

Abriendo y analizando los Diarios de Sesiones del Parlamento uruguayo con R

Daniela Vázquez

[@d4tagirl](#)

Daniela

Economista

Daniela

Data Scientist (pasado: Idatha, Equifax)

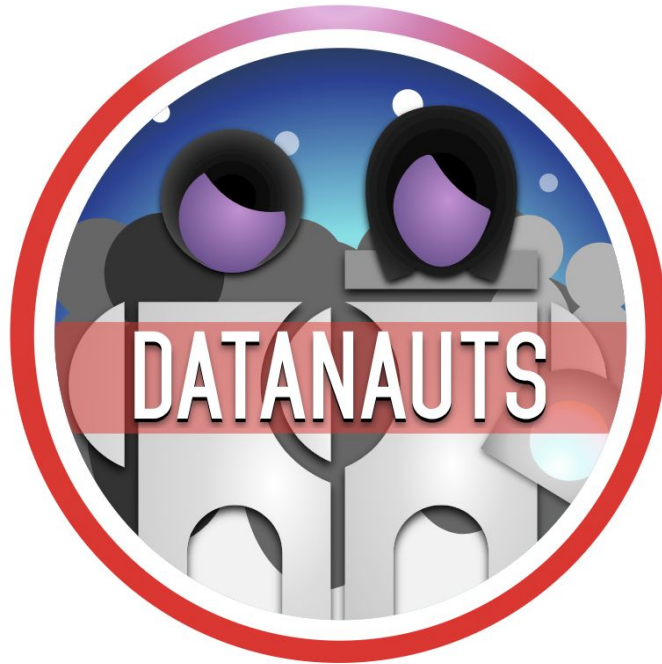
Daniela

R-Ladies



Daniela

NASA Datanaut



Daniela

Pinto :)

Contacto

Twitter: [@d4tagirl](#)

Slack:

- [R-Ladies Santa Fé](#) (~ 1 mes)
- [All the R-Ladies](#) (en inglés, pero con canales en español)

Blog: [dv.uy](#)



Daniela Vázquez @d4tagirl · Apr 9

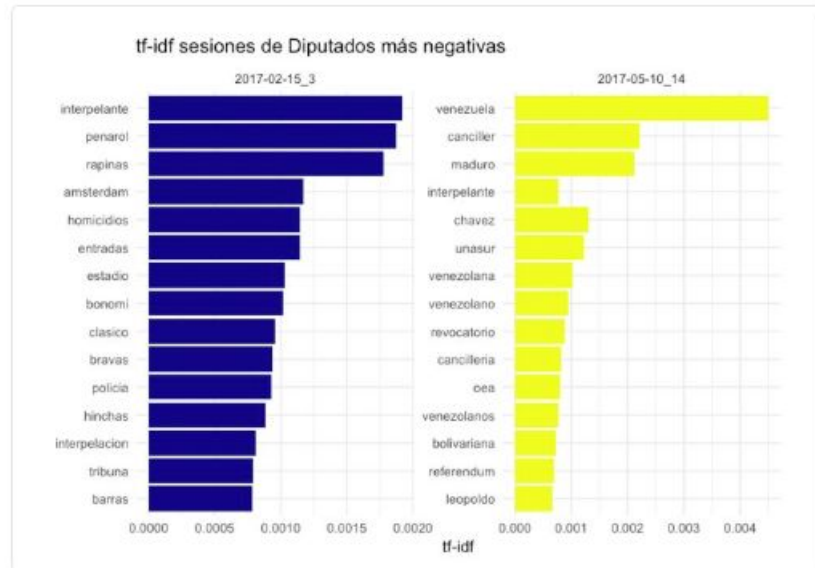
Nuevo #rstatsES post 🧐

Querés saber de qué hablaron los parlamentarios uruguayos desde 2017?
Averiguelo acá! dv.uy/parlamento ☀️

- web ➡️ pdf ➡️ texto
- análisis de sentimiento 💪
- tf-idf para saber de qué se habló en cada sesión 🧑🏻

#tidytext #opendata 🚀

🌐 Translate Tweet



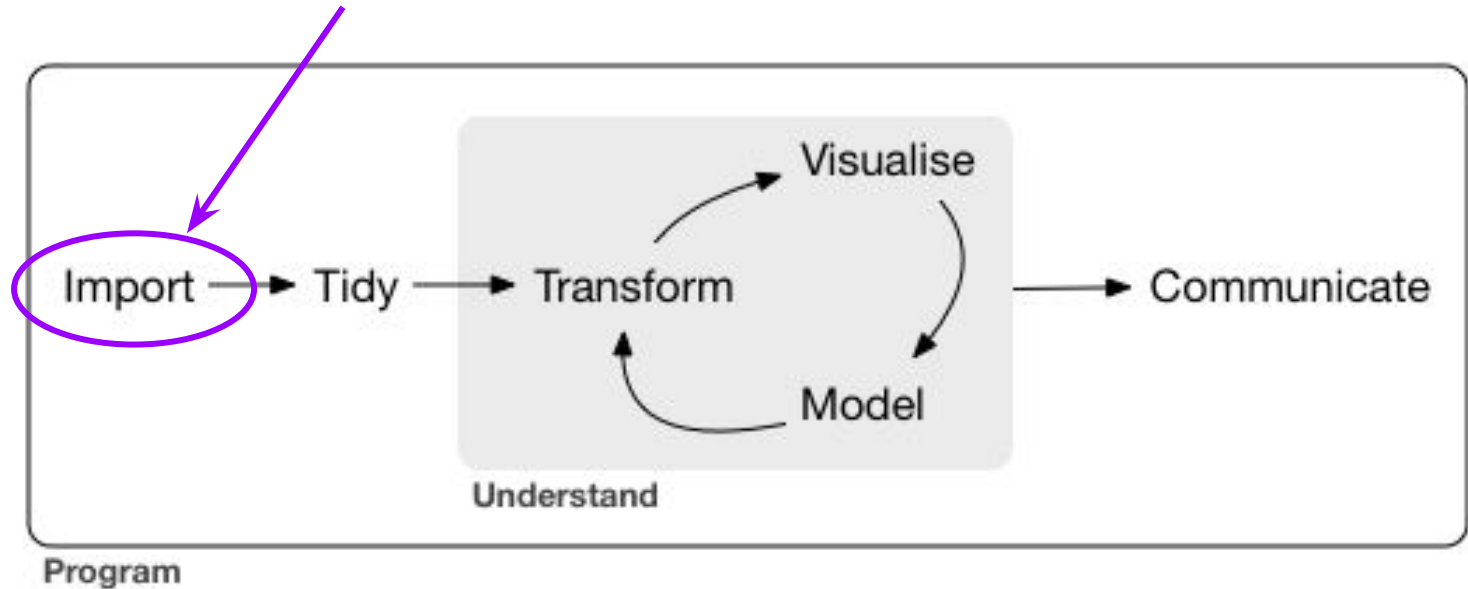
9

54

102




Obtención de datos



Obtención de datos

Situación en Uruguay: Avanzando con las iniciativas de datos abiertos, pero falta.

Problemas resueltos:

★ Descargar los archivos de las Sesiones de forma sistemática: *web scraping*. →  : robotstxt, rvest

★ *Extraer el texto* contenido en los archivos en formato pdf.
└→  : pdftools

Daniela Vázquez @d4tagirl · Apr 4

🚀 Nuevo [#rstatsES](#) blog post 🤖

"Scrapeando las sesiones parlamentarias de Uruguay" 🇺🇾

[dv.uy/scrap-parlam](#)

De qué se trata?

🌐 Scrapear la web usando rvest y robotstxt

📄 Extraer texto de archivos pdf con pdftools

🔍 Datos abiertos y transparencia

🌐 Translate Tweet



💬 3

🔄 20

❤️ 43



Scraping: robotstxt

Paquete de rOpenSci.

Normas *de etiqueta* que es recomendable seguir.

Archivo `robots.txt`: establece para cada sección del sitio si este uso es adecuado.

```
robotstxt::get_robotstxt ("https://parlamento.gub.uy")
```

```
robotstxt::paths_allowed ("https://parlamento.gub.uy/documentosyleyes/documentos/diarios-de-sesion")
```

```
## [1] TRUE
```

Scraping: selección del contenido

Las páginas web son archivos `html` que el navegador interpreta y lo transforma en lo que nosotros vemos.

Forma más intuitiva que encontré para seleccionar el contenido del `html` al que quiero acceder:

[Inicio](#) | [Documentos y Leyes](#) | [Documentos](#) | [Diarios de Sesiones](#)

Búsqueda de Documentos y Leyes

Constitución de la República

Leyes

Códigos

Documentos

Discursos

Diarios de Sesiones

Cuerpo

Cámara de Representantes ▲

Legislatura

-Legislatura Actual- XLVIII (2015-2020) ⬆

Las fechas deben estar entre '15-02-2015' y '14-02-2020' para ésta legislatura

Desde

01-01-2017

hasta

31-03-2018

Nro. de Sesión

Nro. de Diario

Texto

Buscar todas las Palabras 

Filtrar

Limpiar filtros

Scraping: rvest

```
pdfs <- url %>%
```

```
  rvest::read_html () %>%      # lee html
```

```
  rvest::html_nodes (".views-field-DS-File-IMG a" ) %>%
```

```
    # extrae nodo, el que encontré antes! La "a" indica que es  
hipervínculo
```

```
  rvest::html_attr ("href") %>%
```

```
    # selecciona el atributo href, que es el link (relativo) al pdf
```

```
  purrr::map(~ paste0 ("https://parlamento.gub.uy", .))
```

```
    # Completa el link para que sea absoluto
```

Extracción de textos: pdftools

```
pdfs <- pdfs %>%  
  purrr::map(~ paste0(pdftools::pdf_text(.), collapse = ' ')) %>%  
  # extrae texto del pdf y colapsa todas las páginas en un string  
  purrr::map(~ stringr::stri_trans_general(  
    tolower(.), id = "latin-ascii")) %>%  
  purrr::map(~ stringr::stri_replace_all(  
    ., replacement = "", regex = "\\n")) %>%  
  purrr::map_df(function(pdf) {tibble::tibble(pdf)})  
  # transformo la lista en un dataframe
```


Extracción de textos: pdf tools

pdf
numero 4151 montevideo, miercoles 14 de marzo de 2018 republica oriental del uruguay diario de sesiones camara de representantes 5ª sesion (extraordinaria) preside el senor representante jorge gandini (presidente)
numero 4148 montevideo, martes 6 de marzo de 2018 republica oriental del uruguay diario de sesiones camara de representantes 2ª sesion presiden los senores representantes jorge gandini (presidente)
numero 4147 montevideo, jueves 1º de marzo de 2018 republica oriental del uruguay diario de sesiones camara de representantes 1ª sesion preside el senor representante jorge gandini (presidente)
numero 4146 montevideo, jueves 8 de febrero de 2018 republica oriental del uruguay diario de sesiones camara de representantes 3ª sesion (extraordinaria) presiden los senores representantes prof. jose carlos mahia (preside
numero 4145 montevideo, miercoles 7 de febrero de 2018 republica oriental del uruguay diario de sesiones camara de representantes 2ª sesion (extraordinaria) preside el senor representante prof. jose carlos mahia (presidente) actuan en secret
numero 4144 montevideo, miercoles 20 de diciembre de 2017 republica oriental del uruguay diario de sesiones camara de representantes 1ª sesion (extraordinaria) presiden los senores representantes prof. jose carlos mahia (presi



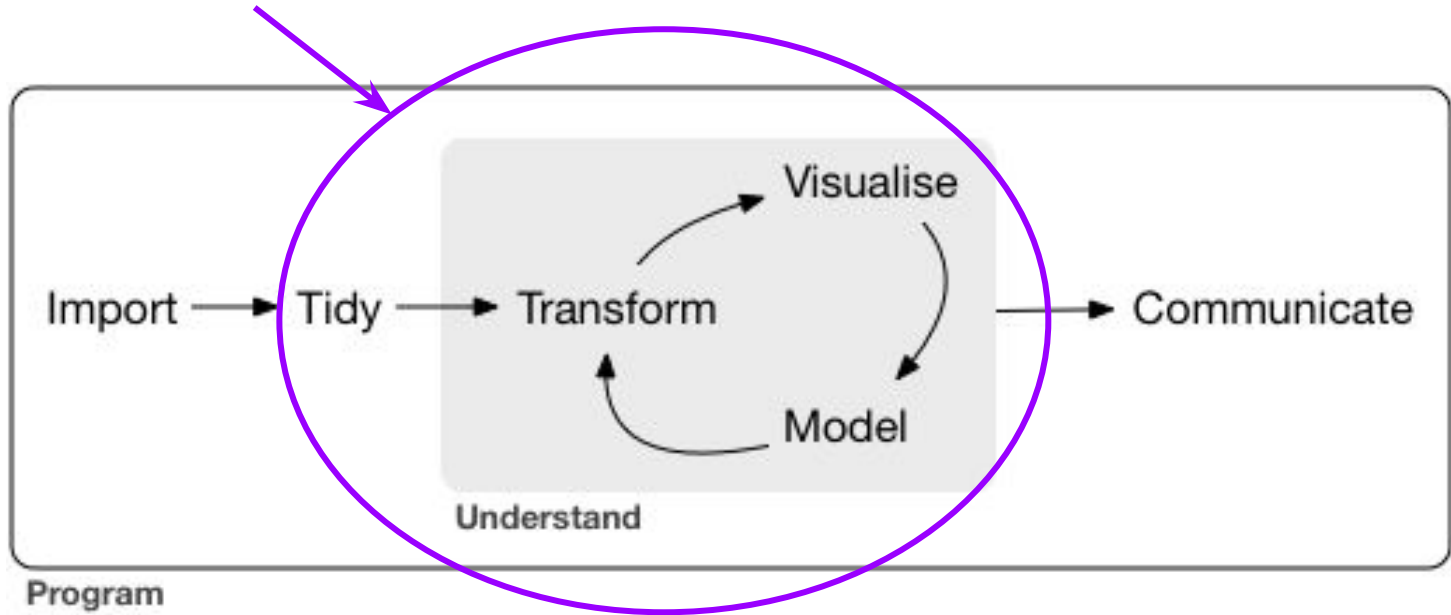
Advertencia!

`pdf tools::pdf_text()` lee los renglones de izquierda a derecha. Relevante para evaluar qué tipo de análisis puedo hacer.

Extracción de textos: pdftools

fecha_sesion	fecha	sesion	pdf
2018-03-06_1	2018-03-06	1	n.º 1 - tomo 578 6 de marzo de 2018 republica oriental del uruguay diario de sesiones de la camara de senadores cuarto periodo de la xlviii legislatura
2018-02-27_57	2018-02-27	57	n.º 57 - tomo 577 27 de febrero de 2018 republica oriental del uruguay diario de sesiones de la camara de senadores tercer periodo de la xlviii legislatura 57
2018-02-21_56	2018-02-21	56	n.º 56 - tomo 577 21 de febrero de 2018 republica oriental del uruguay diario de sesiones de la camara de senadores tercer periodo de la xlviii legislatura 56. ^a
2018-02-07_55	2018-02-07	55	n.º 55 - tomo 577 7 de febrero de 2018 republica oriental del uruguay diario de sesiones de la camara de senadores tercer periodo de la xlviii legislatura 55.
2018-02-06_54	2018-02-06	54	n.º 54 - tomo 577 6 de febrero de 2018 republica oriental del uruguay diario de sesiones de la camara de senadores tercer periodo de la xlviii legislatura 54. ^a s
2017-12-27_53	2017-12-27	53	n.º 53 - tomo 577 27 de diciembre de 2017 republica oriental del uruguay diario de sesiones de la camara de senadores tercer periodo de la xlviii legislatura 53. ^a se

Análisis de datos



Análisis de datos

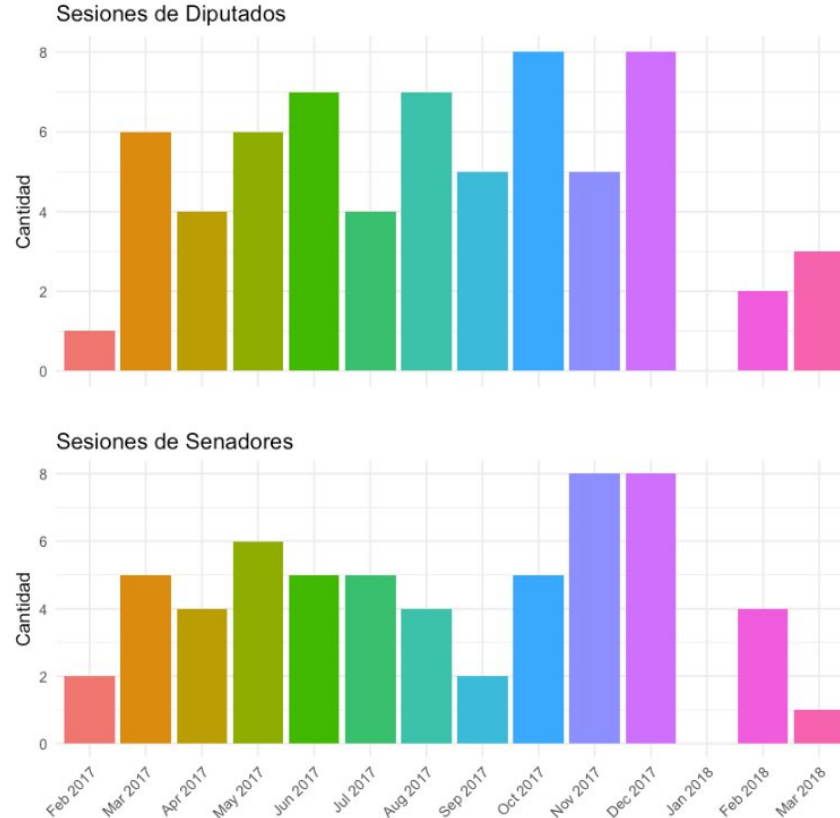
- ★ Llevar los datos a formato *tidy*.



- ★ Exploración de datos:  : tidyverse (esp. ggplot2)

- ★ Análisis de texto y sentimiento:  : tidytext, ggplot2

Frecuencia mensual de sesiones



 `zoo::as_yearmon()`

devuelve mes y año de
cada fecha

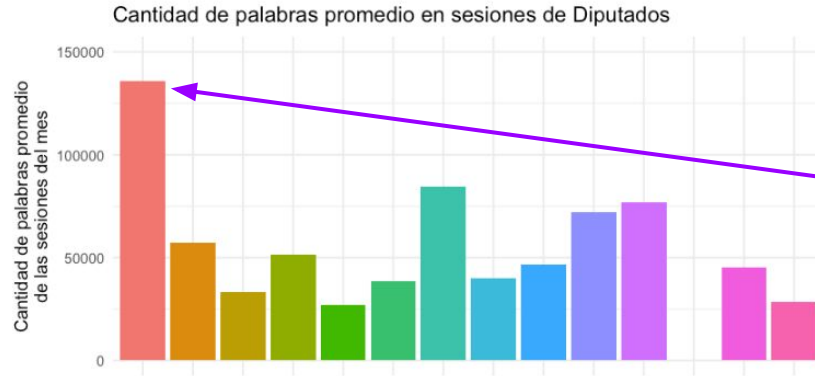
Extensión de las sesiones: tidytext

```
# calculo las palabras por mes
sesion diputados words mes <- diputados %>%
  tidytext::unnest(tokens(word, pdf)) %>%
  mutate(mes = as.yearmon(fecha)) %>%
  count(mes, sort = TRUE) %>%
  ungroup()
```

```
# calculo la cantidad de sesiones por mes
cant sesiones diputados mes <- diputados %>%
  group by(mes = as.yearmon(fecha)) %>%
  summarise(cant_sesiones = n()) %>%
  ungroup()
```

```
# los junto
sesion diputados words mes <-
  left join(sesion diputados words mes, cant sesiones_diputados_mes) %>%
  mutate(palabras_prom_sesion = n/cant_sesiones)
```

Extensión de las sesiones



Más de 15 hs. de sesión:
Interpelación al Ministro
del Interior, por situación
de extrema violencia en el
fútbol.



Análisis de sentimiento: `lexicon`

Una forma de analizar el sentimiento de un texto es como la suma de los sentimientos asociados a las palabras individualmente consideradas.

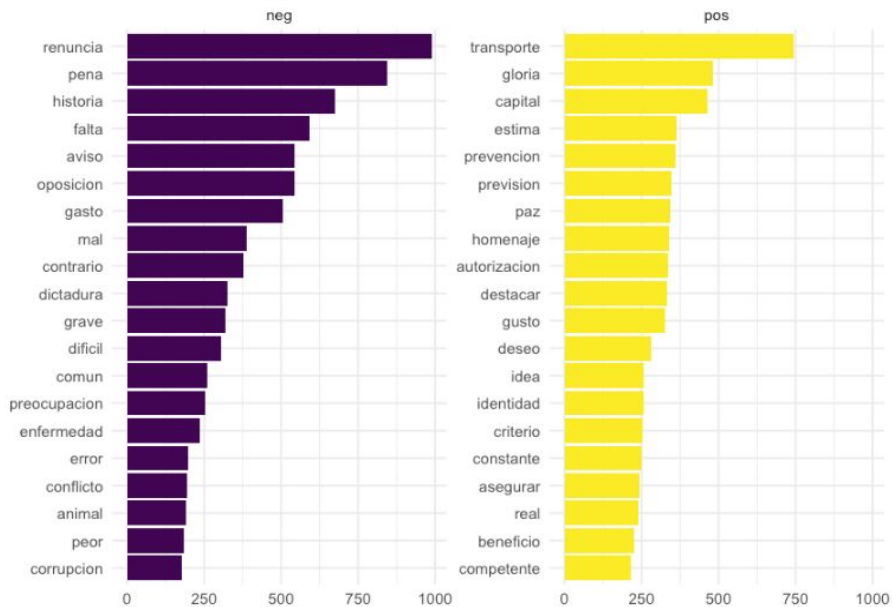
Lexicón usado: Veronica Perez-Rosas, Carmen Banea and Rada Mihalcea (2012)

Limitaciones:

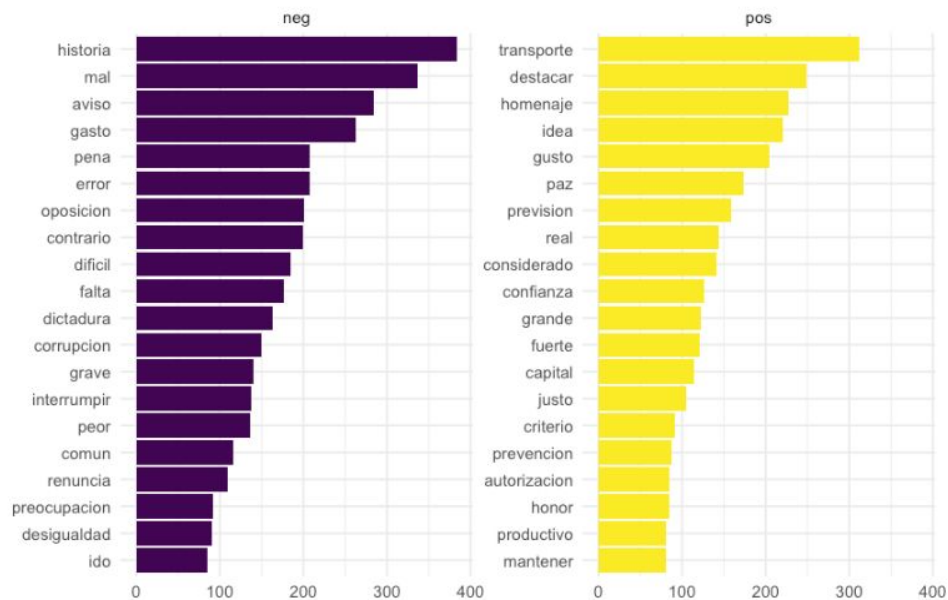
- ★ tiene muy pocos términos (476 palabras positivas de 871 en total);
- ★ la mayoría (si no todos) los adjetivos que considera son masculinos.

Análisis de sentimiento: tidytext + lexicon

Palabras con sentimientos más extremos en sesiones de Diputados

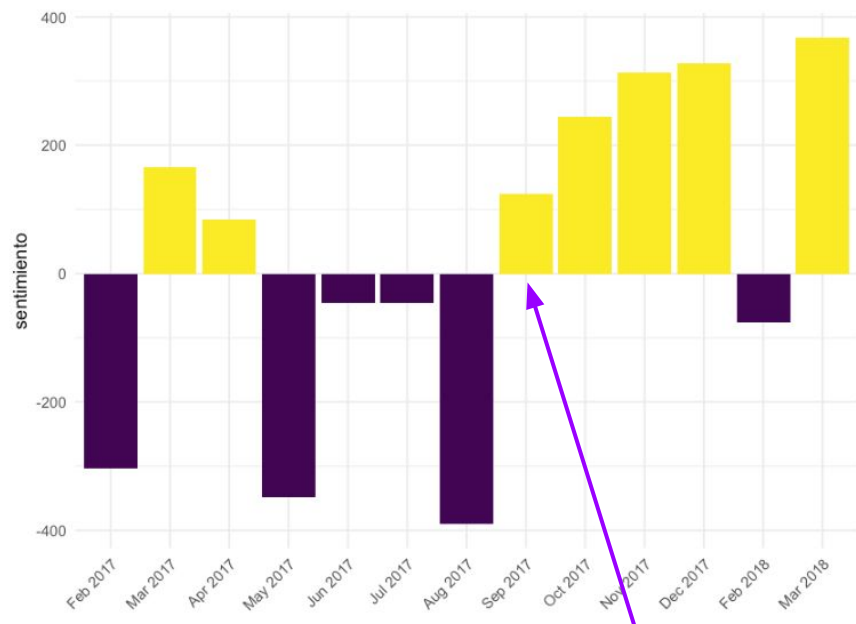


Palabras con sentimientos más extremos en sesiones de Senadores

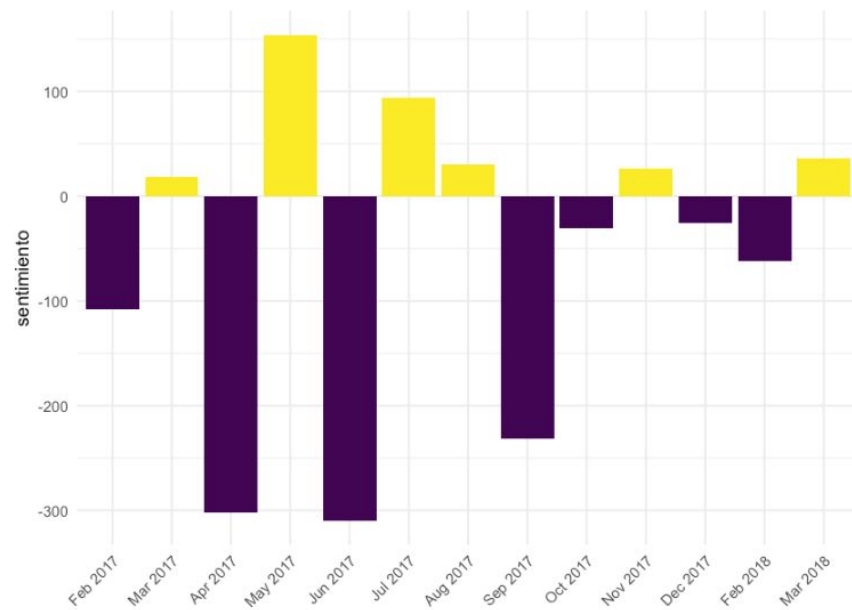


Sentimiento por mes:

Diputados



Senadores



Renuncia del Vice Presidente Raúl Sendic

Sesiones más extremas:

Diputados:

Más positivas

fecha_sesion	neg	pos	sentimiento
2018-03-06_2	158	379	221
2017-12-20_1	336	506	170

Más negativas

fecha_sesion	neg	pos	sentimiento
2017-05-10_14	792	360	-432
2017-02-15_3	788	484	-304

Senadores:

Más positivas

fecha_sesion	neg	pos	sentimiento
2017-05-16_15	52	158	106
2017-11-28_44	243	321	78

Más negativas

fecha_sesion	neg	pos	sentimiento
2017-09-18_32	598	364	-234
2017-06-07_18	458	242	-216

¿De qué se habló en las más negativas? $tf-idf$

term frequency (tf) (frecuencia del término):

qué tan frecuentemente aparece una palabra en la sesión.

inverse document frequency (idf) (inverso de la frecuencia del documento):

reduce el peso a las palabras comúnmente usadas en el total de las sesiones y aumenta el de las que no son muy usadas.

$tf-idf = tf * idf$:

frecuencia del término, ajustada por qué tan raramente es usado.

Qué tan importante es una palabra para una sesión, en el conjunto de sesiones que estoy analizando.

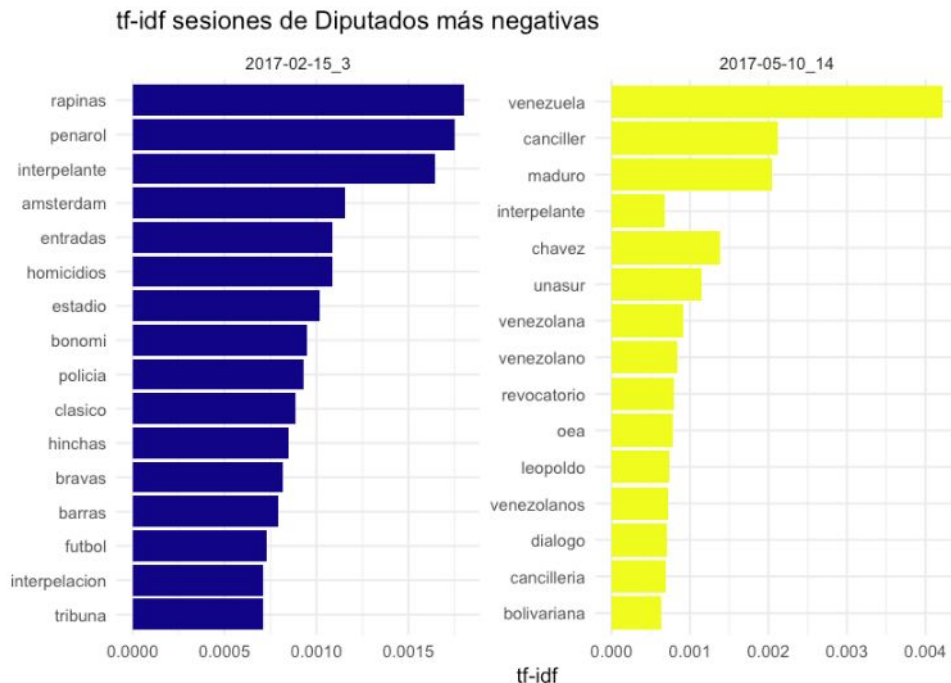
¿De qué se habló en las más negativas? tf-idf

```
sesion_diputados_words <- diputados %>%  
  unnest_tokens(word, pdf) %>%  
  count(fecha, sesion, fecha_sesion, word, sort = TRUE) %>%  
  ungroup()
```

```
diputados_words <- sesion_diputados_words %>%  
  group_by(fecha_sesion) %>%  
  summarize(total = sum(n))
```

```
sesion_diputados_tfidf <- left_join(sesion_diputados_words, diputados_words) %>%  
  tidytext::bind_tf_idf(word, fecha_sesion, n)
```

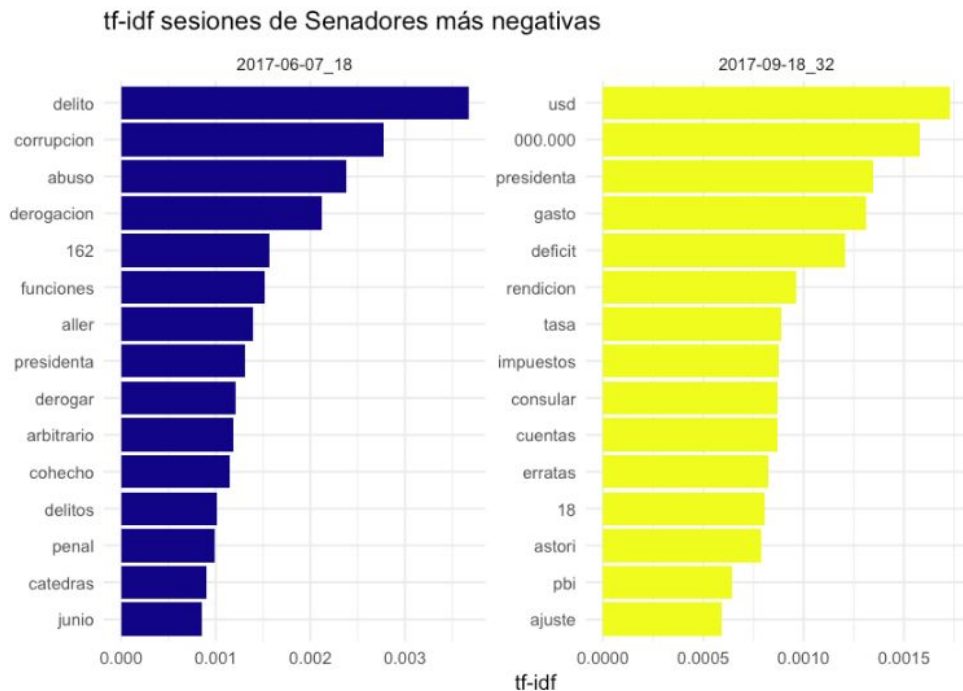
¿De qué se habló en esas sesiones? $tf-idf$



interpelación al Ministro del Interior Eduardo Bonomi por extrema violencia en el fútbol.

Interpelación al canciller Nin Novoa por la posición a adoptar ante la crisis venezolana.

¿De qué se habló en esas sesiones? tf-idf



Derogación de un artículo del Código Penal referido al Abuso de Funciones en casos no previstos especialmente por la ley.

Rendición de Cuentas

Fin!

Artículos detallados y lectura adicional:

- ★ dv.uy/text-parlam

- ★ dv.uy/parlamento

Análisis completo:

- ★ https://github.com/d4tagirl/uruguayan_parliamentary_session_diary

Gracias!

