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

IDENTIFICACIÓN DE PROYECTOS DE INVERSIÓN PARA EL FOMENTO DE LA EQUIDAD DE GÉNERO

Entidad

Departamento Nacional de Planeación

- Dirección de Desarrollo Digital.
- Dirección de Seguimiento y Evaluación de Políticas Públicas.

Sector	Lenguaje	Fuente de datos
Planeación	R	SUIFP, CONPES

Presentación

En el marco del seguimiento a la implementación del documento CONPES 161, equidad de género para las mujeres, se propone crear un clasificador basado en análisis de textos que considere la similitud entre los textos de 32.386 proyectos de inversión y documentos de política pública creados para fomentar la equidad de género. La metodología integra algoritmos de minería de texto con conocimiento experto, obteniendo las palabras clave necesarias para identificar proyectos con enfoque de género y asignando puntajes a los proyectos según sea la frecuencia de estas palabras en los textos descriptivos de los proyectos. El clasificador permite identificar los proyectos de inversión característicos y organizarlos por nivel de similitud, insumo que puede complementarse con conocimiento experto para definir un umbral de clasificación y formular indicadores de inversión de interés nacional.

Within the framework of monitoring the implementation of document CONPES 161, gender equity for women, it is proposed to create a classifier based on text analysis that considers the similarity between the texts of 32,386 investment projects and public policy documents created to promote gender equity. The methodology integrates text mining algorithms with expert knowledge, obtaining the key words necessary to identify projects with a gender focus and assigning scores to projects according to the frequency of these words in the descriptive texts of the projects. The classifier allows to identify characteristic investment projects and organize them by similarity, input that can be complemented with expert knowledge to define a classification threshold and formulate investment indicators of national interest.

Objetivo general

Analizar la inversión realizada por el Gobierno Colombiano para el fomento de la equidad de género, uno de los objetivos de desarrollo sostenible.

Objetivos específicos

1. Identificar las palabras sobre igualdad de género integrando algoritmos de análisis de texto y conocimiento experto.
2. Definir una metodología de asignación automática de puntajes a los proyectos de género, según su formulación sea relevante para la implementación del documento CONPES 161.
3. Brindar las herramientas necesarias para definir indicadores de inversión de recursos públicos para el fomento de la equidad de género.

Metodología

La metodología utilizada para identificar proyectos de inversión relacionados con la implementación del CONPES social 161 puede dividirse en dos grandes fases: (1) la identificación de las palabras alusivas al tema de género y (2) el desarrollo de un algoritmo de clasificación basado en texto. Para la primera fase, se consideró el documento CONPES 161 junto con dos documentos de en los que se evalúa su implementación, además de 36 documentos CONPES de 12 sectores de la economía. Estos últimos se escogieron aleatoriamente entre un grupo de documentos CONPES clasificados por sector y se utilizaron para identificar palabras que no son representativas del tema de género al encontrarse de forma recurrente en documentos de política pública. 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 y saneamiento	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. “participar”, “participado” y “participación”, que se convierten en “participación”), mediante una transformación basada en el algoritmo de lematización de Porter. Para ejemplificar este procedimiento, puede pensarse en una cadena de texto como “4. Fortalecimiento a la participación de las mujeres en política” que quedaría convertida en “fortalecimiento participacion mujer politica".

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 el texto “fortalecimiento participacion mujer politica”, las columnas “fortalecimiento”, “participacion”, “mujer” y “politica” 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 género y no en los demás documentos, es decir, las palabras exclusivas del tema de género. De forma similar, las palabras representativas (mas no exclusivas) del tema de género se identificaron a partir de las frecuencias promedio de cada palabra en los documentos de género 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 género. 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: Etiquetadas como "esenciales": 15 puntos Etiquetadas como "buenas": 5 puntos **Palabras representativas:** Etiquetadas como "esenciales": 10 puntos Etiquetadas como "buenas": 3 puntos	**Palabras representativas:** Todas: Valor del puntaje calculado a partir de los indicadores de similitud. **Palabras exclusivas:** Todas: Puntaje promedio de las primeras 50 palabras.

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 "fortalecimiento", "participacion", "mujer" y "politica" asignan 3, 10, 15 y 5 puntos respectivamente, el texto "fortalecimiento participacion mujer politica" tendría 33 puntos. Nótese que los puntajes pueden asignarse de dos formas: contando si la palabra está presente o contando cuántas veces lo está (frecuencia). Finalmente, se desarrolló una aplicación para facilitar la definición de un umbral de puntos a partir del cual los proyectos se consideran de género, 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 representativas del tema de género se obtuvo una lista de palabras exclusivas y alusivas ordenadas según su puntaje obtenido a partir de los 3 indicadores. En la figura 1 se muestran algunas de las palabras más representativas escogidas por los expertos para identificar proyectos de inversión asociados a la promoción de la equidad de género.

Figura 1: Palabras más representativas escogidas por los expertos para identificar proyectos de inversión relacionados con el fomento de la equidad de género.

Si bien la aplicación desarrollada permitía escoger entre los sistemas de puntuación basados en la aparición o en la frecuencia de las palabras, los resultados consideran solo el sistema de puntuación por aparición, que permite 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 según el fin con el que se quiera usar el clasificador. Así pues, la definición del umbral propuesto (38 puntos), que se presenta por defecto en la aplicación, agrupa 279 proyectos de inversión con inicio entre 2017 y 2019 dentro de los proyectos enfocados al fomento de la equidad de género, entre los cuales 93 provienen del PGN, 21 del SGR y 165 de entidades territoriales. En la tabla 3, se presenta el título de los 5 proyectos clasificados con mayor puntaje. Finalmente, en la figura 2 se muestra la interfaz gráfica de la aplicación.

N°	Proyecto
1	Implementación de los enfoques de género e interseccionalidad en la gestión pública a nivel nacional.
2	Implementación del enfoque de género en los planes, programas y proyectos a nivel nacional.
3	Fortalecimiento a la participación de las mujeres en los diferentes escenarios sociales, culturales, económicos, políticos y productivos en todo el departamento del Guaviare.
4	Prevención atención, sensibilización y reconocimiento de los derechos de las mujeres, Cundinamarca.
5	Implementación de estrategias de asistencia técnica para el fortalecimiento institucional en asuntos de género a nivel nacional.

Tabla 3: Proyectos con mayor puntaje asignado.

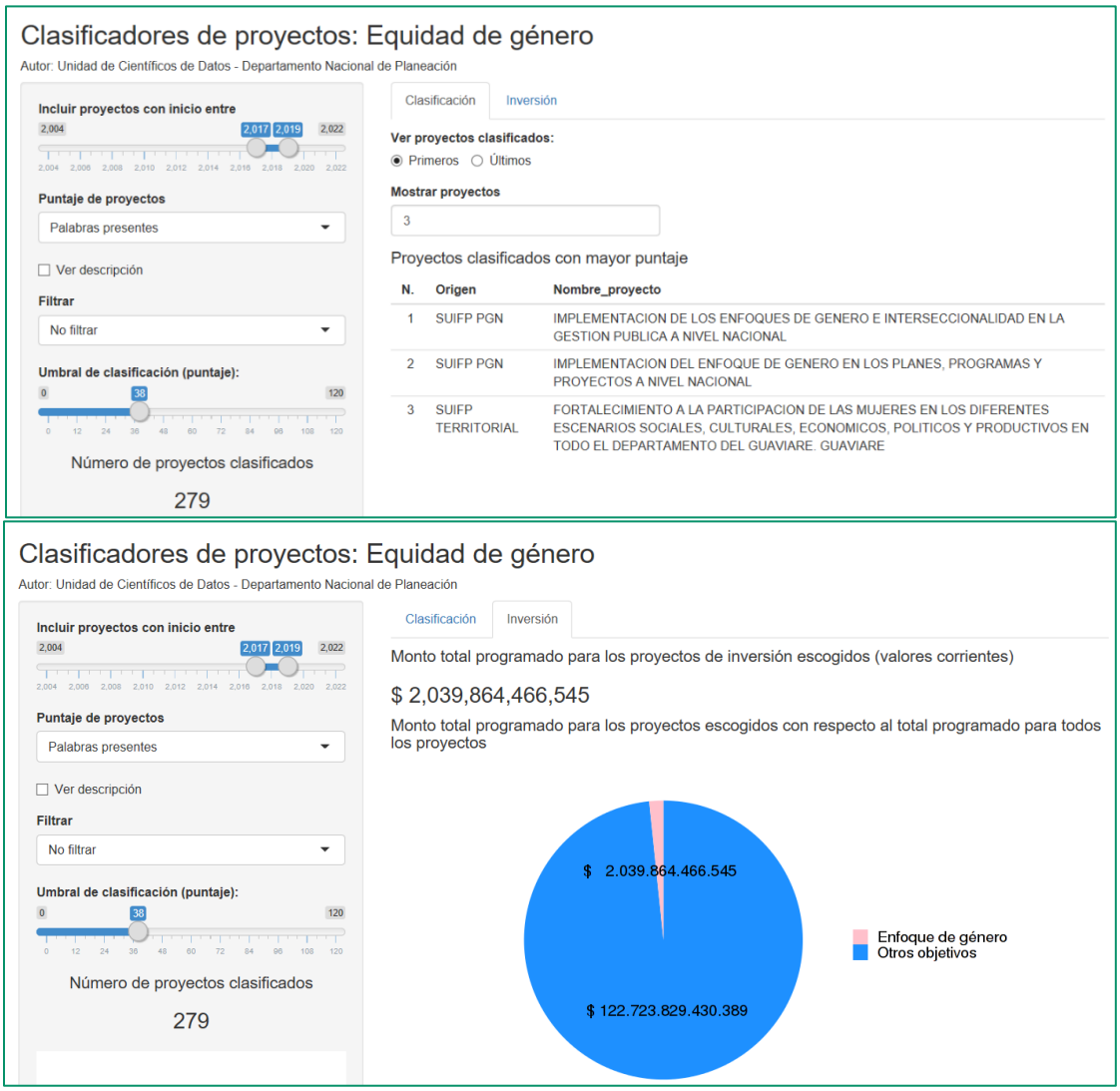

Figura 2: Interfaz gráfica de los módulos "Clasificación" (superior) e inversión (inferior) del aplicativo creado.

Conclusiones

1. El análisis brinda la información necesaria para estimar los recursos destinados a promover la equidad de género a nivel nacional.
2. El procedimiento permitió identificar proyectos de inversión cuyo fin es promover la equidad de género.
3. El algoritmo, complementado con el conocimiento experto, permite asignar puntajes a los proyectos de inversión para determinar su nivel de relación con el tema de género.
4. Con la aplicación desarrollada, los expertos pueden redefinir fácilmente algunos criterios de clasificación, en caso de considerarse pertinente.

Socialización

El proyecto se socializó con la Dirección de Seguimiento y Evaluación de Políticas Públicas y con consultores de ONU Mujeres.