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

RASTREO DE RECURSOS DESTINADOS A REDUCIR LA BRECHA DE GÉNERO EN EL MARCO DE LA ECONOMÍA DEL CUIDADO

Entidad

Departamento Nacional de Planeación

- Dirección de Desarrollo Digital.
- Dirección de Desarrollo Social.

Sector	Lenguaje	Fuente de datos
Inversión y finanzas públicas	R	CONPES, SUIFP, APC

Presentación

El artículo 2 de la Ley 1413 de 2010 define la economía del cuidado en función del trabajo no remunerado que se realiza en el hogar, relacionado con el mantenimiento de la vivienda, los cuidados a otras personas del hogar o la comunidad y el mantenimiento de la fuerza de trabajo remunerado. Estudios como la Encuesta Nacional de Uso del Tiempo (ENUT) han visibilizado la importancia de abordar la economía del cuidado con un enfoque de género en Colombia, pues son las mujeres quienes más asumen la realización de estos trabajos no remunerados. La formulación de política pública en el marco de la economía del cuidado, sin embargo, todavía debe enfrentar algunos retos para que sea acertada y basada en datos, uno de los cuales comprende la cuantificación de la inversión pública realizada en el tema de Economía del Cuidado con enfoque de género. Este proyecto presenta una metodología que facilita la identificación de proyectos de inversión formulados en este marco, mediante la integración de conocimiento experto y algoritmos de minería de texto. Los resultados se presentan mediante un aplicativo que muestra de manera visual y agregada la inversión, que puede ser analizada para distintas vigencias (2012 a 2019, con corte a junio) y para proyectos de inversión financiados con distintas fuentes, como el Sistema General de Regalías (SGR), el Presupuesto General de la Nación (PGN) y la cooperación internacional.

Article 2 of Law 1413 of 2010 defines the care economy in terms of unpaid work performed at home, related to the maintenance of housing, care of other people at home or in the community and the maintenance of the paid work force. Studies such as the National Time Use Survey (Encuesta Nacional de Uso del Tiempo, ENUT) have highlighted the importance of studying the care economy with a gender approach in Colombia, since it is women who most perform these unpaid jobs. The formulation of public policy in the framework of the care economy, however, must still face some challenges to be accurate and based on data, one of which includes the quantification of public investment in the field of Care Economy with a gender approach. This project presents a methodology that facilitates the automatic identification of projects formulated within this framework through the integration of expert knowledge and text mining algorithms. The results are presented through an application that shows the investment in a visual and aggregated manner, which can be analyzed for different periods (2012 to June 2019) and for investment projects with funding from different sources, such as the General System of Royalties (SGR), the General Budget of the Nation (PGN) and international cooperation.

Objetivo general

Facilitar la identificación y caracterización de los programas y proyectos enmarcados en la economía del cuidado con enfoque de género, a través de la aplicación de técnicas de minería de texto sobre las bases de SUIFP (Sistema Unificado de Inversiones y Finanzas Públicas) y APC (Agencia Presidencial de Cooperación Internacional).

Objetivos específicos

1. Identificar palabras clave que faciliten la identificación de programas y proyectos en el marco de la economía del cuidado con enfoque de género, mediante la integración de algoritmos de análisis textual y conocimiento experto.
2. Asignar un puntaje de relevancia a cada proyecto, con respecto al tema de economía del cuidado con enfoque de género, a partir de la relación entre los textos que describen cada proyecto y las palabras clave identificadas.
3. Desarrollar una herramienta que brinde apoyo a la Subdirección de Género de la Dirección de Desarrollo Social en temas de política pública de cuidado, mediante la presentación estructurada de los resultados en un tablero de visualización interactivo.

Metodología

El desarrollo de este proyecto se realizó en diferentes fases, en concordancia con los objetivos específicos expuestos en este documento. En la primera fase, se identificaron las palabras clave de economía del cuidado, es decir, aquellas palabras que se utilizan con mayor frecuencia cuando se habla de este tema. Dicha identificación se realizó siguiendo una metodología utilizada previamente en tres proyectos al interior de la Unidad de Científicos de Datos (UCD), siempre en el marco de rastreo de recursos en proyectos de inversión: unos sobre paz, otros sobre postconflicto y víctimas, y otros sobre reducción de la brecha de género. Las palabras utilizadas en este último proyecto (palabras sobre el tema de género) también se utilizaron como insumo para el análisis aquí presentado.

La metodología de identificación de palabras clave requiere de un insumo conformado por dos grupos de documentos: unos que hagan referencia al tema de interés (economía del cuidado) y otros que hagan referencia a otros temas. En este caso, se trabajó únicamente con documentos de política pública, para garantizar un cierto nivel de homogeneidad en el lenguaje de los documentos. El grupo de documentos (en adelante, corpus) de economía del cuidado fue escogido por la Subdirección de Género e involucraba 6 documentos: los documentos CONPES 155 y 156, los lineamientos de Política de Cuidado en Colombia, el resumen del documento de la OIT sobre “el trabajo de cuidados y los trabajadores del cuidado”, el borrador de la política nacional de infancia y adolescencia y el documento de “Propuesta para un modelo solidario y corresponsable de cuidados en Uruguay”. Por otro lado, el corpus de referencia (otros documentos) estaba conformado por 36 documentos CONPES de 12 sectores de la economía (3 por sector), los cuales se escogieron aleatoriamente de un grupo de 500 documentos CONPES etiquetados a mano para un proyecto realizado por la UCD en 2018. La tabla 1 presenta los documentos CONPES que constituyeron el corpus de referencia.

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 para el corpus de referencia, por sector económico.

Escogidos los documentos, se realizó la lectura automática utilizando R, obteniendo una cadena de texto para el contenido de cada 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. Para este proyecto no se realizó la remoción de tildes, pues la inclusión de palabras como "género" (que sin tilde se relacionaría con "generar") era esencial en este análisis. Se resalta, sin embargo, que algunos proyectos con esta palabra podrían no ser identificados cuando el texto original se encuentre sin tilde. Esta limpieza se complementó con una igualación de palabras similares en significado (por ejemplo: "participar", "participado" y "participamos", se convierten en "participar"), mediante un proceso de lematización realizado con un diccionario. Para ejemplificar este procedimiento, puede pensarse en una cadena de texto como "4. Programa de atención de personas con discapacidad", que quedaría "programar atender persona discapacidad".

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 "programar atender persona discapacidad", las columnas "programar", "atender", "persona" y "discapacidad" 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 economía del cuidado y no en los demás documentos, es decir, las palabras exclusivas del tema de economía del cuidado. De forma similar, las palabras representativas (mas no exclusivas) del tema se identificaron a partir de las frecuencias promedio de cada palabra en los documentos de economía del cuidado 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. Con una agregación ponderada 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. La figura 1 presenta un resumen de los indicadores construidos para la puntuación de las palabras.

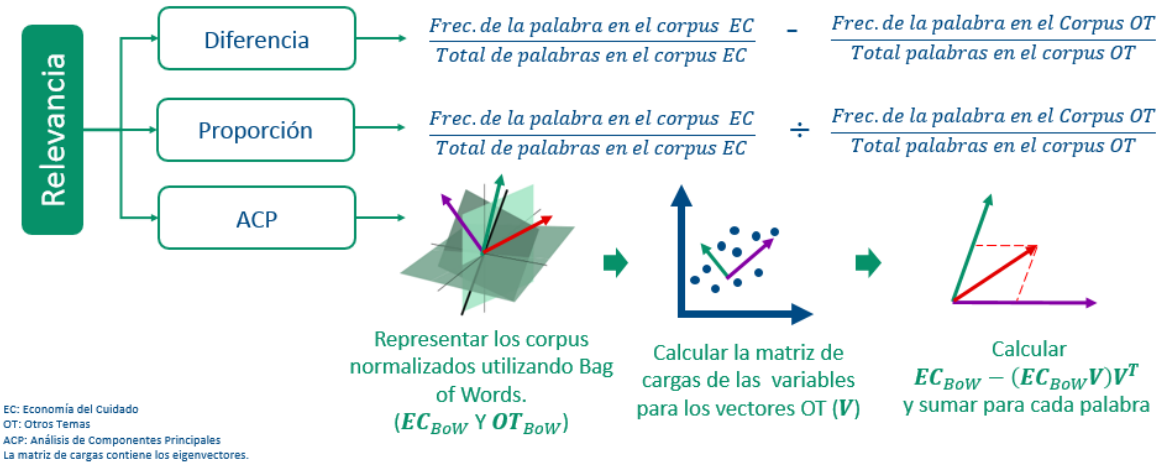

Figura 1. Medidas de relevancia para identificar palabras clave.

Para la segunda fase, se identificaron sinónimos de las palabras encontradas utilizando el algoritmo Word2Vec, definiendo un umbral de similitud semántica a partir del cual una palabra no observada en los documentos, pero sí en los proyectos de inversión, se pudiera considerar como sinónima. Así, se envió una lista con 1067 palabras representativas y 70 palabras exclusivas al equipo de la Subdirección de Género, el cual calificó las palabras como "esenciales", "buenas" o "descartadas" según el potencial que tuvieran para identificar un proyecto sobre el tema de interés. Adicionalmente, la Subdirección de Género revisó la clasificación que se tenía para las palabras de género (obtenidas en un proyecto anterior) de forma tal que los criterios de clasificación utilizados para ambos grupos de palabras fueran consistentes. Habiendo clasificado las palabras, se asoció a cada tipo de palabra una cantidad determinada de puntos, detallada en la tabla 2. Se dio una ponderación más alta a las palabras de economía del cuidado considerando que son menos comunes y que es preferible identificar proyectos sobre economía del cuidado que no tengan enfoque de género antes que proyectos con enfoque de género que no estén enmarcados en la economía del cuidado.

Puntuación de palabras de Economía del Cuidado	Puntuación de palabras de Equidad de Género
Palabras exclusivas: Etiquetadas como "esenciales": 10 puntos Etiquetadas como "buenas": 7 puntos **Palabras representativas:** Etiquetadas como "esenciales": 10 puntos Etiquetadas como "buenas": 7 puntos	**Palabras exclusivas:** Etiquetadas como "esenciales": 5 puntos Etiquetadas como "buenas": 3 puntos **Palabras representativas:** Etiquetadas como "esenciales": 5 puntos Etiquetadas como "buenas": 3 puntos

Tabla 2: Puntajes asignados a las palabras clasificadas por los expertos.

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 "programar", "atender", "persona" y "discapacidad" asignan 0, 0, 3 y 10 puntos respectivamente, el texto "programar atender persona discapacidad" tendrá 13 puntos. Bajo esta lógica, los puntajes pueden asignarse de dos formas: (1) contando si la palabra está presente o (2) contando cuántas veces lo está (frecuencia), por lo que se realizó la puntuación con ambos mecanismos. Adicionalmente, se asignó un tercer puntaje dividiendo el obtenido de la asignación 2 (frecuencia) entre el número de palabras contenidas en el texto del proyecto de inversión.

Finalmente, se desarrolló una aplicación para facilitar la definición de un umbral de puntos a partir del cual los proyectos se consideran de economía del cuidado con enfoque de género. El aplicativo permite visualizar los proyectos clasificados en función del umbral escogido por el usuario, observar el monto total de recursos invertido en el tema con respecto al total y generar reportes automáticamente. Además, permite al usuario filtrar la información según fuente de recursos: para los proyectos de SUIFP, pueden aplicarse filtros para analizar recursos del PGN o del SGR, así como es posible filtrar por sector y por vigencia (año de ejecución de los recursos); y para los proyectos de Cooperación Internacional, es posible filtrar por estado del proyecto (en ejecución o terminado) y por año de inicio del proyecto,

Resultados

En el proceso de extracción de las palabras más representativas del tema, se obtuvo una lista de palabras exclusivas y alusivas ordenadas según su puntaje obtenido a partir de los 3 indicadores. En la figura 2 se muestran algunas de las palabras más representativas escogidas por los expertos para identificar proyectos de inversión asociados a la economía del cuidado con enfoque de género.

Figura 2: Palabras más representativas de los documentos de política pública de cuidado.

En cuanto al monto de inversión, no se tiene un resultado definitivo (pues depende del umbral escogido), pero un análisis que podría realizarse con el aplicativo desarrollado sería el siguiente: “Con un umbral de 200 puntos, se clasifican 736 proyectos de inversión, financiados por el SGR en la vigencia de 2018, dentro de los proyectos enfocados a la economía del cuidado con enfoque de género. Estos proyectos representan una inversión de $207 mil millones de pesos, equivalente a un 0,204% del monto total programado para la misma vigencia”. En la figura 3 se muestra una captura de pantalla con la interfaz gráfica de la aplicación.

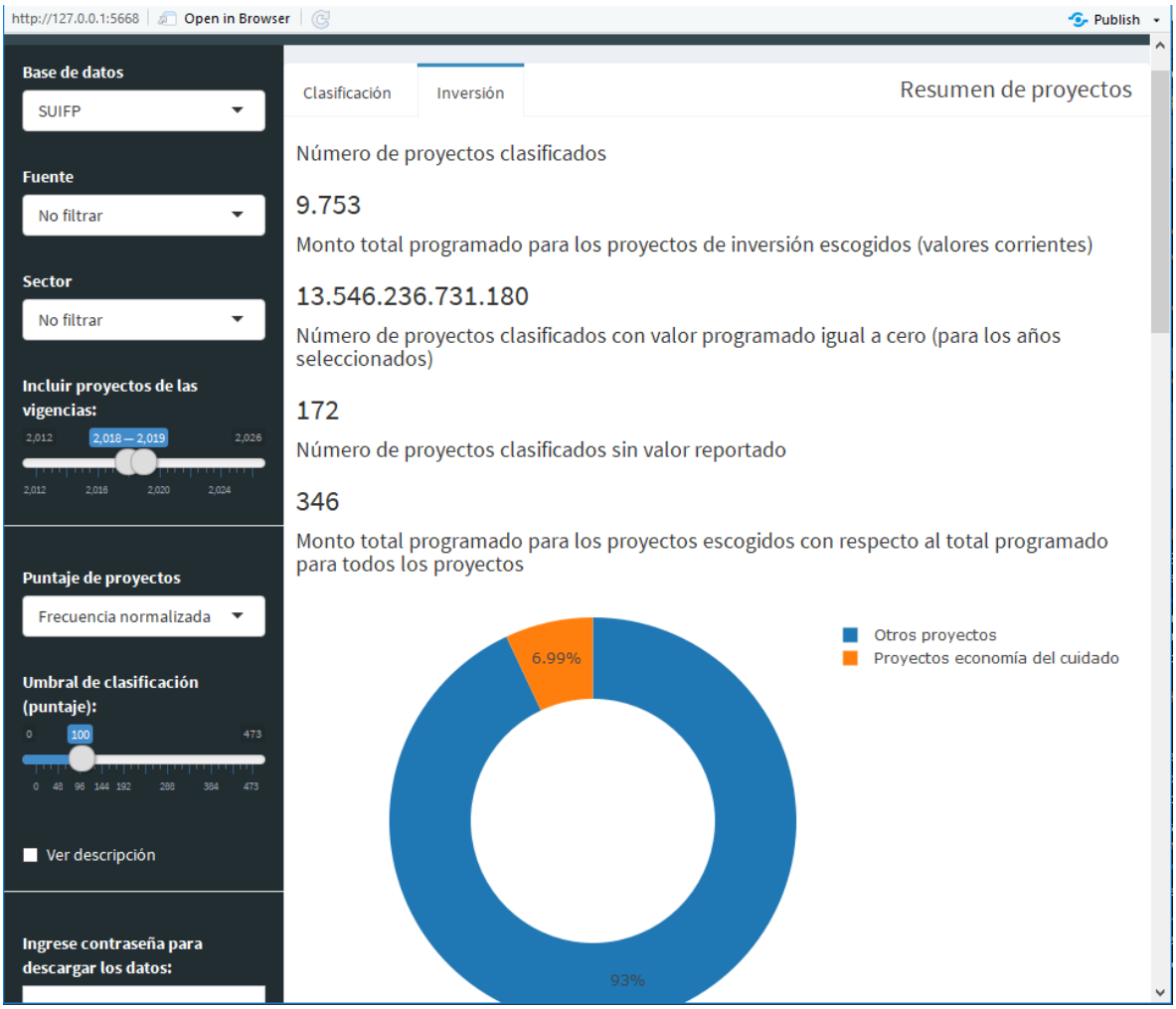

Figura 3: Interfaz gráfica del aplicativo desarrollado

Conclusiones

1. El análisis brinda la información necesaria para estimar los recursos destinados a la economía del cuidado con enfoque de género, por sector, por fuente de recursos y por vigencia.
2. El procedimiento facilita la identificación de proyectos de inversión formulados en el marco de la economía del cuidad con enfoque 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 interés.
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 Subdirección de Género de la Dirección de Desarrollo Social.