speech: extracción, disponibilización y análisis de discursos parlamentarios en Uruguay

Elina Gómez , Nicolás Schmidt

Palabras clave: discursos parlamentarios - ocr - minería de texto

Introducción

speech es un paquete que permite convertir diarios de sesiones del parlamento uruguayo en formato PDF (URL o archivo local) a bases de datos ordenadas en la que cada fila es la intervención de cada uno/a de los/las legisladores/as que interviene en esa sesión. Se presentarán tres ejes de trabajo entorno al paquete y potencialidades.

Ejes de trabajo:

- Extracción:

El paquete desarrollado incluye un conjunto de funciones que permiten extraer el texto de las menciones parlamentarias incluidas en los diarios de sesión a partir de técnicas de OCR (reconocimiento óptico de caracteres), de forma ordenada (desagregada o agrupada) en formato data.frame y recuperando información anexa como el nombre de el/la legislador/a, sexo, número de legislatura, fecha, cámara a la que pertenece el documento. Asimismo, dado que los diarios de sesión muchas veces son imágenes escaneadas y que en el proceso de OCR se puede perder o dañar la información recuperada, el paquete provee otro conjunto de funciones que ayudan a mejorar estos problemas (*speech_check()* , *speech_legis_replace()*), así como indicadores que dan cuenta la calidad de la recuperación. Por su parte, a partir de la integración con otro paquete (*puy*) que contiene información sobre políticos/as uruguayos/as, es posible anexar integrar fácilmente información sobre el partido político al que pertenece.

- Disponibilización:

El paquete se encuentra en *CRAN* para su instalación y uso por parte de usuarios/as de R, sin embargo para ampliar la disponibilidad se ha desarrollado una aplicación Shiny que permite la descarga de bases de datos de menciones a partir de introducir URL y/o archivos locales que contengan diarios de sesiones en formato PDF. La misma permite descargar las menciones de forma agrupada o desagregada, y contiene todas las variables anexas para favorecer el análisis de los datos textuales.

- Análisis:

Por último, se encuentran en desarrollo varias líneas de análisis a partir de la información que se obtiene con el paquete *speech*. Por un lado, la clasificación automática de temas a partir de la aplicación de modelos y algoritmos de aprendizaje automático, que permiten un análisis agregado a partir de las variables que recupera el paquete. Por otro lado, se proyecta desarrollar, a partir de una base de datos histórica y acumulada de intervenciones parlamentarias construida a partir del paquete, una aplicación que permita el procesamiento y visualización interactivas mediante técnicas de minería de texto, e incluyendo filtros dinámicos según variables de agregación recuperadas.

Elina Gómez
UMAD (FCS) - UdelaR
elina.gomez@cienciassociales.edu.uy

Nicolás Schmidt
UMAD (FCS) - UdelaR
nicolas.schmidt@cienciassociales.edu.uy