Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

GENERACIÓN DE RESULTADOS PARA DESARROLLO DE TABLERO DE CONTROL CON LOS RESULTADOS DE LOS ALGORITMOS DE PROCESAMIENTO NATURAL DEL LENGUAJE PARA LAS RELATORÍAS DE TALLERES CONSTRUYENDO PAÍS

Entidad

- Dirección de Desarrollo Digital y Dirección General del DNP.
- Alta consejería presidencial para las regiones – Presidencia de la república

Sector	**Lenguaje**	**Fuente de datos**
Planeación	R.	Documentos con las relatorías de los Talleres Construyendo País

Presentación

Los Talleres Construyendo País son encuentros con las comunidades que ha venido realizando el presidente Iván Duque en donde se hacen mesas de trabajo para conocer las necesidades que tiene el municipio que se está visitando y se establecen promesas desde el Gobierno Nacional junto con las gobernaciones y alcaldes para dar atención a los puntos que se acuerden. Cada Taller tiene un documento que resume los puntos y estrategias a realizar, así como las entidades responsables de hacerlas cumplir y los plazos estimados para dar lugar a la realización de actividades en cada punto correspondiente.

Al estar realizando talleres cada sábado, es necesario contar con una herramienta que analice los textos que se encuentran en estas relatorías y posteriormente una plataforma que visualice los resultados de manera que se puedan conocer por municipio o por sectores cuales son los municipios que presentan necesidades específicas y de esta forma dar una idea general de las necesidades agregadas por departamentos o regiones.

The Workshops so-called "Construyendo país" are meetings between Colombian communities and its President Iván Duque where panel discussion is held to know the needs of the visited municipalities and where promises by the National Government together with the governors and mayors are established. Each Workshop has a document that summarizes the points and strategies to be carried out, as well as the responsible entities and estimated deadlines for the completion of activities at each corresponding point.

These workshops are held every Saturday; so, in order to focus the governmental efforts it becomes necessary to have a tool that analyzes the texts in those documents and a platform that allows to visualize the municipalities organized results (by municipality or sectors) in order to identify its specific needs and in this way give a general idea of the added needs by departments or regions.

Objetivo general

Realizar un análisis del contenido y las temáticas más relevantes tratadas en los Talleres Construyendo País y generar un tablero de control que permita visualizar los análisis realizados por la UCD para los actores interesados en conocer de manera rápida y concisa lo escrito en estos documentos.

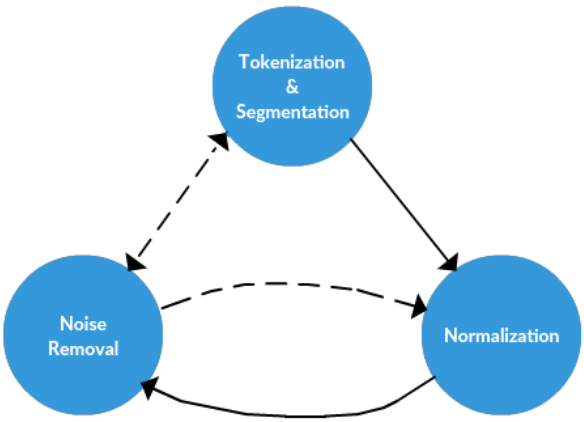
Objetivos específicos

1. Depurar la información contenida en los documentos Talleres Construyendo País eliminando sus conectores lógicos y/o palabras vacías (“stopwords”), la puntuación, los números y las palabras que no sean relevantes para el análisis en términos de significado.
2. Generar las tablas de frecuencias de las palabras en el documento completo y por sectores en cada uno de los municipios visitados (unigramas, bigramas y trigramas).
3. Generar un tablero de control que permita visualizar los análisis obtenidos del procesamiento de lenguaje natural realizado.

Metodología

Inicialmente, todos los documentos resumen de los Talleres “Construyendo país” son llevados al formato estándar definido por la UCD en el documento denominado “Lineamientos_Doc_TPais.docx” con el fin de garantizar la homogeneidad de los archivos pdf entregados por la Alta Consejería Presidencial para las Regiones.

En la segunda etapa los documentos en formato estándar son preprocesados bajo el marco que se presenta en la Ilustración 1:



El marco de preprocesamiento de datos de texto.

Eliminación de ruido

En esta parte del marco se realizan tareas de normalización de texto que a menudo tienen lugar antes de la tokenización. Las tareas de eliminación de ruido realizadas son:

- eliminar encabezados de archivos de texto, pies de página
- eliminar HTML, XML, etc. marcado y metadatos
- extraer datos valiosos de otros formatos, como JSON

Tokenización

La segmentación de texto o análisis léxico es un paso que divide cadenas de texto más largas en piezas más pequeñas o tokens. Los trozos de texto más grandes pueden ser convertidos en oraciones, las oraciones pueden ser tokenizadas en palabras, etc. El procesamiento adicional se realiza después de que una pieza de texto ha sido apropiadamente concatenada.

Nota: la segmentación se usa para referirse al desglose de un gran trozo de texto en partes más grandes que las palabras (por ejemplo, párrafos u oraciones), mientras que la tokenización se reserva para el proceso de desglose que se produce exclusivamente en palabras.

Normalización

La normalización generalmente se refiere a una serie de tareas relacionadas destinadas a poner todo el texto en igualdad de condiciones. Las actividades realizadas fueron:

- eliminar conectores lógicos y/o palabras vacías (“stopwords”)
- eliminar la puntuación, los números y las palabras que no sean relevantes para el análisis en términos de significado

Una vez se tienen los documentos “normalizados” se desarrolla un dashboard que involucra un análisis de frecuencia de palabras de bigramas (secuencia de dos palabras que aparecen en el texto) y trigramas y un análisis sectorial en cada municipio. Se realiza un análisis por cada uno de los sectores a través de la medida conocida como tf-idf la cual permite aumentar el número de veces que una palabra aparece en un documento. Este análisis permite conocer que palabras están altamente asociadas a un sector y disponer de una visión sectorial en cada municipio donde se desarrollan los talleres construyendo país.

El valor tf-idf aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero se pondera por la frecuencia de la palabra en todo el corpus (conjunto de todos los documentos, en este caso en todos los talleres). Para llevar a cabo un análisis se calcula una matriz de términos frecuencias (en las filas los documentos, en las columnas los términos) y se calculan las siguientes dos medidas. La frecuencia relativa de cada término en el documento (**tf**), la suma por fila de este término es 1.

La frecuencia inversa de documento (idf) es una medida de si el término es común o no, en la colección de documentos. Se obtiene dividiendo el número total de documentos por el número de documentos que contienen el término, y se toma el logaritmo de ese cociente:

$$idf(t, D) = \log\left(\frac{|D|}{|\{d \in D: t \in d\}|}\right)$$

El producto de la frecuencia del término (df) y el inverso de la frecuencia del documento (idf), constituye la medida conocida como tdf-idf.

Para ilustrar el cálculo de esta medida suponga que en total se disponen de 6 documentos (D = 6), si los 6 documentos contienen un término el valor idf será igual a cero, es decir, idf(t,D) = log(1) = 0, por tanto el término está presente en todos los documentos y como consecuencia este término no permite discriminar

adecuadamente los documentos. Al contrario si un término aparece únicamente en un documento, es decir idf= log(6/1) = 1,79 o en dos documentos idf = log(6/2) = 1,098, entonces estos términos no discriminan bien el documento d.

Resultados

Para la visualización de esta plataforma, se contó como insumo los archivos generados por la UCD luego de preprocesar todos los datos y generar los correspondientes unigramas, bigramas y trigramas tanto por fecha de realización, como por departamento y municipio obteniendo los siguientes resultados:

El tablero de control se ajusta automáticamente ante filtros como por ejemplo cuando se selecciona El Socorro:

También es posible filtrar desde el mapa georreferenciado para conocer el porcentaje de pertenencia de términos sobre el total, por ejemplo:

Conclusiones

1. El tablero de control funciona adecuadamente y permite realizar consultas inmediatas de los resultados del Procesamiento de Lenguaje Natural generado.
2. El procesamiento realizado a los documentos de los talleres podría ser automático cuando los documentos enviados por la Dirección General tengan un formato unificado.

Socialización

Este tablero ha sido socializado con la Dirección General y con la Alta Consejera para las Regiones en Presidencia de la República con el objetivo de conocer de primera mano qué necesidades pueden surgir para nutrir este ejercicio tanto desde la parte visual, como desde el análisis de frecuencias.