# Lab Nº2

Sebastian Urbina

**Lab #2 Assignment ~ Conduct Text Analysis**

**1. Collect some texts. Compare them in a number of ways.**

I will conduct textual analysis on official presidential speeches from presidents of Chile that differ in their ideologies. By web scraping the speeches from the national library's archives, I gathered substantive textual data from Gabriel Boric (current president) and Sebastian Piñera (former president) - representatives of contrasting political leanings.

An initial notable difference lies in the speech lengths themselves. Boric's speech spanned 46 pages, while Piñera's was shorter at 36 pages. This disparity in length could potentially signify several factors.

Rhetorical Style: Progressive leaders like Boric may favor a more expansive, rhetorical approach to connect with the public, whereas conservative voices like Piñera might tend toward more concise policy/governance messaging.

Breadth of Agenda: The greater length of Boric's speech could indicate a wider-ranging set of priorities and proposals reflecting his activist roots and vision for social transformation.

Audience Engagement: President Boric may have tailored his lengthier speech to galvanize his base with inspirational language, while Piñera's relatively shorter speech suggests a more measured tone appealing to centrist and business constituencies.

However, length alone does not necessarily determine ideological substance. A thorough analysis of word frequencies, topics, sentiments, and linguistic choices is required to unpack the substantive contrasts in these speeches emanating from their different worldviews.

**2.You will likely want to have them be "bags of words." Prepare the text through removing upper case, white space, punctuation, and consider stemming the words, if appropriate for you purpose.**

```
# Loading Stopwords in spanish
spanish_stopwords <- stopwords("spanish")
```

```r
stopwords_df <- tibble(word = spanish_stopwords)
```

```r
paper_words <- data_frame(file = paste0("/Users/saurbina/Documents/pdf-txt/",
                                        c("boric_v1.txt", "pinera_V1.txt"))) %>%
  mutate(text = map(file, read_lines)) %>%
  unnest() %>%
  group_by(file = str_sub(basename(file), 1, -5)) %>%
  mutate(line_number = row_number()) %>%
  ungroup() %>%
  unnest_tokens(word, text) %>%
  anti_join(stopwords_df, by = "word")
```

**3. Generate relative word frequencies for each bag of words, and compare them to each other.**

**Boric over Piñera**

| | names | boric_v1.x | pinera_V1.x | boric_v1.y | pinera_V1.y |
|---|---|---|---|---|---|
| 2618 | discurso | 162 | 1 | 0.004467487 | 0.0001309586 |
| 646 | año | 182 | 7 | 0.005019028 | 0.0009167103 |
| 5609 | personas | 125 | 5 | 0.003447135 | 0.0006547931 |
| 4885 | mil | 121 | 8 | 0.003336826 | 0.0010476689 |
| 6287 | quiero | 112 | 9 | 0.003088633 | 0.0011786276 |
| 7689 | ustedes | 84 | 4 | 0.002316475 | 0.0005238345 |
| 6826 | sabemos | 81 | 4 | 0.002233743 | 0.0005238345 |
| 386 | además | 127 | 15 | 0.003502289 | 0.0019643793 |
| 4682 | manera | 74 | 4 | 0.002040704 | 0.0005238345 |
| 5628 | pesos | 55 | 0 | 0.001516739 | 0.0000000000 |
| 6459 | recursos | 69 | 4 | 0.001902818 | 0.0005238345 |
| 4965 | momento | 48 | 0 | 0.001323700 | 0.0000000000 |
| 1366 | chile | 266 | 46 | 0.007335503 | 0.0060240964 |
| 2842 | ello | 104 | 12 | 0.002868016 | 0.0015715034 |
| 90 | 2023 | 49 | 1 | 0.001351277 | 0.0001309586 |
| 840 | así | 63 | 4 | 0.001737356 | 0.0005238345 |
| 6224 | pueblo | 53 | 2 | 0.001461585 | 0.0002619172 |
| 762 | aquí | 48 | 1 | 0.001323700 | 0.0001309586 |
| 4894 | millones | 81 | 8 | 0.002233743 | 0.0010476689 |
| 1744 | conjunto | 52 | 2 | 0.001434008 | 0.0002619172 |
| 6984 | ser | 122 | 17 | 0.003364404 | 0.0022262965 |
| 4604 | logramos | 41 | 0 | 0.001130660 | 0.0000000000 |

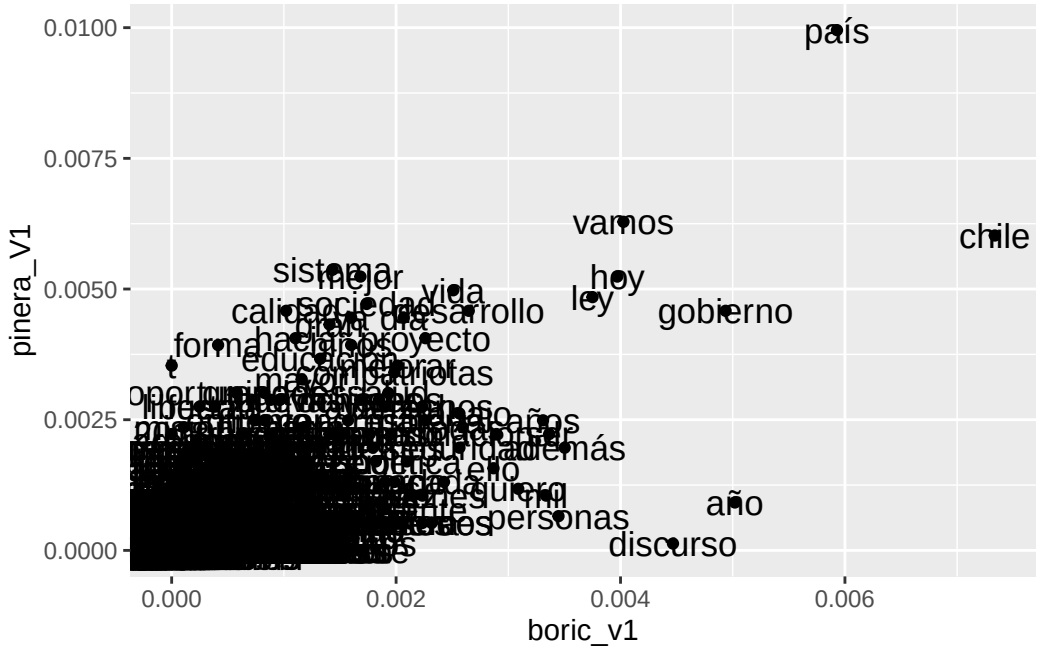
| | | | | | |
|---|---|---|---|---|---|
| 1169 | cada | 88 | 10 | 0.002426783 | 0.0013095862 |
| 3947 | importante | 69 | 6 | 0.001902818 | 0.0007857517 |
| 7818 | vez | 78 | 8 | 0.002151012 | 0.0010476689 |
| 6668 | respecto | 54 | 3 | 0.001489162 | 0.0003928759 |
| 6977 | sepan | 38 | 0 | 0.001047929 | 0.0000000000 |
| 1562 | cómo | 47 | 2 | 0.001296123 | 0.0002619172 |
| 5433 | parlamentarios | 42 | 1 | 0.001158237 | 0.0001309586 |
| 4909 | ministra | 37 | 0 | 0.001020352 | 0.0000000000 |

| | boric.over.pinera |
|---|---|
| 2618 | 0.004336528 |
| 646 | 0.004102318 |
| 5609 | 0.002792342 |
| 4885 | 0.002289157 |
| 6287 | 0.001910005 |
| 7689 | 0.001792640 |
| 6826 | 0.001709909 |
| 386 | 0.001537910 |
| 4682 | 0.001516869 |
| 5628 | 0.001516739 |
| 6459 | 0.001378984 |
| 4965 | 0.001323700 |
| 1366 | 0.001311406 |
| 2842 | 0.001296513 |
| 90 | 0.001220318 |
| 840 | 0.001213521 |
| 6224 | 0.001199668 |
| 762 | 0.001192741 |
| 4894 | 0.001186074 |
| 1744 | 0.001172091 |
| 6984 | 0.001138107 |
| 4604 | 0.001130660 |
| 1169 | 0.001117197 |
| 3947 | 0.001117067 |
| 7818 | 0.001103343 |
| 6668 | 0.001096286 |
| 6977 | 0.001047929 |
| 1562 | 0.001034205 |
| 5433 | 0.001027279 |
| 4909 | 0.001020352 |

**Piñera over Boric**

| | names | boric_v1.x | pinera_V1.x | boric_v1.y | pinera_V1.y | boric.over.pinera |
|---|---|---|---|---|---|---|
| 5378 | país | 215 | 76 | 0.0059290718 | 0.009952855 | -0.004023783 |
| 7076 | sistema | 52 | 41 | 0.0014340081 | 0.005369303 | -0.003935295 |
| 1180 | calidad | 37 | 35 | 0.0010203519 | 0.004583552 | -0.003563200 |
| 4798 | mejor | 61 | 40 | 0.0016822018 | 0.005238345 | -0.003556143 |
| 7245 | t | 0 | 27 | 0.0000000000 | 0.003535883 | -0.003535883 |
| 3496 | forma | 15 | 30 | 0.0004136562 | 0.003928759 | -0.003515102 |
| 7095 | sociedad | 63 | 36 | 0.0017373559 | 0.004714510 | -0.002977154 |
| 3771 | hacia | 40 | 31 | 0.0011030831 | 0.004059717 | -0.002956634 |
| 3699 | gran | 51 | 33 | 0.0014064310 | 0.004321634 | -0.002915203 |
| 7699 | va | 58 | 34 | 0.0015994705 | 0.004452593 | -0.002853122 |
| 4493 | libertad | 9 | 21 | 0.0002481937 | 0.002750131 | -0.002501937 |
| 7834 | vida | 91 | 38 | 0.0025095141 | 0.004976427 | |
| 5297 | oportunidades | 21 | 23 | 0.0005791186 | 0.003012048 | |
| 2518 | día | 75 | 34 | 0.0020682808 | 0.004452593 | |
| 3569 | fuerza | 14 | 21 | 0.0003860791 | 0.002750131 | |
| 2763 | educación | 48 | 28 | 0.0013236997 | 0.003666841 | |
| 5134 | niños | 58 | 30 | 0.0015994705 | 0.003928759 | |
| 7732 | vamos | 146 | 48 | 0.0040262534 | 0.006286014 | |
| 4926 | misión | 4 | 18 | 0.0001103083 | 0.002357255 | |
| 85 | 2018 | 0 | 17 | 0.0000000000 | 0.002226296 | |
| 3701 | grandes | 30 | 23 | 0.0008273123 | 0.003012048 | |
| 4760 | mayor | 42 | 25 | 0.0011582373 | 0.003273965 | |
| 3592 | futuro | 32 | 22 | 0.0008824665 | 0.002881090 | |
| 2391 | desarrollo | 96 | 35 | 0.0026473995 | 0.004583552 | |
| 5177 | nueva | 35 | 22 | 0.0009651977 | 0.002881090 | |
| 761 | aquellos | 8 | 16 | 0.0002206166 | 0.002095338 | |
| 6205 | proyecto | 82 | 31 | 0.0022613204 | 0.004059717 | |
| 4830 | mensaje | 7 | 15 | 0.0001930395 | 0.001964379 | |
| 6120 | promover | 7 | 15 | 0.0001930395 | 0.001964379 | |
| 2969 | enfrentar | 27 | 19 | 0.0007445811 | 0.002488214 | |

```
7834      -0.002466913
5297      -0.002432930
2518      -0.002384312
3569      -0.002364052
2763      -0.002343142
5134      -0.002329288
7732      -0.002259760
4926      -0.002246947
85        -0.002226296
3701      -0.002184736
4760      -0.002115728
3592      -0.001998623
2391      -0.001936152
5177      -0.001915892
761       -0.001874721
6205      -0.001798397
4830      -0.001771340
6120      -0.001771340
2969      -0.001743633
```




We can observe significant differences between both speeches. The words most characteristic of Boric, in relation to former President Piñera, are 'discurso' (speech), 'año' (year), 'personas' (people), 'mil' (thousand), and 'quiero' (I want). Conversely, the words most reminiscent of

Piñera, relative to President Boric, are 'país' (country), 'sistema' (system), 'calidad' (quality), and 'mejor' (better).

**4. Articulate what differences (if any) you notice and whether this comports with a theory of why these bags of words should be similar or different.**

The differences in word frequencies reveal contrasting rhetorical styles and priorities between the two speeches. President Boric's speech demonstrates a more personal approach, with higher frequencies of words like 'discurso' (speech), 'año' (year), 'personas' (people), 'mil' (thousand), and 'quiero' (want). This suggests a focus on directly addressing and connecting with the people.

In contrast, ex-President Piñera's speech leans towards a more technocratic and policy-oriented rhetoric, evident from the higher relative frequencies of words such as 'país' (country), 'sistema' (system), 'calidad' (quality), and 'mejor' (better). His language seems geared towards discussing national issues, systems, and improvements.

Boric's personal style aims to forge a closer connection with the public by using more relatable language about daily life and aspirations. Piñera, on the other hand, adopts a more detached, institutional approach focused on the functioning of the country as a whole.

These contrasting word frequencies shed light on the differing tones and priorities of the two leaders. While Boric employs a more grassroots-oriented discourse, Piñera maintains a broader, system-level perspective in his speech. Furthermore, the political ideologies reflect different styles and ways of seeing their speeches.

**5. Run statistical tests of association between the bags of words (correlation, cosine similarity, regression or Chi-squared), and explain what they indicate.**

**Cosine**

```
# Cosine Correlation
cosine(t.t)
```

```
           boric_v1 pinera_V1
boric_v1  1.0000000 0.7108636
pinera_V1 0.7108636 1.0000000
```

First we will implement a cosine similarity measure. After removing stop words and stemming the words, the speeches of Obama and Trump exhibit a cosine similarity of 0.71, indicating a high degree of similarity. One possible hypothesis is that inaugural speeches are solemn

occasions, providing valuable insights into the institutionalization of the situation rather than being highly politicized. It is a highly protocol-driven scenario.

```
# Pearson Correlation
cor(t.t, method="pearson")
```

```
           boric_v1 pinera_V1
boric_v1  1.0000000 0.6726576
pinera_V1 0.6726576 1.0000000
```

Now if we compar the last cosine measure with pearson correlation we see that the speeches have a Pearson correlation coefficient of 0.67. This is an expected result because Cosine similarity tends to give higher values than Pearson correlation, given that the Pearson correlation coefficient represents the angular separation between two normalized data vectors measured from the mean, while the cosine similarity measures the angular separation of two data vectors measured from zero.

On the other way we can do a chi-test to see if the speeches are independet. After stop-words and with stemmed words, a Chi-squared test suggests that president Boric and ex president Piñera speeches are not independent. A p-value of less than 2.2e-16 indicates an extremely significant result. It suggests that there is a very low probability of observing the observed data under the null hypothesis, assuming no association between the variables being tested. In other words, it provides strong evidence against the null hypothesis and suggests that there is a significant relationship between the variables.

```
ctable <- table(t.t)
chisq.test(ctable)
```

```
    Chi-squared test for given probabilities

data:  ctable
X-squared = 361872, df = 100, p-value < 2.2e-16
```

Finally, we will conduct a regression to model how ex-president piñera relative frequency words predict president Boric speech.

```
================================================
                        Dependent variable:
                    ---------------------------
```

```
                            boric_v1.y
-----------------------------------------------
pinera_V1.y                   0.510***
                               (0.006)

Constant                      0.0001***
                              (0.00000)

-----------------------------------------------
Observations                    7,967
R2                              0.452
Adjusted R2                     0.452
Residual Std. Error      0.0002 (df = 7965)
F Statistic          6,582.101*** (df = 1; 7965)
===============================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

By examining the relative word frequencies used by ex-President Piñera and the words he avoided, we can potentially predict President Boric's word choices with up to 45% greater accuracy. This supports the theory that the rigid, institutionalized format of these official speeches constrains leaders from expressing too much overt political ideology. Instead, they tend to adopt contrasting tones and priorities based on their personal styles and intended audiences.

**6. Do one more big thing– either a sentiment analysis of the bags of words; rerun your analysis but using bigrams and/or trigrams; consider the role of negation words ("not," "no", etc.) on your earlier analysis; run a parts of speech tagger; look at the temporal unfolding of your words; or do a topic modelling exercise. For whichever thing you choose, explain what you are doing and whatever you find makes sense in some way theoretically.**

We will conduct a topic modeling analysis of President Boric's speech, incorporating bigrams to provide additional context to our topics. Our approach will utilize Latent Dirichlet Allocation (LDA), one of the most widely used topic modeling methods. In LDA, each document comprises multiple words, and each topic consists of various words associated with it. The objective of LDA is to discern the topics to which a document pertains, guided by the words it contains.

```
     Topic 1        Topic 2    Topic 3    Topic 4     Topic 5    Topic 6
[1,] "además"       "chile"    "país"     "chile"     "chile"    "chile"
[2,] "vamos"        "país"     "gobierno" "todas"     "país"     "gobierno"
[3,] "nacional"     "vamos"    "además"   "nacional"  "hoy"      "discurso"
```

```
 [4,] "quiero"       "discurso" "ley"      "seguridad" "año"      "vamos"
 [5,] "ello"         "año"      "personas" "mil"       "discurso" "año"
 [6,] "ustedes"      "vida"     "discurso" "hoy"       "todas"    "país"
 [7,] "compatriotas" "ley"      "años"     "cada"      "quiero"   "vida"
 [8,] "discurso"     "gobierno" "chile"    "proyecto"  "mundo"    "ley"
 [9,] "millones"     "acuerdo"  "vamos"    "trabajo"   "familias" "mil"
[10,] "sabemos"      "trabajo"  "política" "años"      "si"       "ser"
      Topic 7      Topic 8    Topic 9      Topic 10
 [1,] "año"        "chile"    "gobierno"   "año"
 [2,] "personas"   "país"     "hoy"        "ser"
 [3,] "ser"        "vida"     "ser"        "plan"
 [4,] "importante" "ustedes"  "vamos"      "chile"
 [5,] "política"   "plan"     "año"        "nacional"
 [6,] "pesos"      "trabajo"  "mil"        "personas"
 [7,] "ello"       "acuerdo"  "además"     "años"
 [8,] "quiero"     "recursos" "años"       "política"
 [9,] "ley"        "todas"    "trabajo"    "seguridad"
[10,] "millones"   "mundo"    "importante" "vez"
```

The emergence of a topic (Topic #1) related to social rights and the government's commitment to them aligns with President Boric's left-leaning, progressive agenda. Simultaneously, the presence of a security-focused topic (Topic #5) suggests that both leaders recognized the importance of addressing law and order, potentially reflecting concerns over social unrest or crime during their respective tenures.

The topic on national development and government structure (Topic #9) points to a shared priority – outlining a vision for the country's future and the role of the administration in achieving it. This overarching topic likely encompasses economic policies, institutional reforms, and long-term planning.

### 7. Extra credit: do some wordclouds of your texts

```
set.seed(1234)
#| warning: false
#| echo: false
# Generate word cloud for each speech

#Boric Wordcloud
suppressWarnings({wordcloud(df.t.t$names, df.t.t$boric_v1, min.freq = 20,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))})
```


```
#Piñera Wordcloud
#| warning: false
#| echo: false

suppressWarnings({wordcloud(df.t.t$names, df.t.t$pinera_V1, min.freq = 15,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))})
```


esfuerzo
mejorar
forma
mayor
niños
día
desarrollo
dos
sistema
crear
2018
ser
va
hacia
gobierno
mejor
hoy
país
vida
1
chile
salud
además
ley
t
calidad
educación
vamos
futuro
grandes
misión
sociedad
gran
nueva
junio
toda
proyecto
años
muchas