Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento
Nacional de Planeación

ESTUDIO DE COMPARACIÓN DEL CONPES DE REACTIVACIÓN ECONÓMICA CON LOS PLANES DE DESARROLLO TERRITORIAL (PDT) 2020-2023

Entidad
Departamento Nacional de Planeación
- Dirección de Desarrollo Digital.
- Dirección de Descentralización y Desarrollo Regional

Sector
Territorial

Lenguaje
Python

Fuente de datos
Documentos PDF del CONPES de reactivación económica y los Planes de Desarrollo Territorial

Contenido

1. Presentación

El documento CONPES de reactivación económica busca definir la estrategia para que la economía colombiana supere los problemas causados por la pandemia del COVID-19. En este contexto, la Dirección de Descentralización y Desarrollo Regional busca realizar un análisis de la vocación de inversión en municipios y departamentos comparando el CONPES de reactivación con los Planes de Desarrollo Territorial de las entidades colombianas. Adicionalmente, buscan comparar sectores económicos con los PDT y encontrar términos clave de distintos temas en cada Plan. La Unidad de Científicos de Datos está trabajando de manera conjunta con la DDDR para facilitar este trabajo con metodologías de procesamiento y análisis de texto de manera automatizada.

The CONPES document regarding economic reactivation defines Colombia's economic strategy to overcome the problems caused by the COVID-19 pandemic. Within this context, the Decentralization and Regional Development Directorate is studying the municipalities' and departments' investment vocation by comparing the reactivation CONPES with Colombia's Territorial Development Plans. Furthermore, they intend to compare different economic sectors with the PDT's and find keywords from different socioeconomic subjects in each Plan. The Data Scientists Unit in DNP is working alongside the DDDR to implement automated text processing and analysis techniques.

2. Objetivos del proyecto

2.1. General

Medir la inclusión de políticas de reactivación, crecimiento económico y de temas socioeconómicos en los Planes de Desarrollo Territorial de municipios y departamento de Colombia.

2.2. Específicos

1. Buscar términos clave que sean característicos del CONPES de reactivación económica en los PDT
2. Buscar términos clave que sean exclusivos del CONPES de reactivación económica
3. Comparar el CONPES de reactivación económica con los PDT con distintas medias de similitud y distancia
4. Buscar términos clave de distintos temas socio económicos dentro de los PDT
5. Comparar los textos de sectores económicos con los PDT con distintas medidas de similitud y distancia

3. Metodología

La metodología consiste en dos fases principales. En la primera se compara el CONPES de reactivación económica con otros 24 escogidos aleatoriamente, dos para cada sector, para identificar palabras exclusivas del CONPES de reactivación y palabras relevantes de este CONPES que pueden encontrarse en los otros documentos. La segunda parte muestra el análisis hecho sobre los PDT al compararlos con el CONPES de reactivación, con distintos sectores económicos y con la búsqueda de términos clave de temas socioeconómicos en los documentos.

3.1. Definición de los términos exclusivos y relevantes del CONPES de reactivación

El desarrollo de la metodología en esta fase consiste en 4 puntos: (i) definición y obtención de los CONPES que serán utilizados como comparación del CONPES de reactivación; (ii) extracción de los textos de los CONPES; (iii) limpieza y preprocesamiento de los textos; (iv) vectorización de los textos y obtención de los términos clave del CONPES de reactivación económica.

- ***Definición y obtención de los documentos CONPES***

El documento principal es un fragmento del borrador del CONPES de reactivación económica que fue entregado por miembros de la DDDR. Los CONPES de comparación se escogieron de forma aleatoria, de manera que hubiera dos para cada sector. Los sectores de cada CONPES son: transporte; cultura, deporte y recreación; educación; vivienda; agua potable y saneamiento básico; agricultura; inclusión social y reconciliación; ambiente y desarrollo sostenible; salud y protección social; minas y energía; telecomunicaciones; comercio, industria y turismo. Los documentos de comparación se obtuvieron de la página web de CONPES del DNP:
https://www.dnp.gov.co/CONPES/documentos-conpes/Paginas/documentos-conpes.aspx

- ***Extracción de los textos***

Antes de aplicar técnicas de minería de textos sobre los documentos de interés, es necesario extraer los textos que se encuentran en varios archivos (en este caso PDF) a un formato codificado que pueda ser leído desde Python. Este proceso puede ser demorado, por lo cual primero se extrae y guarda el texto de los archivos PDF en archivos de texto plano (.txt), de tal manera que pueda ser leído con más rapidez desde Python al momento de ser trabajados. Todos los documentos PDF de los CONPES se encuentran digitalizados, no escaneados, por lo cual no es necesario utilizar reconocimiento óptico de caracteres (OCR), lo cual puede generar errores dependiendo de la calidad de la imagen. Esto facilita la extracción de textos y su guardado en archivos de texto plano.

- ***Limpieza y preprocesamiento de los textos***

Para aplicar los métodos de análisis sobre los textos es necesario quitar los elementos que se consideran como ruido, es decir, quitar términos que no aportan información importante tales como preposiciones, artículos,

adjetivos y en ocasiones números, nombres y apellidos. Adicionalmente, para el preprocesamiento de los textos se deja todo el texto en minúsculas, se quitan las tildes y signos de puntuación, esto se hace con el objetivo de dejar el texto lo más uniforme posible.

- ***Vectorización y búsqueda de términos clave***

Sobre el texto limpio preprocesado de todos los documentos CONPES se construye una matriz base de términos y documentos (*Bag of Words*) M, cuyas columnas corresponden a todas las palabras que aparecen en todos los documentos, cada fila corresponde a un documento CONPES y el elemento ij-ésimo a la frecuencia de aparición de la palabra j en el documento i. Esta matriz se crea tanto para palabras como para bigramas (dúos de palabras) y trigramas (tríos de palabras). A partir de la matriz *Bag of Words* se siguen los siguientes pasos para obtener los términos clave exclusivos y relevantes:

a) Se escogen los términos más frecuentes que aparecen únicamente en el CONPES de reactivación económica. Estos son los términos exclusivos.
b) Se separa la matriz M en dos: la correspondiente al texto del documento CONPES de reactivación económica (MR), por un lado, y a la de los CONPES de comparación (MC), por el otro.
c) Aplicando análisis de componentes principales, se establece el espacio base de MC, con su correspondiente matriz de proyección UC. Se puede interpretar este espacio como el de "documentos CONPES de distintos temas".
d) Se encuentra la componente de MR que se proyecta a la base de MC. Algebraicamente, esto corresponde a $MRC = MR\,UC\,UC^T$. Cada fila de MRC se puede interpretar como la parte de cada documento que es común al espacio de "documentos CONPES".
e) Se encuentra la proyección de MR al espacio nulo de UC: $MR_0 = \mathrm{MR} - \mathrm{MRC}$. Cada fila de esta matriz se puede interpretar como la parte de cada documento que es exclusiva de "documento del CONPES de reactivación económica".
f) Los términos (columnas de MR_0) que tienen mayor valor esperado se escogen como pertenecientes predominantemente al CONPES de reactivación económica. Estos son los términos relevantes que podrían buscarse luego en los PDT.

- ***Generación de los archivos consolidados***

Luego de definir las palabras, bigramas y trigramas exclusivos y relevantes del CONPES de reactivación económica, se exportan los resultados a un Excel con los 50 términos más relevantes para cada ngrama mencionado.

3.2. Análisis de los Planes de Desarrollo Territorial

En esta sección se muestra las metodologías de análisis de PDT, con métodos de comparación de texto automatizados y búsqueda de términos clave en los documentos. Se presenta en 7 fases: (i) definición y obtención de los PDT que se van a analizar en formato PDF; (ii) extracción de los textos, donde se usa OCR para los PDF que se encuentran escaneados; (iii) limpieza y preprocesamiento de los textos; (iv) vectorización de los textos y cálculos de medidas de similitud; (v) generación de tablas de frecuencias de términos clave; (vi) revisión de la calidad textos extraídos de los PDT; (vii) generación de archivos de Excel con los resultados

- ***Obtención de los documentos CONPES y PDT.***

El documento principal es un fragmento del borrador del CONPES de reactivación económica que fue entregado por miembros de la DDDR. Los PDT de comparación se obtuvieron del repositorio de la UCD. Estos han sido recolectados a lo largo del año y en el momento se cuenta con 991 planes de desarrollo municipales y 32 departamentales.

- ***Extracción de los textos***

Para aplicar cualquier tipo de procesamiento en los PDT y el CONPES de reactivación, es necesario extraer el texto de los archivos PDF a un formato codificable en Python. Como el proceso de extracción puede ser demorado, se extraen los textos y se guardan en archivos de texto plano (".txt"), desde donde se pueden leer con más rapidez con Python. La UCD ha desarrollado la librería ConTexto, la cual permite extraer textos desde PDF digitalizados con métodos convencionales, pero también se puede extraer texto desde documentos escaneados utilizando reconocimiento óptico de caracteres (OCR). Uno de los desafíos que se presentan es que el OCR en algunos casos no siempre extrae el texto de manera precisa debido a la calidad de la imagen (texto escaneado), por lo cual se pueden introducir caracteres ruidosos. Asimismo, es posible que haya textos que no se pueden extraer satisfactoriamente.

- ***Limpieza y preprocesamiento de los textos***

Para aplicar los métodos de análisis sobre los textos es necesario quitar los elementos que se consideran como ruido, es decir, quitar términos que no aportan información importante tales como preposiciones, artículos, adjetivos y en ocasiones números, nombres y apellidos. Debido a las imprecisiones del OCR algunos textos van a tener palabras inexistentes y se recomienda también eliminarlas en caso de ser recurrentes. Adicionalmente, para el preprocesamiento de los textos se deja todo el texto en minúsculas, se quitan las tildes y signos de puntuación, con el objetivo de dejar el texto lo más uniforme posible.

Durante el preprocesamiento también se implementaron las metodologías de *stemming* y lematización. El primero consiste en la reducción de palabras similares o relacionadas entre sí a una única raíz y el segundo a reducir las palabras a una que represente sus formas flexionadas. Por ejemplo, las palabras "jugando", "jugarán", "jugaron", "jugarían" podrían transformarse a "juga" con la metodología de *stemming* y a "jugar" con la metodología de lematización.

Por esta razón, y para no tener que volver a correr los scripts de preprocesamiento, se guardaron tres tipos de textos de PDT preprocesados en archivos ".txt": (1) con el preprocesamiento sin *stemming* ni lematización, (2) con el preprocesamiento con *stemming* y (3) con el preprocesamiento con lematización.

- ***Vectorización y cálculo de medidas de similitud***

La vectorización de textos consiste en representar los textos preprocesados en formatos numéricos para poder realizar comparaciones entre las representaciones numéricas de los textos. En este proyecto se comparó el CONPES de reincorporación con el resto de PDT, luego de hacer la vectorización, y también se comparó cada uno de 12 textos cortos de sectores económicos, enviados por la DDDR, con cada PDT. Los sectores son los siguientes:

1. Agricultura, ganadería, caza, silvicultura y pesca
2. Explotación de minas y canteras
3. Industrias manufactureras
4. Suministro de electricidad, gas, vapor y aire acondicionado; Distribución de agua; evacuación y tratamiento de aguas residuales, gestión de desechos y actividades de saneamiento ambiental
5. Construcción
6. Comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas; Transporte y almacenamiento; Alojamiento y servicios de comida
7. Información y comunicaciones
8. Actividades financieras y de seguros
9. Actividades inmobiliarias
10. Actividades profesionales, científicas y técnicas; Actividades de servicios administrativos y de apoyo
11. Administración pública y defensa; planes de seguridad social de afiliación obligatoria; Educación; Actividades de atención de la salud humana y de servicios sociales

12. Actividades artísticas, de entretenimiento y recreación y otras actividades de servicios; Actividades de los hogares individuales en calidad de empleadores; actividades

En la comparación del CONPES con el resto de textos se incluyeron vectorizaciones TF-IDF y Doc2Vec, además de metodologías de lematización y *stemming* en el preprocesamiento, mientras que en la comparación de cada sector económico con los PDT se utilizó únicamente TF-IDF y sin lematización ni *stemming*. La vectorización es necesaria para poder representar los textos en una manera numérica para poder compararlos luego. En el caso de TF-IDF, se puede hacer tanto con unigramas (palabras) como bigramas (dos palabras) o ngramas (n palabras), dependiendo del ejercicio que se quiere realizar. La metodología de TF-IDF crea un indicador de frecuencias por ngrama para cada documento, el cual pondera con una mayor puntuación los términos que aparecen en pocos documentos. En este proyecto la vectorización de textos con TF-IDF se hizo con unigramas, bigramas y trigramas para incluir términos que tienen más sentido cuando se componen de varias palabras. Es decir, se creó una vectorización TF-IDF que incluye únicamente palabras, otra que incluye palabras y bigramas y una última que incluye palabras, bigramas y trigramas. Adicionalmente, al hacer la vectorización se tuvieron en cuenta los 20.000 términos más frecuentes del conjunto de todos los textos, para eliminar aquellos considerados menos importantes.

Doc2Vec es un algoritmo que usa un modelo de redes neuronales que permite generar representaciones vectoriales de oraciones, párrafos o documentos enteros. Es una ampliación del modelo Word2Vec, el cual aplica para generar vectores de palabras. En este proyecto se generaron dos vectorizaciones con Doc2Vec. En la primera se definieron vectores con un tamaño de 100 elementos para cada documento y el modelo se entrenó con 20 épocas (el número de veces que se actualizan los parámetros en el modelo de redes neuronales). Para el segundo modelo los vectores se definieron con 300 elementos y el entrenamiento tuvo 40 épocas, por lo que se demora más en correr y es, en teoría, más preciso. Ambas vectorizaciones Doc2Vec incluyen únicamente los términos que aparecen al menos 5 veces en los documentos.
Llevar los textos a un espacio vectorial n-dimensional permite aplicar cálculos matemáticos convencionales tales como similitud coseno o distancia euclidiana, entre otros. En este proyecto se calcularon las medidas de similitud coseno y distancia euclidiana del CONPES de reactivación con el resto de PDT con cinco vectorizaciones: tres vectorizaciones de TF-IDF, una con palabras, otra con palabras y bigramas y una última con palabras, bigramas y trigramas; dos vectorizaciones Doc2Vec, una con los tamaños de los vectores con 100 elementos y 20 iteraciones (épocas) de la red neuronal y otra con los vectores de 300 elementos y 40 iteraciones de la red neuronal.

Es importante recordar que en el preprocesamiento de los textos se aplicaron técnicas de *stemming* y lematización, por lo cual se tienen tres tipos de textos preprocesados: uno con el preprocesamiento que no aplica estas dos técnicas, otra con *stemming* y una última con lematización. Por lo tanto, los resultados de similitud coseno y distancia euclidiana se calculan con las cinco vectorizaciones mencionadas arriba para cada uno de los tres tipos de preprocesamiento. Es decir, se calcularon en total 30 medidas de comparación del CONPES de reactivación con los PDT.

En el caso de la comparación de los sectores económicos con los PDT, se calcularon las medidas de similitud coseno y distancia euclidiana con las mismas tres vectorizaciones TF-IDF mencionadas arriba, con palabras, bigramas y trigramas, pero no con Doc2Vec. Por lo tanto, cada uno de los 12 sectores económicos tiene 6 medidas de similitud.
El indicador de la similitud coseno muestra qué tanto se parece un documento con otro. En este caso, entre más grande sea el indicador, habrá una similitud mayor entre el CONPES de reactivación con el texto del PDT. Por su lado, la distancia euclidiana muestra qué tanto difieren los textos estudiados. Entre menor sea la distancia, más similitud tendrán.

- ***Tablas de frecuencias de términos clave***

Por parte de la DDDR se escogieron términos clave de diversos temas para ser buscados dentro de los PDT. Estos incluyen los términos exclusivos y relevantes del CONPES de reincorporación (definidos en el Entregable 1) y aquellos relacionados con los siguientes temas: "Crecimiento verde", "Enfoque territorial", "Desarrollo rural", "Educación", "Primera infancia y adolescencia", "Equidad", "Equidad de género", "Salud", "Empleo y trabajo", "Derechos", "Paz", "Víctimas", "Transparencia y lucha contra la corrupción", "Participación ciudadana", "Seguridad", "Justicia", "Asuntos étnicos", "Ordenamiento territorial", "Adulto mayor", "Discapacidad", "Agua y saneamiento básico", "Cultura", "Recreación y deporte Innovación y competitividad", "Recursos", "Vivienda", "Planeación y seguimiento", "Sector minero-energético".

Los términos clave se buscan en cada párrafo y los resultados muestran el número de párrafos que contienen el término dentro del PDT. Se escogen, si posible, las raíces de los términos para ampliar la búsqueda. Por ejemplo, el término "sostenib" se puede encontrar en un documento si contiene las palabras "sostenible", "sostenibles" o "sostenibilidad" (antes de la búsqueda todos los textos se convierten a minúsculas y se eliminan tildes). Adicionalmente, los términos con más de una palabra se separan por el carácter "|". Esto significa que los términos que separa este carácter se buscan en cada párrafo del documento y en caso de que todos se encuentren en el párrafo se sumará a la frecuencia total. Por ejemplo, se considera que "desarrollo|sostenib" se encuentra en un párrafo si ese párrafo contiene texto como "desarrollo sostenible", "desarrollos sostenibles", "desarrollo y sostenibilidad", etc. Los resultados de la tabla de frecuencia se interpretan como el número de párrafos que contienen el término clave de interés.

- ***Revisión de la calidad textos extraídos de los PDT***

Con el fin de verificar la calidad del texto extraído de los PDT (el texto del CONPES se extrajo con éxito), se creó una tabla que cuenta el número de veces que aparecen las palabras *"de", "la", "con", "y", "a"* en cada PDT por página. Al considerar que estas palabras son esenciales dentro de un texto escrito en español, un PDT con un nivel bajo de estas palabras por página tiene una probabilidad alta de ser de mala calidad. Si bien depende de la manera como se escribió el documento, deberían tener alrededor de 10 palabras por página como mínimo.

- ***Generación de los archivos consolidados***

Los resultados con las distintas medidas de comparación se exportaron a tablas de archivos de Excel para poder ser evaluados con facilidad. En la sección de Resultados se explicará a profundidad cada uno de estos archivos

4. Resultados

Los resultados se separan en cinco: (1) los resultados de la comparación del CONPES de reactivación con los PDT con medidas de distancia y similitud, (2) la comparación de los 12 sectores económicos con cada PDT con las medidas de distancia y similitud, (3) la búsqueda de términos clave relevantes y exclusivos del CONPES de reactivación en los PDT, (4) la búsqueda de términos clave por tema, enviados por la DDDR, en los PDT y (5) la verificación de calidad del texto de los PDT.

4.1. Comparación con medidas de similitud y distancia

Hay tres archivos de Excel con los resultados de las medidas de comparación del texto del CONPES con cada PDT: uno para los resultados en los cuales no se utilizó *stemming* ni lematización en el preprocesamiento, otro en el cual se usó *stemming* y el último con lematización. Son los siguientes:

- **Consolidado_Medidas.xlsx**: preprocesamiento sin *stemming* ni lematización
- **Consolidado_Medidas_Lematizacion.xlsx**: preprocesamiento con lematización
- **Consolidado_Medidas_Stemming.xlsx**: preprocesamiento con *stemming*

Cada archivo de Excel tiene cinco pestañas, que corresponden al método de vectorización de los textos:

- **TFIDF**: contiene los resultados con la vectorización TF-IDF para palabras

- **TFIDF_Palabras_Bigramas**: contiene los resultados con la vectorización TF-IDF para palabras y bigramas
- **TFIDF_Palabras_Trigramas**: contiene los resultados con la vectorización TF-IDF para palabras, bigramas y trigramas
- **Doc2Vec**: contiene los resultados con la vectorización Doc2Vec con 100 elementos en cada vector y 20 épocas
- **Doc2Vec_2**: contiene los resultados con la vectorización Doc2Vec con 300 elementos en cada vector y 40 épocas

Cada pestaña contiene cinco columnas:

- **Codigo**: código DIVIPOLA de la entidad (municipio o departamento)
- **Nombre**: nombre de la entidad
- **Similitud Coseno**: medida de similitud coseno entre el PDT de la entidad con el CONPES de reactivación
- **Distancia Euclidiana:** medida de distancia euclidiana entre el PDT de la entidad con el CONPES de reactivación

En la *Figura 1* se muestra la pestaña con los resultados de las comparaciones con la vectorización TF-IDF del archivo de Excel "Consolidado_Medidas.xlsx**"**. También se ven en la imagen las distintas pestañas con las vectorizaciones.

Figura 1: Ejemplo de resultados de comparación en Excel

	A	B	C	D	E
1	Codigo	Nombre	Priorizado	Similitud Coseno	Distancia Euclidiana
2	5001	MEDELLÍN	Si	0,358938525	1,132308682
3	5002	ABEJORRAL	No	0,340373655	1,148587258
4	5004	ABRIAQUÍ	No	0,33790324	1,150736077
5	5021	ALEJANDRÍA	No	0,304572446	1,179345203
6	5030	AMAGÁ	No	0,217400025	1,251079514
7	5031	AMALFI	No	0,214058137	1,253747872
8	5034	ANDES	No	0,295442658	1,187061365
9	5036	ANGELÓPOLIS	No	0,335119195	1,1531529
10	5038	ANGOSTURA	No	0,446744877	1,051907907
11	5040	ANORÍ	Si	0,264202879	1,213092842
12	5044	ANZÁ	No	0,286636271	1,194456972
13	5051	ARBOLETES	No	0,240280321	1,232655409
14	5055	ARGELIA	No	0,346986352	1,142815513
15	5059	ARMENIA	No	0,34296154	1,146331941
16	5079	BARBOSA	No	0,39639578	1,098730377
17	5086	BELMIRA	No	0,250263167	1,224529978
18	5088	BELLO	Si	0,334093063	1,154042405
19	5091	BETANIA	No	0,414564738	1,082067708
20	5093	BETULIA	No	0,169632577	1,288695016
21	5101	CIUDAD BOLÍVAR	No	0,434052704	1,063905349

TFiDF | TFiDF_Palabras_Bigramas | TFiDF_Palabras_Trigramas | Doc2vec | Doc2vec_2

Fuente: elaboración propia

4.2. Comparación de sectores económicos con cada PDT

Los resultados de la comparación de cada sector económico con los PDT se encuentran en 12 archivos de Excel, uno para cada sector. Tienen el siguiente nombre:

- **Sectores_Consolidado_Medidas_#.xlsx**

Donde "#" representa el número del sector económico, definido a continuación:

1. Agricultura, ganadería, caza, silvicultura y pesca
2. Explotación de minas y canteras
3. Industrias manufactureras
4. Suministro de electricidad, gas, vapor y aire acondicionado; Distribución de agua; evacuación y tratamiento de aguas residuales, gestión de desechos y actividades de saneamiento ambiental
5. Construcción
6. Comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas; Transporte y almacenamiento; Alojamiento y servicios de comida
7. Información y comunicaciones
8. Actividades financieras y de seguros
9. Actividades inmobiliarias
10. Actividades profesionales, científicas y técnicas; Actividades de servicios administrativos y de apoyo
11. Administración pública y defensa; planes de seguridad social de afiliación obligatoria; Educación; Actividades de atención de la salud humana y de servicios sociales
12. Actividades artísticas, de entretenimiento y recreación y otras actividades de servicios; Actividades de los hogares individuales en calidad de empleadores; actividades

Cada archivo tiene las siguientes tres pestañas:

- **TFIDF**: contiene los resultados con la vectorización TF-IDF para palabras
- **TFIDF_Palabras_Bigramas**: contiene los resultados con la vectorización TF-IDF para palabras y bigramas
- **TFIDF_Palabras_Trigramas**: contiene los resultados con la vectorización TF-IDF para palabras, bigramas y trigramas

Cada pestaña contiene cinco columnas:

- **Codigo**: código DIVIPOLA de la entidad (municipio o departamento)
- **Nombre**: nombre de la entidad
- **Similitud Coseno**: medida de similitud coseno entre el PDT de la entidad con el sector económico correspondiente
- **Distancia Euclidiana:** medida de distancia euclidiana entre el PDT de la entidad con el sector económico correspondiente

4.3. Frecuencia de términos relevantes y exclusivos del CONPES de reactivación en los PDT

Las tablas de frecuencia de los términos del exclusivos y relevantes del CONPES de reactivación se encuentran en el siguiente archivo:

- **Frecuencia_Terminos_Exclusivos_Relevantes.xlsx**

Tiene dos pestañas:

- **Exclusivos**: los términos exclusivos que se buscaron en los PDT y que fueron verificados por la DDDR
- **Relevantes**: los términos relevantes que se buscaron en los PDT y que fueron verificados por la DDDR

Cada pestaña contiene las siguientes columnas:

- **Codigo**: código DIVIPOLA de la entidad
- **Nombre**: nombre de la entidad
- **Términos clave**: una columna para cada uno de los términos clave que muestra el número de párrafos donde se encuentra el término o los términos (cuando se busca más de un término).
- **suma_frecuencias**: suma total de las frecuencias de términos clave
- **indicador_max_min**: indicador que normaliza la suma de frecuencias entre 1 y 0, donde 1 es el PDT con mayor número de párrafos con términos clave y 0 el PDT con menor número.

En la *Figura 2* se ve el archivo de Excel con una muestra de la tabla de frecuencias de términos clave:

Figura 2: Tabla de frecuencias de términos clave – exclusivos y relevantes

	A	B	C	D	E	F	G	H
1	codigo	nombre	acceso|financiamiento	aislamiento|preventivo	analisis|riesgo|desastre	articulacion|institucional	bioeconomia	bioseguridad
2	52001	PASTO	2	9	0	19	0	26
3	5	ANTIOQUIA	0	5	0	6	0	2
4	54001	CÚCUTA	1	5	0	7	0	8
5	91	AMAZONAS	0	2	0	2	0	25
6	25181	CHOACHÍ	0	3	0	0	0	4
7	5001	MEDELLÍN	0	2	0	21	0	3
8	25430	MADRID	1	1	0	4	0	2
9	25785	TABIO	0	1	0	1	0	4
10	76834	TULUÁ	0	0	0	7	0	0
11	70	SUCRE	1	0	0	4	0	14
12	76130	CANDELARIA	0	4	1	2	0	0
13	15176	CHIQUINQUIRÁ	0	0	0	0	0	7
14	5670	SAN ROQUE	0	1	0	0	0	14
15	73026	ALVARADO	1	0	2	1	0	6
16	94	GUAINÍA	0	6	0	9	0	5
17	5361	ITUANGO	0	2	0	3	0	7
18	68705	SANTA BÁRBARA	0	2	0	0	0	6
19	73	TOLIMA	2	4	0	5	1	6
20	52788	TANGUA	0	0	0	3	0	0
21	68498	OCAMONTE	0	2	0	1	0	6

Fuente: elaboración propia

4.4. Frecuencia de términos por tema

El archivo de Excel con los términos clave por cada tema enviado por la DDDR es el siguiente:

- **Frecuencia_Terminos_Clave_Tema.xlsx**

Tiene un total de 27 pestañas, una por tema que contiene los términos clave. Las pestañas son las siguientes:

- "Crecimiento verde", "Enfoque territorial", "Desarrollo rural", "Educación", "Primera infancia y adolescencia", "Equidad", "Equidad de género", "Salud", "Empleo y trabajo", "Derechos", "Paz", "Víctimas", "Transparencia y lucha contra la", "Participación ciudadana", "Seguridad", "Justicia", "Asuntos étnicos", "Ordenamiento territorial", "Adulto mayor", "Discapacidad", "Agua y saneamiento básico", "Cultura", "Recreación y deporte", "Innovación y competitividad", "Recursos", "Vivienda", "Planeación y seguimiento", "Sector minero-energético"

Cada pestaña tiene las siguientes columnas:

- **Codigo**: código DIVIPOLA de la entidad
- **Nombre**: nombre de la entidad
- **Términos clave**: una columna para cada uno de los términos clave que muestra el número de párrafos donde se encuentra el término o los términos (cuando se busca más de un término).

4.5. Revisión de la calidad textos extraídos de los PDT

El archivo de Excel con estos resultados es el siguiente:

- **Revision_PDT_palabras_basicas.xlsx**

Tiene una sola pestaña con las siguientes columnas:

- **Codigo**: código DIVIPOLA de la entidad
- **Nombre**: nombre de la entidad
- **Frecuencia de palabras básicas por página**: muestra de la frecuencia total de las palabras “de”, “la”, “con”, “y”, “a” dividida por página.

En caso de observar alguna inconsistencia con los resultados, sería bueno revisar este archivo, dado que pudo haberse dado por la mala calidad de la extracción de los textos. En la *Figura 3* se observa una imagen con esta tabla.

Figura 3: Tabla con frecuencia de palabras básicas por página de los PDT

	Código	nombre	Frecuencia de palabras básicas por página
2	5631	SABANETA	0
3	73275	FLANDES	0
4	19256	EL TAMBO	0,010638298
5	25175	CHÍA	0,353448276
6	85430	TRINIDAD	0,379084967
7	54347	HERRÁN	0,408256881
8	25513	PACHO	0,444444444
9	86757	SAN MIGUEL	0,5
10	66045	APÍA	0,542986425
11	13244	EL CARMEN DE BOLÍVAR	0,611764706
12	25	CUNDINAMARCA	0,648093842
13	81794	TAME	0,794212219
14	76109	BUENAVENTURA	0,815286624
15	15114	BUSBANZÁ	0,818181818
16	15806	TIBASOSA	0,824884793
17	95015	CALAMAR	0,987730061
18	68101	BOLÍVAR	1
19	5789	TÁMESIS	1,093283582
20	25436	MANTA	1,104166667

Sheet1

Fuente: elaboración propia

5. Conclusiones y recomendaciones

1. Los resultados pueden utilizarse como insumo para otros proyectos relacionados con los PDT
2. A partir de las tablas de frecuencias se pueden crear indicadores y sumas con todos o algunos de los términos buscados. Dado que las tablas están en un formato de Excel, se pueden hacer varios cálculos y crear otros resultados que puedan ser utilizados como insumos en otros proyectos.
3. Se recomienda revisar y validad los indicadores de las metodologías de comparación hechas con distintas vectorizaciones, para escoger la que mejor se adecúa a las necesidades de la DDDR
4. Los resultados de las comparaciones con procesamiento de lenguaje natural tienen que ser revisadas con criterio experto para validar si corresponden al análisis deseado

6. Socialización

Los resultados fueron enviados a los miembros de la DDDR.