MinaR los discuRsos pResidenciales

Juan Pablo Ruiz Nicolini , Camila Higa , Lucas Enrich

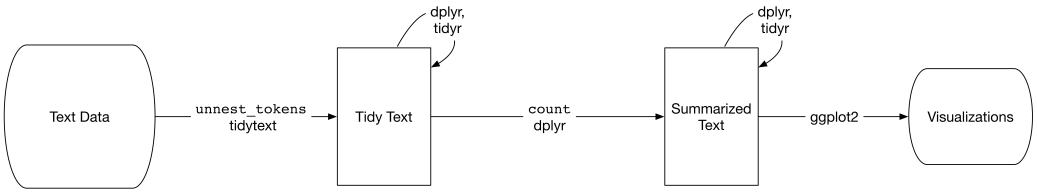
Palabras clave: discurso - text minning - política

Abstract

El primero de marzo de cada año las cámaras de Diputados y Senadores de la Nación Argentina se reúnen en asamblea para dar comienzo al año legislativo. Cada año el presidente de turno encabeza dicho acto con un discurso[1]. Éstos suelen girar en torno a los ejes de gobierno o promesas y objetivos del año. Es notorio que estos mensajes tiene un estilo y contenido marcado por quien ejerce el gobierno. En este trabajo partimos de una gran cantidad de texto contenido en cada uno de los discursos presidenciales desde el retorno de la democracia en 1983 para poder encontrar tanto patrones comunes como diferencias entre los mandatarios a lo largo del tiempo.

Hacer uso de la *minería de texto* de los discursos de los presidentes como estrategia de investigación es de utilidad para un rápido y eficiente análisis exploratorio del gran volúmen de información contenida en los mismos. Dentro del ecosistema de `R` este campo ha ido creciendo sostenidamente. Liberias como **tm** y **topicmodels** son herramientas poderosas para el procesamiento, manipulación y modelado de la información contenida en el texto. Siguiendo la filosofía de **tidyverse** Sigle y Robinson (2016) desarrollaron **tidytext**, que hace mucho más facil una primera introducción a esta técnica de investigación y su integración con otras como **ggplot2** para la visualización.

Un flujo de trabajo como el descripto más arriba puede ilustrarse siguiendo el esquema propuesto por Silge y Robinson (2020):



1. Descargamos los archivos con el texto de 37 discursos emitidos por 8 presidentes. Desde el primero de Alfonsín en la transición a la democracia (1984) hasta el último con el que Alberto Fernández dio inicio a las sesiones legislativas (2020). Entre todos suman alrededor de 365_{mil} palabras con un promedio de 9,880 y picos mínimo de 2,846 (Carlos Menem en 1996) y máximo de 26,189 (Cristina Fernández de Kirchner en 2013).

2. Con esa información construimos una única base de datos siguiendo el principo *datos de texto ordenados* (*tidy text*) propuesto por Sigle y Robinson (2016) como extensión de los *datos ordenados* (*tidy*) de Wickham (2014):

- Cada variable debe tener su propia columna.
- Cada observación debe tener su propia fila.
- Cada valor debe tener su propia celda.

Sigle y Robinson (2016) definen entonces a los *datos de texto ordenados* cuando están en una tabla compuesta por "un *token* por fila". *Un token es una unidad de texto significativa, como una palabra* (o un *bigrama*), *que estamos interesados en usar para el análisis, y la tokenización es el proceso de dividir el texto en tokens*[2].

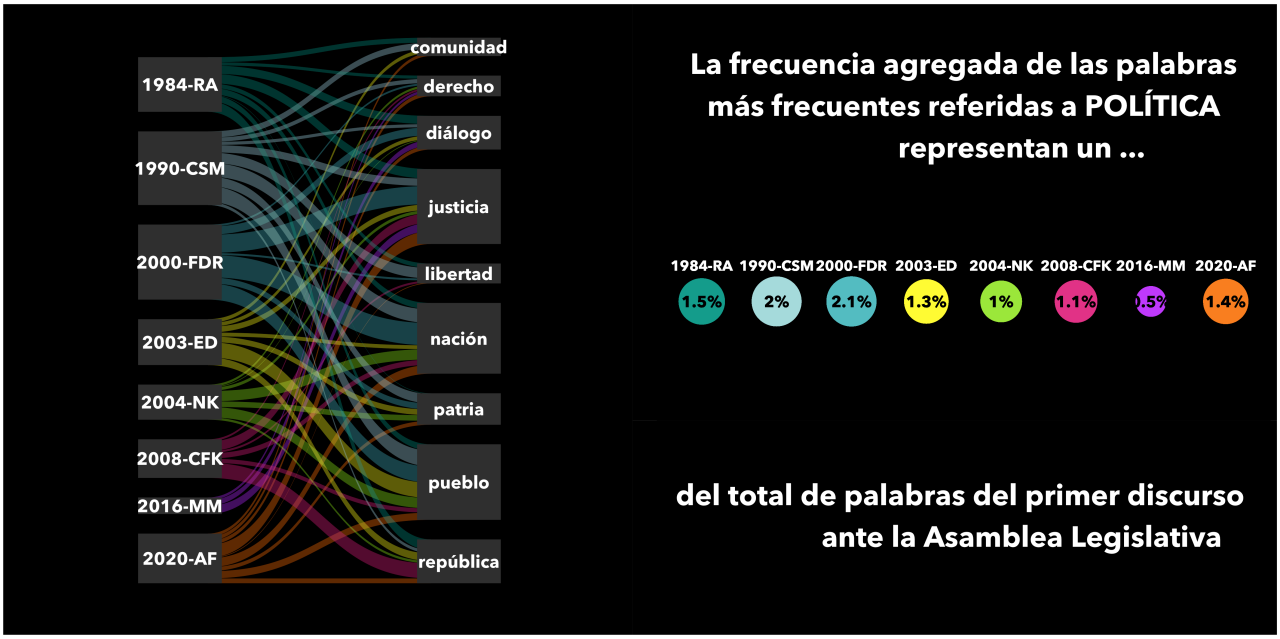
[1] La fuente original de todos los discursos puede consultarse en línea en https://www.hcdn.gob.ar/secparl/dgral_info_parlamentaria/dip/documentos/mensajes_presidenciales.html

[2] Traducción propia de *The tidy text format* (Silge and Robinson 2016).

3. Trabajamos con **dplyr** para calcular frecuencias de palabras y **tidytext** para calcular palabras de mayor importancia comparada entre discursos (*tf_idf*). Por último **ggplot2** (y extensiones como **patchwork** y **ggforce**) para las visualizaciones.

Ejemplo: primeros discursos presidenciales

Inspirados en un trabajo del equipo de datos de la Universidad de Berkeley (California)[3] calculamos la frecuencia con la que los presidentes utilizaron determinadas palabras relacionadas con un tópico específico, en este caso etiquetadas como *Política*.



Referencias

datascience@berkeley. 2019. "The Language of Data: Analyzing the State of the Union," January. https://datascience.berkeley.edu/blog/trump-state-of-the-union-analysis/.

Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *Journal of Open Source Software* 1 (3). The Open Journal: 37. doi:10.21105/joss.00037.

———. 2020. *Text Mining with R*. O'Reilly. https://www.tidytextmining.com/.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software, Articles* 59 (10): 1–23. doi:10.18637/jss.v059.i10.

Juan Pablo Ruiz Nicolini
Universidad Torcuato Di Tella
juan.ruiznicolini@mail.utdt.edu

Camila Higa
menta Comunicación
chiga1226@gmail.com

Lucas Enrich

[3] *The Language of Data: Analyzing the State of the Union* (datascience@berkeley 2019).

Universidad Nacional de La Matanza
lucas.a.enrich@gmail.com