwell on this prospect with every satisfaction which an ardent love for my country can inspire since there is no truth more thoroughly established than that there exists in the oeconomy and course of nature an indissoluble union between virtue and happiness between duty and advantage between the genuine maxiof an honest and magnanimous policy and the solid rewards of public prosperity and felicity since we ought to be no less persuaded that the propitious smiles of heaven can never be expected on a nation that digards the eternal rules of order and right which heaven itself has ordained and since the preservation of the sacred fire of liberty and the destiny of the republican model of government are justly considered as deeply perhaps as finally staked on the experiment entrusted to the hands of the american people

text <chr>
besides the ordinary objects submitted to your care it will remain with your judgment to decide how far an exercise of the occasional power delegated by the fifth article of the constitution is rendered expedient at the present juncture by the nature of objections which have been urged against the system or by the degree of inquietude which has given birth to them instead of undertaking particular recommendations on this subject in which i could be guided by no lights derived from official opportunities i shall again give way to my entire confidence in your discernment and pursuit of the public good for i assure myself that whilst you carefully avoid every alteration which might endanger the benefits of an united and effective government or which ought to await the future lessons of experience a reverence for the characteristic rights of freemen and a regard for the public harmony will sufficiently influence your deliberations on the question how far the former can be more impregnably fortified or the latter be safely and advantageously promoted
to the preceeding observations i have one to add which will be most properly addressed to the house of representatives it concerns myself and will therefore be as brief as possible when i was first honoured with a call into the service of my country then on the eve of an arduous struggle for its liberties the light in which i contemplated my duty required that i should renounce every pecuniary compensation from this resolution i have in no instance departed and being still under the impressions which produced it i must decline as inapplicable to myself any share in the personal emoluments which may be indispensably included in a permanent provision for the executive department and must accordingly pray that the pecuniary estimates for the station in which i am placed may during my continuance in it be limited to such actual expenditures as the public good may be thought to require
having thus imparted to you my sentiments as they have been awakened by the occasion which brings us together i shall take my present leave but not without resorting once more to the benign parent of the human race in humble supplication that since he has been pleased to favour the american people with opportunities for deliberating in perfect tranquility and dispositions for deciding with unparellelled unanimity on a form of government for the security of their union and the advancement of their happiness so his divine blessing may be equally conspicuous in the enlarged views the temperate consultations and the wise measures on which the success of this government must depend
whereas it is the duty of all nations to acknowledge the providence of almighty god to obey his will to be grateful for his benefits and humbly to implore his protection and favor and whereas both houses of congress have by their joint committee requested me to recommend to the people of the united states a day of public thanks giving and prayer to be observed by acknowledging with grateful hearts the many signal favors of almighty god especially by affording them an opportunity peaceably to establish a form of government for their safety and happiness
now therefore i do recommend and assign thursday the th day of november next to be devoted by the people of these states to the service of that great and glorious being who is the beneficent author of all the good that was that is or that will be that we may then all unite in rendering unto him our sincere and humble thanks for his kind care and protection of the people of this country previous to their becoming a nation for the signal and manifold mercies and the favorable interpositions of his providence which we experienced in the course and conclusion of the late war for the great degree of tranquillity union and plenty which we have since enjoyed for the peaceable and rational manner in which we have been enabled to establish constitutions of government for our safety and happiness and particularly the national one now lately instituted for the civil and religious liberty with which we are blessed and the means we have of acquiring and diffusing useful knowledge and in general for all the great and various favors which he hath been pleased to confer upon us

The next thing would be to split them into words. However, we `Stemming` and `Lemmatizing`. For example, consider the word `apple` and `apples`. Both means the same word, but if we simply split the texts based on these speeches, these two words might appear. Therefore, we need a method to convert each of the word into their basic forms, so that we do not have unnecessary different forms of the words.

1. **Stemming** is heuristic way to convert each word to their base form, by removing `-ing`, `-er`, `-s` etc. from the end of a word. Hence, it is not always greatly satisfying, however, it is very fast.
2. **Lemmatizing**, on the other hand is a methodical and grammatical way to convert each word to their base form, by carefully mapping them to an existing dictionary, which is created through the understanding of parts of speeches of the words. It has a great performance, at the cost of taking more time to compute.

We are going to use `textstem` library for computing this lemmatization.

```
lemma_text <- textstem::lemmatize_strings(speech_df$text)   # perform lemmatization
speech_df$text <- lemma_text
```

However, it is possible that the word “five” is converted into “5”, and since we do not want to deal with numbers as a word, we would like to remove them.

Then, we split the speeches into words, and then compute term document frequency, considering all speeches of a president as a single document.

```
term_df <- speech_df %>%
  mutate(text = str_replace_all(text, pattern = "[^a-zA-Z\\s]", replacement = " ") ) %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words, by = "word") %>%
  count(president, word, sort = TRUE)

dim(term_df)
```

```
[1] 179062      3
```

```
term_df[1:100, ]
```

president <chr>	word <chr>	n <int>
Lyndon B. Johnson	president	1394
Barack Obama	applause	1324
Bill Clinton	people	1016
Lyndon B. Johnson	people	1011
Ronald Reagan	people	941
Barack Obama	people	931
Andrew Jackson	government	884
Lyndon B. Johnson	american	860
Lyndon B. Johnson	nation	830
Barack Obama	american	821

1-10 of 100 rows Previous 1 2 3 4 5 6 … 10 Next

Since different words are not comparable with their respective counts, for example, the word “people” would appear a lot often than the word “facebook” in the speeches. So, we need to divide these term frequencies by the word’s respective total counts to convert it into some sort of ratio.

```
term_df <- term_df %>%
  left_join(term_df %>% group_by(word) %>% summarise(total = sum(n)), by = "word") %>%
  mutate(tf = n / total) %>%
  select(president, word, tf)
```

To make a time series over all the time span of US presidency, we require to add rows to the above dataframe containing every combination of all existing words and all presidents. For this, we shall use the `pres_df` dataframe from our previous post.

```
all_words <- unique(term_df$word)
pres_df <- readRDS('../datasets/US-president.Rds')

term_df <- term_df %>%
  full_join(expand.grid(president = pres_df$president, word = all_words), by = c("president", "word")) %>%
  replace_na(list(tf = 0)) %>%
  left_join(pres_df, by = "president")

dim(term_df)
```

```
[1] 1282677       5
```

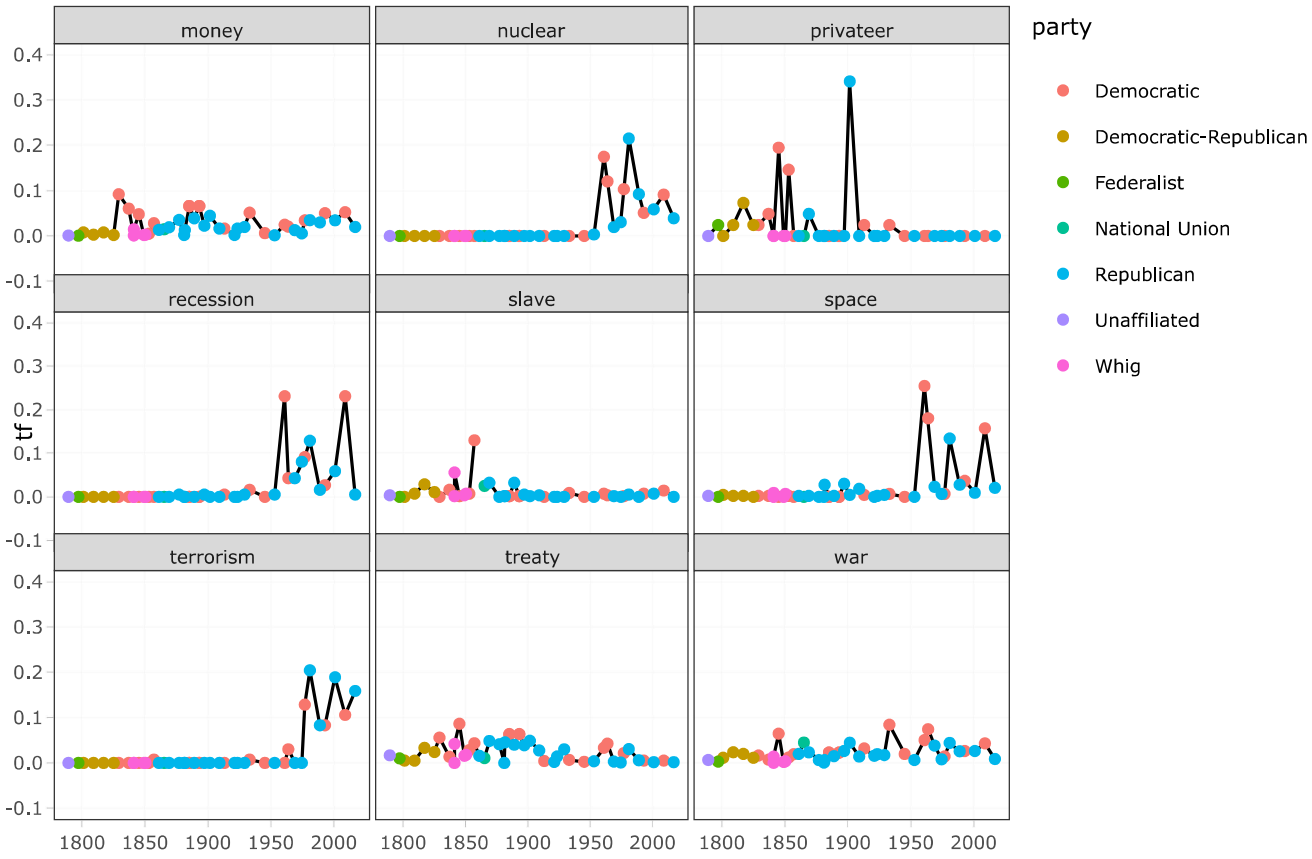
Now as you see, the `term_df` is quite large in size. Each of the time series can now be extracted easily from this dataframe.

For example, let us see how the following words evolved over time.

1. Privateer.
2. Space.
3. Terrorism.

4. Slave
5. War.
6. Money.
7. Treaty.
8. Recession.
9. Nuclear.

```
check_words <- c("privateer", "space", "terrorism", "slave", "war", "money", "treaty", "recession", "nuclear")

p <- ggplot(term_df %>% dplyr::filter(word %in% check_words), aes(x = date)) +
  geom_line(aes(y = tf)) +
  geom_point(aes(y = tf, color = party)) +
  facet_wrap(~ word) +
  theme_bw() +
  ylim(-0.1, 0.4) +
  xlab("")

ggplotly(p)
```



There are some interesting patterns.

1. There are some words like “war”, “treaty” which does not evolve much over time.
2. There are some words like “privateer”, “slave” which were mostly addressed during 1850-1900s.
3. There are some words like “space”, “nuclear”, “terrorism” etc. which were mostly addressed in last 70 years, from 1950 onwards.
4. It is surprising the the word “nuclear” shows its first peak at 1961, but not in 1945 or within 1950s, when the atomic bombing on Hiroshima and Nagasaki happened.

4 Mean Shift Detection (CUSUM Algorithm)

Now that we have the frequencies of all the words, when the distribution of words changes, and also when the vocabularies changes to detect the linguistic shifts or changepoints. A very simple algorithm is to consider mean shifts.

Let, $y_1, y_2, \dots y_T$ be a time series, then the mean shift based on CUSUM (CUmulative SUM) at time t is simply given by the difference of the averages of precedent and proceeding part of the time series.

$$MS_t = \frac{1}{T-t}\sum_{i=(t+1)}^{T} y_i - \frac{1}{t}\sum_{i=1}^{t} y_i$$

However, here we tweak this formula a bit to incorporate only local level changes, for instance, we consider a window length of w, and consider only observations from $(t-w+1)$ to $(t+w)$ for computation of the means.

Therefore, here we take;

$$MS_t = \left| \frac{1}{w}\sum_{i=(t-w+1)}^{t} y_i - \frac{1}{w}\sum_{i=(t+1)}^{(t+w)} y_i \right|$$

To compute this easily, we consider a matrix with rows as the words, and each column being the dates of the presidency.

```
term_mat <- term_df %>% cast_sparse(word, date, tf)

# order the columns in order of time
term_mat <- term_mat[, sort(colnames(term_mat))]
```

Now, we write a function that computes this MS_t series and then apply it to every row of the `term_mat` matrix. However, before that, we need to standardized each series,

$$y^*_w(t) = \frac{y_w(t) - \bar{y}_w}{s_w}$$

where $y_w(t)$ is the frequency of the word w at time t, and $\bar{y}_w$ and s_w are respectively the mean and standard deviation of those time series of frequency.

```
# ms function
ms <- function(x, w = 3) {
  # w = window length
  x <- (x - mean(x, na.rm = T))/sd(x, na.rm = T)   # standardize x, the series
  pre <- stats::filter(x, rep(1/w, w), sides = 1)
  post <- lead(rev(stats::filter(rev(x), rep(1/w, w), sides = 1)))

  return(abs(pre - post))
}

ms_mat <- apply(term_mat, 1, FUN = ms)
ms_mat <- t(ms_mat)
colnames(ms_mat) <- colnames(term_mat)

dim(ms_mat)
```

```
[1] 27291    45
```

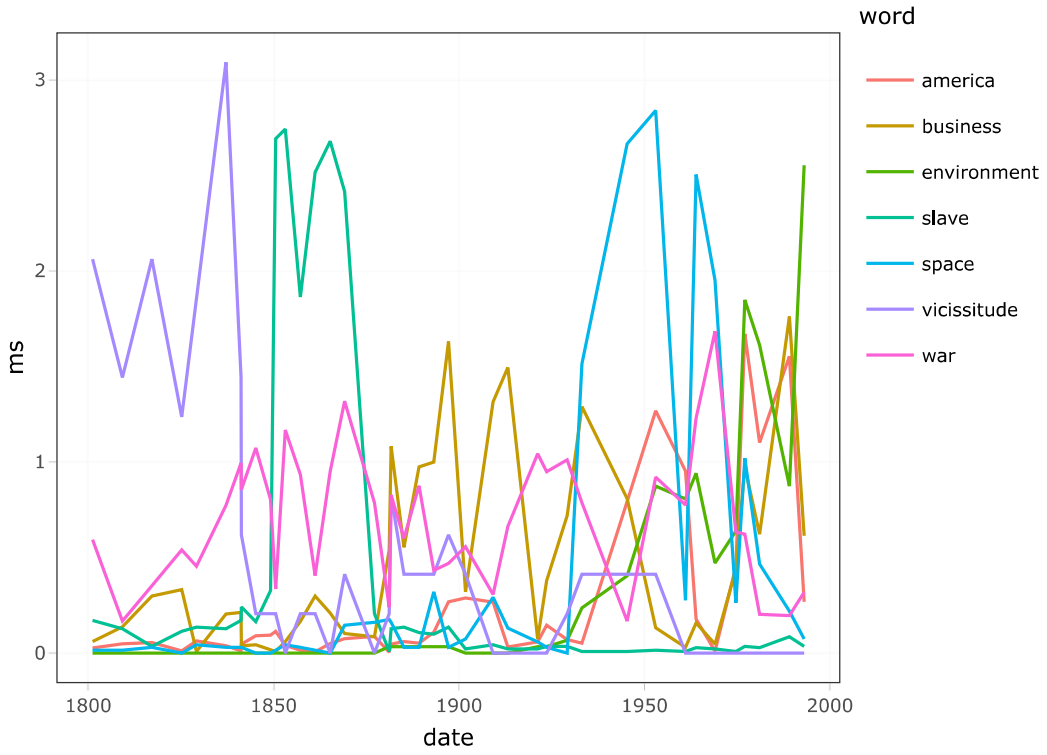
Now, since the window length is chosen to be 3, the first 2 columns and last 3 columns of the MS matrix will be `NA` values, hence we can simply remove them to free up some spaces.

```
ms_mat <- ms_mat[, -c(1:2, (dim(ms_mat)[2]-2):dim(ms_mat)[2] )]
dim(ms_mat)
```

```
[1] 27291    40
```

Now for each of the words, we can extract the plots of the mean shift series.

```
tmp <- as_tibble(ms_mat, rownames = "word") %>%
  pivot_longer(-word, names_to = "date", values_to = "ms") %>%
  mutate(date = as.Date(date))

check_words <- c("space", "war", "america", "vicissitude", "slave", "business", "environment")

p <- ggplot(tmp %>% dplyr::filter(word %in% check_words), aes(x = date)) +
  geom_line(aes(y = ms, color = word)) +
  theme_bw()

ggplotly(p)
```



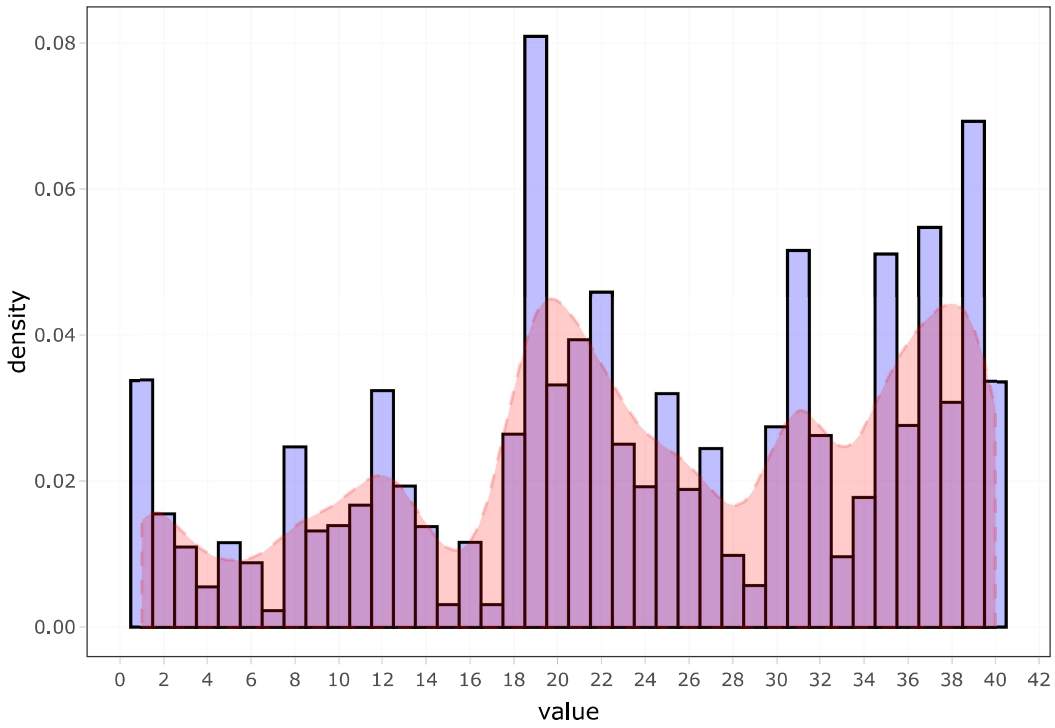
For each word, the estimate of the changepoints can now be readily extracted from the date where maximum for each mean shift series occurs. Also, since each of the series is normalized, we can use a quantile $z_{1-\alpha}$, to check whether that maximum value indeed is significant. In this regard, we assume a gaussian distribution and only take the changepoints which has mean shift higher than 1.96, corresponding to a 5% level of significance (of both sides, as we are considering change in absolute value).

```
cp <- apply(ms_mat, 1, FUN = function(x) {
  if (max(x) < qnorm(0.975)) {
    return(NA)
  }
  else {
    which.max(x)
  }
})

head(cp, 25)
```

```
 president   applause     people government   american     nation    america
        32         40         NA         19         NA         NA         NA
      time        law      unite      world      power      peace    country
        NA         19         25         NA          2         32         NA
       job      child        war     soviet     mexico   congress    vietnam
        NA         37         NA         39          7         NA         33
      hope        tax     public       bank
        34         36         NA          4
```

Now, we see an histogram of these changepoints, to understand where most of these changepoints actually cluster up.

```
p <- ggplot(as_tibble(cp) %>% drop_na(), aes(x = value)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill = "blue", color = "black", alpha = 0.25) +
  geom_density(fill = "red", alpha = 0.2, linetype = "dashed", color = "red") +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 20)) +
  theme_bw()

ggplotly(p)
```



5 Conclusion

We kind of see that there are 5 segments where the linguistic shifts occur, namely the 5 peaks of the kernel density estimate of the changepoints. However, the first peak is very close to the starting point, which is possibly a misleading peak due to truncation errors.

The most prominent peaks correspond to the index 12, 20, 31 and finally 38. Next, we extract the corresponding timelines as well as the president's name.

```
cp_clust <- as.Date(colnames(ms_mat)[c(12, 20, 31, 38)])
pres_df %>% inner_join(tibble(date = cp_clust), by = "date")
```

date <date>	party <chr>	president <chr>
1853-03-04	Democratic	Franklin Pierce
1885-03-04	Democratic	Grover Cleveland
1945-04-12	Democratic	Harry S. Truman
1981-01-20	Republican	Ronald Reagan

4 rows

So, we finally see that there are 4 era of linguistic patterns in US president's speech, as follows:

1. George Washington (1789) - Millard Fillmore (1853), signifying possibly the end of Whig party and establishment of the ideals related to Democratic and Republican thoughts.
2. Franklin Pierce (1853) - Chester A. Arthur (1885), ending the depression of finance market throughout 1882-85.
3. Grover Cleveland (1885) - Franklin D. Roosevelt (1945), signifying possibly the end of World War 2.
4. Harry S. Truman (1945) - Jimmy Carter (1981), signifying possibly the end of Vietnamese war and enhancement of global relations.

5. Ronald Reagan (1981) - Donald Trump (present)