Montevideo, 12 de Agosto de 2023

Facultad de Ciencias Sociales

UdelaR

Recuperación y análisis de texto con R Educación Permanente
FCS Consigna de Trabajo Final 2023

Consigna de Trabajo Final

Soc. Gervasio Riveiro

Introducción y fundamentación del trabajo

Para realizar la entrega de la consigna final del curso de Recuperación y análisis de texto con R de Educación Permanente. Opte para trabajar como tema la cuestión Indígena en Uruguay desde los charrúas en el discurso del orden parlamentario.

Los motivos que me llevaron a optar por este tema en principio se conectan con estudios de compañeros cercanos de la Facultad de Humanidades que están trabajando el tema desde diversas aristas. Considero interesante también, el aporte sociológico para profundizar en una mirada sobre los discursos que ocupan la agenda estatal y ciudadana con respecto a este tema.

La cuestión indígena en Uruguay es un tema de relevancia histórica, cultural, sociológica y política. La situación de los charrúas, una de las poblaciones indígenas originarias de Uruguay, ha generado debates y reflexiones sobre la identidad nacional y los derechos humanos. La consideración de un punto de vista sociológico y político sobre esta temática es esencial para comprender cómo las dinámicas sociales y políticas han influido en la percepción y el tratamiento de las identidades y derechos de los pueblos indígenas en el país.

Parte de los nuevos discursos académicos que han circulado en la sociedad Uruguaya plantean la idea de la invisibilización de la presencia indígena en Uruguay. Siendo una cuestión históricamente marginada en la narrativa oficial. A pesar de las adversidades, las comunidades indígenas en Uruguay persisten en la resistencia conformando una conciencia colectiva y reivindicadora en cara a los derechos humanos.

En el caso de la sociología, podemos ver un potencial como herramienta que ayude a mostrar la profundidad de los relatos alternativos a los dominantes, sobre la historia y sobre la construcción de identidades en Uruguay. Dando lugar a posibles análisis que incluyan la perspectiva indígena histórica, pero también en el hoy y el presente en la resistencia de sus comunidades.

A nivel de antecedentes solamente mencionare autores como Mónica Michelena, Gonzalo Figueredo y Mónica Sans, entre otros. Nos presentan distinto material que trabaja la cuestión indígena y las transformaciones en las formas de abordaje de esta cuestión desde la academia, pero también desde el ámbito político e histórico en sus relatos.

Selección de muestra y casos

Un primer acercamiento trae consigo desde este curso es la posibilidad de trabajar los discursos parlamentarios presentes en los diarios de sesión, siendo materia prima privilegiada para explorar cómo estas cuestiones han sido abordadas y debatidas en el ámbito político Uruguayo, reflejando las tensiones y transformaciones en la sociedad.

Como muestra se presenta una base de datos que conforma 44 sesiones parlamentarias en su formato de diario de sesión, de las mismas se extraen los discursos sistematizados de los parlamentarios declarantes a través de la técnica de scraping aplicada en la web del parlamento. Esta herramienta fue desarrollada por Nicolás Schmidt, politólogo del departamento de ciencias políticas de UdelaR. Todas estas sesiones tienen como temporalidad desde la fecha actual hasta principios de marzo del año Dos mil veinte.

En cuanto a la selección de las sesiones, se realizó una búsqueda de los diarios de sesión donde se mencione la palabra indígena y charrúa. De la base de datos construida se tomó la variable speech, refiriendo a las intervenciones parlamentarias. Haciendo un conteo de palabras podemos encontrar que palabras y en qué contexto, se asocian con las palabras que elegimos para centralizar el trabajo.

Sobre el código

Se instalaron las siguientes librerías:

- Rvest
- Dplyr
- Quanteda
- Readtext
- Stringr
- ggplot2
- quanteda.textstats
- quanteda.textplots

Dentro del código hay instrucciones y explicaciones de las librerías utilizadas y sobre la utilidad de las funciones aplicadas. Se destaca la utilización de SPEECH y PUY para la construcción de las bases de datos.

```
#Instalación de las librerías
remotes::install_github("Nicolas-Schmidt/speech")
remotes::install_github("Nicolas-Schmidt/puy")
```

Luego se aplicó una función de speech para obtener el scraping de cada diario de sesión

```
url <- "https://parlamento.gub.uy/documentosyleyes/documentos/diarios-de-sesion/6204/IMG"

sesion39_40 <- speech::speech_build(url)

sesion39_40 <- speech::speech_build(file = url, compiler = TRUE, quality = TRUE)
```

Es importante aclarar que cada sesión tiene como formato en su nombre el luego de la palabra **sesión Nº de sesión_Nº de diario.**

Una vez conjugada toda la base de datos con todos los diarios de sesión sistematizados, se hace una limpieza de los datos obteniendo un data frame quitando palabras que no son relevantes.

```
dfm_sesiones <- quanteda::dfm(quanteda::tokens(sesiones$speech,
                                               remove_punct = TRUE,
                                               remove_numbers = TRUE),
                              tolower=TRUE,
                              verbose = FALSE)%>%
  quanteda::dfm_remove(pattern = c(quanteda::stopwords("spanish"),
                              tolower(sesiones$legislator),"presi*"),
                       min_nchar=3)%>%
  quanteda::dfm_trim(min_termfreq = 6)%>%
  dfm_tfidf()%>%
  quanteda::dfm_group(groups = sesiones$party)
```

sacopalabras= c(quanteda::stopwords("spanish"),tolower(sesiones$legislator),"presidente", "señor", "vos", "al menos", "creo", "senadora", "senador", "ahora", "nº", "diputado")

Luego aplicamos funciones para visualizar datos. Una de ellas nos da una nube de conceptos que se repiten y que pueden ser mayormente asociados a distintos partidos políticos.

```
quanteda.textplots::textplot_wordcloud(dfm_sesiones, min.count = 7,max_words = 500,
      random.order = FALSE,colors = RColorBrewer::brewer.pal(8,"Dark2"),comparison = T)

quanteda::topfeatures(dfm_sesiones,50)
```

```
quanteda::topfeatures(dfm_sesiones,50)
    señor         ley        país    nacional         hoy    gobierno         ser
     2810        1644        1507        1374        1369        1316        1315
     años      señora    proyecto       decir     uruguay       hacer    artículo
     1230        1183        1118        1092        1085        1052         994
    puede        hace presupuesto        bien       vamos      quiero      muchas
      970         941         929         927         910         880         880
      año  ministerio         así       parte    comisión         dos     derecho
      877         839         838         834         827         802         801
     sino        tema        vida     trabajo        creo       poder     gracias
      793         792         791         780         759         749         747
 personas       tener    diputado   situación    política      social        pido
      745         731         716         711         711         704         700
  momento      frente       ahora  importante       gente    libertad         vez
      698         698         683         668         659         655         648
     solo
      646
```

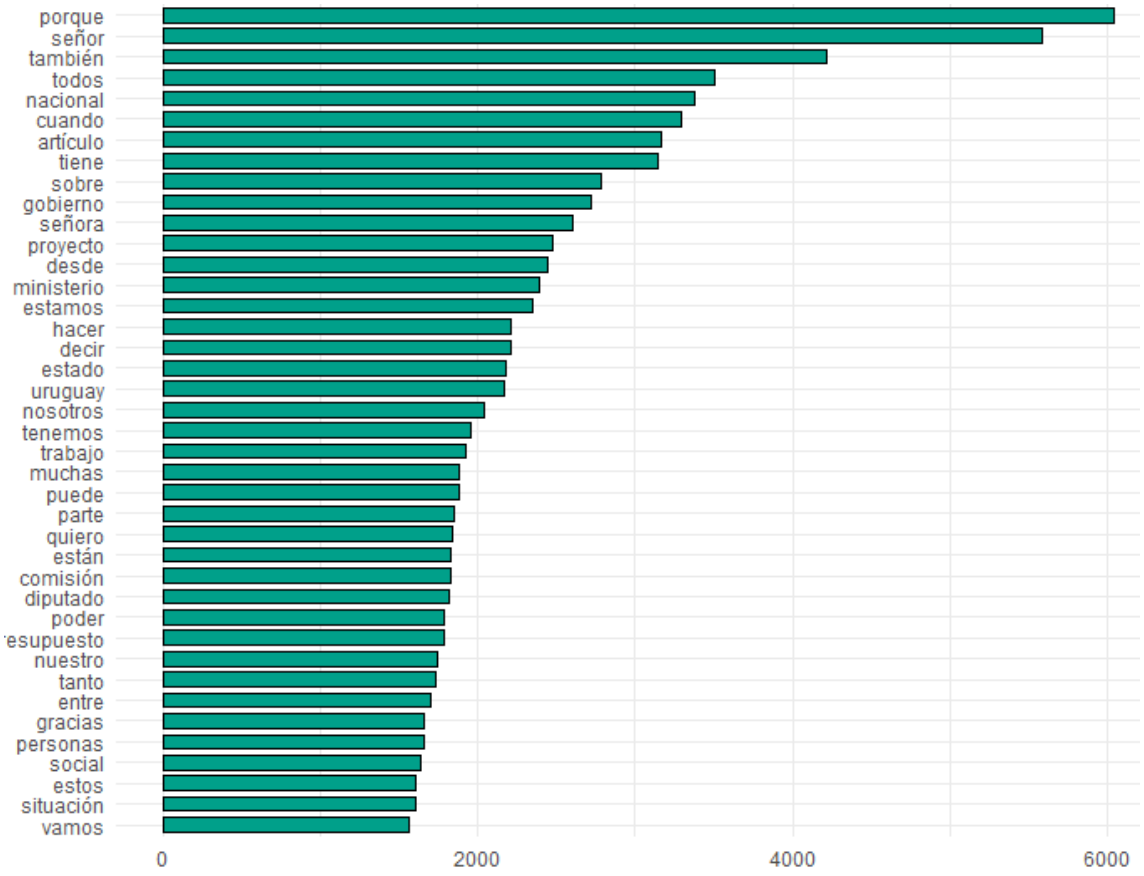
Entendiendo que se puede mejorar mucho la depuración de datos, podemos encontrar algunas palabras relevantes como derecho, vida, proyecto, libertad, dentro de las 50 primeras palabras. En la primera imagen también aparecen nombres y la asociación que hay en cada partido en torno al debate.

Con ggplot aplicamos una función para visualizar la presencia de las palabras en gráfico de barras

```
library(ggplot2)


sesiones$speech %>%
  minchar(., min = 4) %>%
  tibble::enframe() %>%
  tidytext::unnest_tokens(word, value) %>%
  dplyr::count(word, sort = TRUE) %>%
  dplyr::mutate(word = stats::reorder(word, n)) %>%
  dplyr::filter(!stringr::str_detect(word, "presidente") ) %>%
  .[1:40,] %>%
  ggplot(aes(word, n)) +
  geom_col(col = "black", fill = "#00A08A", width = .7) +
  labs(x = "", y = "") +
  coord_flip() +
  theme_minimal()
```

Obteniendo como resultado este gráfico:



Con la librería quanteda, pudimos hacer una correlación de la palabra indígena y charrúa a través de esta función:

```
quanteda.textstats::textstat_simil(dfm_tfidf(dfm_sesiones),selection = "indígena",
                                   method = "correlation",margin = "features")%>%
  as.data.frame()%>%
  dplyr::arrange(-correlation)%>%
  dplyr::top_n(20)
```

Aquí se muestran las 20 primeras palabras que se correlacionan con mayor intensidad en el caso de **Indígena.**

```
Selecting by correlation
          feature1 feature2 correlation
1            facto indígena   0.9986188
2            drama indígena   0.9876583
3        horizonte indígena   0.9808165
4           relata indígena   0.9799579
5         uniforme indígena   0.9759001
6           muerto indígena   0.9759001
7        favorecen indígena   0.9759001
8           cuidan indígena   0.9759001
9    diputada.ñora indígena   0.9759001
10   diputado.ñora indígena   0.9759001
11   levantamiento indígena   0.9759001
12           vicio indígena   0.9759001
13        montagno indígena   0.9759001
14            sexo indígena   0.9759001
15    apartamiento indígena   0.9759001
16         aceptan indígena   0.9759001
17         pegarle indígena   0.9759001
18        muriendo indígena   0.9759001
19       infelices indígena   0.9759001
20          máxime indígena   0.9759001
```

Aquí se muestran las 20 primeras palabras que se correlacionan con mayor intensidad en el caso de **Charrúa.**

```
Selecting by correlation
         feature1 feature2 correlation
1       eléctrica  charrúa   0.9955796
2            --en  charrúa   0.9955796
3            data  charrúa   0.9955796
4    fideicomisos  charrúa   0.9955796
5          lafluf  charrúa   0.9954250
6             utu  charrúa   0.9944694
7       genocidio  charrúa   0.9943278
8     tecnológica  charrúa   0.9940203
9      rechazamos  charrúa   0.9937721
10       quehacer  charrúa   0.9937721
11       cerrados  charrúa   0.9937721
12  argumentación  charrúa   0.9923451
13           nube  charrúa   0.9921924
14     concretado  charrúa   0.9921924
15       diplomas  charrúa   0.9921924
16      comparten  charrúa   0.9921924
17      violentas  charrúa   0.9921924
18     anunciamos  charrúa   0.9921924
19     recorrimos  charrúa   0.9921924
20          sello  charrúa   0.9921924
```

Con la siguiente función presentamos el contexto en el que fueron dichas estas palabras

```
kwic = quanteda::kwic(quanteda::tokens(sesiones$speech,
                                       remove_punct = TRUE,
                                       remove_numbers = TRUE),
                      pattern = quanteda::phrase(c("indígena")),
                      window = 20)

DT::datatable(kwic)
```

pre	keyword	post	pattern
agradezco Me hizo acordar a un poema solo tomará veinte segundosen el cual un joven se encuentra con una anciana	indígena	explotada como casi todos los ancianos de la América Latina y en un momento dice Hubiera querido hablar con ella	indígena
que provocó el desbande de miles de indios misioneros Muchos de ellos vinieron a la Banda Oriental constituyendo el elemento	indígena	principal mucho más numeroso que los charrúas que vivirá en nuestra campaña en la época de Artigas Son por otra	indígena

pre	keyword	post
tenis de mesa rugby y halterofilia Además se realizan salidas didácticas deportivas en ese sentido se ha ido al Estadio	Charrúa	al Estadio Centenario y al Campeón del Siglo También han ido a encuentros a partidos de rugby y de fútbol
toda la población de la campaña Se e hereda de su abuelo paterno como ya lo habíamos en	charrúa	conviviendo con una india doptó como hijo propio y heredó todos sus bienes una te llevar adelante no

Por último quedan guardadas en formato Excel y Rdata las bases de datos utilizadas. Aquí dejo las funciones que utilicé para las mismas, y la función para combinar dichas bases.

#Combine todas las bases de datos en una sola

sesiones <- rbind(sesion1_4382, sesion11_4336, sesion13_4397, sesion15_15, sesion15_4274, sesion16_17, sesion18_18, sesion19_19, sesion19_4344, sesion2_4261, sesion21_21, sesion22_4406, sesion24_4408, sesion25_4409, sesion27_4470, sesion28_4412, sesion28_4471, sesion29_29, sesion29_30, sesion29_4288, sesion3_3, sesion3_4328, sesion3_4384, sesion31_31, sesion31_4415,

sesion36_37, sesion36_4295, sesion38_4442, sesion39_40, sesion4_4388, sesion44_4303, sesion48_4307, sesion49_4443,

sesion51_4376, sesion52_4377, sesion54_4379, sesion58_4317, sesion6_4449, sesion6_6, sesion7_4391, sesion7_4450, sesion7_7,

sesion8_4328, sesion8_4333, sesion8_8)

```r
#funcion para guardar cada base de datos como Rdata en la carpeta Rdata
#*Cada sesion esta guardada con el formato en el nombre ej: sesion(nºsesion_nºdiario)
#*la base de datos sesiones es una combinación de cada base de cada diario de sesion

save(sesion3_4384, file = "E:\\Mis Documentos\\GitHub\\EntregaR\\EntregaDB\\Rdata\\sesion3_4384.Rdata")



#*guardé también cada base de datos en la carpeta Excel carpeta en formato xlsx

install.packages("xlsx")
library(xlsx)

write.Rdata(sesion8_8,"E:\\Mis Documentos\\GitHub\\EntregaR\\EntregaDB\\Excel\\sesion8_8.Rdata")
```

Para concluir, el trabajo de procesamiento de los datos, la elección del tema a trabajar y las técnicas que utilicé, hicieron que el trabajo fuese de mi agrado. Considero que es una línea de investigación a trabajar sumamente interesante, y donde las ciencias sociales tenemos mucho para aportar desde este tipo de análisis.