TextMining - Discursos Presidenciales tidy way

alpoza

6 de septiembre de 2017

## R Markdown

require(tidyverse)  
require(tidytext)  
require(ggrepel)  
require(tm)

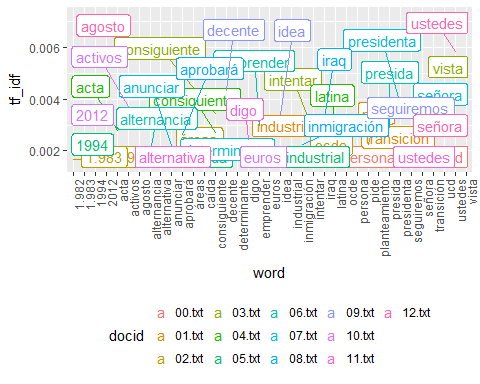
filenames <- list.files(".",pattern="\*.txt")  
discursos\_raw = map(filenames,~readLines(.x,encoding="UTF-8")) %>%   
 map2(filenames, ~tibble(text=.x, docid=.y)) %>%   
 map\_df(rbind)  
  
# Replaces varios  
discursos\_raw = discursos\_raw %>%   
 filter(text != "") %>%   
 mutate(text = gsub("SS.SS", "Sus\_Señorías", text)  
 ,text = gsub("SS. SS", "Sus\_Señorías", text)  
 )  
discursos\_raw

## # A tibble: 1,978 x 2  
## text  
## <chr>  
## 1 X0  
## 2 Muchas gracias, señor Presidente. Con su venia. Señoras y señores Diputados  
## 3 El cambio político realizado en nuestro país ha sido profundo y sincero. Pe  
## 4 Se trata, por consiguiente, de saber realizar el cambio social con sinceri  
## 5 El cambio político se verificó en torno a un eje de sensatez consistente en  
## 6 Con esta voluntad, me permito invitar a Sus\_Señorías. a avanzar en la defi  
## 7 La cuestión, en la realidad, una vez más, consiste en averiguar si, a uno y  
## 8 En síntesis, éste es el cuadro y en este gran marco se inscribe la gran tar  
## 9 Y tenemos voluntad, fortaleza y experiencia política para serlo. La oportun  
## 10 Nos enfrentamos con una situación nueva porque iniciamos una nueva legislat  
## # ... with 1,968 more rows, and 1 more variables: docid <chr>

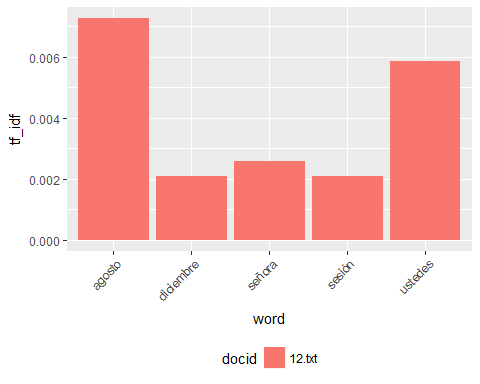
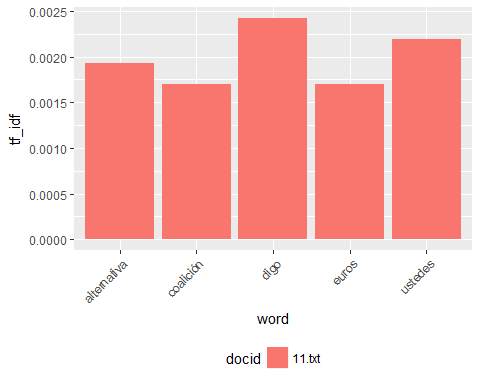
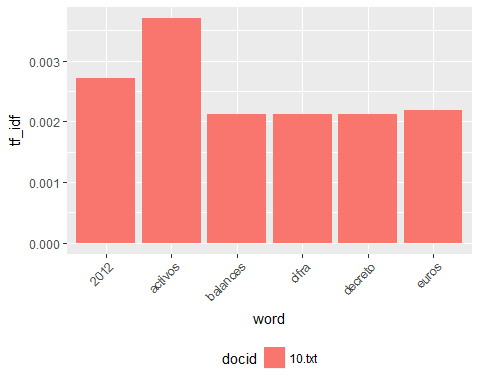
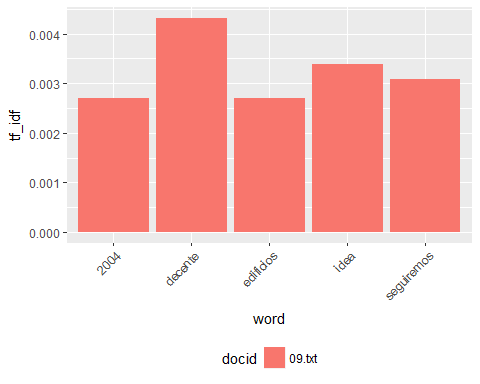
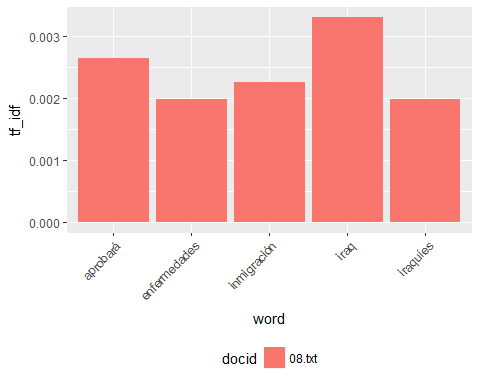
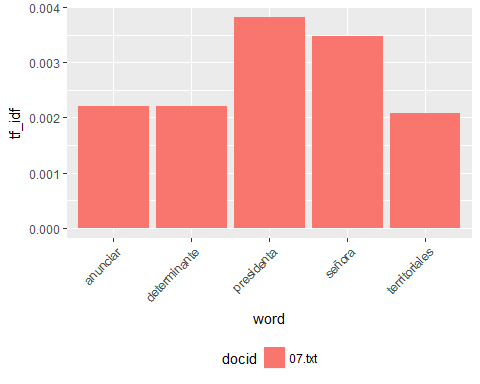
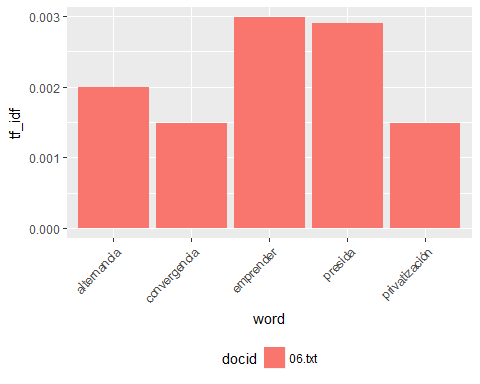
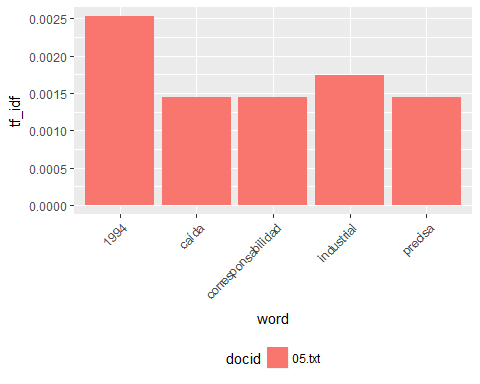
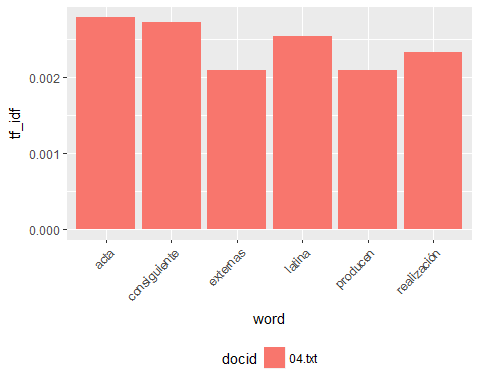
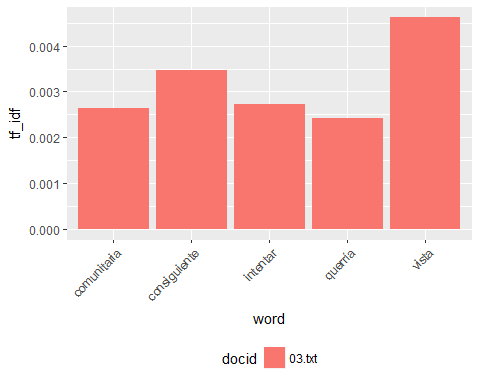
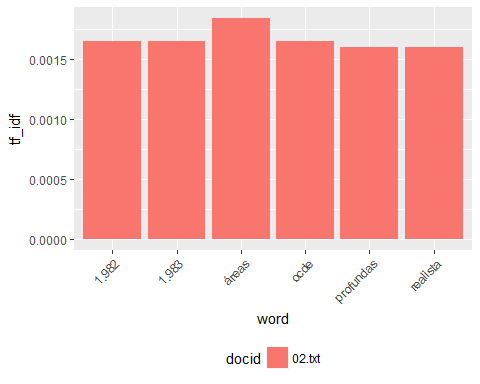
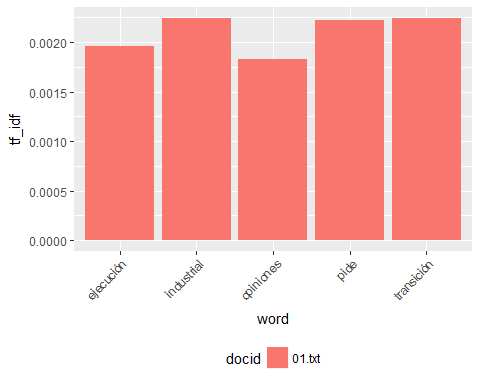
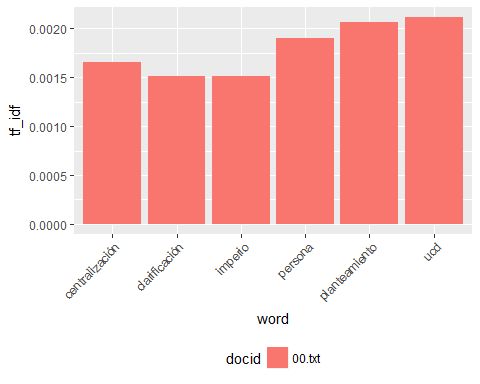
discursos = discursos\_raw %>%   
 unnest\_tokens(word, text) %>%   
 count(docid, word, sort = TRUE) %>%  
 ungroup()  
  
stop\_words = data.frame(word=tm::stopwords("spanish"))  
add.stops=c(  
 "vez", "sino", "cada", "ello", "así", "sólo", "digo",   
 "que", "que,", "señorías","gobierno",  
 "españa","españoles","país","ser","hacia","años",  
 "debe","cualquier","año","manera","todas","mayor",  
 "parte","presidenta","ustedes","vista","señora","hecho","sus")  
add.stops=c("kk")  
  
discursos = discursos %>%   
 filter(!word %in% stopwords("spanish")) %>%   
 filter(!word %in% add.stops)  
  
discursos\_tfidf = discursos %>%   
 bind\_tf\_idf(word, docid, n) %>%   
 arrange(desc(tf\_idf))  
  
discursos\_tfidf %>% arrange(desc(n))

## # A tibble: 26,342 x 6  
## docid word n tf idf tf\_idf  
## <chr> <chr> <int> <dbl> <dbl> <dbl>  
## 1 11.txt gobierno 82 0.016935151 0.00000000 0.0000000000  
## 2 03.txt política 69 0.012858740 0.00000000 0.0000000000  
## 3 11.txt españa 67 0.013837257 0.00000000 0.0000000000  
## 4 08.txt gobierno 65 0.016722408 0.00000000 0.0000000000  
## 5 06.txt gobierno 64 0.012398295 0.00000000 0.0000000000  
## 6 11.txt señorías 63 0.013011152 0.00000000 0.0000000000  
## 7 11.txt españoles 59 0.012185048 0.08004271 0.0009753242  
## 8 00.txt gobierno 59 0.009525347 0.00000000 0.0000000000  
## 9 09.txt españa 56 0.011809363 0.00000000 0.0000000000  
## 10 12.txt gobierno 54 0.019128587 0.00000000 0.0000000000  
## # ... with 26,332 more rows

discursos\_tfidf %>%   
 group\_by(docid) %>%   
 top\_n(3) %>%   
 ungroup() %>%   
 arrange(docid) %>%   
 ggplot(aes(word,tf\_idf, color=docid)) +   
 geom\_label\_repel(aes(label=word)) +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +  
 theme(legend.position="bottom")



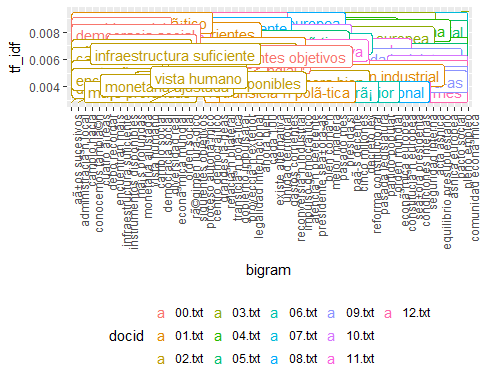
plots <- discursos\_tfidf %>%  
 group\_by(docid) %>%   
 top\_n(5) %>%   
 ungroup() %>%   
 mutate(bigram = reorder(word, tf\_idf)) %>%  
 split(.$docid) %>%  
 map(~ ggplot(.x, aes(word,tf\_idf)) +   
 geom\_bar(stat="identity", aes(fill=docid)) +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 theme(legend.position="bottom")  
 )  
  
walk(plots, print)



discursos\_bigram\_raw = discursos\_raw %>%   
 unnest\_tokens(bigram, text, token = "ngrams", n = 2)  
  
add.bigramsstops=c("señora presidenta", "sus señorías")  
discursos\_bigram\_raw = discursos\_bigram\_raw %>%   
 filter(!bigram %in% add.bigramsstops)  
   
discursos\_bigram = discursos\_bigram\_raw %>%   
 separate(bigram, c("word1", "word2"), sep = " ") %>%   
 filter(!word1 %in% stopwords("spanish")) %>%  
 filter(!word2 %in% stopwords("spanish")) %>%   
 filter(!word1 %in% add.stops) %>%  
 filter(!word2 %in% add.stops) %>%   
 unite(bigram, word1, word2, sep = " ")   
  
discursos\_bigram = discursos\_bigram %>%   
 count(docid, bigram, sort = TRUE) %>%  
 ungroup() %>%   
 bind\_tf\_idf(bigram, docid, n) %>%   
 arrange(desc(tf\_idf))  
  
discursos\_bigram

## # A tibble: 19,620 x 6  
## docid bigram n tf idf tf\_idf  
## <chr> <chr> <int> <dbl> <dbl> <dbl>  
## 1 03.txt comunidad econã³mica 12 0.006475985 1.466337 0.009495977  
## 2 07.txt pleno empleo 6 0.003594967 2.564949 0.009220908  
## 3 07.txt seã±ora presidenta 10 0.005991612 1.466337 0.008785722  
## 4 05.txt pacto social 5 0.003272251 2.564949 0.008393159  
## 5 03.txt ãnica europea 6 0.003237992 2.564949 0.008305287  
## 6 07.txt equilibrio presupuestario 5 0.002995806 2.564949 0.007684090  
## 7 04.txt acta ãnica 7 0.005177515 1.466337 0.007591982  
## 8 04.txt condiciones internas 4 0.002958580 2.564949 0.007588608  
## 9 04.txt seguridad interna 4 0.002958580 2.564949 0.007588608  
## 10 08.txt constituciã³n europea 4 0.002814919 2.564949 0.007220125  
## # ... with 19,610 more rows

discursos\_bigram %>%   
 group\_by(docid) %>%   
 top\_n(3) %>%   
 ungroup() %>%   
 mutate(bigram = reorder(bigram, tf\_idf)) %>%  
 ggplot(aes(bigram,tf\_idf, color=docid)) +   
 geom\_label\_repel(aes(label=bigram)) +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +  
 theme(legend.position="bottom")



plots <- discursos\_bigram %>%  
 group\_by(docid) %>%   
 top\_n(5) %>%   
 ungroup() %>%   
 mutate(bigram = reorder(bigram, tf\_idf)) %>%  
 split(.$docid) %>%  
 map(~ ggplot(.x, aes(bigram,tf\_idf)) +   
 geom\_bar(stat="identity", aes(fill=docid)) +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 theme(legend.position="bottom")  
 )  
  
walk(plots, print)

