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

ANÁLISIS Y CLASIFICACIÓN DE LAS PETICIONES, QUEJAS, RECLAMOS, SUGERENCIAS Y DENUNCIAS (PQRSD) RADICADAS EN EL DNP

Entidad

Departamento Nacional de Planeación

- Dirección de Desarrollo Digital.
- Programa Nacional de Servicio al Ciudadano (PNSC).

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 las peticiones, quejas, reclamos, sugerencias y denuncias (PQRSD) es una tarea de gran valor para las entidades del gobierno, ya que estas constituyen una comunicación con numerosos actores, entre los cuales se encuentran los ciudadanos. Por ley, todas las entidades públicas deben responder a las PQRSD en determinados tiempos, los cuales dependen de la clasificación que esta adquiera al momento de ser radicada por el solicitante (petición, consulta, queja, reclamo, etc.). Esto obliga a cada entidad a clasificar las PQRSD desde el momento en que ingresan a su sistema, actividad que actualmente se realiza de manera manual y que puede significar hasta un día de los tiempos de respuesta. En este trabajo, se brindan las bases para realizar un modelo de aprendizaje de máquina que permita clasificar las PQRSD de forma automática y mejorar los tiempos de respuesta del DNP. Para ello, se plantean de cinco enfoques de limpieza y preprocesamiento de texto, cuatro formas de vectorización y cinco modelos de predicción, permitiendo identificar las principales ventajas y desventajas que trae para el análisis cada uno de los enfoques y métodos utilizados. Los resultados presentan un bajo poder de predicción y muestran que es de gran importancia estandarizar el registro de las PQRSD en el sistema de información, o bien, segmentar el análisis en función del canal de atención por el cual se recibe la PQRSD para, eventualmente, mejorar la capacidad de predicción del modelo. De manera complementaria, se realiza un análisis no supervisado en el que se identifican categorías que pueden predecirse con mayor precisión en trabajos posteriores.

The analysis of petitions, complaints, claims, suggestions and denunciations (PQRSD) is a task of great value to government entities, as these represent communication with several actors, including citizens. By law, all public entities must respond to PQRSDs within certain times, depending on the classification it gets when delivered by the petitioner (petition, consultation, complaint, claim, etc.), what obliges each entity to classify PQRSDs from the moment they enter its system, an activity that is currently performed manually and which can mean up to one day of the response times. In this work, we provide the bases to carry out a machine learning model to classify the PQRSD automatically and to improve the response times of the DNP. For this purpose, we explore five approaches of text cleaning and preprocessing, four vectorization forms and five prediction models, allowing to identify the main advantages and difficulties that each one of the approaches and methods used brings for the analysis. The obtained results present a low prediction power and highlight the importance of standardizing the PQRSD registration in the information system, or segmenting the analysis according to the channel of attention through which the PQRSD is received to improve the prediction capacity of the model. Furthermore, an unsupervised analysis is carried out, suggesting categories that could be predicted with greater accuracy in future projects.

Objetivo general
Identificar patrones en las PQRSD que llegan al DNP para clasificarlas según su tipo de documento.

Objetivos específicos
1. Realizar la limpieza y preprocesamiento de los asuntos registrados para cada PQRSD.
2. Realizar un análisis descriptivo y exploratorio del contenido de los textos.
3. Representar los textos en un lenguaje matemático que represente apropiadamente la información contenida en cada uno.
4. Entrenar un modelo de aprendizaje de máquina que permita clasificar las PQRSD de forma automática.
5. Identificar el tipo de limpieza, vectorización y el modelo que permita obtener la mejor capacidad predictiva en la clasificación de las PQRSD

Metodología
La metodología utilizada para analizar las PQRSD puede dividirse en cinco grandes fases: (1) el preprocesamiento de los textos, (2) la separación de las PQRSD entre "SISBÉN" y "NO SISBÉN", (3) la representación vectorial (matemática) de los textos, (4) el desarrollo de un algoritmo de clasificación supervisado basado en texto y (5) la identificación de temas representativos de las PQRSD.

Preprocesamiento o limpieza del texto
Inicialmente, se realizó la lectura de los asuntos contenidos en la base de datos, obteniendo una cadena de texto por PQRSD. La limpieza de las cadenas de texto obtenidas consistió en la transformación del texto a minúsculas y en la remoción de números, signos de puntuación y demás caracteres que no fueran letras; también se removieron conectores, preposiciones y palabras que no agregan significado al texto, entre las cuales se incluyeron nombres, apellidos y zonas geográficas. Al texto resultante de este tratamiento se le denominó "texto limpio".

Separación de PQRSD de SISBÉN y NO SISBÉN
Considerando la alta frecuencia del término "SISBÉN" en las PQRSD, se decidió separarlas en dos grupos tomando como referencia el contenido de las mismas, pues algunas estaban relacionadas con temas del SISBÉN, mientras otras contenían temas concernientes al DNP directamente. Dicha separación fue realizada con base en algunas palabras clave como "SISBÉN", "punto de corte", "programas sociales", "ser pilo paga", "verificación puntaje", "modificación puntaje", "cambio puntaje", "régimen subsidiado" y "sistema salud", las cuales fueron elegidas y validadas por el Programa Nacional del Servicio al Ciudadano (PNSC).

Transformaciones alternativas del texto
Para contar una mayor gama de alternativas en el modelamiento, se consideraron realizaron otras transformaciones al texto limpio (dejando el mismo texto limpio como una alternativa). La primera transformación consistió en realizar el proceso conocido como *stemming*, el cual reduce cada palabra a su raíz (por ejemplo, "jugar" y "jugamos" se reducen a "jug") y permite cuantificar de manera más precisa la ocurrencia de un concepto en el texto; al resultado de este tratamiento se le denominó "texto *stemming*".

Otro enfoque consistió en realizar la corrección ortográfica del texto resultante del primer tratamiento ("texto limpio"), para lo cual se toma la palabra que se quiere corregir, se calcula la distancia de Levenshtein entre ella y un listado de palabras ordenado por probabilidad de ocurrencia, y se toma aquella más probable entre las que tienen la menor distancia. En caso de que todas las palabras tengan una distancia de Levenshtein mayor a 2, la palabra no se modifica. El listado de referencia para realizar la corrección está compuesto por las 10.000 formas (palabras) más frecuentes del español según la RAE, junto con palabras propias del problema como "SISBÉN", "RUV" y "Gesproy", a las cuales se les asignó la mayor probabilidad de ocurrencia. De este tratamiento se obtuvo la representación denominada "texto corregido".

Tomando como base el texto corregido, se efectuó también un proceso de lematización, que consiste en obtener el lema de cada palabra, siendo esta una forma inflexiva del término y permitiendo así la eliminación de formas verbales conjugadas, plurales, entre otras (por ejemplo, "jugar, jugamos y jugaríamos" se convierten en "jugar"). Al texto resultante se le denominó "texto lematizado". El último enfoque consistió en la realización de *stemming* (ver primer párrafo de esta sección) sobre el texto corregido, enfoque con el cual se obtuvo el denominado "texto corregido con *stemming*".

Representación vectorial (matemática) de los textos

Una vez realizada la separación de las PQRSD en las dos categorías, se determinó una representación matemática de los textos mediante cuatro métodos. Inicialmente se usó el modelo de *Bag of Words*, un método que permite representar textos indicando la frecuencia de sus palabras a través de una matriz con textos en sus filas y palabras en sus columnas, donde cada posición representa el número de veces que la palabra (de la columna) aparece en el texto (de la fila).

Como segunda opción, se hizo uso de la ponderación conocida como *Term Frequency - Inverse Document Frecuency,* que es una medida numérica que expresa cuán relevante es una palabra en un texto y que permite normalizar las frecuencias de la matriz de *Bag of Words* en función de la relevancia de cada palabra respecto al agregado de las PQRSD, obteniendo una nueva representación vectorial de cada texto.

La tercera técnica utilizada fue Doc2Vec, a partir de la cual se cuantifica la similitud entre términos mediante la observación de las palabras que se encuentran antes y después de cada una, y que se integra con la medida de la similitud coseno entre textos vectorizados para representar las palabras en un espacio n-dimensional, donde el número de dimensiones deseadas es un parámetro del algoritmo que define el usuario. El último método usado fue el *hashing*, a partir del cual se asigna una codificación única a cada texto y que permite obtener una representación de los textos similar (en cuanto a estructura) a la que se obtiene con Doc2Vec.

Cabe mencionar que la longitud de los vectores (es decir, el número de dimensiones) fue definida *a priori* con un valor de 100 dimensiones, tanto para Doc2Vec como para *hasging*, tomando en cuenta que esto facilitaría el procesamiento a nivel computacional y que 100 dimensiones suelen ser suficientes para representar una gran parte de la información contenida en los datos, especialmente teniendo textos no muy extensos como los de las PQRSD. Nótese también que estas conversiones fueron realizadas cuando ya se había hecho la separación entre las PQRSD que son sobre SISBÉN y las que no, de forma tal que no fuera el tema de SISBÉN el que causara la mayor separación de los puntos en el espacio al representar las PQRSD de manera vectorial.

Modelo de clasificación

En cuanto a la parte predictiva, se decidió en compañía del PNSC realizar un modelo de clasificación que permitiera identificar automáticamente el tipo de documento (Petición, Queja, Reclamo, Sugerencia, Denuncia de corrupción, Consulta, Petición entre autoridades o Solicitud de información). Este modelo se entrenó inicialmente con las PQRSD sobre SISBÉN, que se catalogaron como más importantes. Para ello, se dividió el conjunto de datos en un 80% de entrenamiento (para ajustar el modelo) y un 20% de prueba (para estimar y validar su capacidad de predicción). En este caso, se consideraron 5 modelos como posibles candidatos para realizar la clasificación: los K vecinos más cercanos (KNN, por sus siglas en inglés), regresión logística, redes neuronales, *Random Forest* y máquinas de soporte vectorial (SVM, por sus siglas en inglés). Así, se tuvieron al final una gran cantidad de alternativas (5 x 4 x 5 = 100) como se muestra en la figura 1. Inicialmente, la metodología consideraba la realización de los 100 modelos y la escogencia del mejor, sin embargo, fue necesario realizar cambios en la metodología a medida que se fueron realizando dichos modelos, como se explica en la sección de resultados.

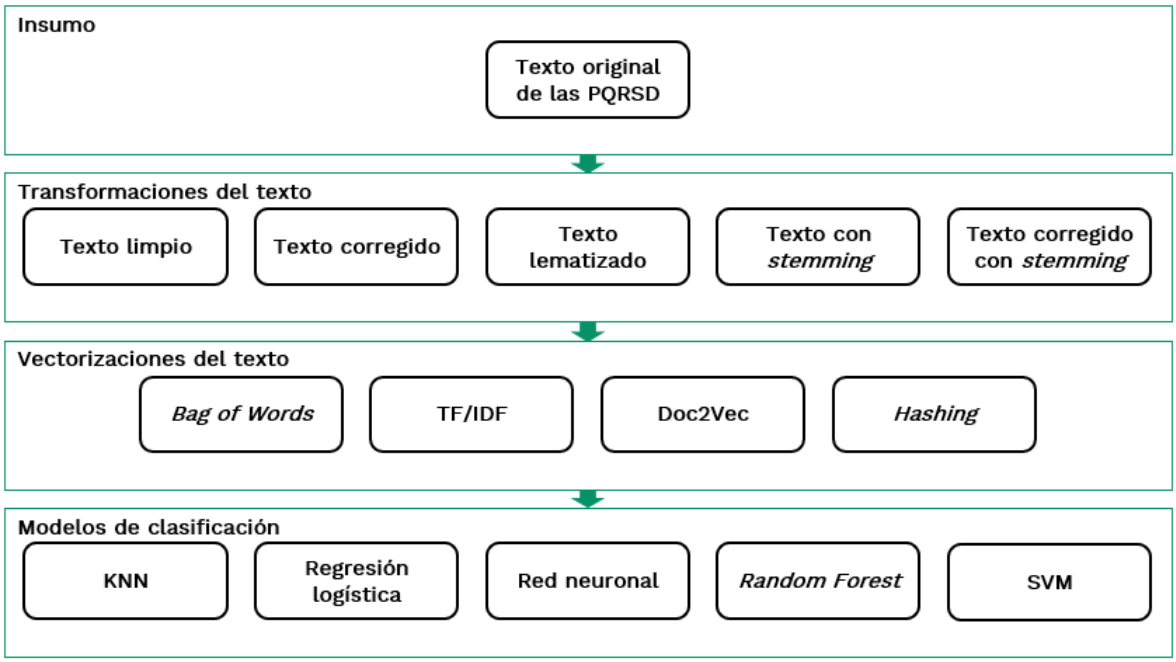

Figura 1: Alternativas para la realización de un modelo de clasificación por tipo de documento.

Una consideración adicional que se realizó antes de realizar el modelamiento fue que las clases de la categoría "tipo de documento" se encontraban desbalanceadas, es decir, la mayoría de las PQRSD correspondían a algunas pocas clases, mientras que las otras representaban un muy bajo porcentaje del total: de hecho, entre peticiones, consultas y solicitudes de información se recogen el 97,2% de las PQRSD, mientras que las otras cinco categorías representan el 2,8% restante. Este problema se muestra de manera gráfica en la figura 2. Como esto es un problema para la estimación de los parámetros en la mayoría de los modelos propuestos, se balancearon las clases mediante un proceso de muestreo, para reducir las categorías con frecuencias muy altas, y de sobremuestreo, para aumentar artificialmente el número de observaciones de aquellas con frecuencias pequeñas. Habiendo balanceado las clases, se pasó a realizar el ajuste de los modelos al conjunto de datos de entrenamiento.

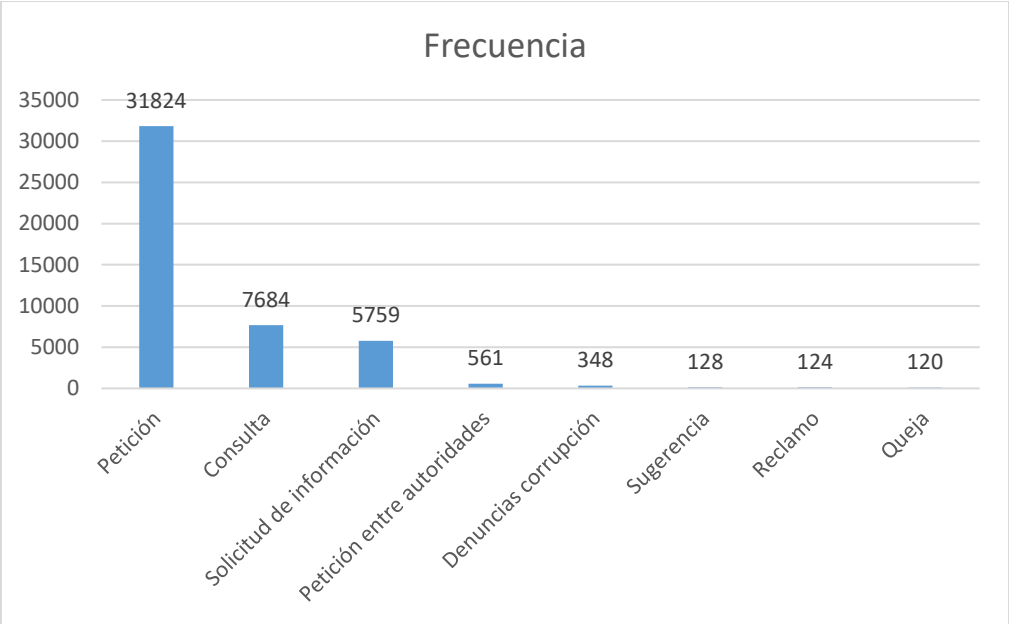

Figura 1: Frecuencias de PQRSD por tipo de documento. Peticiones, consultas y solicitudes de información representan el 97% de las PQRSD que se radican en el DNP.

Identificación de temas representativos de las PQRSD (análisis no supervisado)

De manera paralela al análisis supervisado, se realizó una identificación de grupos de PQRSD con temas similares. Esta clasificación se conoce como no supervisada porque no hay categorías predefinidas para clasificar, sino que se espera obtener unas categorías con significado a partir de los datos y de su estructura. La utilidad de estas técnicas radica en que brindan información desagregada sobre cuáles son las peticiones más frecuentes, además de que permiten etiquetar los datos con categorías que pueden tener una mayor capacidad de predicción.

Para la realización de este análisis, se utilizaron solamente las versiones de texto limpio y texto corregido con *stemming*, vectorizadas utilizando Doc2Vec, con 100 dimensiones. Sobre los datos transformados, se aplicó un algoritmo de *k-means,* que permite agrupar los datos en un número *k* de grupos o clústeres, utilizando valores de *k* de 2 a 10 y midiendo la suma de las varianzas internas de cada clúster (*within sum of squares*), lo cual suele tomarse como referencia para escoger el número de clústeres, sin embargo, el aumento en el número de clústeres no implicaba reducciones significativas en la suma de varianzas interna, por lo que se prefirió escoger los clústeres de manera gráfica, observando la separación de los puntos, coloreados por clúster, al proyectarlos sobre las dos o tres componentes principales. Para nominar los clústeres, se realizó el conteo de sus palabras más frecuentes, de sus bigramas más frecuentes y una prueba de representatividad estadística de palabras (*keyness*) con la función "*textstat_keyness*" del paquete "quanteda", comparando cada clúster contra el resto.

De manera alternativa para la clasificación no supervisada, se probó un *clustering* jerárquico, que arrojaba resultados muy poco favorables debido a las PQRS que eran muy distintas del resto, resultando clústeres de muy pocas observaciones. También se programó de manera empírica un algoritmo con el siguiente pseudocódigo:

```
Definir k = número de clústeres
Realizar k-1 veces:
   Definir C = el clúster con mayor número de PQRSD.
   Vectorizar las PQRSD del clúster C utilizando Doc2Vec.
   Crear 2 clústeres realizando k-means sobre las PQRSD vectorizadas.
   Crear una lista L que contenga las 10 palabras más frecuentes de cada
   clúster.
   Calcular la frecuencia relativa de aparición de cada palabra en cada
   clúster.
   Para cada palabra, calcular la diferencia entre las frecuencias relativas
   de aparición en cada clúster.
   Definir la palabra P = aquella de la lista L con mayor diferencia
   porcentual de aparición.
   Crear dos nuevos clústeres, dividiendo el clúster C entre los que tienen
   la palabra P y los que no.
```

Obtenidos los clústeres con este algoritmo, se realizó el mismo proceso de denominación que con la técnica de *k-means*, basada en palabras y bigramas frecuentes y en palabras representativas.

Aplicación interactiva

Finalmente, se decidió desarrollar una aplicación para que los usuarios pudieran visualizar los resultados del análisis realizado, incluyendo un módulo para clasificar PQRSD donde se implementó el modelo que arrojó mejores resultados y otro donde se presentan los clústeres identificados. Allí es posible realizar filtros por tipo de PQRSD, tiempo de respuesta, canal de recepción, "SISBÉN" o "NO SISBÉN" y clúster, así como es posible agregar nuevas *stopwords* al análisis y generar reportes descriptivos de frecuencias y nubes de palabras. Adicionalmente, la aplicación cuenta con un módulo para analizar grupos de PQRSD como redes de palabras, donde los nodos son palabras y sus conexiones se relacionan con el número de veces que una palabra aparece en un texto cuando la otra está presente.

Resultados

Modelo de clasificación

De los 100 posibles modelos propuestos en la metodología, se ajustaron 43 dada la baja capacidad de predicción que se observó durante la realización del análisis. Inicialmente, se ajustó un modelo de KNN para todas las posibles combinaciones de transformaciones y vectorizaciones del texto y se evaluó cada modelo midiendo en los datos de prueba el error de clasificación, definido como el número total de PQRSD clasificadas correctamente dividido entre el total de PQRSD. Dado que todos los modelos tenían una baja capacidad de predicción, se inspeccionaron las matrices de confusión, pues generalmente la medida de error no es suficiente para evaluar un modelo de clasificación, especialmente cuando se tienen muchas clases (8 clases, en este caso). Observando las matrices de confusión, se encontró también un desempeño muy bajo de todos los modelos, con altas tasas de falsos positivos y negativos en la predicción de todas las categorías (o todas menos dos, en el mejor de los casos). Por este bajo desempeño, se consideró suficiente reportar el error de clasificación, ya que sintetiza la información de interés. El proceso de entrenamiento y prueba se repitió con la regresión logística para las 20 combinaciones. Como se siguió observando un mal desempeño, en los modelos restantes solo se utilizó el texto lematizado con vectorización Doc2Vec, que fue el que mostró mejores resultados entre los modelos entrenados. Sin embargo, la capacidad de predicción de los modelos restantes (Random Forest, redes neuronales y SVM) entrenados con esta versión del texto fue también muy baja, como se evidencia en la tabla 1.

Modelo	Vectorización	Tipo de texto	Precisión
KNN	BoW	Limpio	39,10%
		Stemming	40,52%
		Corregido	40,87%
		Lematizado	42,40%
		Corregido con stemming	41,32%
	TF-IDF	Limpio	39,15%
		Stemming	40,57%
		Corregido	40,89%
		Lematizado	42,50%
		Corregido con stemming	41,39%
	Hashing	Limpio	38,12%
		Stemming	39,43%
		Corregido	39,81%
		Lematizado	41,47%
		Corregido con stemming	40,31%
	Doc2Vec	Limpio	40,27%
		Stemming	41,38%
		Corregido	41,68%
		Lematizado	43,56%
		Corregido con stemming	42,47%
Regresión Logística	BoW	Limpio	40,29%
		Stemming	41,78%
		Corregido	41,68%
		Lematizado	43,59%
		Corregido con stemming	42,57%
	TF-IDF	Limpio	40,23%
		Stemming	41,68%
		Corregido	41,91%
		Lematizado	43,63%
		Corregido con stemming	42,52%
	Hashing	Limpio	39,23%
		Stemming	40,53%
		Corregido	40,89%
		Lematizado	42,61%
		Corregido con stemming	41,68%
	Doc2Vec	Limpio	41,45%
		Stemming	42,38%
		Corregido	42,78%
		Lematizado	44,67%
		Corregido con stemming	43,58%
Random Forest	Doc2Vec	Lematizado	45,89%
Redes Neuronales	Doc2Vec	Lematizado	43,56%
SVM (kernel radial, c=1)	Doc2Vec	Lematizado	26,89%

Tabla 1: Desempeño de los modelos realizados para predecir el tipo de documento.

Dadas las dificultades presentadas en el modelamiento, se consultó al PNSC sobre la metodología con la cual se obtuvieron las etiquetas de tipo de documento, lo cual motivó una reunión con el Centro de Servicio al Ciudadano (CSC). En dicha reunión, el CSC explicó que las etiquetas de tipo de documento asignadas a las PQRSD no eran acertadas e indicó, además, que la clasificación que se asigna a cada PQRSD puede variar a lo largo del proceso de respuesta, por lo que no era conveniente automatizar su clasificación.

Identificación de temas representativos de las PQRSD (análisis no supervisado)
Los clústeres obtenidos para las PQRSD de NO SISBÉN y las de SISBÉN mediante la aplicación de *k-means* se ilustran en las figuras 2 y 3. Las palabras que se usaron para nominar los clústeres se pueden consultar en el módulo "palabras clave por clúster" de la aplicación desarrollada. Los resultados del algoritmo alternativo de filtros por palabras representativas no se incluyen en la aplicación, siendo que los encontrados con *k-means* se consideraron mejores para sintetizar el contenido de las PQRSD. Esto se debió, principalmente, a que el algoritmo alternativo de *clustering* resultó agrupando en función de palabras que son propias de los canales de recepción, como se observa en la figura 4.

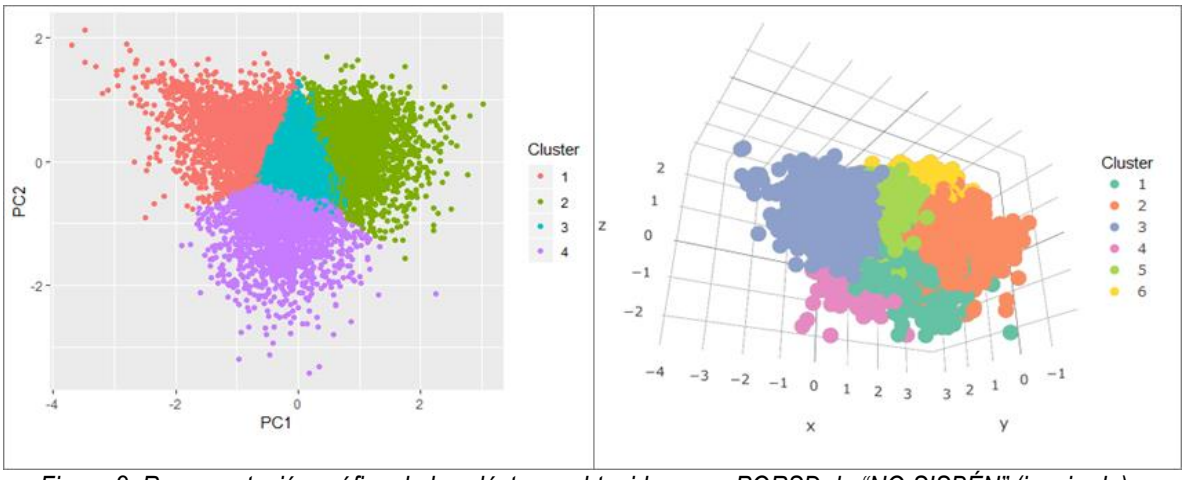

Figura 2: Representación gráfica de los clústeres obtenidos para PQRSD de "NO SISBÉN" (izquierda) y "SISBÉN" (derecha).

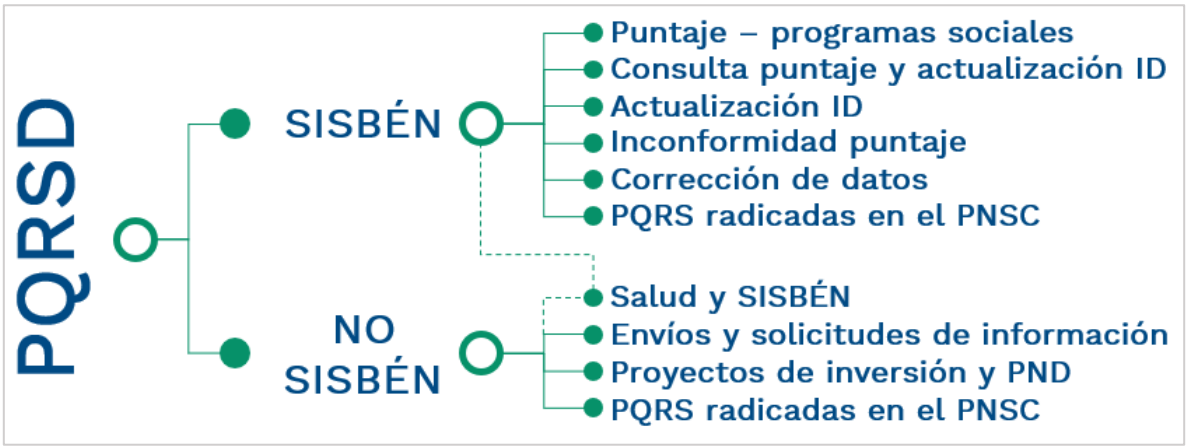

Figura 3: Nominación asignada a los clústeres a partir de sus palabras representativas y de sus monogramas y bigramas más frecuentes.

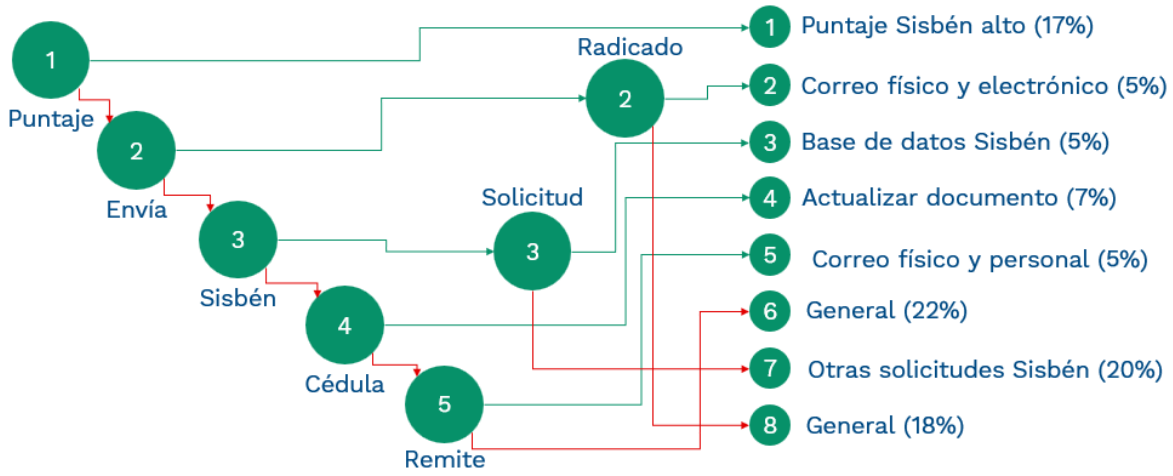

Figura 4: Resultados de la agrupación de PQRSD mediante el algoritmo de identificación y filtrado de palabras representativas. Las palabras junto a los círculos grandes corresponden a aquellas identificadas en cada iteración, las flechas verdes indican el subgrupo que contiene la palabra y las rojas el que no.

Aplicación interactiva

En las figuras 5 y 6 se muestran dos capturas de pantalla de la aplicación que se desarrolló para visualizar los resultados del proyecto y para facilitar la realización de análisis descriptivos.

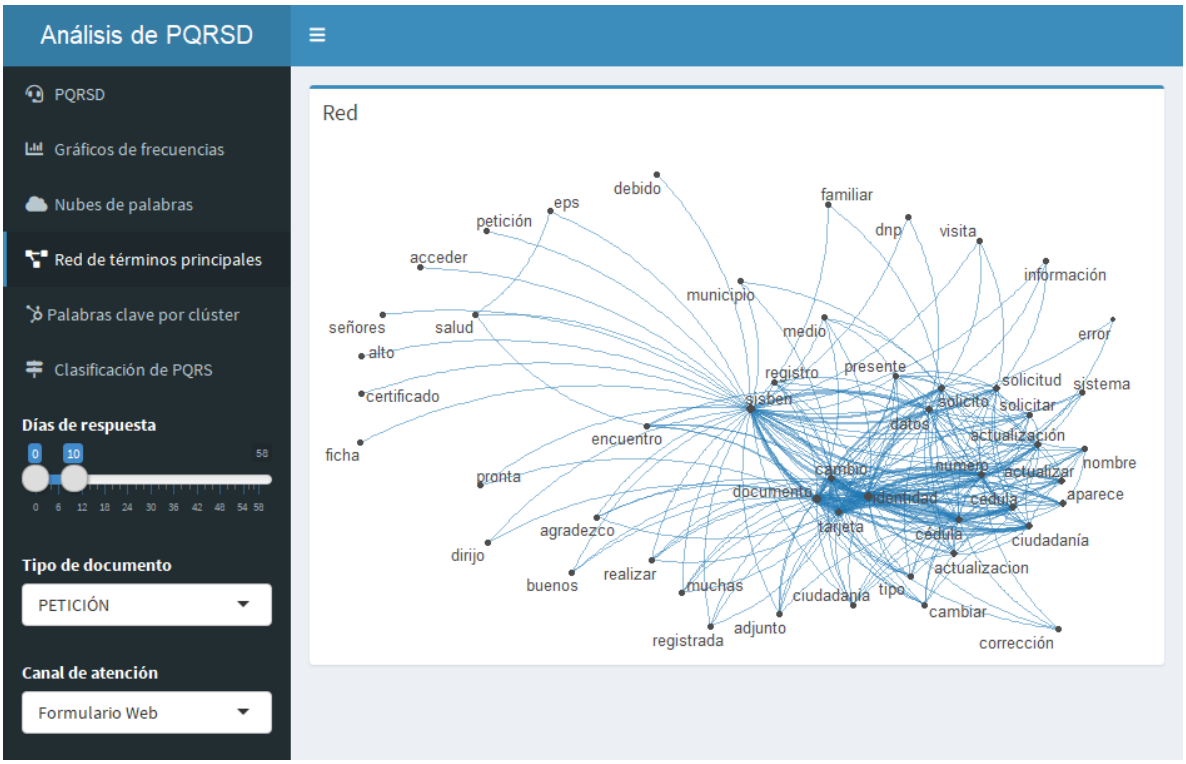

Figura 5: Interfaz de la aplicación desarrollada en el módulo de "Red de términos principales"

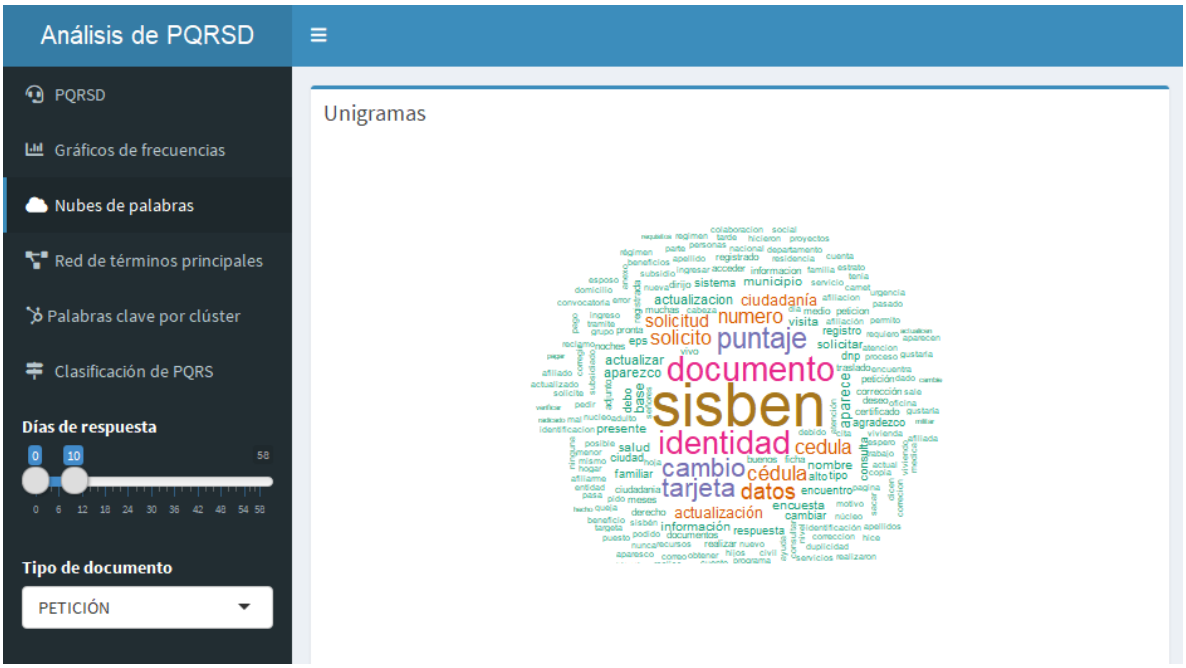

Figura 6: Interfaz de la aplicación desarrollada en el módulo de "Nubes de palabras"

Conclusiones

1. Automatizar la clasificación de PQRSD por tipo de documento no es viable dado que las etiquetas que se les asignan en ORFEO no corresponden con la clasificación que deberían tener en realidad.
2. No es necesario ni oportuno automatizar la clasificación de PQRSD por tipo de documento, pues esto no apoya la resolución de ningún problema existente en el DNP.
3. Si se deseara realizar análisis similares o modelos de clasificación en un futuro, sería importante estandarizar el registro de las PQRSD en el sistema de información. De forma alternativa, podría segmentarse el análisis por canal de recepción, ya que existen términos que son propios del modo en que se registran las PQRSD cuando se radican por uno u otro medio.
4. Los clústeres encontrados permiten identificar temas sobre los cuales llegan numerosas solicitudes, como inconformidad con el puntaje de SISBÉN o actualización de documento en la base de datos del SISBÉN. En torno a estos temas, podrían explorarse otras soluciones tecnológicas que faciliten la atención de PQRSD en el DNP.

Socialización

Los resultados del proyecto se socializaron con el Programa Nacional de Servicio al Ciudadano (PNSC) y con el Centro de Servicio al Ciudadano (CSC) del DNP.