Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

ANÁLISIS DE DOCUMENTOS DE GASTO PÚBLICO

Entidad
Departamento Nacional de Planeación

- Dirección de Desarrollo Digital - Unidad de Científicos de Datos
- Dirección General

Sector
Planeación

Lenguaje
Python

Fuentes de datos
Planes Nacionales de Desarrollo
Bibliografía de Gasto Público

Contenido

1. Presentación

Un diagnóstico de los términos relevantes en los documentos y estudios relacionados con gasto público permite entender las principales problemáticas en torno al gasto en Colombia y si estas se han mantenido en el tiempo o han evolucionado, con el fin de plantear estrategias de solución y poder enfocar las políticas públicas del gasto a ello.

Este documento presenta los resultados obtenidos del análisis de términos relevantes identificados en los Planes Nacionales de Desarrollo y documentos de gasto público de forma gráfica, así como la tendencia en el tiempo de algunos términos de interés, redes de términos, nubes de palabras, la clasificación de los documentos de gasto público en diversas categorías y asimismo los resultados obtenidos del análisis de recomendaciones de los documentos de gasto público. Los hallazgos se dividen en tres partes principales, primero la comparación entre las recomendaciones y un análisis de similitud entre ellas, segundo, frecuencias de las recomendaciones que mayor similitud presentan con otras recomendaciones y tercero, un análisis de las recomendaciones y su persistencia a través de los años.

A diagnosis of the relevant terms in documents and studies related to public spending allows us to understand the main problems surrounding spending in Colombia and whether these have been maintained over time or have evolved, in order to propose solution strategies and be able to focus spending public policies to it.

This document presents the results obtained from the analysis of relevant terms identified in the National Development Plans and public expenditure documents graphically, as well as the trend over time of some terms of interest, term networks, word clouds and classification of the public expenditure documents in various categories and also the results obtained from the analysis of the recommendations of the public expenditure documents. The findings are divided into three main parts, first the comparison between the recommendations and an analysis of similarity between them, second, frequencies of the recommendations that present the greatest similarity with other recommendations and third, an analysis of the recommendations and their persistence through the years.

2. Objetivos del proyecto

2.1. General

Realizar un análisis de texto de los diferentes planes nacionales de desarrollo (PND) y otros documentos relacionados con el gasto público mediante la identificación de términos relevantes, gráficas descriptivas y otros, con el propósito de mejorar el entendimiento del contenido de los documentos e identificar posibles soluciones a las problemáticas que se presenta en torno al gasto en Colombia.

2.2. Específicos

1. Determinar los términos más relevantes dentro de los documentos de gasto público.
2. Generar gráficas descriptivas del contenido de los documentos que permitan mejorar el entendimiento del contenido de estos.
3. Identificar las recomendaciones más frecuentes y persistentes en el tiempo.
4. Realizar una clasificación de los documentos en diversas categorías de acuerdo con su contenido.

3. Metodología

3.1. Análisis descriptivo y exploratorio de los datos

Como primera aproximación a los documentos y recomendaciones es necesario verificar la calidad de los insumos y de esta forma confirmar la viabilidad de la metodología propuesta.

- ***Resumen de los datos disponibles***

Los documentos que se tienen a disposición se encuentran especificados en la Tabla 1:

Tabla 1: documentos disponibles

Tipo	Número de Documentos
Bases de PND	3
PND completos	2
PND por tomos	4
Bibliografía de gasto público	140

Fuente: elaboración propia

El análisis realizado es independiente para los Planes Nacionales de Desarrollo (PND) y para los demás documentos de bibliografía de gasto público, sin embargo, la metodología empleada es la misma para ambos casos.

Para la identificación de la frecuencia y relevancia de las recomendaciones se tiene a disposición un archivo Excel con la información de los diferentes documentos de bibliografía relacionada con gasto público. A continuación, se lista la información que contiene el archivo:

- Nombre del archivo
- Número (número secuencial de identificación del documento)
- Autores
- Año
- Institución
- Título
- Sector
- Diagnóstico
- Recomendaciones

Esta información se encuentra para los 146 documentos de gasto público.

El número de recomendaciones por conjunto de documentos anuales se encuentra en la Figura 1. Dado que un documento puede contener varias recomendaciones, se procedió a realizar una separación de estas obteniendo un total de 542 recomendaciones.

Figura 1: Número de recomendaciones analizadas

Fuente: elaboración propia

3.2. Lectura automática de los documentos

La lectura de los documentos se realizó de forma automática para aquellos que fue posible, sin embargo, algunos no eran leídos de forma correcta y presentaban muchas variaciones en comparación con el texto original, razón por la cual fue necesario realizar la lectura mediante OCR (Reconocimiento óptico de caracteres) para algunos documentos. La lectura mediante OCR es un proceso que permite el reconocimiento de textos que se encuentran almacenados en imágenes (por ejemplo, documentos escaneados), a través de la identificación de caracteres o símbolos que son convertidos en cadenas de texto que pueden ser manipuladas e interpretadas como cualquier otro texto en formato plano, dependiendo de la calidad del insumo esta metodología permite que la extracción de texto sea mejor. Cuando los textos son leídos se guardan automáticamente en formato de texto plano y quedan dispuestos para los siguientes pasos.

3.3. Preprocesamiento y limpieza de los textos

Con los textos ya leídos el paso a seguir es la limpieza. Dado que algunos de los documentos se encuentran en inglés es indispensable realizar la traducción de estos antes de continuar, con el fin de estandarizar el idioma de todos los documentos y poder hacer un análisis en conjunto.

Posterior a la traducción de documentos se realiza la limpieza del texto, en esta etapa se dejan los archivos en formato plano de forma tal que luego puedan ser analizados de forma óptima, este procedimiento es el mismo para las recomendaciones contenidas en el archivo Excel. El proceso de limpieza consiste en eliminar caracteres especiales, números, signos de puntuación y palabras vacías o *stopwords* (palabras cuyo significado por sí solas es nulo, como preposiciones, conectores o artículos, las cuales no aportan valor al análisis), al igual que palabras en encabezados o pies de página que se repiten con frecuencia dentro del texto.

3.4. Identificación de términos relevantes

Para la identificación de términos clave se procedió a realizar una representación numérica de los textos. Teniendo en cuenta que se tienen dos grupos de documentos similares, documentos referentes a los Planes Nacionales de Desarrollo y documentos referentes a gatos público, se hizo una comparación de términos permitiendo extraer aquellos más relevantes y términos únicos de cada grupo de documentos de interés, teniendo como referencia los documentos del otro grupo temático.

- ***Gráficos descriptivos***

Una vez procesados los textos se procedió a generar cuatro tipos de gráficos que facilitan el entendimiento del contenido de estos, se generaron nubes de palabras, gráficos de barras, gráficos de coocurrencias y gráficos de dispersión léxica.

3.5. Clasificación de documentos

Inicialmente para la clasificación de los documentos se planteó usar un modelo de clasificación supervisado. Sin embargo, al no contar con un conjunto de datos etiquetados que permita implementar un modelo de este tipo, se optó por implementar la clasificación mediante el uso de reglas duras establecidas por la Dirección General. Se procedió a realizar la clasificación de documentos de gasto público en 9 categorías diferentes, "*Artículo Científico*", "*Ciclo de Inversión*", "*Revisión de Gasto*", "*Evaluación de Resultados*", "*Propuestas, planeación y recomendaciones*", "*Caracterización del Gasto*", "*Política Fiscal*" y "*Otros*".

3.6. Análisis de recomendaciones

- ***Extracción de verbos de las recomendaciones***

Una vez se tiene el texto de las recomendaciones limpio, a través de un análisis de entidades que permite identificar la categoría léxica de los términos dentro de un texto, como son los verbos, sujetos, instituciones y otros; se extrajeron los verbos de cada una de las recomendaciones, teniendo en cuenta que estos permiten identificar la división entre dos oraciones o recomendaciones y facilitan un análisis más específico. Por ejemplo, para la recomendación*:*

"estimular competencia más eficaz entre las eps incorporando elementos de prevención y calidad en cálculo de la upc", el algoritmo identifica las palabras *estimular e incorporando* como verbos. Con esta información se consolida una tabla con la información ya disponible de las recomendaciones y una nueva columna con los verbos.

Dado que dentro de cada una de las recomendaciones puede existir más de un verbo como se puede observar en el ejemplo anterior, cada uno de ellos se deja en una fila distinta. Con esta información consolidada se identifican los cuatro términos que suceden al verbo para tener un resumen de la oración y poder identificar información valiosa. Para el ejemplo anterior, el verbo *estimular* está sucedido por "*competencia más eficaz entre",* por tanto, en una nueva columna de la tabla se tiene*: estimular competencia más eficaz entre.*

- ***Similitud entre recomendaciones***

A través de Word2Vec, una técnica para el procesamiento del lenguaje natural que utiliza un modelo de red neuronal para identificar asociaciones entre palabras de un texto, se realiza una representación vectorial de las recomendaciones con el fin de poder compararlas. Una vez se encuentran vectorizadas, a través de la similitud coseno, una función matemática simple que indica el nivel de similitud semántica entre las oraciones representadas por dichos vectores, se genera una matriz con el cálculo de las similitudes entre cada una de las recomendaciones como la que se puede observar en la Tabla 2. En las filas y columnas se encuentra el número identificador de la recomendación y al interior de la matriz la similitud calculada por el algoritmo. El valor de similitud puede tomar un valor entre 0 y 1, en cuanto más cercano a 1 sea mayor será la similitud entre los textos, por esto se puede observar que la diagonal de la matriz, es decir, la comparación entre las recomendaciones consigo mismas toman el valor de 1.

Tabla 2: ejemplo matriz de similitudes

	0	1	2	3	4
0	1	0.696808	0.645124	0.672958	0.7335
1	0.696808	1	0.783771	0.546708	0.680953
2	0.645124	0.783771	1	0.780797	0.658375
3	0.672958	0.546708	0.780797	1	0.669622
4	0.7335	0.680953	0.658375	0.669622	1

Fuente: elaboración propia

Teniendo en cuenta que la matriz es simétrica, es decir que la diagonal superior e inferior tienen la misma información, nos quedamos con la diagonal inferior con el fin de facilitar el análisis y no duplicar datos. Asimismo, se filtran las similitudes dejando solo aquellas superiores a 0.9, esto nos permite observar las recomendaciones que mayor similitud presentan con otras. Una vez se realizan estos pasos se obtiene una tabla como la presentada en la Tabla 3. Las columnas *fila* y *col* indican el número de fila y columna correspondiente a la matriz de similitudes obtenida anteriormente, lo cual se traduce a un número identificador de cada recomendación, la columna *similitud* indica el nivel de la similitud entre cada par de recomendaciones, asimismo se encuentra el texto completo de la *recomendación*, la *recomendación resumida* que corresponde al verbo con las cuatro palabras que le suceden, el *número* que identifica el documento de bibliografía y el campo *año* que indica el año de publicación del documento asociado a la recomendación para cada elemento del par analizado, con el número 1 se identifica la información correspondiente a la recomendación asociada a la columna fila y con el número 2 la correspondiente a la columna. Para el caso de la Tabla 3 se observan los textos, número y año correspondiente a la fila, identificado con el número 1.

Tabla 3: Ejemplo tabla comparaciones con mayor similitud

Fila	Col	Similitud	Recomendación	Recomendación resumida 1	Número 1	Año 1
1387	528	0,90002	privilegiar principio de competencia en procesos de contratación directa	privilegiar principio competencia procesos contratación	1021	2017
620	216	0,90003	complementarse con un fortalecimiento de los mecanismos de fiscalización y control por parte de la ugpp y la dian	complementarse fortalecimiento mecanismos fiscalización control	1037	2018
1229	1222	0,90005	reformar el sistema tributario para aumentar los ingresos incrementar la eficiencia y fortalecer la equidad	reformar sistema tributario aumentar ingresos	1122	2015

Fuente: elaboración propia

4. Resultados

A través del desarrollo metodológico descrito en la Sección 3, se obtuvieron los resultados que se presentan a continuación.

4.1. Términos relevantes

En la Tabla 4 se presentan los 10 bigramas más relevantes en los documentos del PND de 2014-2018, teniendo como referencia el corpus de documentos relacionados con gasto público.

Tabla 4. Términos relevantes PND 2014-2018

Bigramas
Gobierno Nacional
Entidades territoriales
Movilidad social
Conflicto armado
Pueblos indígenas
Pueblo Rrom
Sistema nacional
Siguientes acciones
Niños niñas
Buen gobierno

Fuente: elaboración propia

- ***Nubes de palabras***

Las nubes de palabras son una representación visual que muestra la frecuencia relativa de las palabras y bigramas dentro de un texto. Aquellos términos que son más frecuentes (es decir, que aparecen más dentro de los documentos) se presentan con un mayor tamaño y se encuentran en el centro de la nube. A medida que los términos son menos frecuentes, pierden relevancia y por consiguiente tamaño en la representación visual. De manera ilustrativa, en la Figura 2 se pueden observar las nubes de palabras y bigramas para los documentos de los PND entre 2014 y 2018.

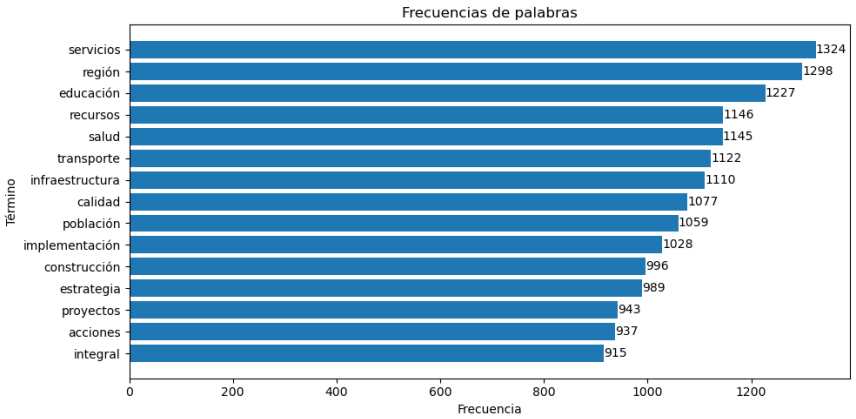
Figura 2: Nube de palabras(izquierda) y nube de bigramas(derecha) PND 2014-2018

Fuente: elaboración propia

- ***Gráficos de barras***

Al igual que las nubes de palabras, los gráficos de barras muestran la frecuencia de los términos, en este caso los gráficos de barras de realizaron para los 15 términos más frecuentes de los documentos y se puede observar el número de veces que aparecen dentro de los documentos. En la Figura 3 se puede observar un ejemplo de los gráficos de barras para el PND 2014-2018.

Figura 3: Gráfico de barras con las 15 palabras más frecuentes del PND 2014-2018.



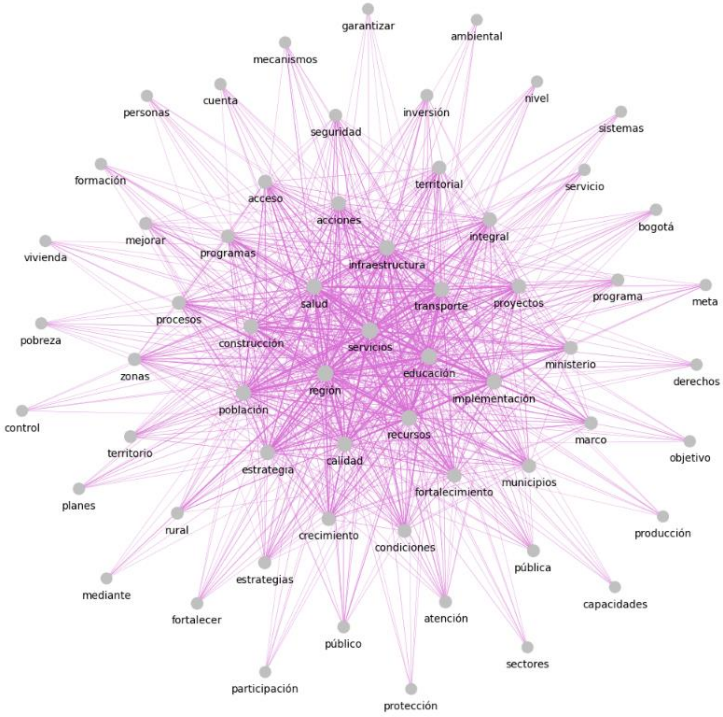
Fuente: elaboración propia

- ***Coocurrencias y redes de términos***

Las coocurrencias muestran la aparición conjunta de dos o más términos dentro de un mismo documento, esta información queda consignada en una matriz denominada matriz de coocurrencias. Con esta matriz se pueden construir redes de términos en las cuales el tamaño de cada punto es proporcional a la frecuencia de aparición de ese

término y el grosor de las líneas entre puntos es proporcional a la cantidad de veces que dos términos aparecen juntos en un documento. En la Figura 4 se puede observar un ejemplo de la red de términos para los PND entre 2014 y 2018.

Figura 4: Red de términos para el PND 2014-2018.



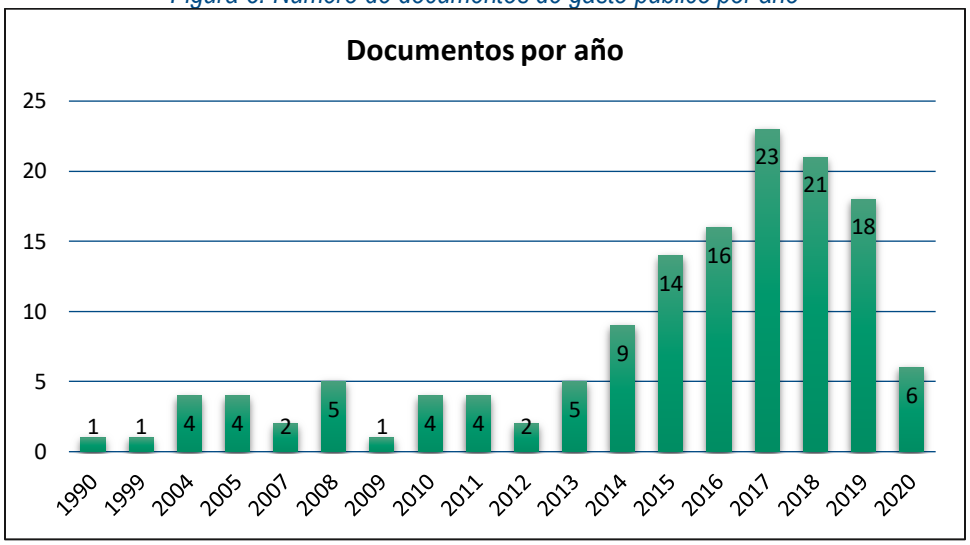
Fuente: elaboración propia

- ***Gráficos de dispersión léxica***

El gráfico de dispersión léxica es una representación visual de la aparición de términos dentro de un texto o grupo de documentos, el objetivo de este gráfico es mostrar la pérdida o ganancia de relevancia de un término en un documento.

Es importante aclarar que el análisis se hizo por años para los documentos de gasto público, por esta razón los documentos fueron agrupados por año de publicación. En la Figura 5 se puede observar el número de documentos que fueron tenidos en cuenta para cada año.

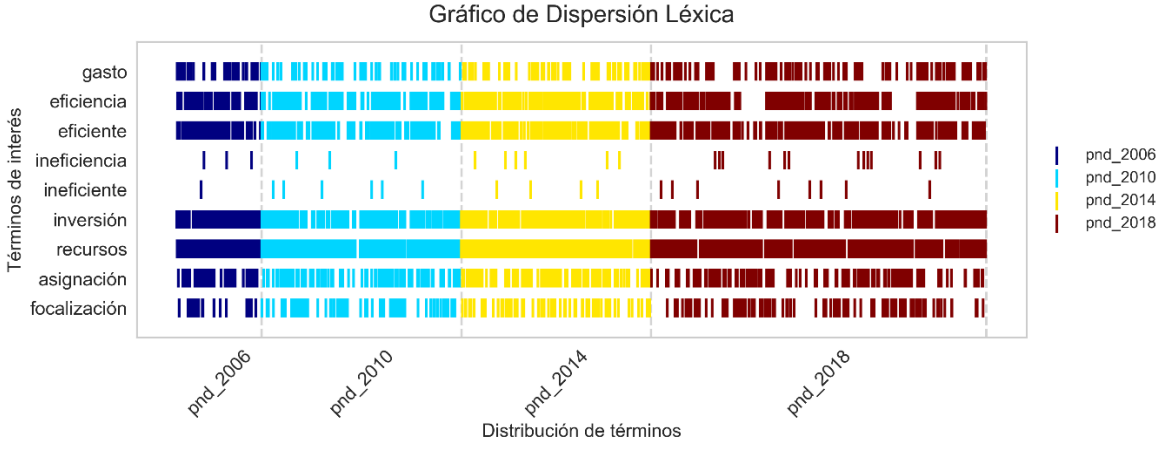
Figura 5. Número de documentos de gasto público por año

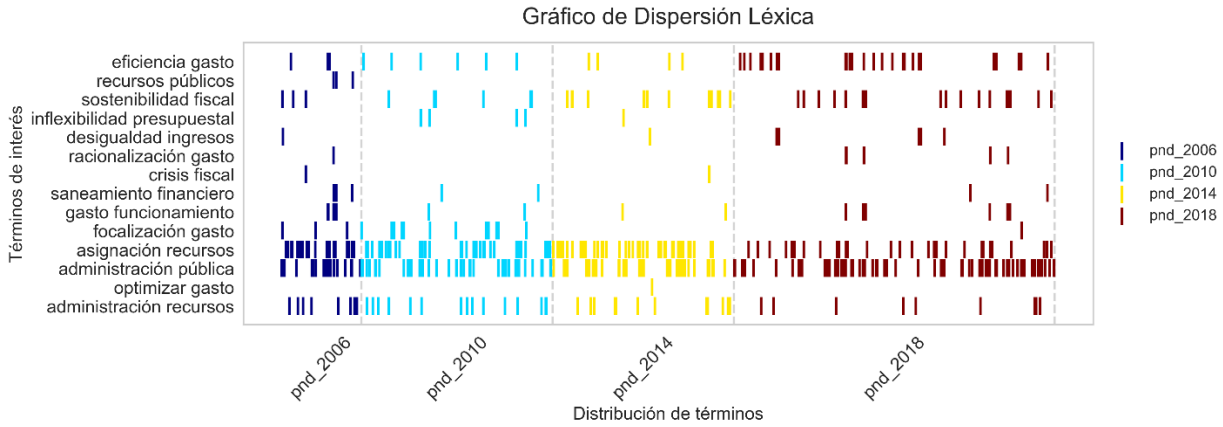


Fuente: elaboración propia

En la Figura 6 se pueden observar los gráficos de dispersión léxica para los términos de interés de los PND de 2006 a 2018, estos términos son unigramas (una palabra) y bigramas (dos palabras).

Figura 6: Gráfico de dispersión léxica PND, unigramas(arriba) y bigramas(abajo)





Fuente: elaboración propia

4.2. Clasificación de documentos

Como se mencionó anteriormente, la clasificación de documentos de gasto público se realizó mediante la implementación de reglas duras de categorización, para esto se utilizaron dos atributos de texto "*Título LARGO*" y "*Diagnósticos*" contenidos en el archivo "*Matriz Diagnósticos.xlsx*", es decir, se realizó el ejercicio de clasificación aplicando las mismas reglas, pero teniendo en cuenta dos textos diferentes para cada documento. En la Tabla 5, se presentan los resultados de la clasificación de documentos utilizando el texto de la columna diagnósticos.

Tabla 5. Número de documentos clasificados por categoría – texto diagnóstico.

Categoría	Número de documentos
Otros	32
Ciclo de Inversión	31
Evaluación de impacto	23
Artículo Científico	18
Caracterización del Gasto	15
Evaluación de Resultados	12
Propuestas, planeación y recomendaciones	9
Revisión de Gasto	4
Política Fiscal	2

Fuente: elaboración propia

4.3. Análisis de recomendaciones

- ***Frecuencias y años de las recomendaciones similares***

Una vez se tiene la información presentada en la Tabla 3, se pueden identificar las recomendaciones con mayor similitud y la frecuencia de aparición de estas, para ello se consolida la información de filas y columnas de la Tabla 3 en una misma tabla y se hace el conteo de su frecuencia de aparición. En la Tabla 6 se presenta un ejemplo de una de las recomendaciones, la columna *Número* que identifica el documento de bibliografía asociado a la recomendación, se observa el texto de la *Recomendación* y la *Recomendación resumida*, que fue la utilizada para el análisis, la columna *Frecuencia* indica el número de veces que se encontró un texto similar (con una similitud mayor a 0.9) para cada una de las recomendaciones en la tabla de comparaciones y las columnas *Años* y *Números*, que corresponden a los años de publicación y números que identifican los documento de bibliografía correspondientes a las recomendaciones más similares.

Tabla 6: Tabla de frecuencia y años de las recomendaciones con mayor similitud

Recomendación	Recomendación resumida	Número	Año	Frecuencia	Años	Números
hacer una reforma institucional al programa que permita reducir costos de coordinación y centrarse en los objetivos del mismo	permita reducir costos coordinación centrarse	1090	2013	7	[2013, 2014, 2015, 2016]	[1010, 1010, 1084, 1084, 1090, 1141, 1141]

Fuente: elaboración propia

5. Conclusiones y recomendaciones

A partir de la metodología desarrollada y de los resultados obtenidos para cumplir los objetivos de este proyecto, se presentan a continuación las principales conclusiones obtenidas por el equipo de la UCD y las principales recomendaciones para un mejor uso y aprovechamiento del proyecto.

1. La etapa de lectura y limpieza de los documentos tomó más tiempo de lo esperado, sin embargo, se recalca la importancia de esta etapa dado que su resultado es el insumo de todas las etapas posteriores.
2. Se identificaron los términos más relevantes y recurrentes en los documentos relacionados con gasto público y los PND.
3. Se generaron diversas gráficas descriptivas de los textos de los documentos, facilitando la comprensión del contenido de estos.
4. Se logró realizar la clasificación de los documentos de gasto público de acuerdo con las reglas establecidas por la Dirección General.
5. El análisis de texto de las recomendaciones mediante la identificación de la categoría léxica de términos, permitió entender en mejor medida cuáles son las principales problemáticas que se presentan en torno al gasto en Colombia y que persisten en el tiempo. Vale la pena mencionar que los resultados del análisis se podrían mejorar si se utilizaran textos de mayor longitud.

6. Socialización

Los resultados del proyecto fueron socializados con la Dirección General.

7. Contacto

Si tiene alguna duda, comentario o sugerencia sobre este proyecto, o si le gustaría conversar con la Unidad de Científicos de Datos sobre la posibilidad de una nueva fase para el mismo, puede comunicarse con nosotros a través del correo electrónico ucd@dnp.gov.co.