Análisis de Planes de Desarrollo Territorial

Unidad de Científicos de Datos

2017

Contenido

1. Análisis Planes de Desarrollo Territorial

El objetivo de este proyecto es encontrar planes de desarrollo territorial similares y construir grupos temáticos de los mismos.

En este proyecto se realizaron las siguientes actividades:

- ✓ Lectura de la información contenida en 106 Planes de Desarrollo Territorial -DPT
- ✓ Depuración de la información (lematización, e identificación de las palabras más frecuentes).
- ✓ Desarrollo de un análisis de k-medias y de caracterización con la matriz de palabras.
- ✓ Cálculo de medidas de similitud con la matriz de palabras.
- ✓ Similitud de palabras con palabras claves preestablecidas.

1.1. Clasificación de temáticas por los Planes de Desarrollo Territorial

Para la lectura de los planes de desarrollo territorial se utilizó el paquete *pdftools* el cual permite leer documentos en pds.

Posteriormente se desarrollaron herramientas para llevar a cabo la depuración, y limpieza de la información textual limpieza (remoción de palabras vacías o *stopwords*, como preposiciones, artículos, adverbios, se eliminaron puntuaciones, símbolos numéricos*)*.

Después de ese ejercicio de limpieza de textos se construyó, con las diferentes áreas técnicas del DNP, un listado de temáticas (sectores básicos, grupos poblacionales, sectores económicos) que son relevantes al momento de analizar el contenido de los PDT y que agrupan distintas palabras asociadas a la temática (para ver las temáticas construidas y sus palabras asociadas ver Anexo No.1); por ejemplo, el sector de educación está compuesta por palabras como: docentes, aulas, estudiantes, colegios, maestros, pruebas SABER, Ministerio de Educación, educación básica, educación media, educación superior, entre otros. Como alguna de estas nociones están constituidas por 2 o más palabras consecutivas, hubo la necesidad de construir un listado de palabras compuestas, lo que en lenguaje de programación se llama *n-gramas* y en el código fue denominado como *grupos* para que en el procesamiento se analizaran conjuntamente. Una vez se realizado el procesamiento del texto, se hizo un conteo de frecuencias de las palabras que componen el PDT, es importante resaltar que para el análisis de frecuencias los *n-gramas* son tomados como una sola palabra.

Terminado el procesamiento del texto y el conteo, el programa arroja como resultado un listado con todas las palabras que componen el PDT y su frecuencia. A este *output* se eliminan las palabras o se suma la frecuencia del término mal escrito al correcto. Con el listado de palabras y frecuencias depurado, y con el fin de hacer un indicador comprable para todos los municipios y departamentos del país, independientemente de la extensión del PDT, se estableció que el análisis se iba a realizar sobre las 2000

palabras más frecuentes, puesto que 2000 palabras representan cerca del 85%[1] del total de palabras de los PDT y, en una buena porción de los planes, estas 2000 tienen frecuencias mayores a 1.

El análisis de las frecuencias se obtiene de la participación relativa de cada término sobre las primeras 2000 palabras, para calcular la participación se divide la frecuencia de cada término sobre la sumatoria de las frecuencias de las 2000 palabras más repetidas y se multiplica por 1000 (frecuencia/∑frecuencia 2000) *1000. Para términos de comparación, las palabras más frecuentes de los planes tienen, en promedio, una participación de 18.

Teniendo las participaciones relativas, lo que se hizo fue construir una base de datos con las temáticas y nociones por temática, con el fin de identificar la importancia de los ejes estratégicos del Plan Nacional de Desarrollo (PND), los sectores básicos, el enfoque poblacional y otros temas relevantes para el desarrollo del país en los PDT. Los resultados se presentan por el total de las temáticas, que no es más que la sumatoria de las participaciones relativas de los conceptos que la integran. Al final se evaluaron 20 temáticas y 5 grupos poblacionales.

Asimismo, a partir del listado de las 2000 palabras más frecuentes se grafica la nube de palabras o W*ordCloud* para cada uno de los municipios y departamentos. En el *WordCloud* el tamaño de la palabra refleja la frecuencia con la que se repite la palabra dentro del PDT, así las palabras de mayor tamaño son las que más se repiten en el texto.

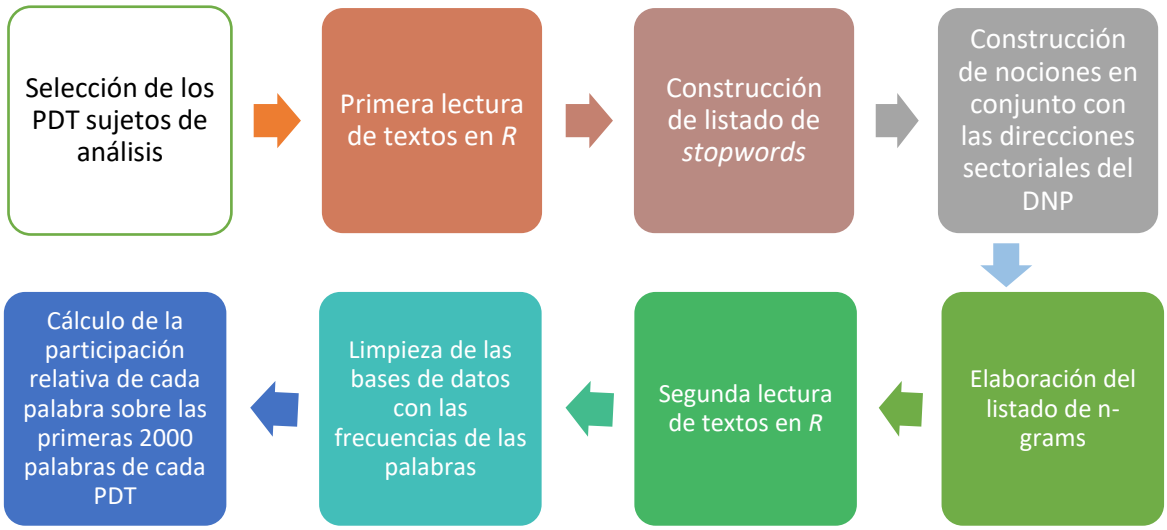

Gráfico 1.
Metodología utilizada para análisis de PDT

Por otra parte, y con el fin de identificar si había diferencias significativas en la importancia que daban los departamentos y municipios a las temáticas y poder agrupar a los entes territoriales según la importancia con la que trataban los temas en su PDT, además de tratar de entender a qué se debían esas diferencias, se aplicó un método de agrupamiento llamado *k-means*. Este método usa iteraciones de agrupamiento

[1] Natagaima, Tolima y Mercaderes, Cauca tienen menos de 2000 palabras.

aleatorios hasta estabilizarse y lograr uno que minimice la inter-varianza dentro de los elementos de un grupo (*cluster)* y maximice la intra-varianza entre los diferentes *clusters*, asegurando así que los grupos sean lo más homogéneos en su interior y lo más disimiles entre ellos. Este método de agrupación utiliza una prueba de diferencias de medias (entre el de cada término del PDT con la de cada término del total de PDT evaluados) para conformar los grupos.

El insumo sobre el cual se realiza la clasificación es la matriz de *matriz de palabras – documentos.* La matriz se encuentra conformada en las **filas** por las palabras que satisfacen los siguientes criterios, i) palabras resultantes de la limpieza explicada anteriormente (eliminación de *stopwords*, puntuaciones, etc.), ii) palabras cuya ocurrencia en los PDT es mayor a 2000 y palabras, iii) Palabras cuya aparición es significativa en todos los documentos (en total se retiene para el análisis 5582) . En las **columnas** se encuentran cada uno de los PDT (106 analizados), los valores de las **celdas** ($n_{i,j}$) corresponden a la frecuencia de ocurrencia de cada palabra en el documento respectivo.

Palabras/Doc	**D_1**	**D_2**	**⋯**	**D_{106}**
P_1	$n_{1,1}$	$n_{1,2}$	⋯	$n_{1,106}$
P_2	$n_{2,1}$	$n_{2,2}$	⋯	$n_{2,106}$
⋮	⋮	⋮	⋱	⋮
P_{5582}	$n_{5582,1}$	$n_{5582,2}$	⋯	$n_{5582,106}$

Tabla 1.1
Matriz palabras/documentos

Las frecuencias son relativizadas con respecto a cada uno de los documentos (en documento se obtiene la frecuencia relativa de cada una de las palabras), la suma de las frecuencias por columna es 1.

Palabras/Doc	**D_1**	**D_2**	**⋯**	**D_{106}**
P_1	$f_{1,1}$	$f_{1,2}$	⋯	$f_{1,106}$
P_2	$f_{2,1}$	$f_{2,2}$	⋯	$f_{2,106}$
⋮	⋮	⋮	⋱	⋮
P_{5582}	$f_{5582,1}$	$f_{5582,2}$	⋯	$f_{5582,106}$
Total	1	1	⋯	1

Tabla 1.2
Matriz de frecuencias relativas palabras/documentos

Posteriormente con la matriz obtenida se lleva a cabo un algoritmo de clasificación no supervisado de k-medias el cual agrupa las palabras de acuerdo a su frecuencia de aparición relativa en los documentos. Las clases construidas se describen de acuerdo a la frecuencia en que ocurren en cada uno los documentos. Por ejemplo, la clase 21 está conformada por las siguientes:

Palabra	Frecuencia	Grupo
Energética	182	21
Actuaciones	171	21
Alcaldías	168	21
Pedagogía	167	21
Aula	166	21
Tolerancia	165	21
Bicicleta	125	21
Cerros	116	21
Crecer	61	21
Maternidad	60	21
Extranjeros	59	21
.	.	.
.	.	.
.	.	.

Tabla 1.3
Palabras clase 21

La aparición de las palabras de la clase 21 es estadísticamente más alta en la clase 21 que el corpus de los PDT. Las palabras de la clase 21 tiene una frecuencia de aparición de 39 veces por cada 10000 palabras en el documento, las mismas palabras tienen una aparición promedio de cerca de 18 veces en todos los PDT, por lo tanto, las palabras que aparecen en esta clase tienen una propensión de utilización de más del doble en Bogotá que en los PDT en general. Los dos promedios (media del grupo y media global) son comparados mediante la prueba estadística *t- student* para evaluar si existen diferencias significativas en los dos promedios.

PDT	Grupo	Valor t	Media grupo	Media global	Nro. Palabras
Bogotá	21	2,117	**0,039**	0,0179	55

Tabla 1.4
Estadísticas grupo 21

Finalmente, las clasificaciones obtenidas se contrastan con las temáticas construidas.

Para poder llevar a cabo estas agrupaciones hubo primero que realizar un proceso llamado lematización a la base de datos de las frecuencias. La lematización es un proceso lingüístico que consiste en regresar las palabras de su formal flexionada (plural, conjugada, en femenino) a su lema, algo así como llevar las palabras a su raíz, a la manera como se encuentra en un diccionario. Así, por ejemplo, estudiando pasa a estudiar; educativos, educativas y educativa pasa a educativo; y en nuestro caso, a las palabras que tenían

problemas de ortografía se lematizó a la palabra correctamente escrita, como en el caso de las palabras sin tilde como educacion o participacion. Esto con el fin de que la agrupación por temáticas sea más ajustada. Una vez se identifica a qué se lematiza cada palabra, se suman en el lema las frecuencias de todas las formas flexionadas y se aplica el proceso de clustering.

De esta manera se identificaron X clusters al analizar los 128 PDT, y uno de los hallazgos más importantes de este análisis es que no existen diferencias significativas en la importancia relativa que dan los entes territoriales a temas como educación, salud, paz, conflicto, agua y saneamiento. Estos son temas que fueron incluidos en todos los PDT y debido a la ausencia de diferencias significativas en la importancia relativa, se podría decir que la inclusión de estos temas no se debe a factores "objetivos", tales como cobertura, resultados en pruebas, incidencia del conflicto, acceso, entro otras. Por el contrario, hubo 29 ET que si incluyeron algunos temas de manera diferenciada (ver Tabla 1), por ejemplo, Santa Marta trata el tema de turismo, y algunos municipios de la Guajira el cuidado prenatal y a madres gestantes.

	Relación Estado-Ciudadano	Desarrollo Territorial	Desarrollo Económico	TICs	Desarrollo Rural	Reconciliación	Diversidad cultural	Minería	Hábitos saludables	Turismo
Antioquia	X	X	X							
Atlántico	X	X								
Boyacá	X	X	X							
Caldas	X		X	X						
Cauca							X		X	
Cundinamarca	X	X	X							
Guajira	X	X	X							
Magdalena			X							
Nariño					X	X				
Quindío			X							
Risaralda	X			X						
Santander	X		X							
Sucre	X	X								
Tunja	X		X							
Valle del cauca			X							
Riosucio						X				
Puerto Rico						X				
Río Iró					X					
Mapiripán					X					
Bojayá						X				
El Charco							X			
Mocoa	X									
El tambo								X		
El Carmen									X	
Mutatá									X	
Dibulla							X		X	
Uribia							X		X	
Manaure									X	
Santa Marta										X

Tabla 1.5 Temáticas diferenciadoras de los PDT

1.2. Indicador de similitud de los PDT con el Plan Nacional de Desarrollo

Para cuantificar la similitud entre los PDT y con el Plan Nacional de Desarrollo con respecto a los términos utilizados se utilizar el coseno entre los vectores asociados a los PDT y el PND. El coseno genera una medida entre cero y uno, en donde 1 significa que los documentos son idénticos y 0 que los documentos difieren de forma máxima en cuanto a los términos utilizados.

Para esto se parte de la utilización de la matriz *documento-término*, la cual contiene la frecuencia de los términos utilizados en cada uno de los PDT (una vez realizada el preprocesamiento del texto: limpieza de *stopwords*, eliminación de puntuaciones, etc.). En las filas se referencia los PDT y en las columnas cada uno de los términos utilizados el Corpus.

Se calcula como medida de similaridad el coseno conformado por cada una de los vectores fila (documentos), esta medida tiene como efecto eliminar el efecto del tamaño de texto de cada uno de los documentos. La similaridad entre el documento i el documento j está dada por el cociente entre el producto punto de la fila i-ésima y j-esima de la matriz de documento-frecuencia sobre sus respectivas normas

$$\text{Similaridad}(\mathrm{D}_1, \mathrm{D}_2) = \frac{D_1 \cdot D_2}{\|D_1\|\|D_s\|}$$

Se calcula la medida de similitud de cada PDT (fila de la matriz documento-término) con el vector de frecuencias de términos del PND. Se muestran los 5 PDT más similares y menos similares (en azul y rojo respectivamente) con el PND en cuanto a la frecuencia de los términos utilizados:

PDT	Índice
Cundinamarca	0,74
Manizales	0,69
Guajira	0,68
Caldas	0,68
Antioquia	0,66
Carmen_del_Darien (Chocó)	0,34
Santa Barbara Iscuande (Nariño)	0,33
Puerto Guzmán (Putumayo)	0,31
Puerto Leguízamo (Putumayo)	0,31
Córdoba (Bolivar)	0,30