METADATOS SOCIALES:
iniciativas, tecnologías, aplicaciones y softwares

Ariel Alejandro Rodríguez García
COORDINADOR

Z666.7
M48

Metadatos sociales : iniciativas, tecnologías, aplicaciones y softwares / Coordinador Ariel Alejandro Rodríguez García. - México : UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información, 2024.

xv, 293 p. - (Metadatos)
ISBN: 978-607-30-8624-0

1. Metadatos. 2. Datos vinculados. 3. Indización - Aspectos sociales. 4. Tecnología de la información - Aspectos sociales. 5. Contenidos generados por los usuarios. I. Rodríguez García, Ariel Alejandro, coordinador. II. ser.

Diseño de cubierta: Mario Ocampo Chávez
Imagen: Kishore Newton - stock.adobe.com

Primera edición: Mayo de 2024

Instituto de Investigaciones Bibliotecológicas
y de la Información
Circuito Interior s/n, Torre II de Humanidades,
pisos 11, 12 y 13, Ciudad Universitaria, C. P. 04510,
Alcaldía Coyoacán, Ciudad de México

ISBN: 978-607-30-8624-0

Publicación dictaminada

Impreso y hecho en México

Contenido

Aplicaciones

Softwares

Metodología para establecer relacionamiento automatizado de patrones comunes en testimonios escritos de víctimas del conflicto armado en Colombia

Fabián Orlando Baena Henao
Universidad de Antioquia, Colombia

Introducción

Desde finales de la década del 50, Colombia ha estado inmersa en una violencia justificada, entre otras causas, por la lucha política entre el Estado y las guerrillas, que se fue agudizando conforme se fueron sumando otros actores bélicos tales como paramilitares, narcotráfico, e incluso el mismo Estado. Según datos compilados por el Centro Nacional de Memoria Histórica,[1] se estima que entre 1958 y 2012 el conflicto armado en Colombia dejó por lo menos 220 000 muertos.

Los vestigios de la guerra en Colombia han generado la victimización de otro número considerablemente alto de personas a causas que guardan relación, entre las que se encuentran la desaparición forzada, la violencia sexual, el reclutamiento de menores, y el desplazamiento forzado, que, según datos tomados a

1 Grupo de Memoria Histórica, "BASTA YA! Colombia: Memorias de guerra y dignidad. Informe General Grupo de Memoria Histórica".

2013 del Registro Único de Víctimas, asciende a más de 4 744 046 personas;[2] estas cifras y formas pueden aumentar dado que muchos datos se encuentran en subregistros no oficiales, así como personas que no fueron reportadas.

Las experiencias recopiladas de las víctimas, de manera particular los testimonios escritos, contienen gran cantidad de información a través de patrones clave de espacio, tiempo, género, individuos, grupos, entre muchos otros; con dichos patrones, es posible establecer conexiones que pueden ayudar a entender la configuración del conflicto en factores tales como sucesos, motivos o modos, sólo por mencionar algunos, además de facilitar el esclarecimiento de la verdad y evitar la repetición. Ahora bien, procesar cada testimonio implica un considerable esfuerzo en términos de recursos y tiempo, dado que, para establecer relaciones entre ellos, se involucra el trabajo de profesionales que lo interpreten, para encontrar patrones representados a través de frecuencias de palabras con sentido semántico dentro de un corpus, que de acuerdo con la definición de Parodi,[3] este último término consiste en focalizar datos observables a modo de evidencia científica.

Para el investigador, establecer un relacionamiento entre los testimonios es una tarea dispendiosa cuando se presentan grandes volúmenes de información, pues al ser tradicionalmente una tarea manual, poco repetitiva y en su mayoría subjetiva, los datos generalmente no se encuentran etiquetados con estándares de organización para su consulta. Por tanto, es en este punto donde la Ingeniería y las Ciencias Sociales convergen, proponiendo metodologías para dar solución a problemas en el estudio de fenómenos sociales, tales como el procesamiento de lenguaje natural a través del uso de algoritmos,[4] y así lograr la identificación de

2 *ibidem*.

3 Giovanni Parodi, "Lingüística de Corpus: Una introducción al ámbito".

4 Manuel Montes *et al*., "Minería de texto empleando la semejanza entre estructuras semánticas".

elementos representativos en los textos, lo que facilita la recuperación de información y la construcción de nuevo conocimiento.[5]

Fases de la investigación

Para establecer la relación entre testimonios escritos, que en este caso particular corresponden a las narraciones de hechos violentos por parte de las víctimas del conflicto armado en Colombia, la investigación se desarrolla en 3 fases consecutivas como se describe a continuación y se muestra en la figura 1:

Figura 1. Fases del relacionamiento automático de textos

Fuente: Elaboración propia.

La fase uno establece la documentación de los procedimientos y la construcción del corpus de testimonios de víctimas del conflicto armado colombiano. La segunda fase se subdivide a su vez en tres etapas, pues se desarrollan algoritmos de PLN y ML que gestionan los datos del corpus previamente consolidado para identificar los elementos relevantes de cada documento y que apuntan a ser comunes entre los demás y así establecer relaciones entre ellos.

5 Erika T. Duque, "Metodología para la extracción de metadatos semánticos de textos en español…".

La fase final corresponde al momento en el que se establecen las relaciones entre los testimonios, al identificar la afinidad entre pares de documentos y su aplicación en grafos que muestran visualmente la relación.

Fase 1: Documentación y creación de corpus

Con el fin de tener un acercamiento a los diferentes enfoques que guardan relación con el objeto de investigación, se realizaron rastreos documentales en bases de datos bibliográficas tales como Scopus, Science Direct, Springer, Redalyc, Scielo y Google Académico. Se utilizó como punto de partida en la estrategia de búsqueda frases que identifican los temas de impacto tales como "corpus analysis", "natural language processing" y "text mining", asociados a los campos de la Ingeniería, Ciencias de la Computación y Ciencias Sociales. Posteriormente, luego de identificar casos de aplicación, se ajustaron los criterios de búsqueda a técnicas de ML que se acercaban más a los objetivos planteados, es allí donde los resultados arrojaron investigaciones que daban cuenta de la aplicación de "text similarity" y "clustering".

La etapa de documentación, además de identificar los casos de aplicación, también consideró las herramientas y técnicas para el procesamiento de datos textuales, lo que dio paso desarrollar los algoritmos en el lenguaje de programación Python, referenciado por la comunidad de programadores como ideal para realizar el análisis de los datos de la investigación; por ello, de manera particular, librerías como NLTK toolkit, Pandas y Numpy, resultaron fundamentales en el trabajo.

Por su parte, la construcción del corpus de testimonios escritos, que a una primera vista puede representar una tarea simple cuando es considerada sólo como el almacenamiento de documentos, es decir, la unificación de archivos de texto en una misma ubicación conformando una base de datos textual, realmente es una actividad que reviste mayor complejidad desde la misma consecución de los testimonios, pues aunque sólo por mencionar algunos, los procesos explícitos en la Ley de Reparación de

Víctimas[6] o la Jurisdicción Especial para la Paz JEP[7] consideran registrar los testimonios de las víctimas, estos tienen limitaciones para su uso público al tratarse de información sensible, tanto por los datos que consignan, como por el derecho a la privacidad y confidencialidad; todo ello sumado a los trámites para la autorización de uso.

Así pues, aunque se sabe que son varias las instituciones que cuentan con gran número de documentos que recogen las voces de las víctimas (entre ellas las ya mencionadas), se presentan barreras de uso, por lo que fue necesario rastrear testimonios que cumplieran con las características del corpus y estuvieran disponibles públicamente. Así, el sitio web del Centro Nacional de Memoria Histórica, en la sección Podcast,[8] el cual aloja series de testimonios de víctimas del conflicto armado colombiano, formó pieza clave en esta fase. Sin embargo, dado que los testimonios se encuentran en formato de grabaciones sonoras, particularmente podcast, fue necesario realizar la transcripción de tales archivos a texto; para lo cual, los archivos fueron extraídos en formato audio y posteriormente transcritos en procesadores de texto utilizando herramientas de conversión automática de audio a texto (Speech to Text STT). Los archivos resultantes fueron finalmente revisados y ajustados en un proceso de revisión manual cuando se encontraban caracteres no coincidentes o fuera de contexto.

Fase 2: Preprocesamiento de corpus

En esta fase se desarrollaron actividades centrales en el procesamiento de la información en varias etapas, que van desde la consolidación de los elementos a identificar, hasta la implementación de algoritmos que realizan trabajos de tratamiento, limpieza y marcado

6 Colombia, Congreso de la República, Ley de reparación de víctimas.

7 Colombia, Unidad para la Atención y Reparación Integral a las Víctimas, "Jurisdicción especial para la paz".

8 Colombia, Centro Nacional de Memoria Histórica, Podcasts.

de los elementos representativos para establecer relaciones de los datos en el corpus.

Dimensiones de relacionamiento
Las dimensiones corresponden a los elementos centrales sobre los que se establecen las relaciones, es decir, los temas que agrupan y representan coincidencia entre cada testimonio. Al comprender la naturaleza de los testimonios, se definen 4 dimensiones: Entidades, Afectaciones, Temporalidad, Georreferenciación.

- *Entidades*. Esta dimensión comprende nombres de personas, grupos o instituciones sobre los que recaen las acciones, tales como grupos armados, víctimas, victimarios, entre otros. Entre las unidades que componen esta dimensión se encuentran palabras como "civiles, familia, compañero, hermano, guerrillero".
- *Afectaciones*. Esta dimensión corresponde a las marcas que dejan los vestigios de los hechos violentos en las víctimas, las cuales pueden ser de tipo económicas, sociales, físicas o psicológicas. Entre las unidades que integran esta dimensión se pueden encontrar palabras como "terror, estrés, psicosis, asesinato".
- *Temporalidad*. En esta dimensión se enmarcan las fechas, épocas o periodos. Se pueden ubicar palabras tales como "viernes, marzo, año".
- *Georreferenciación*. Esta dimensión comprende referencias geográficas o de espacio, tales como ciudades, barrios, sitios, parajes, entre otros. Algunas de las palabras que se encuentran en esta dimensión pueden ser "Medellín, bodega, acera, vía pública".

Para cada una de las dimensiones se realizó un trabajo de marcado manual en un documento al cual, a partir de la lectura de varios testimonios, se fueron agregando palabras y frases que identificaran el elemento o dimensión. Las unidades que identifican las dimensiones pueden ser complementadas con nuevas palabras y

frases cuando el investigador lo considere, pues adicionar nuevas unidades que describan las dimensiones aporta a la precisión de la metodología, haciendo más precisa la descripción de cada testimonio. Es de resaltar que la duplicidad de palabras en la identificación de las dimensiones no implica afectación en la metodología, es decir, no representa alteraciones de precisión si alguna palabra o frase ya se encuentra registrada previamente, pues esta etapa aplica un algoritmo que descarta duplicados, lo que facilita el trabajo sólo con palabras y frases únicas.

Limpieza y preprocesamiento del corpus

Dado que los testimonios hacen parte de narraciones escritas de las víctimas que corresponden a sujetos individuales, los archivos que las condensan no cuentan una estructura definida, es decir, el texto se encuentra desestructurado, por lo que se hace necesario establecer condiciones de igualdad tales como la conversión de caracteres de mayúsculas a minúsculas, pues en el procesamiento de texto, una palabra que semánticamente sea idéntica a otra, pierde tal propiedad al representarse con caracteres diferentes ($M \neq m$).

Asimismo, es preciso eliminar el ruido que puedan generar las denominadas palabras vacías o *stopwords*, las cuales representan poca relevancia o no ofrecen información al proceso, entre ellas se pueden resaltar los elementos más comunes del lenguaje, como lo son las preposiciones y artículos, ya que de omitir este paso, se puede incurrir en falsos positivos, es decir, establecer el relacionamiento de dos testimonios que no tienen ninguna conexión en sus dimensiones sólo por coincidir con elementos del tipo referenciado. A su vez, en este proceso, se expulsan tanto los símbolos o caracteres especiales como lo son asteriscos, símbolos de número, guiones, etcétera, como espacios en blanco extra que ocupan caracteres innecesarios en el texto.

Identificación de dimensiones

En esta etapa, el algoritmo individualiza los elementos de cada testimonio, es decir, divide en tokens los elementos del corpus y coteja cada unidad con las unidades de las dimensiones

previamente establecidas, que al encontrar una coincidencia la separa para posteriormente construir un documento con el total de unidades por cada dimensión. El resultado final de esta etapa es que, a partir de un testimonio, se genera un nuevo documento con las unidades separadas (palabras y frases) que hacen parte de cada dimensión.

Fase 3: Relacionamiento de testimonios

La fase final del proceso consiste en establecer las relaciones entre pares de testimonios que cuenten con elementos comunes; éstas se dan por la comparación entre coincidencias de las unidades de cada testimonio en sus dimensiones. Ello quiere decir que, si un par de testimonios cuentan con elementos comunes en una misma dimensión, de manera automática se establece un vínculo entre ambos.

Ahora bien, si la relación de un testimonio está dada en más de un par, es importante determinar la distancia que existe entre pares de documentos, o dicho de otra forma, se considera importante determinar el nivel de semejanza de un testimonio comparado con los otros; para hallar esta similitud, el algoritmo utiliza la distancia de coseno, la cual es una medida que compara las distancias a través de valores trigonométricos considerando las frecuencias de los elementos que componen los testimonios.

Finalmente, dado que se cuenta con la relación de pares de documentos y una distancia entre cada uno de ellos, es posible representar gráficamente en un plano tales valores, lo que hace más perceptible la interpretación de los datos y sus relaciones, como se muestra en la figura 2.

Aplicación

La identificación de patrones comunes en los testimonios escritos de las víctimas del conflicto armado en Colombia permite aportar herramientas que facilitan el entendimiento de la configuración

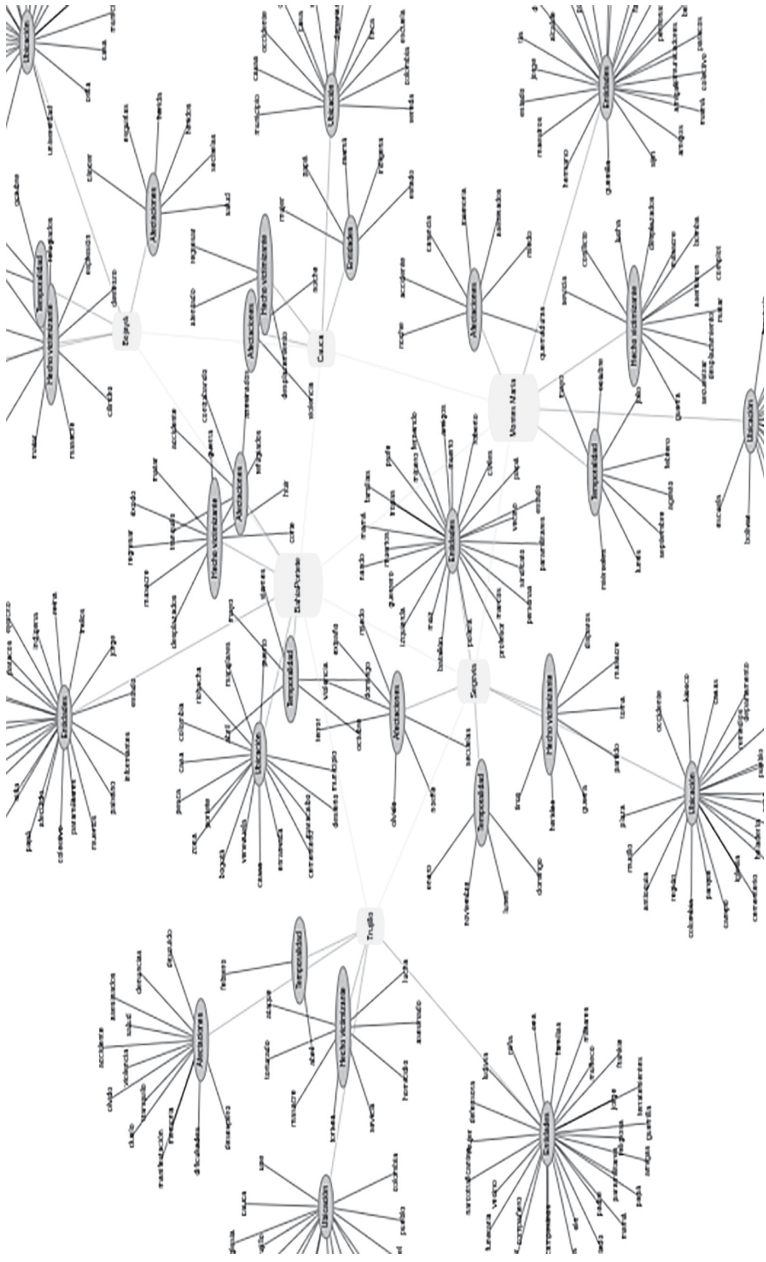

Figura 2. Grafo de relacionamiento entre textos por dimensiones

Fuente: Elaboración propia.

de la guerra en Colombia, así como los actores y su participación, todas éstas enmarcadas en procesos actuales del posconflicto como la Comisión de la Verdad, La Justicia Especial Para la Paz y la Búsqueda de Personas Desaparecidas. Asimismo se convierte en herramienta fundamental para el análisis de testimonios escritos de las víctimas, ya que facilita el procesamiento y análisis de textos desde un lenguaje natural.

El procesamiento del lenguaje natural se plantea como un ejercicio en el que la estructuración (o preprocesamiento) de textos se convierte en paso fundamental para aplicación de técnicas de minería de texto y posteriores análisis de información y extracción de eventos, que facilitan la identificación de patrones que permiten el relacionamiento, tanto entre unidades sintácticas como entre documentos de un corpus.

El procesamiento de lenguaje natural es un campo de la ingeniería y la lingüística que tiene aplicación en todas las áreas del conocimiento, pero que, de manera particular, beneficia las Ciencias Sociales, entre muchos otros casos, al encontrar, como en el caso de esta investigación, patrones comunes y semejanzas entre textos a través del etiquetado automático. Se hace entonces relevante para los profesionales de las Ciencias Sociales, la habilitación y ampliación de un corpus, así como etiquetado de texto y ontologías, que permitan facilitar las tareas y generar nuevo conocimiento por medio de técnicas computacionales como las utilizadas en este trabajo.

Referencias bibliográficas

Colombia. Centro Nacional de Memoria Histórica. Podcasts. Centro Nacional de Memoria Histórica. https://centrodememoriahistorica.gov.co/podcasts (Consultado el 21 de octubre de 2021).

Colombia. Congreso de la República. Ley de reparación de víctimas. Pub. L. No. Ley 1448 de 2011 (2011).

Colombia. Unidad para la Atención y Reparación Integral a las Víctimas. "Jurisdicción especial para la paz". JEP. Jurisdicción especial para la paz. https://www.jep.gov.co/JEP/Paginas/Jurisdiccion-Especial-para-la-Paz.aspx (Consultado el 21 de octubre de 2021).

Duque Bedoya, Erika Teresa. "Metodología para la extracción de metadatos semánticos de textos en español utilizando procesamiento de lenguaje natural: subaplicación para la identificación de contextos espaciales y temporales en textos que describan interacciones entre actores". Medellín: Universidad EAFIT. Departamento de Informática y Sistemas. 2009. http://hdl.handle.net/10784/1261.

Grupo de Memoria Histórica. "BASTA YA! Colombia: Memorias de guerra y dignidad. Informe General Grupo de Memoria Histórica". Bogotá: Centro Nacional de Memoria Histórica. 2013.

Montes y Gómez, Manuel, Alexander Gelbukh y Aurelio López López. "Minería de texto empleando la semejanza entre estructuras semánticas". *Computación y sistemas* 9, núm. 1 (2005): 63-81.

Parodi, Giovanni. "Lingüística de Corpus: Una introducción al ámbito". *RLA. Revista de Lingüística Teórica y Aplicada* 46, núm. 1 (2008): 93-119. https://doi.org/10.4067/S0718-48832008000100006.

Metadatos sociales: iniciativas, tecnologías, aplicaciones y softwares. Instituto de Investigaciones Bibliotecológicas y de la Información/UNAM. La edición consta de 100 ejemplares. Coordinación editorial, Sergio J. Sepúlveda H., revisión especializada: Marcos Emilio Bustos Flores; corrección de pruebas: Carlos Ceballos Sosa, Marcos Emilio Bustos Flores; formación editorial, Mario Ocampo Chávez. Fue impreso en papel cultural de 90 g en Editorial Albatros, Av. Benito Juárez M 26 L 14, Col. El Molino Tezonco, c.p. 09960, CdMx. Se terminó de imprimir en mayo de 2024.