MinaR los discuRsos pResidenciales

Juan Pablo Ruiz Nicolini , Camila Higa , Lucas Enrich

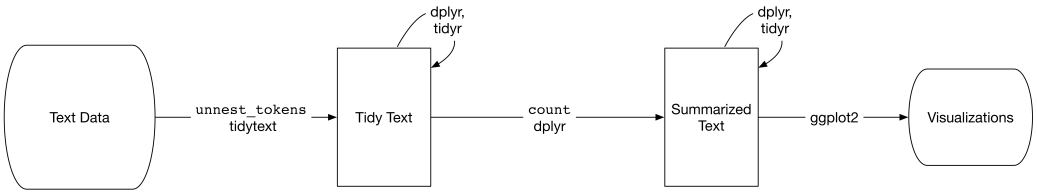
Palabras clave: discurso - text minning - política

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 primero en 1854 por Justo Jose De Urquiza hasta el último en 2020 por Alberto Fernandez

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. Se encuentran digitalizados los 114 discursos emitidos por los 31 presidentes que dieron lugar a la apertura de las sesiones legislativas. Debe mencionarse que no hay un discurso por año debido, principalmente, a las interrupciones institucionales, cuando el congreso no sesionó. Entre todos suman alrededor de 1,358,792 palabras con un promedio de 11,919 y picos mínimo de 99 (Miguel Juarez Celman 1890) y máximo de 43,135 (Ramon Castillo en 1942).

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].

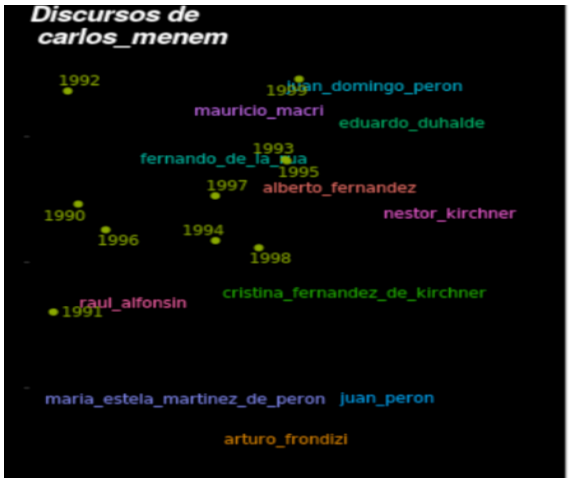
[1] La fuente original de todos los discursos puede consultarse en línea en https://www.hcdn.gob.ar/secparl/dgral_info_parlamentaria/dip/documentos/mensajes_presidenciales.html. Los mismos fueron posteriormente digitalizados mediante proceso de OCR y se encuentran disponibles para descarga desde `R` el paquete `{polAr}`(Ruiz Nicolini 2020).

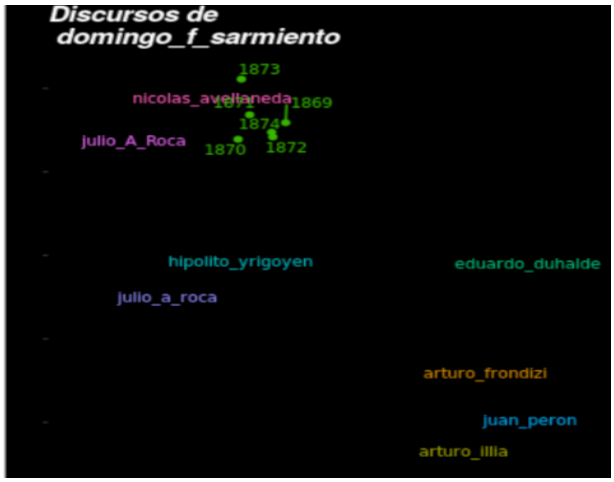
[2] Traducción propia de *The tidy text format* (Silge and Robinson 2016).

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

Ejemplo: Trayectoria en el discurso

Mediante la combinación de las técnicas de *TF-IDF* y *Principal Component Analysis (PCA)* es posible embeber númericamente los discurso mediante el uso de palabras ponderado por el largo del discurso y así visualizar los presidentes de mayor y menor variabilidad discursiva, en comparación con los promedios de los demás.





Las limitaciones de técnicas basadas en la frecuencia de las palabras independientemente del orden (como son *Bag of Words* y *TF-IDF*), estan vinculadas, por un lado, con el vocabulario y la semantica, por el otro. Así, para trabajar mejor con textos históricos (el más antiguo en este caso tiene 166 años), es recomendable usar tecnicas que hagan uso del contexto como puede ser *Doc2Vec*.

Referencias

Ruiz Nicolini, Juan Pablo. 2020. "PolAr: Argentina Political Analysis." https://github.com/electorArg/polAr.

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