Text Mining and Topic Modeling: with special application to central banks

Gustavo Rojas-Matute

1/25/2021

Introduction

Text mining, data and analysis have been becoming in power tools in economics. For instance, creating a news-based uncertainty index or sentiment analysis in forecasting. In monetary policy, text mining could help us to understand if the central bank pays more attention to inflation than unemployment or vice-verse.

The goal of this tutorial is to learn how to extract, store, clean and analyze data from publications of the Federal Reserve.

What we need

We will need the following packages:

```
library(tidyverse)
library(pdftools)
library(stringi)
require(topicmodels)
library(quanteda)
library(tidytext)
library(xml2)
```

Getting start with a simple Text Extraction

Let's start by extracting a PDF text from the Federal Reserve web. This is a document from the Federal Open Market Operations Committee published on July 3rd of 1996. Since this is a PDF document, we use the pdf_text function:

```
omc0796 <- pdf_text("https://www.federalreserve.gov/monetarypolicy/files/FOMC19960703meeting.pdf")
```

Maybe you want to explore the document by typing:

```
writeLines(as.character(omc0796))
```

But our main interest is not the complete text, but the analysis of the words. This is why to need to organize the document. There are different approaches using corpus or dfm, but here we will use data frame.

```
corpus_fed <- data_frame(word=stri_trans_tolower(unlist(stri_extract_all_words(omc0796))),
                         count=1)
corpus_fed[1:10,1]
```

```
## # A tibble: 10 x 1
##    word
```

```
##    <chr>
##  1 meeting
##  2 of
##  3 the
##  4 federal
##  5 open
##  6 market
##  7 committee
##  8 july
##  9 2
## 10 3
```

As we can see, our data frame organizes every word of the document, including numbers, stop words and punctuation. This something we will tackle later.

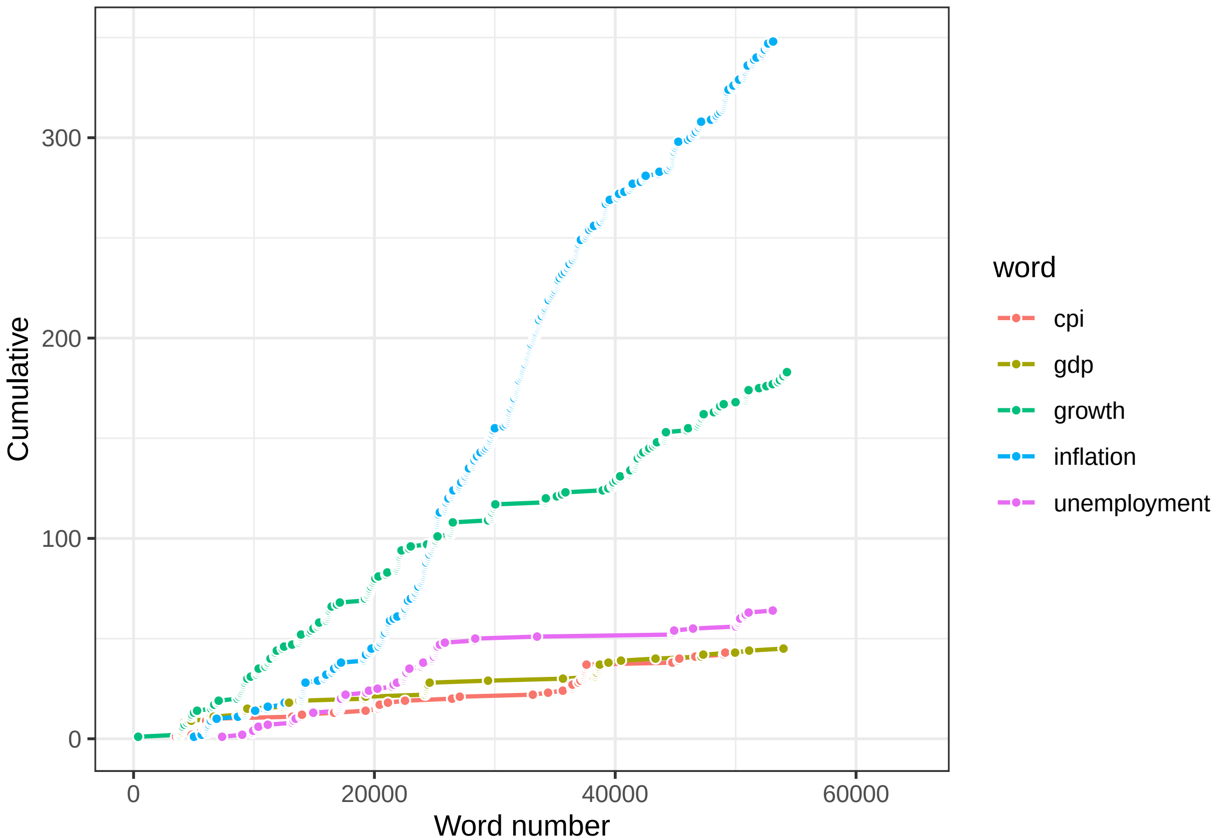
But with this data frame, we can count the number of words and estimate the cumulative words. Suppose we want to investigate: "inflation", "cpi", "growth", "gdp", "unemployment". Here is an example with the first ten rows:

```
corpus_fed$word_number <- 1:nrow(corpus_fed)
cumsum_corpus_fed <- mutate(group_by(corpus_fed, word), cumsum=cumsum(count))
filter_fed <- filter(cumsum_corpus_fed, word %in% c("inflation", "cpi", "growth", "gdp", "unemployment")

cumsum_corpus_fed[1:10,]
```

```
## # A tibble: 10 x 4
## # Groups:   word [10]
##    word      count word_number cumsum
##    <chr>     <dbl>       <int>  <dbl>
##  1 meeting       1           1      1
##  2 of            1           2      1
##  3 the           1           3      1
##  4 federal       1           4      1
##  5 open          1           5      1
##  6 market        1           6      1
##  7 committee     1           7      1
##  8 july          1           8      1
##  9 2             1           9      1
## 10 3             1          10      1
```

Now we are able to plot:

```
gg_fed <- ggplot(filter_fed,
                 aes(x=word_number, y=cumsum))
gg_fed <- gg_fed + geom_line(aes(color=word), size=0.75) +
  geom_point(aes(fill=word), shape=21, color="white", size=1.5) +
  scale_x_continuous(limits=c(1, nrow(corpus_fed)))+
  theme_bw() +
  ylab("Cumulative") +
  xlab("Word number")
gg_fed
```



Organizing and storing with Tibble

Another approach to organize the text in a data frame is using tibble (also the function dtm - document to matrix). We can organize the original document omc0796 using tibble function as follows:

```
library(dplyr) # Tibble is a data frame
text_df <- tibble(line = 1:115, text = omc0796)
```

Cleaning using Tidytext

As mentioned above, in any text there are always characters that are not informative such as stop words (the, in, of, etc), numbers and punctuation.

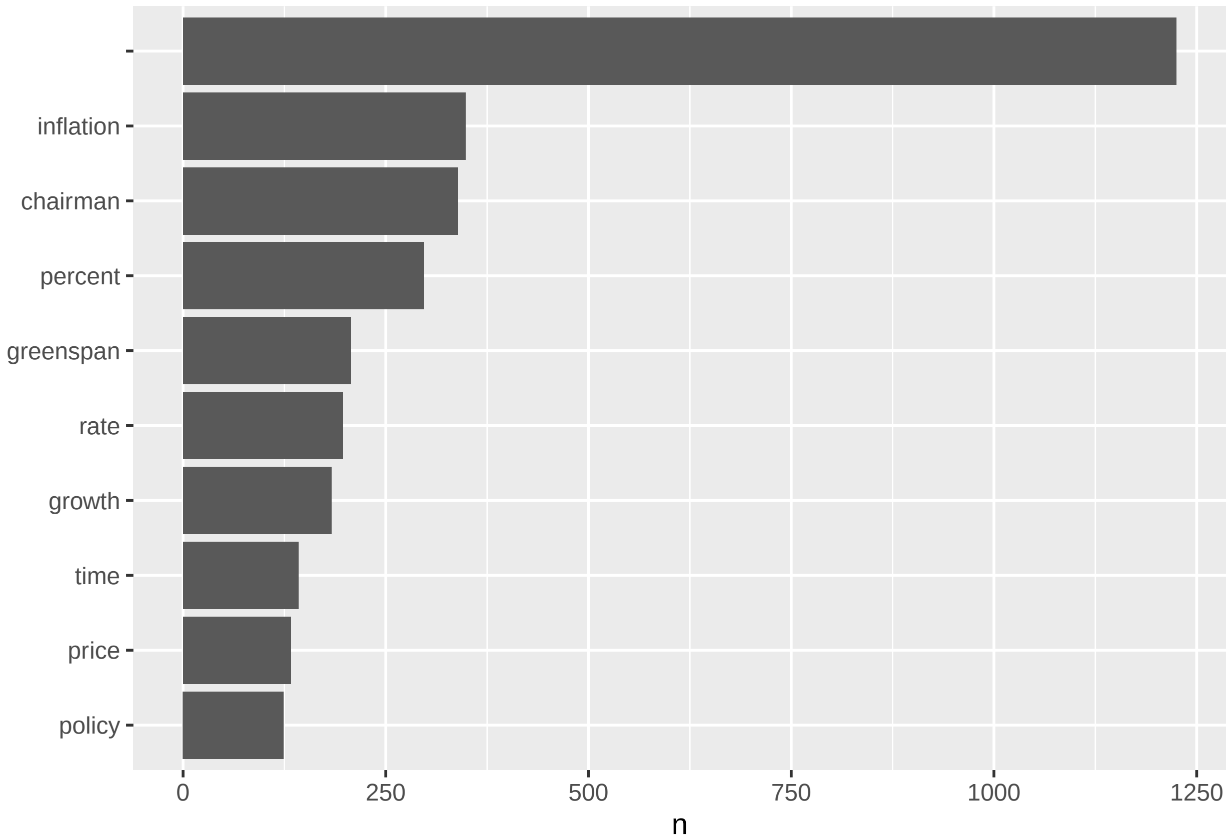
It is possible to remove them:

```
library(tidytext) ## this is equivalent to data_frame (corpus_fed)

fed_df <- text_df %>% unnest_tokens(word, text) ## unnest_tokens convert to lowercase

## Remove stop words

data(stop_words)

clean_text_fed <- fed_df %>%   anti_join(stop_words) %>% mutate(word = gsub("[[:digit:]]","", word)) #R
```

Frequency of words

Now that we have cleaned the data, we can plot a frequency of words:

```
library(ggplot2)

clean_text_fed %>%
  count(word, sort = TRUE) %>%
  filter(n > 100) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word)) +
  geom_col() +
  labs(y = NULL)
```



Word Cloud

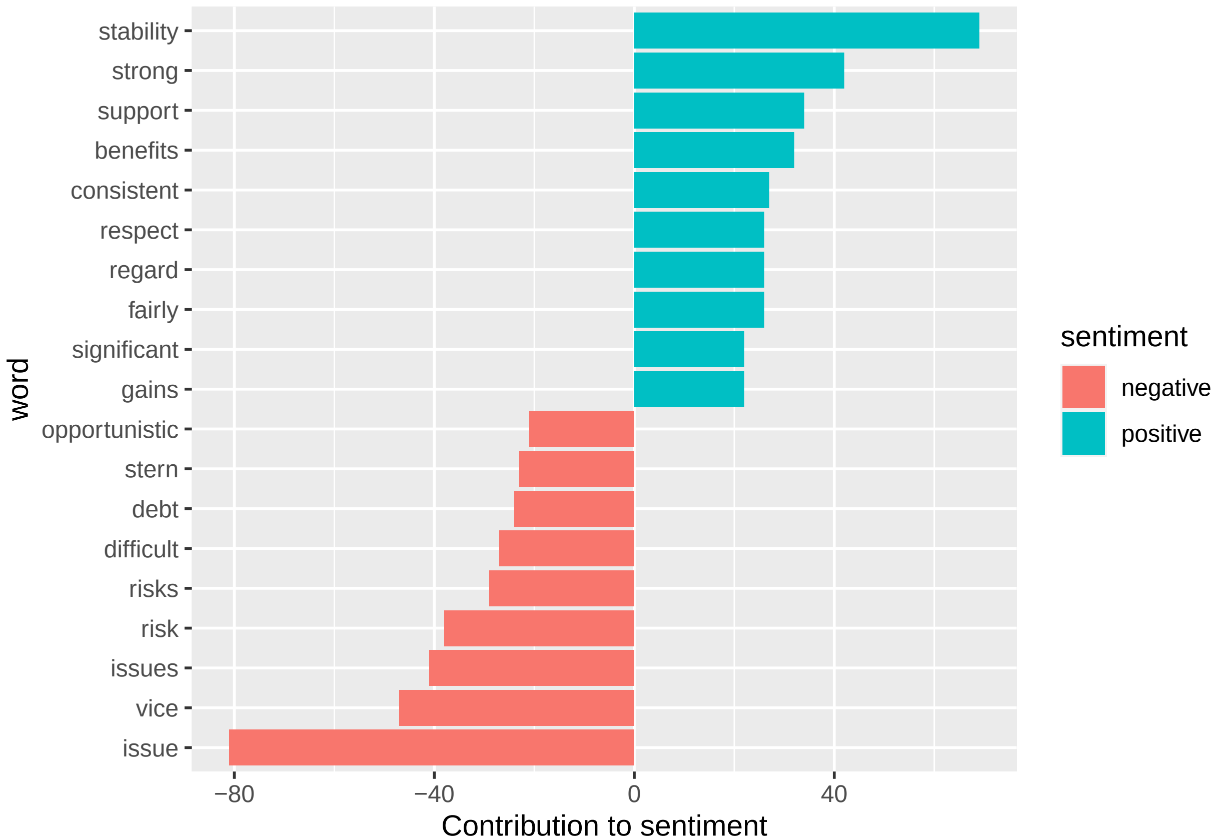
We can also plot a word cloud:

```
library(wordcloud)

clean_text_fed  %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 100))
```

Sentiment Analysis

Finally, we can conduct sentiment analysis and the contribution of every word to the sentiment.

```r
bing <- get_sentiments("bing")

sentiment_fed <- clean_text_fed  %>%
  inner_join(bing) %>%
  count(word, sentiment, sort = TRUE)

sentiment_fed %>% # sentiment contribution
  filter(n > 20) %>%
  mutate(n = ifelse(sentiment == "negative", -n, n)) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col() +
  coord_flip() +
  labs(y = "Contribution to sentiment")
```



Working with multiples documents

Now it is time to learn how to work with multiple docummemt at the same time. There are several examples on how to work with data sets in R, but here we will focused on how to extract and analyze press realeses from the Federal Reserve.

First, lets build a vector of the links we want to analyze.

```
## List of documents

links<-c("https://www.federalreserve.gov/newsevents/pressreleases/monetary20190130a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20190320a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20190501a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20180131a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20180321a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20180502a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20180613a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20180801a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20180926a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20181108a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20181219a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20170201a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20170315a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20170503a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20170614a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20170726a.htm",
         "https://www.federalreserve.gov/newsevents/pressreleases/monetary20170920a.htm",
```

```
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20171101a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20171213a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20160127a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20160316a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20160427a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20160615a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20160727a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20160921a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20161102a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20161214a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20150128a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20150318a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20150429a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20150617a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20150729a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20150917a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20151028a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20151216a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20140129a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20140319a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20140430a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20140618a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20140730a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20140917a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20141029a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20141217a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20130130a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20130320a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20130501a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20130619a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20130731a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20130918a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20131030a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20131218a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20120125a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20120313a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20120425a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20120620a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20120801a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20120913a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20121024a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20121212a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20110126a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20110315a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20110427a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20110622a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20110809a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20110921a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20111102a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20111213a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20100127a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20100316a.htm",
        "https://www.federalreserve.gov/newsevents/pressreleases/monetary20100428a.htm",
```

```r
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20100623a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20100810a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20100921a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20101103a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20101214a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20090128a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20090318a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20090429a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20090624a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20090812a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20090923a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20091104a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20091216a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20080122b.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20080130a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20080318a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20080430a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20080625a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20080805a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20080916a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20081008a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20081029a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20081216b.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20070131a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20070321a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20070509a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20070618a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20070807a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20070817b.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20070918a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20071031a.htm",
          "https://www.federalreserve.gov/newsevents/pressreleases/monetary20071211a.htm"
)
```

What we need:

```r
library(rvest)
library(purrr)
library(stringr)
```

Since these documents are in html format, the process is similar to web scraping, link by link. Here is a useful approach to do it.

```r
out <- vector("character", length = length(links))
for(i in seq_along(links)){
  derby <- read_html(links[i])
  out[i] <- derby %>%
    html_node("body") %>%
    html_text() %>%
    stri_trim()
}
```

A simple way to organize the documents using dfm and tidy:

```r
myFOMC_td <- tidy(quanteda::dfm(out, verbose = FALSE))
```

If we want, it is possible to analyze directly some topics. For instance, suppose we are interested in analyze "inflation" and "employment". Here is an example on how to count both words among the different texts (links) we have:

```
inflation <-  myFOMC_td %>%
  filter(term %in% c("inflation"))
inflation[1:10,]
```

```
## # A tibble: 10 x 3
##    document term      count
##    <chr>    <chr>     <dbl>
##  1 text1    inflation     9
##  2 text2    inflation     9
##  3 text3    inflation     9
##  4 text4    inflation    13
##  5 text5    inflation    13
##  6 text6    inflation    12
##  7 text7    inflation     9
##  8 text8    inflation     9
##  9 text9    inflation     8
## 10 text10   inflation     8
```

```
employment <-  myFOMC_td %>%
  filter(term %in% c("employment"))
employment[1:10,]
```

```
## # A tibble: 10 x 3
##    document term       count
##    <chr>    <chr>      <dbl>
##  1 text1    employment     3
##  2 text2    employment     4
##  3 text3    employment     3
##  4 text4    employment     4
##  5 text5    employment     3
##  6 text6    employment     3
##  7 text7    employment     3
##  8 text8    employment     3
##  9 text9    employment     3
## 10 text10   employment     3
```

If we want to plot the evolution of the counts of both "inflation" and "employment" in the time, we need to create a vector with the dates the press relaeses were published:

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

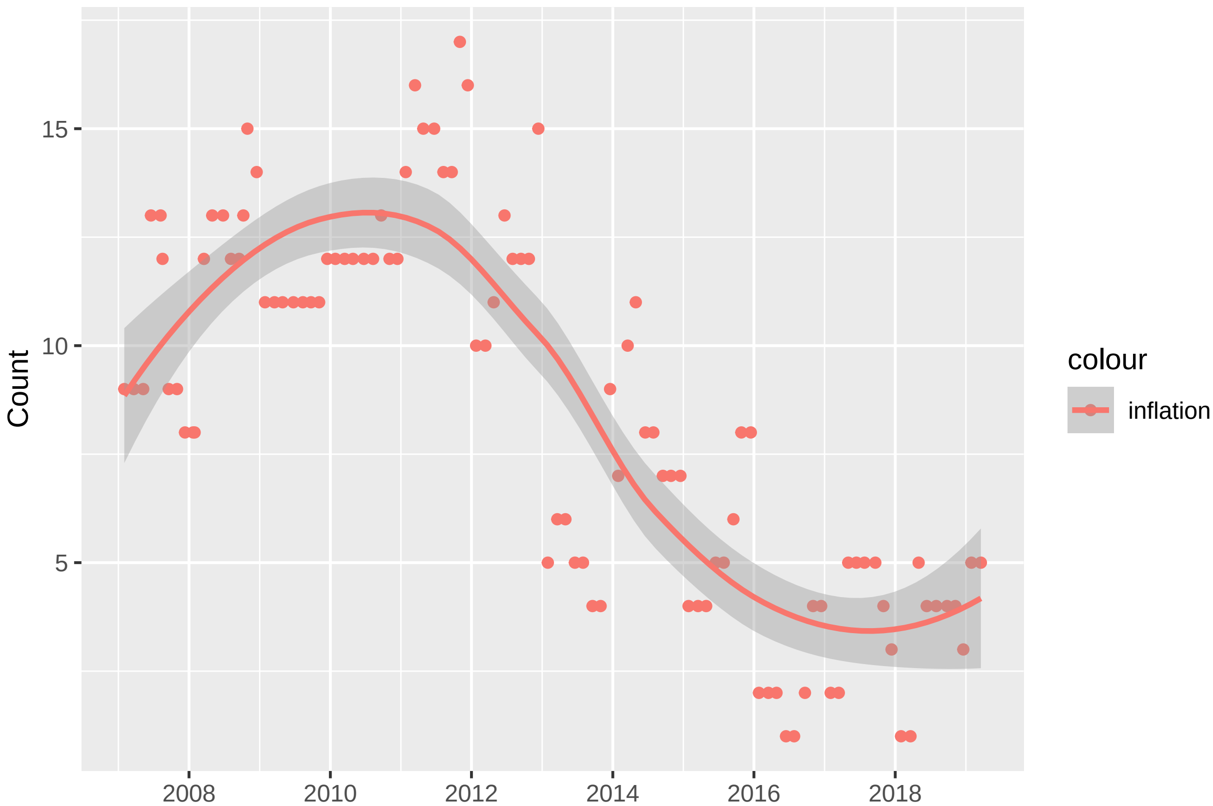
```
statement.dates<-NULL
year<-NULL
for(i in seq(from=1, to=length(links))) {
  statement.dates[i]<-(str_extract(links[i],"[[:digit:]]+"))
  year[i]<-substr(statement.dates[i],1,4)
}
```

```r
reports<-data.frame(year,statement.dates, links)

reports %<>% mutate_if(is.factor, as.character)%>% arrange(statement.dates)

time_FOMC <- as.Date(reports$statement.dates, format = "%Y%m%d")
```

Now, let's plot them:

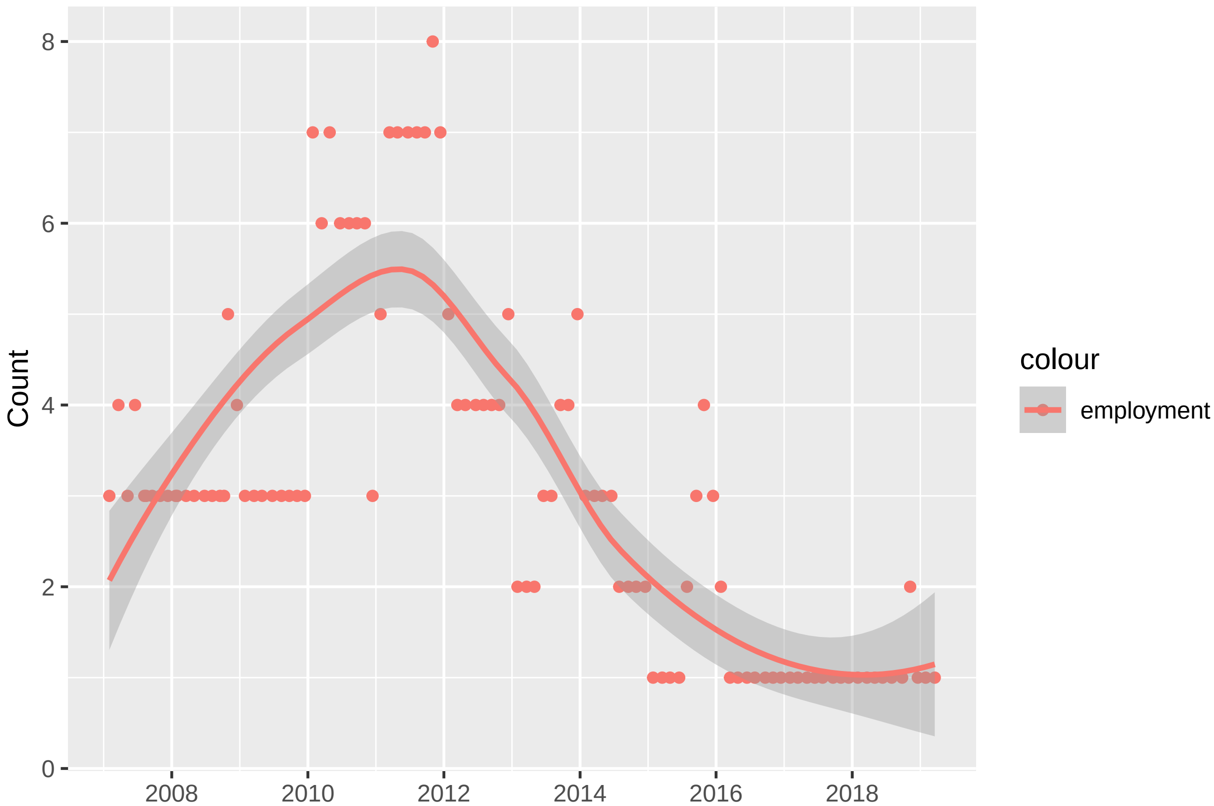
```r
data_FOMC <- data.frame(time_FOMC[1:101], inflation$document, inflation$count, employment$count[1:101])

  ggplot(data =  data_FOMC, aes(time_FOMC.1.101., inflation.count, colour = "inflation")) +
  geom_point() +
  geom_smooth() +
  xlab("") +
  ylab("Count")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```r
  ggplot(data = data_FOMC, aes(x = time_FOMC.1.101., y = employment.count.1.101., colour = "employment"
    geom_point() +
    geom_smooth() +
    xlab("") +
  ylab("Count")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Sentiment analysis

We can also conduct sentiment analysis. Here we can observe the net sentiment (positive - negative) per text:

```
myFOMC_sent <- myFOMC_td %>% inner_join(get_sentiments("bing"), by = c(term = "word"))

sentiment_per_text<- myFOMC_sent %>%
  dplyr::count(document, sentiment, wt = count) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative) %>%
  arrange(sentiment)

sentiment_per_text[1:10,]
```

```
## # A tibble: 10 x 4
##    document negative positive sentiment
##    <chr>       <dbl>    <dbl>     <dbl>
##  1 text91         26       19        -7
##  2 text85         17       13        -4
##  3 text96         15       11        -4
##  4 text84         18       15        -3
##  5 text86         18       15        -3
##  6 text89         18       15        -3
##  7 text95         14       11        -3
##  8 text88         16       14        -2
##  9 text90         17       15        -2
## 10 text92         19       17        -2
```

Cleaning the data

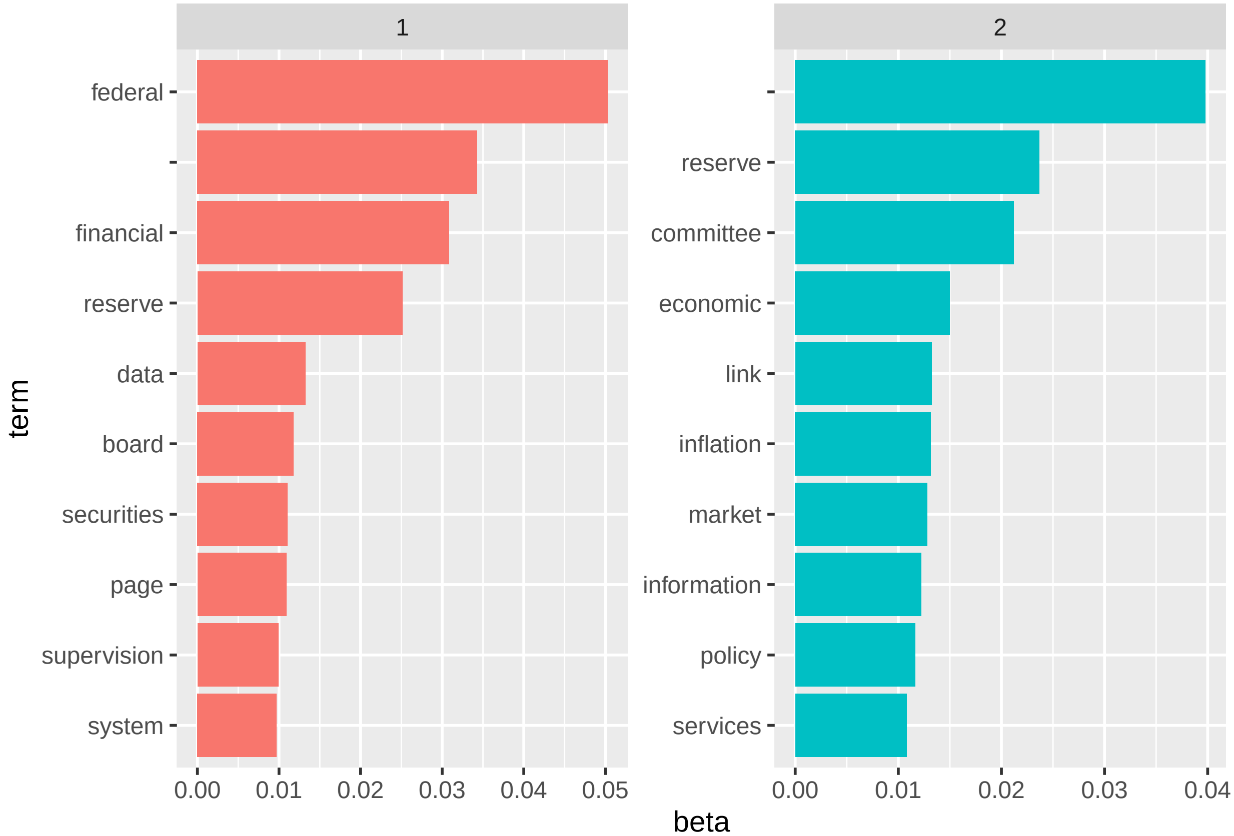
To be more efficient in topic modeling, it is helpful to clean the data by removing uninformative characters such as punctuation and stop words.

```
#cleaning
clean_FOMC <- myFOMC_td %>%
  anti_join(stop_words, by = c(term = "word")) %>%
  mutate(term = gsub("[[:punct:]]","", term)) %>%
  mutate(term = gsub("-","", term)) %>%
  mutate(term = gsub(" ","", term))
```

Topic Modeling

Now that we have cleaned the data, we can conduct topic modeling. We will use the LDA algorithm. We first start the number of topics we want to identify. For instance, by setting k = 2, we want to identify two topics. The following table shows us the words related which topics 1 and 2.

```
myFOMC_dtm <- clean_FOMC %>% cast_dtm(document, term, count)
myFOMC_LDA <- LDA(myFOMC_dtm, k = 2, control = list(seed = 1234))
mytopics<-  tidy(myFOMC_LDA , matrix = "beta")

mytopics_terms <- mytopics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

mytopics_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered()
```

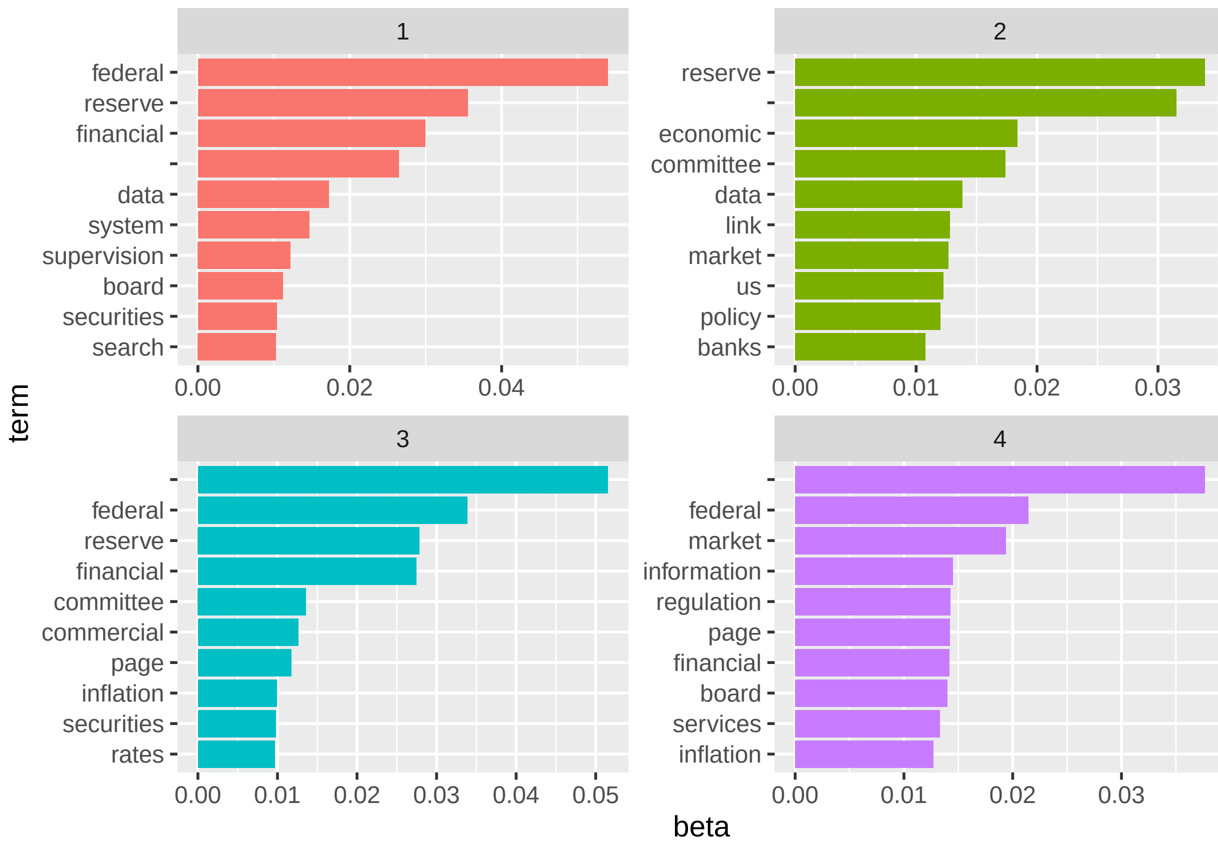



```
beta_spread <- mytopics %>%
  mutate(topic = paste0("topic", topic)) %>%
  spread(topic, beta) %>%
  filter(topic1 > .001 | topic2 > .001) %>%
  mutate(log_ratio = log2(topic2 / topic1))
beta_spread
```

```
## # A tibble: 410 x 4
##    term      topic1  topic2 log_ratio
##    <chr>      <dbl>   <dbl>     <dbl>
##  1 ""       0.0343   0.0398     0.215
##  2 "1"      0.00173  0.00185    0.0928
##  3 "10"     0.000613 0.00134    1.13
##  4 "15"     0.000147 0.00203    3.79
##  5 "17"     0.000851 0.00116    0.442
##  6 "19"     0.000378 0.00166    2.13
##  7 "2"      0.000328 0.00531    4.02
##  8 "20"     0.000933 0.00113    0.281
##  9 "20551"  0.000260 0.00166    2.68
## 10 "3"      0.000808 0.00126    0.645
## # ... with 400 more rows
```

```
myFOMC_dtm <- clean_FOMC %>% cast_dtm(document, term, count)
myFOMC_LDA <- LDA(myFOMC_dtm, k = 4, control = list(seed = 1234))
mytopics<-  tidy(myFOMC_LDA , matrix = "beta")

mytopics_terms <- mytopics %>%
  group_by(topic) %>%
```

```
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

mytopics_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered()
```



Further literature

Benchimol, Kazinnik, and Saadon (2020) have a great working paper on text mining focused on central banks. Julia and Robinson (2020) have great tools fro text mining. ## **References**

Benchimol, Jonathan, Sophia Kazinnik, and Yossi Saadon. 2020. "Text Mining Methodologies with R: An Application to Central Bank Texts."

Julia, Silge, and David Robinson. 2020. *Text Mining with R*. https://www.tidytextmining.com/index.html.