Dirección De Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento
Nacional de Planeación

IDENTIFICACIÓN Y ANÁLISIS DE PALABRAS CLAVE ASOCIADAS A DOCUMENTOS CONPES POR PERIODO DE GOBIERNO

Entidad

Departamento Nacional de Planeación

- Dirección de Desarrollo Digital.
- Grupo CONPES

Sector	Lenguaje	Fuente de datos
Planeación	R	Documentos CONPES y Planes Nacionales de Desarrollo

Presentación

Un diagnóstico de los temas con mayor y menor relevancia en cada período de gobierno permite analizar la evolución histórica de los Planes Nacionales de Desarrollo (PND) y de las políticas públicas, y orientar la construcción de futuros documentos. Sin embargo, esta identificación de manera manual resulta dispendiosa y sujeta a un gran componente subjetivo. Por esta razón, resulta útil un análisis histórico a través de minería de texto de los documentos CONPES de tipo "lineamientos de política", desde 1967 hasta 2019, determinando los términos y temas más relevantes en cada uno de ellos e identificando su cambio de relevancia a través del tiempo para diferentes sectores socioeconómicos.

A diagnosis of the issues with greater and lesser relevance in each government period allows us to analyze the historical evolution of the National Development Plans (PND) and public policies and guide the construction of future documents. However, this identification, if done manually, is very time consuming and subject to a large subjective component. For this reason, a historical analysis through text mining of the CONPES documents of the "policy guidelines" type, from 1967 to 2019, is useful, determining the most relevant terms and topics in each of them and identifying their change of relevance through time for different socioeconomic sectors.

Objetivo general

Identificar los temas relacionados con la política pública del país que han tenido mayor participación y relevancia durante los distintos períodos de gobierno a partir del análisis automático de los documentos CONPES sobre lineamientos de políticas públicas para varios sectores socioeconómicos.

Objetivos específicos

1. Identificar las palabras más representativas de cada sector socioeconómico analizado (Agropecuario; Agua potable y saneamiento básico; Ambiente y desarrollo sostenible; Ciencia, tecnología e innovación; Comercio, industria y turismo; Cultura, deporte y recreación; Defensa; Educación; Estadística; Inclusión social y reconciliación: Interior; Justicia y del derecho; Minas y energía; Relaciones exteriores; Salud y protección social; Tecnologías de la información y las comunicaciones (TIC); Trabajo; Transporte; Vivienda) durante los diferentes periodos de gobierno, mediante un proceso automático.
2. Construir gráficos que ilustren la aparición de los términos de cada sector en los documentos CONPES de cada periodo de gobierno, para determinar la evolución de estos a través del tiempo.
3. Construir un tablero que permita visualizar de forma interactiva los gráficos y resultados obtenidos.

Metodología

Como primera aproximación al análisis de los documentos CONPES a través del tiempo, se realizó un análisis descriptivo y exploratorio que permitió validar la calidad de los insumos a analizar, y asimismo establecer un proceso viable para el análisis requerido.

Lectura de los documentos CONPES

Con la información contenida en la base de datos que contiene todos los documentos CONPES con su respectivo hipervínculo se realizó la descarga automática de estos. Los que no tenían un enlace o hipervínculo asociado se excluyeron del análisis, pues no era posible determinar exactamente dónde se encuentran alojados estos documentos. Sin embargo, como se puede observar en la Tabla 1, son pocos los documentos que presentaron este problema, en comparación con el total de documentos analizados. Posteriormente, se realizó la lectura de los documentos CONPES, extrayendo el texto directamente para los que fue posible y a través de un OCR (reconocimiento óptico de caracteres) para los que fue necesario. El OCR es un proceso que permite el reconocimiento de textos que se encuentran almacenados en imágenes (como, 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. La Tabla 1 resume los resultados de la lectura de los documentos.

Tabla 1: Resultados de la lectura de documentos

Grupo	Cantidad
Documentos CONPES sin enlace (excluidos)	49
Documentos leídos con éxito, sin OCR	1266
Documentos leídos con éxito, con OCR	201
Total	**1516**

Fuente: Elaboración propia

Limpieza de texto

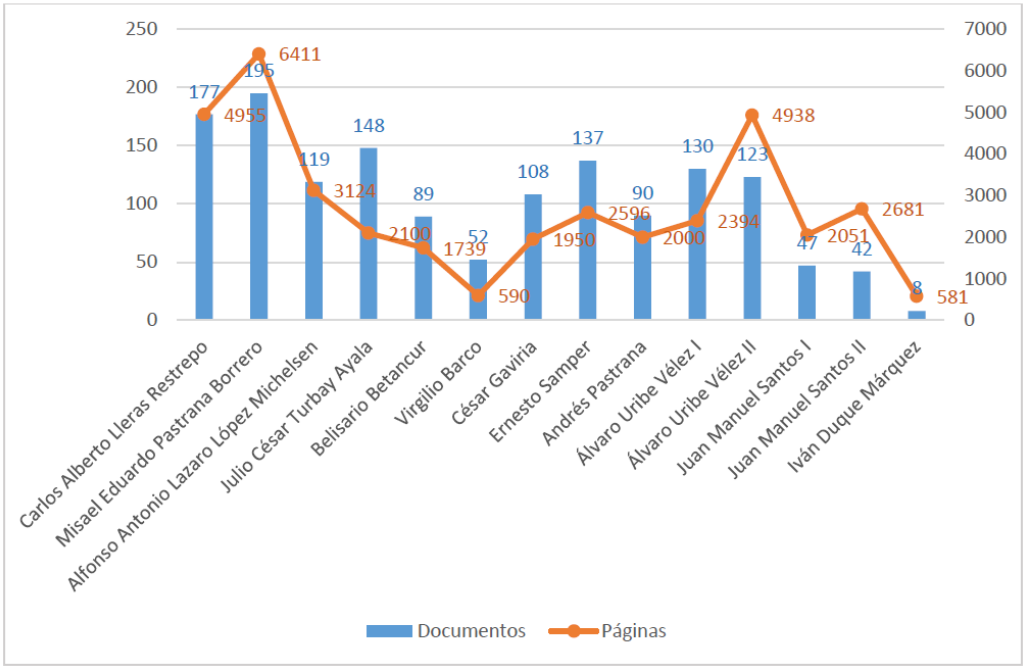
Este proceso consiste en eliminar caracteres especiales, números, signos de puntuación y palabras vacías o *stopwords* (palabras que no son de interés para el análisis, como preposiciones, conectores y artículos). Asimismo, palabras como “plan”, “nacional”, “desarrollo”, “Colombia”, “ministerio”, “país”, entre otras, se excluyeron del análisis ya que se mencionan en la gran mayoría de páginas y documentos, por lo que no son de utilidad para diferenciar unos de otros. Al texto resultante de este proceso se le denomina “texto limpio” y es el que se utiliza como insumo para el análisis descriptivo.

Finalmente, se excluyeron del análisis aquellas páginas que tuvieran 50 palabras o menos. Esto se hizo con el fin de eliminar páginas como la portada, la contraportada y la tabla de contenido de la gran mayoría de documentos CONPES. Como resultado de este procedimiento también se eliminaron páginas con poco texto que pueden corresponder a la parte final de una sección o del documento, o páginas con imágenes que se acompañan de poco texto. Con esta exclusión, hubo dos documentos para los cuales todas sus páginas fueron eliminadas, que fueron:

- CONPES social 43
- CONPES económico 1496

La Figura 1 presenta el número de documentos CONPES (en color azul) y el número total de páginas (en color naranja) que se utilizaron para el análisis, por cada período de gobierno.

Figura 1: Documentos CONPES incluidos en el análisis por cada período de gobierno

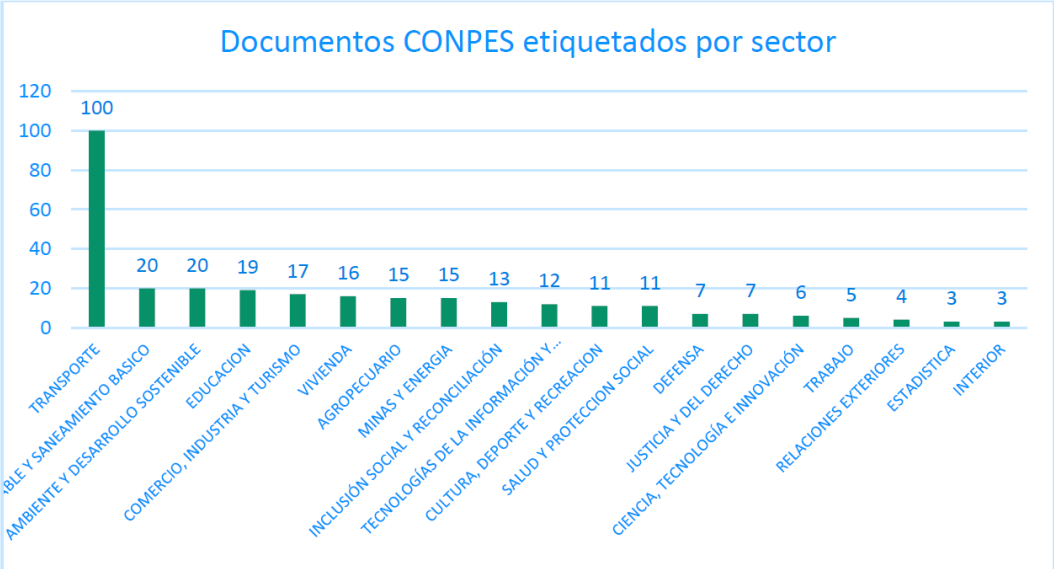


Fuente: Elaboración propia

Identificación de palabras clave

La identificación de palabras clave se realizó a partir de los documentos CONPES indicados por Grupo CONPES como representativos de distintos sectores sociales y económicos. La Figura 2 presenta los sectores que se tuvieron en cuenta para el análisis y el número de documentos representativos de cada sector que fueron utilizados.

Figura 2: Documentos CONPES etiquetados por sector



Fuente: Elaboración propia.

Sobre el texto preprocesado o "limpio" de todos los documentos CONPES, se construyó una matriz base de términos y documentos mediante una técnica llamada BoW (*Bag of Words*), cuyas columnas corresponden a las palabras que aparecen en todos los documentos (el vocabulario del conjunto de documentos), cada fila corresponde a una página de un documento y el elemento (i,j) de la matriz corresponde a la frecuencia o número de veces que aparece la palabra j en la página i. En la Figura 3 se muestra un ejemplo de una matriz de términos y documentos.

Figura 3: Ejemplo de una matriz de términos y documentos (utilizando "Bag of Words")

	un	texto	ejemplo	otro	de	más
un texto ejemplo	1	1	1	0	0	0
otro texto ejemplo	0	1	1	1	0	0
ejemplo de texto	0	1	1	0	1	0
un texto más texto	1	2	0	0	0	1

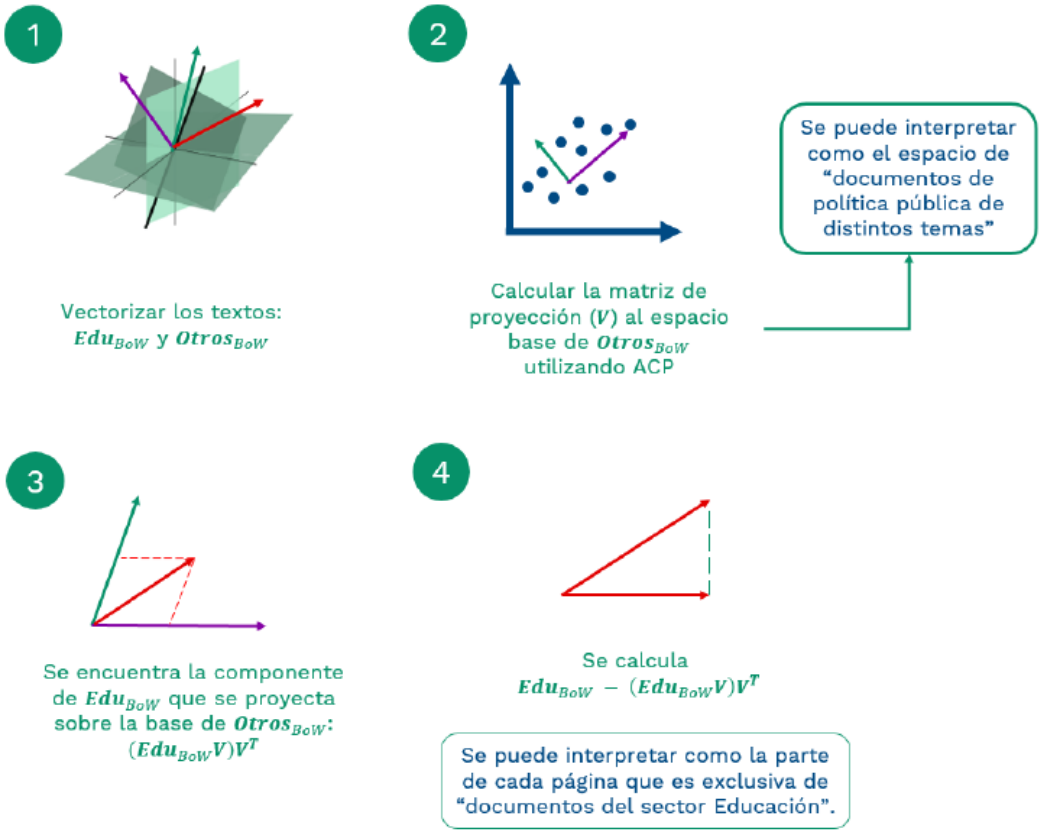
Fuente: Elaboración propia.

Para determinar las palabras que más diferencian los documentos de un sector dado (por ejemplo, el sector educación - Edu) frente a los de otros sectores ($Otros$), se utilizó como punto de partida el siguiente procedimiento:

1. Se identifican y separan los términos que únicamente aparecen en los documentos del sector educación (Edu). Estos se denominan "palabras exclusivas" y su frecuencia de aparición se toma como un "puntaje de exclusividad".
2. Se separa la matriz BoW en dos: la correspondiente a las páginas de documentos del sector educación ($EduBoW$), y la de todas las otras páginas ($OtrosBoW$).
3. Aplicando análisis de componentes principales, se usa una base vectorial que genera el espacio $OtrosBoW$, cuya correspondiente matriz de proyección se denomina V. Este espacio se puede interpretar como el de "documentos de política pública de distintos temas".
4. Se encuentra la componente de $EduBoW$ que se proyecta sobre la base de $OtrosBoW$. Algebraicamente, esto corresponde a $EduOtros = EduBoWVV^T$. Cada fila de $EduOtros$ se puede interpretar como la parte de cada página que es común al espacio de "documentos de política pública".
5. Se encuentra la proyección de $EduBoW$ sobre el espacio nulo de $OtrosBoW$. Esta proyección se calcula como $Edu0 = EduBoW - EduOtros$. Cada fila de esta matriz ($Edu0$) se puede interpretar como la parte de cada página que no puede reconstruirse con la base del espacio "documentos de política pública", es decir, aquella parte del documento que es representativa de los "documentos del sector educación".
6. Se realiza la suma sobre las columnas de $Edu0$ para obtener un "puntaje de representatividad" de cada palabra. Aquellas con mayor puntaje se toman como más representativas del sector educación.

En la Figura 4 se presenta de forma gráfica el procedimiento descrito.

Figura 4: Metodología para identificar palabras representativas de cada sector.



Fuente: Elaboración propia.

Dada la alta dimensionalidad de la matriz $OtrosBoW$, este procedimiento no se pudo seguir de forma exacta, pues computacionalmente tomaba mucho tiempo calcular la matriz de proyección V. Por esta razón fue necesario:

1. Tomar solo aquellos términos del vocabulario que aparecen al menos 5 veces en todos los documentos, descartando así palabras muy poco frecuentes en el conjunto de documentos.
2. Realizar *bootstraping* con muestras de tamaño 100 (cien) y 1000 (mil) iteraciones para calcular los puntajes de "representatividad" de cada palabra.

El *bootstraping* es un proceso de re-muestreo que consiste en extraer muestras de los datos y realizar estimaciones de un parámetro de forma iterativa. Para este caso, esto fue equivalente a no tomar todas las páginas (filas) de la matriz $OtrosBoW$ para calcular la matriz V (y, con ella, $Edu0$ y los puntajes), sino solamente 100 páginas (tamaño de la muestra), de forma que el cálculo fuera mucho más rápido. Por supuesto, al tomar solo 100 páginas para calcular el puntaje se está sujeto a una alta variabilidad, motivo por el cual este procedimiento se repite 1000 veces y el puntaje final estimado es el que resulta de promediar los puntajes obtenidos en cada una de las 1000 iteraciones.

Al realizar *bootstraping* ocurre también que las "palabras exclusivas" pueden ser diferentes en cada iteración. Por ejemplo, la palabra "capacitación" puede no estar en ninguna de las 100 páginas de "otros temas" escogidas en la iteración 1, pero sí en las escogidas en la iteración 2. Por ello, una palabra puede resultar tanto con un puntaje de exclusividad como con un puntaje de representatividad. Para obtener un "puntaje único" para cada palabra se realizó

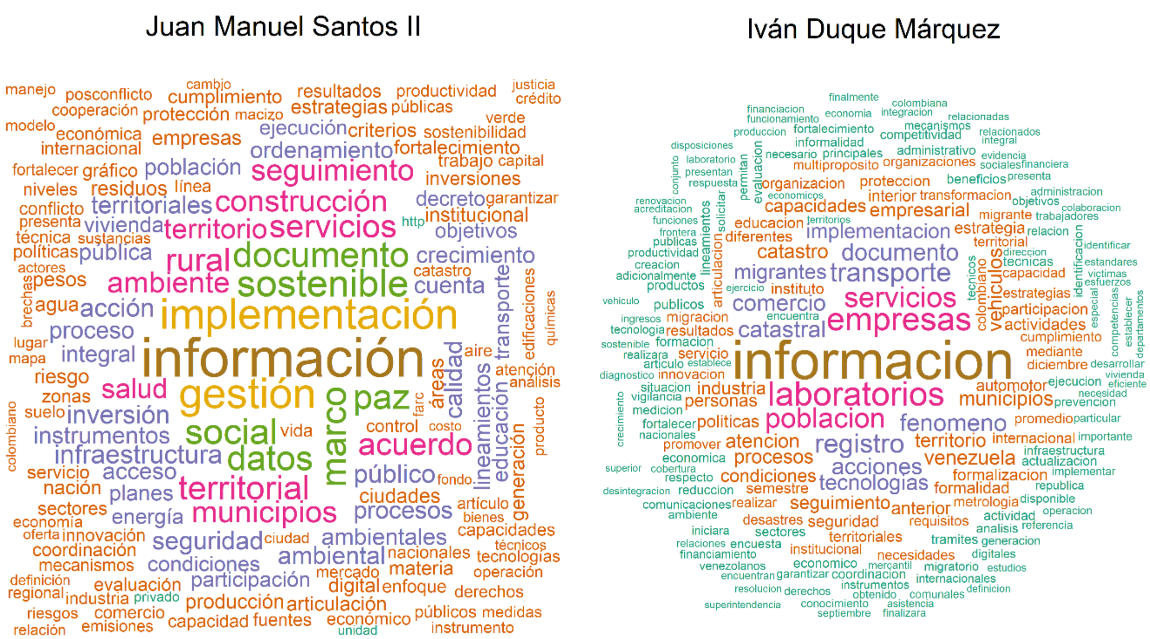
la suma de ambos puntajes, partiendo de que las medidas de exclusividad y representatividad tienen una interpretación similar (frecuencia de aparición o de "no aparición").

Finalmente, se tomaron las 25 palabras de cada sector con mayor "puntaje único" y se contó cuántas veces aparece cada palabra en el total de documentos CONPES publicados en un periodo de gobierno dado. En este caso se toman los documentos con lineamientos de política. Con estas frecuencias se realiza un ranking por frecuencia de aparición, por cada periodo de gobierno, y con este insumo se realizaron los gráficos.

Resultados
A partir del texto limpio se realizó el conteo de frecuencias de palabras y pares de palabras (bigramas). Con estos conteos de frecuencias se construyeron nubes de palabras como las que se presentan en la Figura 5, que corresponden a los dos últimos periodos de gobierno.

En estos resultados se encuentra información consistente con las grandes apuestas de cada gobierno. Por ejemplo, se encuentran frecuencias altas para el término "paz" en el segundo período del expresidente Juan Manuel Santos y frecuencias altas para el término "empresas" en el gobierno del presidente Iván Duque Márquez. Este tipo de análisis permitió validar que la calidad de los insumos es suficiente para que el análisis de tendencias sea viable.

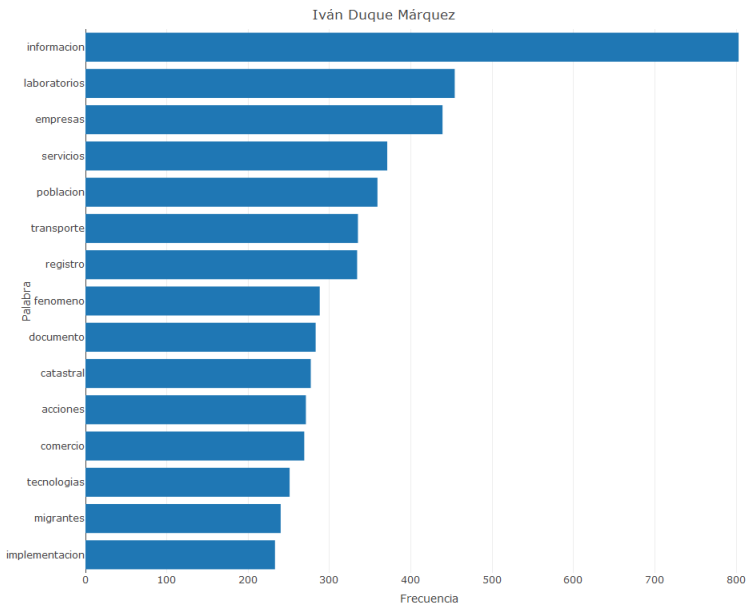
Figura 5: Nubes de palabras para los documentos CONPES de los 2 últimos gobiernos



Fuente: elaboración propia

De manera equivalente, esta información puede presentarse por medio de gráficos de barras, como se encuentra en la Figura 6.

Figura 6: Términos más frecuentes para los documentos CONPES del último periodo de gobierno



Fuente: elaboración propia

Para complementar el análisis, se construyeron grafos de co-ocurrencias de los términos contenidos en los documentos CONPES. Las co-ocurrencias indican el número de veces que aparecen un par de palabras en una misma página, por lo que dan una idea de qué tanta relación existe entre cada par de palabras. La construcción de un grafo con base en estas co-ocurrencias se realiza para presentar gráficamente dicha relación y, en este caso, para agrupar términos que puedan dar idea de temas similares.

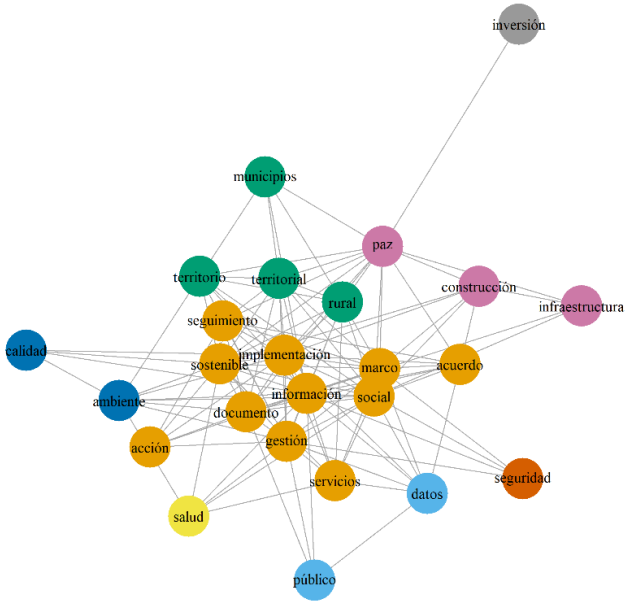
Para la construcción de cada grafo se tomaron los 25 términos con mayor frecuencia y las 150 relaciones más fuertes (valores más altos de co-ocurrencia). Sobre cada grafo se realizó luego una agrupación jerárquica (*clustering* jerárquico) con un corte de 8 grupos. Estos valores se escogieron de forma empírica, tras ensayar con distintos valores y observar las agrupaciones resultantes. Con la aplicación del anterior procedimiento se obtuvieron los grafos presentados en la Figura 7.

Los resultados de este análisis también son consistentes con las grandes apuestas de cada gobierno:

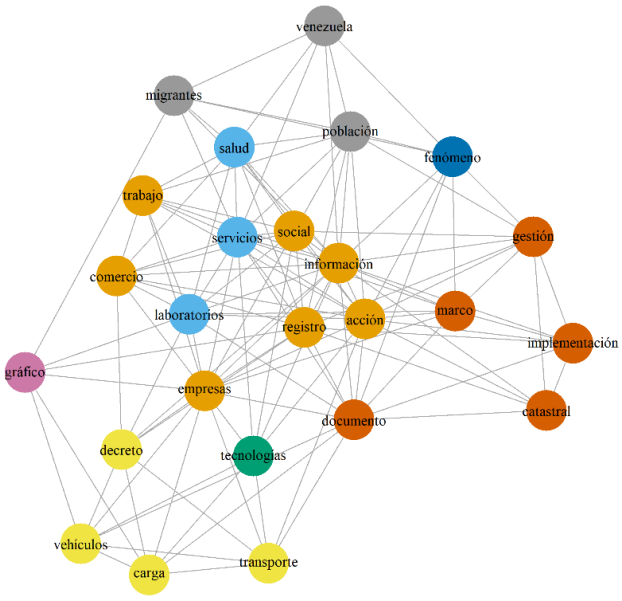
- En el segundo período del expresidente Juan Manuel Santos se encuentran un grupo que se relaciona con paz, territorio, ruralidad y con el plan marco de implementación, así como un grupo sobre gestión de datos e información (podría relacionarse con el CONPES de Big Data y con la política para la adopción e implementación de un catastro multipropósito rural-urbano, por ejemplo).
- En el gobierno del presidente Iván Duque Márquez se identifica un grupo sobre población migrante de Venezuela, otro sobre información catastral y registro y otro sobre transporte de carga.

Figura 7: Grupos de términos para los documentos CONPES de los dos últimos gobiernos

Juan Manuel Santos II



Iván Duque Márquez



Fuente: elaboración propia

Por otra parte, existen palabras representativas dentro de cada periodo de gobierno, pero que parecen no tener sentido al estar separadas, por esta razón se presentan nubes de palabras de bigramas (conjuntos de dos palabras) para cada

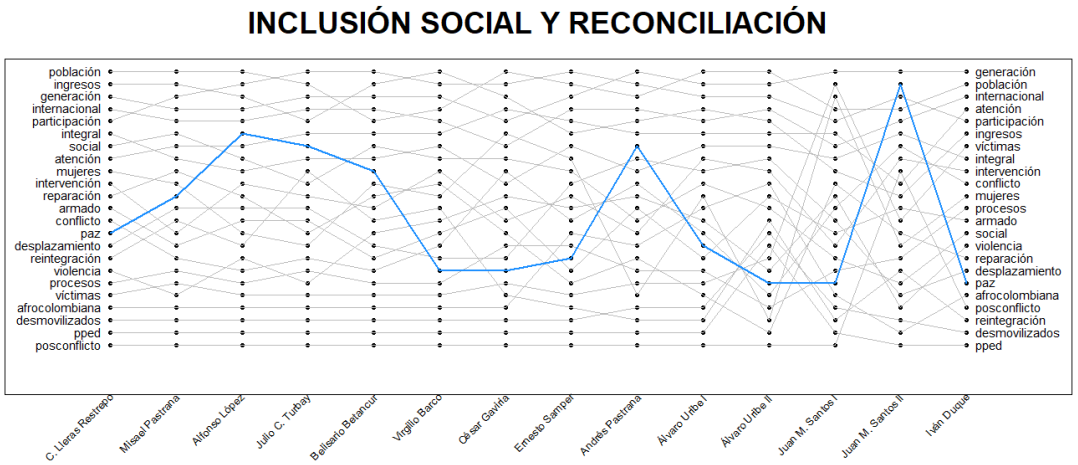
uno de los periodos de gobierno. En la Figura 8 muestra la nube de bigramas para el periodo de gobierno de Andrés Pastrana y el primer periodo de Álvaro Uribe Vélez.

Figura 8: Nubes de bigramas para los documentos CONPES del gobierno de Andrés Pastrana y Álvaro Uribe I

Fuente: elaboración propia

Finalmente, se realizó un análisis para los diferentes sectores, identificando los términos más representativos dentro de cada uno de ellos. Para ello, se tomaron los 25 términos más importantes por sector y se trazó el cambio de relevancia que tuvieron a lo largo del tiempo (medido en periodos de gobierno). En la Figura 9 se observa la evolución de los términos representativos para el sector Inclusión Social y Reconciliación, en este se puede observar que para los periodos de gobierno de Andrés Pastrana y Juan Manuel Santos II, el término "paz" presentó un pico de relevancia, lo cual resulta consistente con los diálogos de paz que se propiciaron durante sus gobiernos.

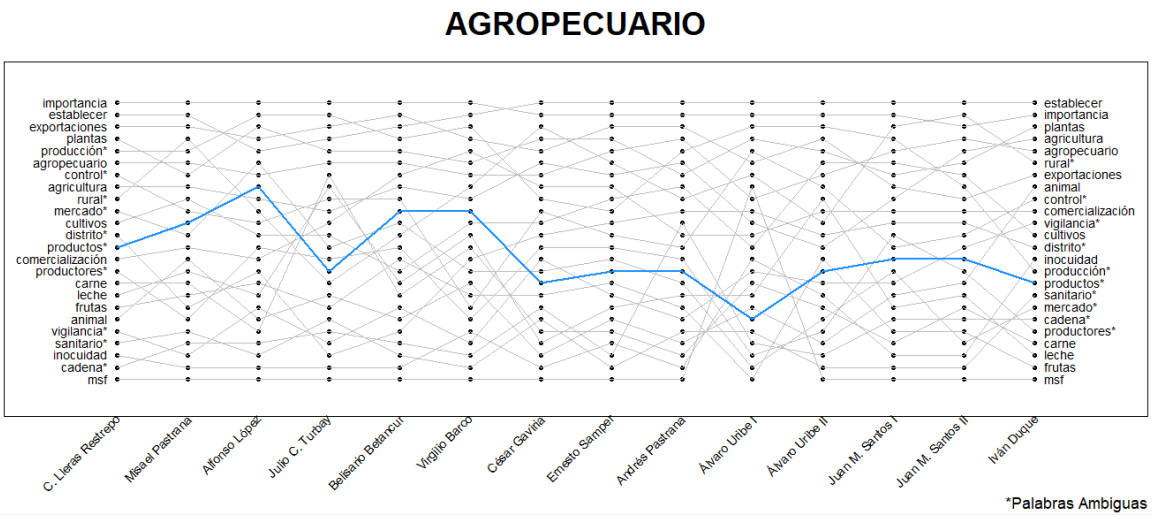
Figura 9: Gráfico de tendencias para el sector Inclusión Social y Reconciliación



Fuente: elaboración propia

Así mismo, y dada la ambigüedad que pueden presentar algunos términos que son representativos en diferentes sectores, se realizó una identificación de algunas palabras que se encuentran presentes en varios sectores con el fin de analizarlos por separado para cada sector, para identificar esto los términos considerados ambiguos se presentan acompañados con el nombre del sector. En la Figura 10 se puede observar resaltado el producto* que indica que es una variable ambigua, lo cual quiere decir que este término tuvo un análisis por separado para diferentes sectores en los que se encontró como relevante.

Figura 10: Gráfico de tendencias con términos ambiguos para el sector Agropecuario



Fuente: elaboración propia

Conclusiones y recomendaciones

1. Con la identificación automática de las palabras representativas de cada periodo de gobierno se puede evidenciar la diferencia del enfoque que dan los gobiernos a sus políticas y la relación de estas con el contexto histórico en el que se desarrollaron.
2. Los análisis por sectores permiten observar la relevancia principal de los términos para los gobiernos de forma más específica, lo cual puede facilitar un seguimiento a los planes de gobierno y su cumplimiento.
3. Con la herramienta de visualización desarrollada se facilita la lectura de los hallazgos del proyecto facilitando la toma de decisiones futuras.
4. Con los resultados de este proyecto se facilitará la labor de búsqueda de documentos CONPES por temas específicos, la identificación de los temas relevantes relacionados con la política pública durante los diferentes periodos de gobierno y de esta forma se podrá orientar la construcción futura de los PND y en general de las políticas públicas del país.

Socialización

Los resultados de este proyecto fueron presentados al grupo CONPES, a quienes se entregó la herramienta de visualización.