Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento
Nacional de Planeación

SIMILITUDES EN COMPETENCIAS DE ENTIDADES PÚBLICAS MEDIANTE ANÁLISIS DE TEXTOS

Entidad

Departamento Nacional de Planeación

- Dirección de Desarrollo Digital.
- Dirección de Descentralización y Desarrollo Regional

Otras entidades

Sector	Lenguaje	Fuente de datos
Planeación	Python	Mapeo general de funciones de entidades provisto por la SDFFDDDR

Presentación

Actualmente, en Colombia, existe una multiplicidad de leyes y actores asumiendo responsabilidades similares entre diferentes niveles de gobierno. Asimismo, la asignación de competencias no necesariamente se acompaña de una asignación de recursos y tampoco tiene en cuenta las capacidades de las entidades territoriales (que son heterogéneas en términos de ingresos y gestión). Se identifica que en Colombia las competencias de los gobiernos subnacionales son supremamente complejas y aunque existen servicios compartidos entre nación – territorio, mecanismos de devolución y asignación de responsabilidades (OECD, 2019) es difícil determinar las responsabilidades y alcances de cada actor.

Adicionalmente, la calidad y la eficiencia del gasto necesitan ser fortalecidos, no solo para procurar la sostenibilidad fiscal, sino para asegurar la entrega de bienes y servicios que genere mayor impacto en el bienestar, especialmente en un contexto de emergencia sanitaria y sus efectos en la economía nacional. Dado lo anterior, se cuenta con la necesidad de ordenar, sistematizar, armonizar y si es el caso replantear el arreglo institucional que sustenta las competencias que tienen los gobiernos subnacionales con el objetivo de afrontar mejor estos desafíos.

Desde el DNP se ha realizado un mapeo de las competencias de acuerdo con la normatividad vigente, en el cual se detalla la competencia, el servicio o bien público asociado y las entidades del nivel nacional o territorial que tienen competencia, especificando la norma que asigna esta competencia (Se incluyen competencias relacionadas con planeación territorial, minería, turismo, cultura, defensa y otros). Esta matriz, de una extensión de más de 2.000 competencias, presenta una información valiosa que permite tener un mapeo general de cómo se distribuyen las funciones entre las entidades del nivel nacional y las entidades territoriales.

Sin embargo, dada su extensión y diversidad entre las diferentes competencias, su análisis resulta complejo con las herramientas actuales. Razón por la cual se plantea realizar un análisis mediante técnicas de minería de texto que permita hacer diferentes lecturas de la información disponible, contribuya a la definición de competencias, identifique los tipos de relaciones entre entidades y niveles de gobierno, entre otras.

Currently, in Colombia, there is a multiplicity of laws and actors assuming similar responsibilities on different levels of the government. Likewise, competencies allocation is not necessarily accompanied by resource allocation, nor does consider the capacities of the territorial entities (which are heterogeneous in terms of income and management). It has been identified that in Colombia the competencies of sub-national governments are supremely complex and although

there are shared services between nation - territory, mechanisms for devolution and allocation of responsibilities (OECD, 2019) it is difficult to determine the responsibilities and scope of each actor.

Additionally, the quality and efficiency of spending need to be strengthened, not only to ensure fiscal sustainability, but also to ensure the delivery of goods and services that generate the greatest impact on well-being, especially in the context of health emergencies and their effects on the national economy. Given the above, there is a need to organize, systematize, harmonize and, if necessary, rethink the institutional arrangement that supports the competencies of subnational governments in order to better address these challenges.

The DNP has carried out a mapping of competencies in accordance with current regulations, which details the competence, the associated service or public good and the entities at the national or territorial level that have competence, specifying the regulation that assigns this competence (This includes competencies related to territorial planning, mining, tourism, culture, defense and others). This matrix, with an extension of more than 2,000 competencies, presents valuable information that allows for a general mapping of how the functions are distributed between the entities at the national level and the territorial entities.

However, given its extension and diversity among the different competencies, its analysis is complex with the current tools. Therefore, an analysis using text mining techniques is proposed, which allows different readings of the available information, contributes to the definition of competencies, identifies the types of relationships between entities and levels of government, among others.

Objetivo general

Explorar la asignación de competencias entre las entidades del nivel nacional y territorial a través de técnicas de análisis de texto, teniendo en cuenta los diferentes sectores, entidades y niveles de gobierno registrados en la base de datos con el propósito de identificar posibles ajustes a la asignación de competencias.

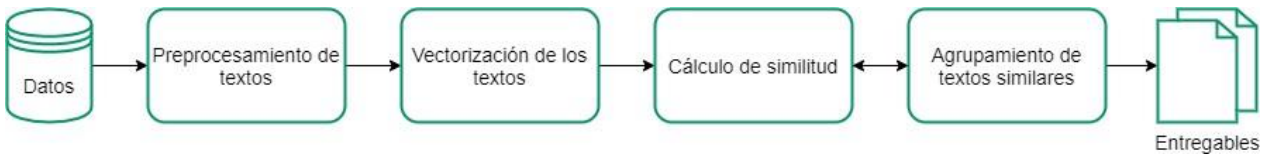
Objetivos específicos

1. Realizar lectura y limpieza de los textos para el posterior análisis de estos.
2. Transformar los textos en una representación numérica que permita realizar el procesamiento y análisis de diferencias y similitudes entre estos.
3. Calcular similitudes de textos entre las diferentes competencias de las entidades y sectores, e identificar posibles grupos de acuerdo con la similitud entre estas.

Metodología

La metodología propuesta para la ejecución de este proyecto está compuesta por cuatro etapas como se muestra en la Figura 1.

Figura 1: Etapas de ejecución del proyecto



Fuente: Elaboración propia.

Datos

Los datos de insumo del proyecto corresponden a un levantamiento de información realizado por la Subdirección de Descentralización y Fortalecimiento Fiscal (SDFF) de la Dirección de Descentralización y Desarrollo Regional (DDDR),

en este se detallan las funciones de entidades a nivel nacional y territorial, el sector al cual pertenecen y otros. Los datos de insumos consisten en una tabla de 9 columnas y 2.788 registros. Las columnas o atributos de los registros son:

- ***Código Competencia***, corresponde a un código de identificación asociado a la descripción de la competencia.
- ***Sector***, corresponde al sector al que pertenece la competencia, este atributo tiene 26 posibles valores.
- ***Competencia (según normativa vigente)***, es el atributo de mayor interés en el proyecto, sobre este se realizará el procesamiento y análisis de textos. Corresponde a la descripción de la competencia de las entidades.
- ***Competencia – resumen***, como su nombre lo sugiere, es un resumen de la información contenida en el atributo "*Competencia (según normativa vigente)".*
- ***Nivel de competencia***, corresponde al alcance de la entidad, puede ser "Territorial" o "Nacional".
- ***Entidad***, corresponde al nombre de la entidad registrada, se tienen 142 valores diferentes. Cabe mencionar que hay algunas entidades genéricas como son "Departamento", "Distrito", "Municipio" y otros.
- ***Norma que otorga competencia a la entidad***, corresponde al número de artículo constitucional o número y año de decreto o ley que otorga la competencia a la entidad.
- ***Otra norma aplicable***, se incluyen aclaraciones o información adicional referente a la norma que otorga la competencia a la entidad. El 95% de los registros no tienen información.
- ***Observaciones***, campo de texto abierto para hacer aclaraciones. Contiene muchos registros sin información. El 96% de los registros no tienen información.

En la Tabla 1 se presenta una muestra de los datos de insumo disponibles, se debe mencionar que el procesamiento y análisis de texto se realizó sobre la descripción contenida en el atributo ***Competencia (según normativa vigente).***

Tabla 1: Muestra de datos de insumo – competencias entidades

Código Competencia	Sector	Competencia (según normativa vigente)	Nivel de competencia	Entidad
AMB_172	Ambiente y Desarrollo Sostenible	Administrar -dentro del área de su jurisdicción- el medio ambiente y los recursos naturales renovables, y propender por el desarrollo sostenible del país.	Territorial	Corporaciones Autónomas Regionales
CIT_398	Comercio, Industria y Turismo	Garantizar la presencia permanente en Aeropuertos, puertos y Terminales de Transporte, de personal capacitado en un segundo idioma, información Turística y conocimientos específicos del turismo de la región en la cual estén prestando sus servicios.	Nacional	Policía Nacional

CU_453	Cultura	Al Ministerio de Cultura, previo concepto favorable del Consejo Nacional de Patrimonio Cultural, le corresponde la declaratoria y el manejo de los bienes de interés cultural del ámbito nacional	Nacional	Ministerio de Cultura

Fuente: Elaboración propia.

Etapa 1: Preprocesamiento de textos.

En esta etapa se realiza el preprocesamiento o limpieza a los textos, con esto se busca estandarizar los textos para facilitar el análisis. En el proceso de limpieza se transformaron todos los textos a minúsculas, se removieron puntuaciones, números, espacios innecesarios y *stopwords*, estos últimos hacen referencia a palabras o términos innecesarios los cuales no aportan valor al análisis, como son artículos, pronombres, preposiciones, adverbios y otros. Los ajustes mencionados anteriormente los denominaremos limpieza básica al texto.

Posteriormente se aplicó una función de *stemming,* este algoritmo evita afectar la frecuencia de una palabra por su conjugación, es decir, toma la raíz de las palabras sin tener en cuenta sufijos de género, número, o conjugaciones de tiempo para una palabra. Por ejemplo, al analizar los términos "familiar", "familias", "familia" y "familiares" se tienen 4 términos diferentes, sin embargo, al aplicar una función de *stemming* se obtendrá una sola raíz, "familia", que representa a estas palabras, de esta manera se reduce el número de términos y facilita la comparación entre textos.

En la Tabla 2 se presenta un ejemplo del proceso de limpieza del texto asociado a la competencia de código AMB_175, en esta se presenta el texto original junto al tengo resultante al aplicar la limpieza básica de texto, y en una última columna el texto resultante al aplicar el proceso completo de limpieza, el cual incluye la función de *stemming*.

Tabla 2: Ejemplo de proceso de limpieza de texto.

Código Competencia	Competencia (según normativa vigente)	Limpieza básica al texto	Limpieza básica al texto + stemming
AMB_175	Promover, cofinanciar o ejecutar, en coordinación con otras entidades públicas, comunitarias o privadas, obras y proyectos de irrigación, drenaje, recuperación de tierras, defensa contra las inundaciones y regulación de cauces o corrientes de agua	promover cofinanciar ejecutar coordinacion entidades publicas comunitarias privadas obras proyectos irrigacion drenaje recuperacion tierras defensa inundaciones regulacion cauces corrientes agua	promov cofinanci ejecut coordinacion entidad public comunitari priv obras proyect irrigacion drenaj recuperacion tierr defens inund regulacion cauc corrient agu

Fuente: Elaboración propia.

Etapa 2: Vectorización de los textos.

Una vez finalizada la etapa de limpieza de textos se procede a la vectorización de estos, de esta manera se busca tener una representación numérica de los textos, generando una matriz que contiene la importancia de las palabras en

los textos, esta contiene n filas correspondientes a cada uno de los textos y las columnas correspondientes a términos o palabras las cuales cambian de acuerdo al vectorizador usado, de esta manera se tiene el insumo que permitirá hacer comparaciones y establecer la similitud entre los textos. En la ejecución del proyecto se utilizaron 3 vectorizadores los cuales se describen brevemente a continuación:

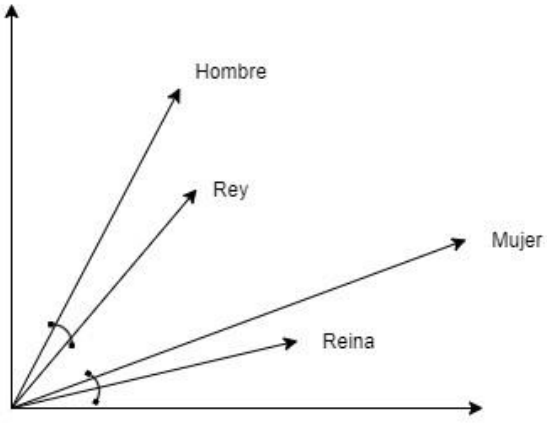
- ***Bag of Words (BOW) o bolsa de palabras,*** este vectorizador genera una columna por cada una de las palabras o términos presentes en los textos y hace un conteo de cuantas veces aparece cada uno de los términos en cada uno de los textos.
- ***TF-IDF (term frequency-inverse document frequency),*** toma como base la matriz generada por el vectorizador Bag of Words, el cual cuenta la frecuencia de aparición de términos en los textos y le agrega una ponderación basada en la aparición de cada término en la totalidad del corpus, es decir, aquellos términos que aparecen con mayor frecuencia en todos los documentos pierden relevancia y aquellos términos que aparecen pocas veces en algunos textos adquieren mayor relevancia en los textos en los que aparecen. Esto permite identificar los términos poco relevantes y aquellos de mayor relevancia en cada texto.
- ***Word2Vec,*** Es una técnica de procesamiento de lenguaje natural basada en redes neuronales, la cual está pre entrenada con el análisis de una gran cantidad y variedad de textos, lo cual le permite identificar diferentes relaciones entre los textos, una de estas siendo la semántica. A modo de ejemplo, al utilizar *Word2Vec* se puede encontrar una relación entre las palabras "Hombre" y "Rey", "Mujer" y "Reina", y "Rey" y "Reina", sin importar que estas no estén escritas de forma similar o no compartan una raíz, estas relaciones no son posibles de identificar con los vectorizadores de *Bag of Words* y *TF-IDF*.

Etapa 3: cálculo de similitud entre textos

Tener una representación numérica de los textos permite que se puedan realizar cálculos y operaciones matemáticas a estos, permitiendo calcular una distancia entre los textos y asociarla a la similitud entre estos. Cabe mencionar que existen diferentes medidas de distancias como medidas de similitud y que los resultados son afectados por el vectorizador que se utilice.

Para este proyecto en particular, se utilizó la medida de similitud coseno, la cual representa la similitud mediante el cálculo del coseno del ángulo que separa los vectores que representan a los textos. La similitud toma un valor que va de 0 a 1, donde 0 representa que los textos están distantes uno del otro y no se parecen, y al tomar el valor de 1 indica que los textos están cercanos por lo tanto son iguales. En la Figura 2 se presenta un ejemplo de la representación vectorial de palabras y cómo sería el cálculo de la similitud coseno.

Figura 2: Representación vectorial y cálculo de similitud coseno.



Fuente: Elaboración propia.

Etapa 4: Agrupamiento de textos similares

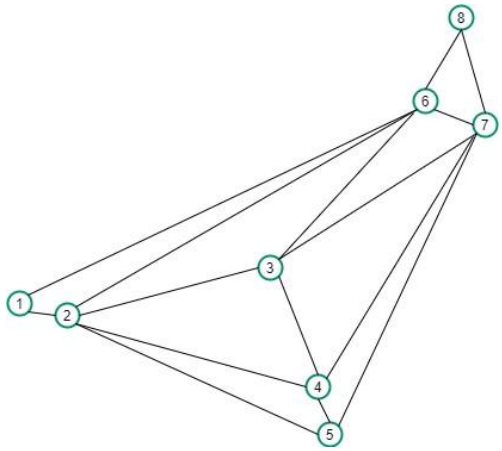
Para el agrupamiento de textos se utilizó la técnica de agrupamiento jerárquico (*hierarchical clustering*), en esta etapa se tienen en consideración las similitudes entre textos analizadas con la similitud coseno y la distancia entre estos. Se tuvieron en cuenta dos aproximaciones al hallar los agrupamientos, la primera consistió en agrupar los textos de las diferentes competencias, y la segunda, consistió en hacer un agrupamiento de las entidades en función de las competencias asociadas a estas. Para la segunda aproximación se tomaron todas las competencias asignadas a cada entidad y se concatenaron los textos para formar un texto consolidado de competencias, estos textos fueron utilizados para realizar las comparaciones y agrupamientos entre entidades.

El agrupamiento jerárquico es un algoritmo iterativo y secuencial, el cual en su configuración ascendente parte de la premisa que cada elemento o texto conforma su propio grupo (*cluster*) y se avanza en la iteración para formar nuevos agrupamientos entre los elementos con una menor distancia, lo cual implica una mayor similitud, hasta tener un solo grupo.

El algoritmo de agrupamiento al ser un modelo no supervisado generará los resultados en base a una configuración previa por parte del usuario para así tener un criterio de "éxito" en la generación de los grupos, en este caso en particular se tienen dos alternativas, la primera, establecer un número n de grupos específicos, es decir, el algoritmo iterará hasta conforma n grupos y en ese momento se detendrá. Dado el planteamiento del proyecto esta aproximación no es viable ya que se quieren encontrar grupos dadas las similitudes de los textos, pero no se tiene una cantidad de grupos definida. La segunda opción, que se utilizó en el proyecto, corresponde a establecer un umbral o distancia máxima para la conformación de grupos, es decir, aquellos textos o grupos que se encuentren a una distancia mayor al umbral definido no podrán conformar un nuevo grupo. Cabe mencionar que la medida de distancia en este algoritmo no corresponde a una distancia física, sino a una aproximación a la representación numérica de los textos y la similitud calculada entre estos, la distancia en este escenario se encuentra normalizada y toma un valor entre 0 y 1, donde 1 representa la mayor distancia de separación entre los elementos.

Para facilitar la comprensión del funcionamiento del algoritmo de agrupamiento jerárquico, en las siguientes figuras se presenta un ejemplo de las iteraciones con un umbral o distancia límite de 0.8.

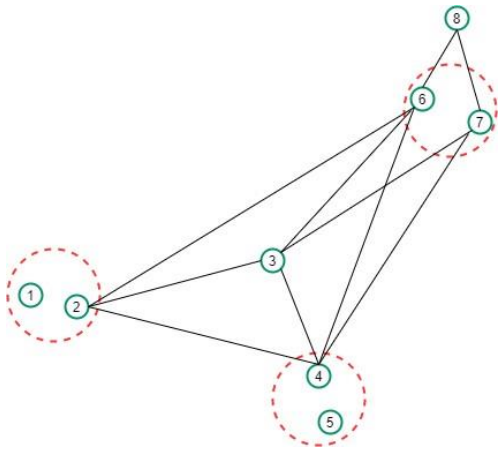
Figura 3: Representación de la vectorización de los textos. Cada nodo representa un texto y los enlaces la distancia de separación entre estos.



Fuente: Elaboración propia.

Al iniciar el algoritmo se calculan las distancias entre los nodos y se inicia el agrupamiento entre los elementos más cercanos, en este ejemplo los nodos más cercanos corresponden en orden a los nodos 4-5, 1-2 y 6-7, conformando tres diferentes grupos como se presenta en la Figura 4, cada vez que se genere un nuevo grupo, al evaluar un nuevo elemento se tendrá en cuenta la distancia del elemento más cercano.

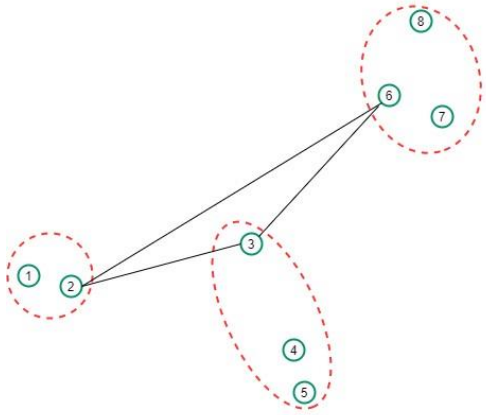
Figura 4: Conformación de grupos, iteración 1,2 y 3.



Fuente: Elaboración propia.

Seguidamente se conformarán los grupos 6-7-8 y 3-4-5.

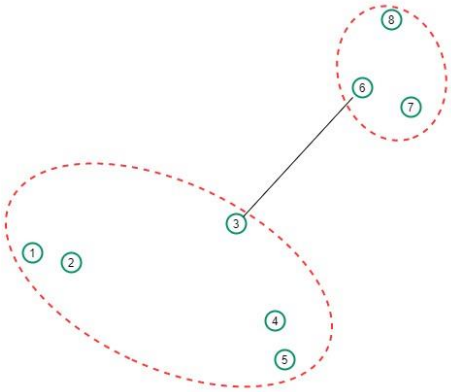
Figura 5: Conformación de grupos, iteración 4 y 5.



Fuente: Elaboración propia.

La siguiente iteración conformará el grupo 1-2-3-4-5, hasta el momento hemos tenido el supuesto que todas las distancias han sido menores al umbral o distancia límite, es decir 0.8. Sin embargo, para la siguiente iteración asumiremos que la distancia entre los nodos 3 y 6 corresponde a 0.9, bajo esta premisa el algoritmo se detendrá y dará como resultado la formación de dos grupos, siendo el primero conformado por los elementos 1-2-3-4-5 y el segundo por 6-7-8.

Figura 6: Conformación de grupos, iteración 6.



Fuente: Elaboración propia.

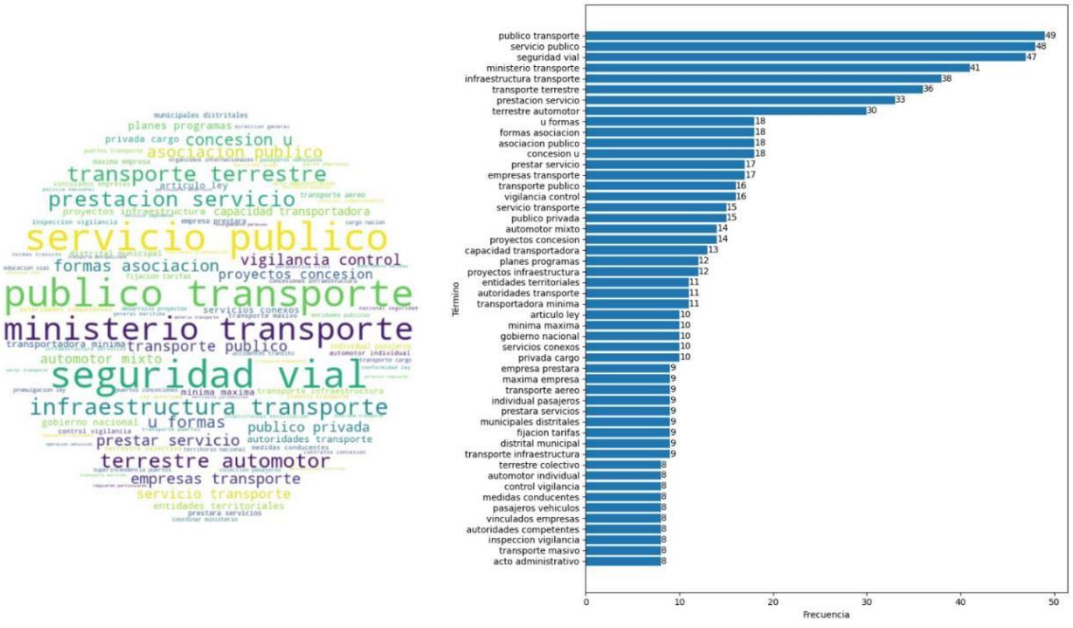
Entregables

Una vez se compararon los resultados obtenidos al usar los 3 vectorizadores planteados en la metodología, se optó por utilizar el vectorizador TF-IDF, ya que presentó una distribución de similitudes un poco más centrada en comparación con los otros vectorizadores, permitiendo identificar competencias con un alto nivel de similitud en sus textos, competencias con textos iguales y textos con baja similitud.

Al finalizar el desarrollo del proyecto se crearon los siguientes entregables filtrando los datos para cada uno de los 26 sectores disponibles, agregando un sector adicional denominado "Todos", el cual contempla el mismo análisis realizado teniendo en cuenta los datos completos sin aplicar ningún filtro previo.

- Dos gráficas descriptivas de los bigramas más frecuentes presentes en los textos, en la Figura 7 se muestran los resultados para el sector Transporte.

Figura 7: Nube de palabras y gráfico de barras de bigramas más frecuentes en el sector Transporte.



Fuente: Elaboración propia.

- Un archivo de Microsoft Excel que contiene las similitudes calculadas entre competencias y un archivo adicional para los resultados obtenidos al realizar el cálculo entre entidades. Dado que algunos sectores cuentan con una un alto número de competencias, lo cual ocasiona un alto número de comparaciones como es el caso del sector transporte, el cual tiene 294 competencias, lo que genera un listado de 86.436 comparaciones, solo se tuvo en cuenta la diagonal superior de la matriz de comparaciones para evitar incluir redundancias y facilitar la consulta de la información, adicionalmente solo se incluyeron en los documentos las comparaciones con un nivel de similitud mayor a 0,5. En la Tabla 3 se presenta un ejemplo de comparación de similitud de competencias del sector Cultura, en donde se evidencia que el texto de una competencia está contenido en el texto de otra, adicional a la información presentada, en el entregable también se incluye información como el nombre de la entidad, nivel de competencia e índices para identificar los registros en el conjunto de datos original.

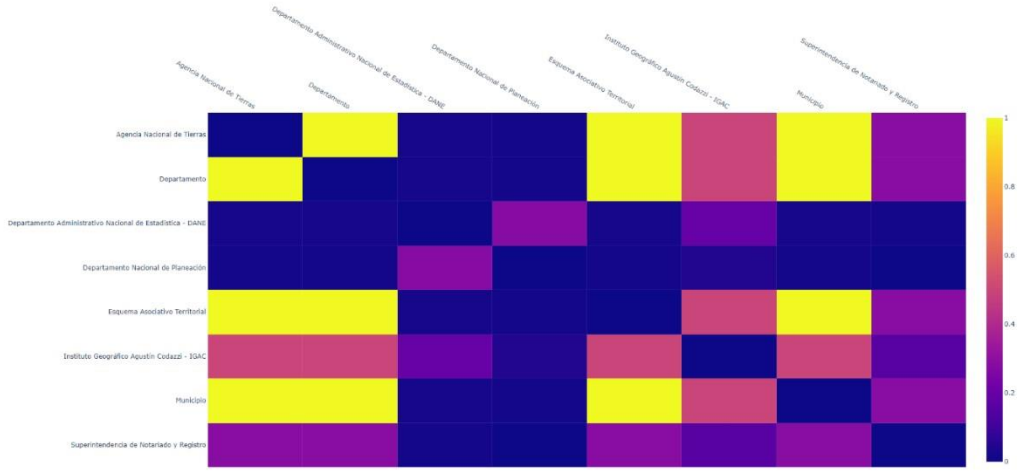
Tabla 3: Ejemplo de comparación de similitud de competencias del sector Cultura.

Cod1	Cod2	Texto1	Texto2	Similitud
CU_481	CU_507	Reunir, organizar incrementar, preservar, proteger, registrar y difundir el patrimonio bibliográfico y documental de la Nación en el ámbito nacional y regional, respectivamente.	La Biblioteca Nacional, y las bibliotecas públicas departamentales son las entidades responsables del depósito legal como mecanismo esencial para el cumplimiento de su misión de reunir, organizar, incrementar, preservar, proteger, registrar y difundir el patrimonio bibliográfico y documental de la Nación en el ámbito nacional y regional, respectivamente.	0,84

Fuente: Elaboración propia.

- Adicional a los listados de comparación entregados en formato de Microsoft Excel, se generaron dos mapas de calor correspondientes a la matriz de comparación de similitudes, uno para la comparación entre competencias y otro para la comparación entre entidades. En la
- Figura 8 se presenta el mapa de calor de similitudes entre entidades para el sector Estadística, en este se identifica que la "Agencia Nacional de Tierras" y la entidad denominada "Esquema Asociativo Territorial" presentan los mismos textos.

Figura 8: Mapa de calor de similitudes entre entidades para el sector Estadística.



Fuente: Elaboración propia.

- En cuanto a los resultados del agrupamiento, se generaron dos archivos en Microsoft Excel, uno para la comparación entre competencias y otro para la comparación entre entidades, teniendo como base la información disponible en el conjunto de original, a esta información se les adicionaron dos columnas a los registros, la primera correspondiente al número de identificación del grupo al que pertenecen y una segunda columna con los 15 bigramas más frecuentes en los textos del respectivo grupo. Para la generación de los grupos en la comparación entre competencias se utilizó un umbral o distancia límite de 0.6 y para el caso de comparación entre entidades se utilizó un umbral de 0.4, estos valores se escogieron teniendo en cuenta los resultados presentados en la Figura 9 y Figura 10 respectivamente, las cuales muestran el número de grupos obtenidos al variar el umbral para la generación de estos, se escogieron los valores basados en el número de grupos generados, sin embargo, se deben evaluar los resultados obtenidos por parte de un experto temático para validar si estos son acordes a lo esperado o si vale la pena realizar un nuevo agrupamiento. Al analizar los resultados presentados en la Figura 9 vale la pena resaltar los sectores "Órganos de Control" y "Hacienda y Crédito Público", en estos se evidencia que al aumentar el valor de "distancia límite" del algoritmo el número de grupos no varía significativamente, lo cual indica que los textos se encuentran distanciados entre sí, por tanto son poco similares, a diferencia de sectores como "Cultura" o "Transporte", en los que se observa que en cada iteración el número de grupos varía, indicando que hay un mayor nivel de similitud entre los textos.

Figura 9: Número de grupos obtenidos en análisis de competencias al variar la distancia límite en agrupamiento jerárquico por sectores.

	competencias										
Número de Clusters - Clustering Jerárquico	distance_threshold										
Sector	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
Ambiente y Desarrollo Sostenible	277	175	140	95	58	20	5	1	1	1	1
Comercio Industria y Turismo	65	51	50	49	49	44	39	28	10	2	1
Cultura	277	110	99	87	63	40	13	3	1	1	1

Planeación	118	95	81	67	54	36	17	6	1	1	1
Transporte	294	195	160	123	70	25	5	1	1	1	1
Educación	246	151	131	103	63	31	14	5	2	1	1
Interior	34	28	27	27	24	21	18	14	8	1	1
Inclusión Social y Reconciliación	148	78	65	52	40	28	19	7	1	1	1
Justicia y del Derecho	126	103	90	75	65	52	32	7	1	1	1
Minas y energía	143	130	121	106	82	50	19	3	1	1	1
Tecnologías de la Información y las Telecomunicaciones	51	28	26	25	25	23	22	15	10	2	1
Trabajo	49	28	28	26	26	25	24	17	6	1	1
Vivienda Ciudad y Territorio	231	156	144	123	69	28	5	2	1	1	1
Ciencia tecnología e Innovación	4	2	2	2	2	2	2	2	2	2	1
Comunicaciones	14	13	13	13	12	12	10	8	6	4	1
Defensa	77	62	60	57	54	50	37	22	8	1	1
Deporte recreación actividad física y aprovechamiento del tiempo libre	56	46	43	38	34	32	29	20	9	2	1
Estadística	38	33	33	33	32	32	26	24	16	3	1
Función Pública	39	26	26	26	26	25	21	12	7	2	1
Hacienda y Crédito Público	33	33	33	33	32	30	30	27	26	6	1
Organización Electoral	1	1	1	1	1	1	1	1	1	1	1
Organos de Control	8	8	8	8	8	8	8	8	5	1	1
Presidencia de la República	105	86	74	67	57	39	22	5	1	1	1
Relaciones Exteriores	41	40	40	39	39	35	31	24	11	2	1
Agropecuario Pesquero y de Desarrollo Rural	161	143	125	104	64	38	11	3	2	1	1
Salud y Protección Social	152	86	72	51	39	22	9	4	2	1	1
Todos	2788	1927	1531	965	410	77	6	1	1	1	1

Fuente: Elaboración propia.

Figura 10: Número de grupos obtenidos en análisis de entidades al variar la distancia límite en agrupamiento jerárquico por sectores.

	entidades										
Número de Clusters - Clustering Jerárquico	distance_threshold										
Sector	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
Ambiente y Desarrollo Sostenible	29	17	11	8	3	2	1	1	1	1	1
Comercio Industria y Turismo	18	14	12	10	8	7	6	4	3	2	1
Cultura	17	10	9	7	6	5	2	2	1	1	1
Planeación	15	11	11	9	7	6	6	2	1	1	1
Transporte	22	17	15	9	6	5	3	3	1	1	1
Educación	14	10	8	8	4	4	3	1	1	1	1
Interior	8	7	6	5	5	5	4	4	3	1	1

Inclusión Social y Reconciliación	17	9	7	7	4	2	1	1	1	1	1
Justicia y del Derecho	19	15	15	13	13	10	5	2	1	1	1
Minas y energía	12	12	12	11	7	5	3	3	1	1	1
Tecnologías de la Información y las Telecomunicaciones	7	5	4	2	2	2	1	1	1	1	1
Trabajo	10	8	8	6	5	5	5	4	2	1	1
Vivienda Ciudad y Territorio	14	9	7	6	4	4	2	1	1	1	1
Ciencia tecnología e Innovación	4	2	2	2	2	2	2	2	2	2	1
Comunicaciones	2	2	2	2	2	2	2	2	1	1	1
Defensa	19	17	17	17	15	12	9	5	2	1	1
Deporte recreación actividad física y aprovechamiento del tiempo libre	6	5	4	2	2	1	1	1	1	1	1
Estadística	8	5	4	4	3	2	2	2	1	1	1
Función Pública	11	8	8	6	3	3	3	2	2	1	1
Hacienda y Crédito Público	8	8	8	8	8	8	8	8	7	3	1
Organización Electoral	8	8	8	8	8	8	8	8	7	3	1
Organos de Control	2	2	2	2	2	2	2	2	2	2	1
Presidencia de la República	9	7	7	7	7	7	6	4	1	1	1
Relaciones Exteriores	4	4	4	4	4	3	3	2	2	1	1
Agropecuario Pesquero y de Desarrollo Rural	14	12	12	12	8	5	3	3	1	1	1
Salud y Protección Social	15	11	7	5	3	2	1	1	1	1	1
Todos	142	96	49	14	9	4	1	1	1	1	1

Fuente: Elaboración propia.

Resultados

Se considera que los resultados obtenidos mediante el análisis de textos permiten tener un mayor entendimiento de la asignación de competencias de las diversas entidades, se realizó el análisis de 26 sectores diferentes, en los que se relacionan 142 entidades, se logró realizar un análisis descriptivo por sectores al generar las gráficas de frecuencia de palabras, identificar textos de competencias que contienen a otras como es el caso de las competencia CU_481 y CU_507 en el sector “Cultura”, similitudes entre entidades como se identifica en el sector “Estadística” en el que la “Agencia Nacional de Tierras” y la entidad denominada “Esquema Asociativo Territorial” presentan los mismos textos, falencias en la asignación de códigos de competencias, se identificaron varias inconsistencias como sucede con los códigos PLA_695 y FP_2037, los cuales corresponden a los sectores “Planeación” y “Función Pública” respectivamente y a pesar de presentar el mismo texto estos presentan códigos diferentes, de igual manera se identifican inconsistencias dentro de un mismo sector, y otros, sin embargo, todo lo anterior debe ser analizado con mayor detenimiento por parte de los expertos temáticos de la SDFF, ya que son ellos quienes puede evaluar la validez y certeza de los resultados obtenidos.

Conclusiones y recomendaciones

1. El desarrollo del proyecto permitió hacer un análisis descriptivo de los sectores, al igual que identificar similitudes entre las diferentes competencias y entidades, para luego formar grupos entre los textos similares, lo anterior facilitará el entendimiento de las competencias y sus relaciones con las entidades.

2. Se recomienda que los resultados se analicen con mayor profundidad y detenimiento por parte de los expertos temáticos de la SDFF para lograr identificar oportunidades de mejoras en la asignación de competencias de las diferentes entidades.
3. Con el proceso de cálculo de similitud se logró identificar inconsistencias en el conjunto de datos suministrado, referentes a la asignación de códigos de las competencias.
4. Para un trabajo futuro se puede implementar un tablero de visualización de información que facilite la consulta y validación de los resultados por parte de la SDFF.

Socialización

Los resultados de este proyecto fueron presentados a la Subdirección de Descentralización y Fortalecimiento Fiscal (SDFF) de la Dirección de Descentralización y Desarrollo Regional (DDDR).