Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

El futuro es de todos

DNP
Departamento
Nacional de Planeación

HERRAMIENTA PARA EL ANÁLISIS DESCRIPTIVO DE PQRSD

Entidad
Departamento Nacional de Planeación
- Dirección de Desarrollo Digital.
- Programa Nacional de Servicio al Ciudadano

Unidad para la Atención y Reparación Integral a las Víctimas de Colombia

Sector	**Lenguaje**	**Fuente de datos**
Planeación	R	ORFEO - Base de datos de PQRSD del Centro de Servicio al Ciudadano del DNP

Presentación
El análisis de grandes cantidades de texto puede ser facilitado hoy en día con herramientas computacionales. Dada la gran cantidad de PQRSD que puede recibir una entidad gubernamental, se creó una herramienta que ayuda con su análisis descriptivo. A esta herramienta se sube una lista de PQRSD y la procesa para visualizar tablas y gráficos de frecuencias de palabras, nubes de palabras, redes de coocurrencia y gráficos de *clusters*. De esta manera, un usuario que usa la herramienta puede observar similitudes y patrones dentro de estas PQRSD.

The analysis of large quantities of written text can be nowadays facilitated with software tools. Given the amount of PQRSD (questions, complaints, suggestions), a program was created in order to help with the PQRSD descriptive analysis. This application takes a list of PQRSD as input and processes it to visualize word and bigram frequency tables and graphs, word clouds, co-occurrence networks and cluster graphs. This way a user can use the application to find similarities and patterns in the list of PQRSD.

Objetivo general
Ayudar con el análisis descriptivo de grandes cantidades de PQRSD que recibe la Unidad para las Víctimas.

Objetivos específicos
1. Preparar el texto de PQRSD para que se pueda ser analizado con estadísticas descriptivas.
2. Crear tablas y gráficos descriptivos que analicen el texto en distintos *clusters* (agrupaciones), los cuales se calculan con una medida de similitud de un texto con otro.
3. Crear una herramienta o aplicación interactiva para que el usuario pueda visualizar las tablas y gráficos calculados en distintos *clusters* con facilidad

Metodología
La metodología se separa en 8 partes: (1) el preprocesamiento del texto; (2) la remoción de stopwords; (3) Reducción de palabras a su raíz o *stemming*; (4) *bag of words*; (5) reducción T-SNE; (6) *clustering*; (7) visualización descriptiva de las PQRSD y (8) creación de la herramienta interactiva.

1. *Preprocesamiento del texto.*

Consiste en pasar el texto de PQRSD ingresado a minúsculas, remover signos de puntuación, números y tildes.

2. *Remoción de stopwords*

Los *stopwords* son palabras que no aportan al análisis descriptivo del texto, por lo que se eliminan del texto ingresado. Estas palabras son usualmente preposiciones, conectores y nombres propios. Por ejemplo, se eliminan palabras como “de”, “a”, “la”, etc.

3. Reducción de palabras a su raíz o *Stemming*

Stemming es la reducción de una palabra a su raíz. Por ejemplo, se reducen distintas conjugaciones del verbo “jugar” a “jug”.

4. *Vectorización del texto con el método bag of words*

El *bag of words* (bolsa de palabras) es una manera de representar matemáticamente los textos. Consiste en crear una matriz de frecuencias donde las filas representan las PQRSD y las columnas son las palabras luego de haber hecho los filtros los pasos anteriores de la metodología. Los valores de la matriz son la frecuencia de las palabras dentro de cada PQRSD.

5. *Reducción T-SNE*

T-SNE es una metodología de reducción de dimensionalidad no lineal. Esta metodología toma la matriz de *bag of words* creada en el paso anterior y crea una representación de la distancia de cada PQRSD en un plano de, en este caso, dos dimensiones.

6. *Clustering*

A partir de las distancias de PQRSD calculadas en el paso (5) se crean hasta 5 *clusters* con la metodología K-medias. De esta manera se crea la posibilidad de agrupar las PQRSD hasta en 5 grupos distintos.

7. *Visualización descriptiva de las PQRSD*

Para cada uno de los *clusters* calculados en el paso (6) se presentan distintas tablas y gráficos descriptivos de las PQRSD. Se muestran tablas y gráficos de frecuencias, nubes de palabras redes de coocurrencia de términos principales y gráficos de *clusters* en un espacio bidimensional T-SNE. Los análisis de frecuencias se hacen tanto para las palabras de cada *cluster* como para los bigramas (secuencia de dos palabras).

8. *Creación de la herramienta interactiva*

La herramienta se ejecuta con el paquete Shiny a partir de R o RStudio y permite al usuario ingresar una lista de PQRSD y visualizar las estadísticas descriptivas por *cluster*. También le permite al usuario ingresar *stopwords* o palabras que considere que no aportan al análisis descriptivo (y que no fueron eliminadas durante el preprocesamiento). Por último, permite descargar la base original con las PQRSD y sus respectivos *clusters*.

Resultados

La aplicación permite visualizar las PQRSD ingresadas, gráficos y tablas de frecuencias de palabras y bigramas, nubes de palabras, redes de coocurrencia y gráficos de clusters. El usuario puede escoger dividir las PQRSD en un máximo de 5 clusters, así como subir stopwords adicionales y descargar las PQRSD originales con sus respectivos clusters.

A continuación se muestran tres ejemplos de gráficas una vez ingresada la lista de PQRSD que se quiere analizar. En la imagen 1 se presentan los gráficos de frecuencia de palabras y bigramas y en la imagen 2 se ilustra la red de términos principales.

Imagen 1. Gráficos de frecuencias de palabras y bigramas.

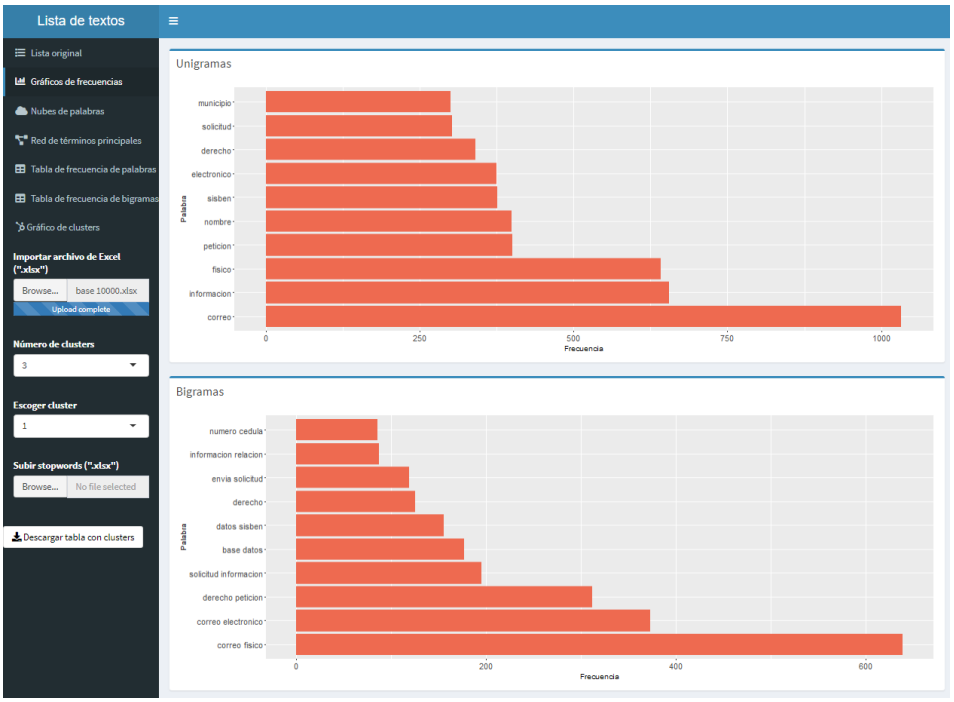
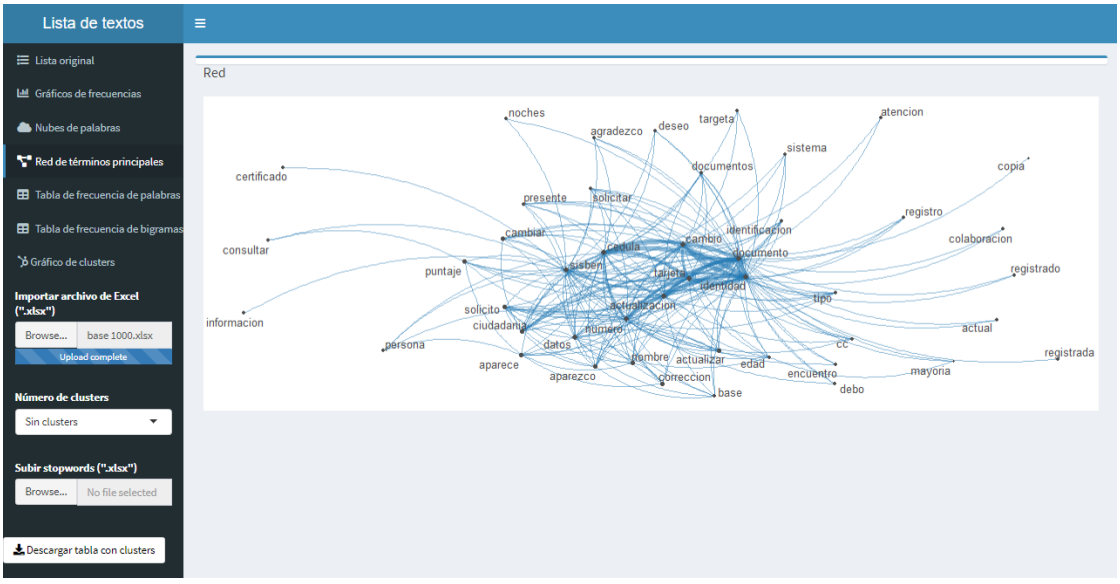


Imagen 2.Red de términos principales



En la imagen 3 se muestra el gráfico de *clusters* con 5 *clusters*.

Imagen 3. Gráfico de clusters

Conclusiones

1. La herramienta creada funciona para hacer análisis descriptivo de grandes cantidades de PQRSD
2. El aplicativo permite visualizar tablas y gráficos de frecuencias, redes de palabras, nubes de palabras y gráficos de *clusters*.

Socialización

El proyecto se presentó a la Unidad para las Víctimas y al Programa Nacional de Servicio al Ciudadano del DNP. La aplicación se encuentra en el servidor vdatascience de la UCD.