Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

CLASIFICACIÓN DE PROYECTOS DE INVERSIÓN EN PROYECTOS ENFOCADOS A VÍCTIMAS DEL CONFLICTO ARMADO

DNP (Dirección de Desarrollo Digital, Grupo de Proyectos Especiales-GPE).

Presentación

En el marco de la Ley de Víctimas y Restitución de Tierras (Ley 1448 de 2011) y en concordancia con los CONPES 3726 y 3712 se propone crear un clasificador basado en análisis de similitud de textos entre 6141 documentos de inversión provenientes del Presupuesto General de la Nación (PGN) y los CONPES anteriormente nombrados, junto con la Guía para el uso de los clasificadores de la "Política pública de protección, prevención, atención, asistencia y reparación para todas las víctimas del conflicto armado". El proyecto clasifica los documentos de inversión que fueron utilizados para realización de esta ley como aquellos dirigidos a la atención de las víctimas del conflicto armado en general.

Objetivo general

Analizar la inversión del Gobierno Colombiano para el cumplimiento de la Ley de Víctimas y Restitución de Tierras (Ley 1448 de 2011), en función de los CONPES 3726 y 3712.

Objetivos específicos

- Crear una herramienta para clasificar proyectos de inversión en el tema de atención a víctimas.
- Dimensionar el gasto monetario en que se incurre para la implementación de la ley 1448.
- Generar reportes automáticos de la clasificación de proyectos de inversión resultante del análisis de texto.

Metodología

La metodología utilizada para identificar proyectos de inversión relacionados con el cumplimiento de la Ley de Víctimas y Restitución de Tierras puede dividirse en dos grandes fases: (1) la identificación de las palabras alusivas al tema de víctimas y (2) el desarrollo de un algoritmo de clasificación basado en texto. Para la primera fase, se consideraron los documentos CONPES 3726 y 3712 y la Guía para el uso de los clasificadores de la Política pública de protección, prevención, atención, asistencia y reparación para todas las víctimas del conflicto armado, además de 36 documentos CONPES de 12 sectores de la economía. Estos últimos se escogieron aleatoriamente y permitieron identificar palabras que no son representativas del tema de víctimas al encontrarse de forma recurrente en documentos CONPES. La tabla 1 presenta los sectores y CONPES escogidos.

Sector	CONPES escogidos			Sector	CONPES escogidos		
Transporte	3916	3900	3857	Inclusión social y reconciliación	3867	3850	3784
Cultura, deporte y recreación	3812	3803	3783	Ambiente y desarrollo sostenible	3716	3700	3697
Educación	3914	3831	3809	Salud y protección social	3887	3755	3622
Vivienda	3897	3859	3848	Minas y energía	3873	3510	3347
Agua potable, saneamiento básico	3798	3780	3715	Telecomunicaciones	3898	3854	3769
Agricultura (agropecuario)	3811	3763	3675	Comercio, industria y turismo	3866	3771	3709

Tabla 1. Documentos CONPES escogidos por sector económico.

En primer lugar, se realizó la lectura de los documentos, obteniendo una cadena de texto por página. 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 distintos a las letras que conforman las palabras; también se removieron conectores, preposiciones y palabras que no agregan significado al texto, entre las cuales se incluyeron zonas geográficas; igualmente, se eliminaron tildes, ya que no son utilizadas en las descripciones

de un gran número de proyectos y se prefirió contar con textos limpios homogéneos para documentos y proyectos. Esta limpieza se complementó con una igualación de palabras similares en significado (e.g. “reparar”, “reparado” y “reparación”, que se convierten en “reparación”), mediante una transformación basada en el algoritmo de lematización de Porter. Para ejemplificar este procedimiento, puede pensarse en una cadena de texto como “4. Asistencia y reparación para todas las víctimas” que quedaría convertida en “asistencia reparacion victima".

Finalizado este proceso, se realizó una vectorización de los textos utilizando el modelo de bolsa de palabras (*bag of words*), que consiste en construir una matriz con palabras en las columnas y páginas en las filas para representar el número de veces que aparece cada palabra en cada página. Por ejemplo, si la página 12 contuviera solamente “asistencia reparacion victima ", las columnas “asistencia”, “reparacion” y “victima” tendrán un valor de 1 en la fila 12 de la matriz, mientras las demás columnas tendrían un valor de 0. Utilizando esta matriz, se identificaron palabras que aparecían solamente en los documentos de víctimas y no en los demás documentos, es decir, las palabras exclusivas del tema víctimas. De forma similar, las palabras representativas (mas no exclusivas) del tema de víctimas se identificaron a partir de las frecuencias promedio de cada palabra en los documentos de víctimas y en los demás documentos. Para estos valores, se calcularon la diferencia y la razón entre los dos valores asociados a cada palabra, obteniendo 2 indicadores sobre qué tan representativa es una palabra para el tema de víctimas. Un tercer indicador se construyó con una técnica basada en la proyección de los documentos vectorizados sobre las componentes principales calculadas a partir de la matriz. A partir de estos tres indicadores, se dio un puntaje único a cada palabra a partir del cual se ordenaron todas las palabras de la más a la menos representativa.

Para la segunda fase (clasificación), se realizó el mismo proceso de limpieza y lematización para los textos de los proyectos y se definieron dos sistemas de puntaje basados (1) en la retroalimentación de las palabras con el experto en el tema y (2) en el puntaje asignado a las palabras identificadas. Estos se muestran en la tabla 2.

Sistema de puntuación 1 (por grupos)		**Sistema de puntuación 2 (por ranking)**
Palabras exclusivas:		Palabras representativas:
Etiquetadas como “esenciales”:	15 puntos	Todas: Valor del puntaje calculado a partir de los indicadores de similitud.
Etiquetadas como “buenas”:	5 puntos	
Palabras representativas:		Palabras exclusivas:
Etiquetadas como “esenciales”:	10 puntos	Todas: Puntaje promedio de las primeras 50 palabras.
Etiquetadas como “buenas”:	3 puntos	

Tabla 2: Sistemas de puntuación definidos para los textos de los proyectos.

A partir de este sistema, se dio una puntuación a cada proyecto en función de los puntajes asociados a las palabras contenidas en sus textos asociados. Por ejemplo, si las palabras “asistencia”, “reparacion” y “victima” asignan 15, 10 y 3 puntos respectivamente, el texto “asistencia victima reparacion victima” tendría 28 puntos. Nótese que los puntajes se asignan si la palabra está presente, sin considerar cuántas veces, pues en caso contrario podría tenerse un texto con muchos puntos solo por ser muy extenso. Finalmente, se desarrolló una aplicación para facilitar la definición de un umbral de puntos a partir del cual los proyectos se consideran de víctimas, en la cual pueden visualizarse los proyectos clasificados en función del umbral escogido por el usuario y generarse reportes automáticamente, sea para todos los proyectos o solo para los del PGN, del SGR o de entidades territoriales. De todas formas, para establecer una propuesta de umbral, se analizó la distribución los puntajes asignados a cada uno de los proyectos de inversión, con el fin de visualizar un eventual punto de corte para separar los proyectos con altos puntajes de la aglomeración de proyectos con bajos puntajes de similitud.

Resultados

En el proceso de extracción de las palabras más sugestivas del tema de víctimas se obtuvo una lista de 203 palabras exclusivas y una lista de 8857 palabras alusivas ordenadas según su puntaje obtenido a partir de los 3 indicadores. En la tabla 3 se muestra el top 10 de las palabras exclusivas con mayor frecuencia, junto con el top 10 de las palabras representativas.

Palabras exclusivas, ordenadas por frecuencia

Palabra	Frecuencia	Palabra	Frecuencia
smlmv [1]	13	pvcai [4]	7
fondelibertad	11	aht [5]	6
sfv [2]	9	perpetrada	6
ged [3]	7	sgsss [6]	6
oxfam	7	asesinados	4

Palabras alusivas con mayor puntaje

Palabra	Palabra
indemnizacion	clasificadores
humanitaria	repeticion
costeo	victimas
victimizantes	ruta
satisfaccion	reubicaciones

1.Salario mínimo legal mensual vigente
2. Subsidio familiar de vivienda
3. Goce efectivo de derechos
4. Población víctima del conflicto armado interno
5. Ayuda humanitaria de transición
6. Sistema de seguridad social en salud

Tabla 3: Principales palabras exclusivas y alusivas al tema de víctimas, identificadas automáticamente.

Si bien la aplicación desarrollada permite escoger entre los sistemas de puntuación explicados en la tabla 2, los resultados presentados consideran únicamente el sistema 1, que permitió obtener mejores resultados al mitigar el efecto de clasificación errada de textos extensos. Igualmente, los resultados que se presentan se obtienen en función del umbral propuesto, por lo que pueden variar tras la retroalimentación final con el GPE. Así pues, la definición del umbral propuesto se deriva de la observación de la figura 1, donde la aglomeración de proyectos se observa para puntajes menores de 30, llevando a escoger este valor como el umbral propuesto, que se presenta por defecto en la aplicación. El clasificador, con este umbral, agrupa 730 proyectos de inversión dentro de los proyectos enfocados a la atención de víctimas, entre los cuales 363 provienen del PGN, 49 del SGR y 318 de entidades territoriales. En la tabla 4, se presenta el título de los 10 proyectos con financiación del PGN clasificados con mayor puntaje. Finalmente, en la figura 2 se muestra la interfaz gráfica de la aplicación.

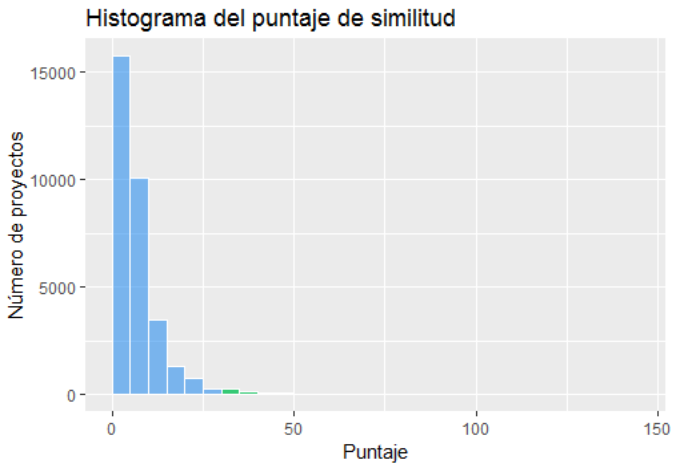

Figura 2: Puntaje de similitud obtenido para los proyectos clasificados (verde) y no clasificados (azul)

N°	Puntaje	Nombre del proyecto
1	145	Asistencia y atención integral a víctimas a nivel nacional.
2	143	Mejoramiento de los canales de atención y comunicación para las víctimas para facilitar su acceso a la oferta institucional.
3	133	Implementación del plan estratégico de tecnología de información para asistencia, atención y reparación integral a las víctimas a nivel nacional.
4	119	Implementación de las medidas de reparación colectiva a nivel nacional.
5	118	Implementación plan estratégico de tecnología de información para la asistencia, atención y reparación integral a las víctimas a nivel nacional.
6	118	Implementación del sistema de alertas tempranas para la prevención de las violaciones masivas de derechos humanos en Colombia.
7	100	Consolidación del sistema de alertas tempranas para la prevención de violaciones de DDHH y DIH a nivel nacional.
8	99	Desarrollo proceso de diseño e implementación del programa nacional de derechos humanos y memoria histórica en Colombia.
9	98	Adquisición de inmueble, diseño y adecuación para laboratorios de análisis de restos óseos nacional.
10	95	Apoyo a entidades territoriales a través de la cofinanciación para la asistencia, atención y reparación integral a las víctimas del desplazamiento forzado a nivel nacional.

Tabla 4: Proyectos del PGN con mayor puntaje de asociación al tema de víctimas.

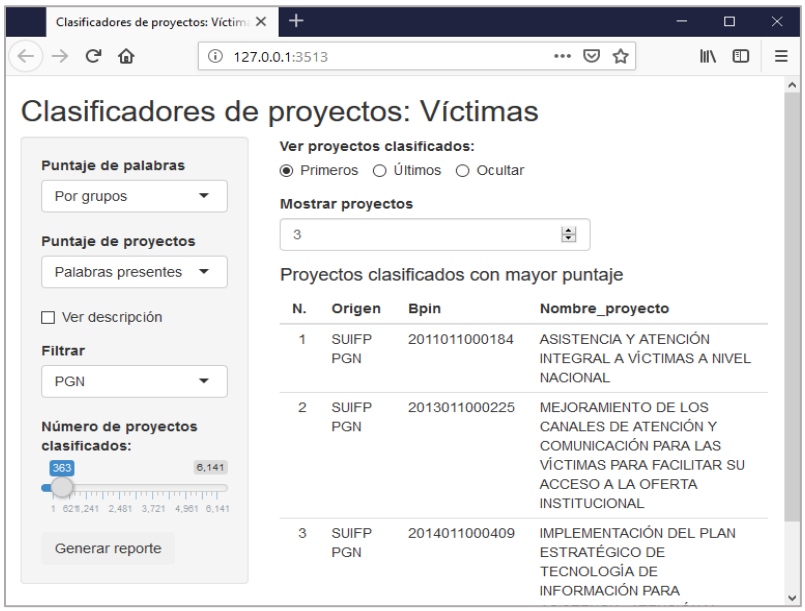

Figura 5: Interfaz gráfica de la aplicación desarrollada.

Conclusiones

El procedimiento desarrollado permitió identificar de forma automática 730 proyectos de inversión enfocados a la atención de víctimas, entre los cuales 363 son financiados con recursos del PGN. El algoritmo basado en texto desarrollado a un muy bajo costo por la Unidad de Científicos de Datos, complementado con el conocimiento experto del GPE, permitió asignar puntajes a los proyectos para determinar su nivel de relación con el cumplimiento de la Ley de Víctimas y Restitución de tierras. Además, el algoritmo permite identificar nuevos proyectos enfocados a la atención de víctimas de forma automática. Igualmente, la clasificación facilita notablemente la estimación de los recursos destinados a atención de víctimas, ya que para ello basta integrar los resultados con los montos asignados a cada proyecto. Finalmente, se entrega una aplicación con la cual los expertos pueden redefinir fácilmente algunos criterios de clasificación, en caso de considerarse pertinente.