# AMALIA: R, Shiny y minería de texto para el análisis masivo de archivos de la dictadura uruguaya

*Elina Gómez*

**Palabras clave:** minería de texto - shiny - memoria histórica

## Introducción

**AMALIA** es una aplicación que surge como una iniciativa en el marco del proyecto CRUZAR (*Sistema de Información de Archivos del Pasado Reciente*), y que a través de técnicas de minería de texto, busca aportar en el análisis masivo e interactivo de documentos de la dictadura uruguaya (1973-1985) que han sido digitalizados y convertidos a texto mediante técnicas de OCR. La misma ha sido desarrollada utilizando el lenguaje R y se encuentra en su versión de prueba. Se nutre de más de 100000 imágenes que conforman el denominado *Archivo Berruti*, y permite realizar búsqueda de términos y palabras, analizar las inter-conexiones entre los mismos, así como el contexto en que son mencionadas en los archivos.

## Estructura general:

- Buscador:

El buscador parte de un listado de palabras validadas previamente a partir de su inclusión en los diccionarios pre-definidos y permite evaluar tanto la frecuencia de aparición de un término como la co-ocurrencia entre diferentes palabras en las unidades de agregación. Así también plantea la posibilidad de analizar su contexto de mención en el texto bruto.

- Explorador:

El explorador permite realizar un análisis partiendo de lo general a lo particular ya que es posible seleccionar un sub-conjunto de documentos y explorar las temáticas que incluye a partir de las frecuencias de términos, nubes de palabras, co-ocurrencia, redes de palabras y asociaciones. También es posible dirigir el análisis seleccionando los diccionarios de interés.

- Analizador:

El analizador plantea un análisis centrado en el contexto en que se menciona una determinada palabra o conjunto de palabras previamente validadas por los diccionarios, a partir de la frecuencia, redes de términos y asociación entre las palabras que forman parte de dicho contexto.

## Paquetes utilizados:

- *shiny* , *shinythemes*, *shinyWidgets*, *shinycssloaders* , *wordclouds2* , *DT* para diseñar la estructura de la aplicación, formato, visualizaciones.
- *dbplyr* para hacer conexión con la base en Postgres y optimizar las búsquedas SQL.
- *quanteda* ( *quanteda.textmodels*, *quanteda.textplots*, *quanteda.textstats*) para visualizacione y cálculos de co-ocurrencias y distancias entre términos.

- Otros: *dplyr* , *ggplot2*, *seededlda*

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