Construcción de un modelo de análisis computacional para la clusterización automática de los documentos de políticas nacionales CONPES 2010-2020

Jeny Carolina Hernández Marín

Trabajo de grado presentado para optar al título de Bibliotecóloga

Asesores

Fabian Baena Henao, Bibliotecólogo

Jaider Ochoa Gutiérrez, Bibliotecólogo y Magíster (MSc) en Gestión de Ciencia, Tecnología e Innovación, Administración y gestión de empresas

Universidad de Antioquia

Escuela Interamericana de Bibliotecología

Bibliotecología

Medellín, Antioquia, Colombia

2022

Cita	(Hernández Marín, 2022)
Referencia **Estilo APA 7 (2020)**	Hernández Marín, J., (2022). *Construcción de un modelo de análisis computacional para la clusterización automática de los documentos de políticas nacionales CONPES* 2010-2020. [Trabajo de grado profesional]. Universidad de Antioquia, Medellín, Colombia.

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

CRAI Escuela Interamericana de Bibliotecología

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes

Decano/Director: Dorys Liliana Henao Henao

Jefe departamento: Camilo García Morales

Dedicatoria

A Camilo, Jakobo y Violeta por su amorosa paciencia y compañía.

Agradecimientos

A mi familia, por estar siempre ahí.
A Fabián y Jaider, gracias por creer en mí y por acompañarme en este proceso.
A mis amigas y amigos, por todo lo aprendido en el camino.
Y a cada una de las personas que hicieron de mi paso por la universidad, una etapa maravillosa.

Tabla de contenido

Lista de figuras

Lista de tablas

Resumen

El crecimiento exponencial de la información digital ha llevado a las ciencias de la información a adaptar sus procesos de organización, búsqueda, análisis y tratamiento de la información y es por ello por lo que se hace necesario vincular herramientas de otras áreas de conocimiento como el procesamiento de lenguaje natural, la lingüística de texto y el análisis computacional para la revisión de grandes volúmenes de información.

Este proyecto pretende diseñar un modelo de análisis computacional para la clusterización automática de los documentos de políticas públicas CONPES 2010-2020, que permita el análisis de la información digital que contienen y que facilitará la selección, evaluación y tratamiento de diversos aspectos de las políticas públicas como población, políticas, estrategias, actores y planes de acción.

Palabras clave: procesamiento de lenguaje natural, análisis lingüístico, análisis computacional, políticas públicas, conpes.

Abstract

The exponential growth of digital information has led Librarianship to adapt its organization, search, analysis and information processing processes and that is why it is necessary to link tools from other areas of knowledge such as natural language processing, text linguistics and computational analysis for reviewing large volumes of information.

This project aims to design a computational analysis model for the natural language processing of CONPES 2010-2020 public policy documents, which allows the automated analysis of the digital information they contain and which will facilitate the selection, evaluation and treatment of various aspects of public policies. such as population, policies, strategies, actors and action plans.

Keywords: natural language processing, linguistic analysis, computational analysis, public policies, conpes.

Introducción

El presente informe da cuenta de los aspectos contextuales, conceptuales y metodológicos que orientan el proyecto de práctica investigativa *Construcción de un modelo de análisis computacional para la clusterización automática de los documentos de políticas nacionales CONPES 2010-2020*. A partir de este proyecto se buscó desarrollar un modelo que permitiera el acceso, limpieza y agrupación de los textos de los documentos de políticas nacionales CONPES y el proceso que se desarrolló para el cumplimiento de los objetivos se estructura en cuatro partes:

a) Identificación de modelos y requerimientos para el Procesamiento de Lenguaje Natural
b) Análisis de los componentes y requerimientos del procesamiento de lenguaje natural.
c) Desarrollo del algoritmo de análisis computacional en Python para el diseño del modelo de Procesamiento de Lenguaje Natural.
d) Ejecución del modelo de análisis computacional en prueba sobre los documentos de políticas nacionales CONPES para validar su funcionalidad.

A sí pues, en una primera parte se presentan los elementos contextuales y conceptuales que orientaron el desarrollo del proyecto. Seguidamente, se abordan los aspectos metodológicos detallando el enfoque, las técnicas, métodos e instrumentos. En un tercer momento se presentan los análisis y resultados de la recolección de la información, el análisis de la revisión documental y el desarrollo del modelo. Finalmente, se da lugar a la triangulación de los resultados y a las respectivas conclusiones y recomendaciones.

1 Justificación

Las políticas públicas tienen un rol decisivo en la transformación productiva, educativa, tecnológica y social del país. Torres-Melo y Santander mencionan que son:

> El reflejo de los ideales y anhelos de la sociedad, expresan los objetivos de bienestar colectivo y permiten entender hacia dónde se quiere orientar el desarrollo y cómo hacerlo, evidenciando lo que se pretende conseguir con la intervención pública y cómo se distribuyen las responsabilidades y recursos entre los actores sociales. (2013, p. 15)

Las políticas nacionales dan transparencia de la inversión económica y la intervención del estado, su revisión y evaluación favorece la eficacia y la eficiencia en la actuación pública al vincular la formulación de las políticas y su implementación con la gestión de los recursos de la nación. En Colombia, los documentos de políticas públicas son gestionados y almacenados por el Consejo Nacional de Política Económica y social (CONPES) que "es la máxima autoridad nacional de planeación y se desempeña como organismo asesor del Gobierno en todos los aspectos relacionados con el desarrollo económico y social del país" (Colombia. Ministerio de Ciencia, Tecnología e Innovación [Minciencias], 2021), y se encuentran clasificados por tipología y año de aprobación.

Para la comunidad académica, especialmente para las instituciones públicas como la Universidad Antioquia, la revisión de los documentos CONPES es de gran relevancia ya que permiten analizar el potencial de integración en las respuestas colaborativas presentes en las herramientas de planeación en Colombia (Herrera-Kit et al., 2021), por esta razón en un trabajo interinstitucional con la Universidad Externado de Colombia, se pretende a partir de un inventario detallado de los documentos de planeación nacional CONPES indagar acerca de los tipos y la frecuencia con la que sectores diversos han coordinado acciones a lo largo del tiempo, a fin de obtener una base de contraste que permita evidenciar los patrones, prácticas y dinámicas

significativas. Entendiendo la importancia del análisis de las políticas públicas nacionales en la universidad pública y para la bibliotecología ya que estas dan cuenta de la forma en la que se toman decisiones sobre los problemas humanos que son de interés público. Sánchez & Liendo (2020) sustentan que:

> la política pública se convierte en un campo de análisis fundamental para las ciencias sociales en la medida en que configura la posibilidad de operacionalizar la acción del Estado como un proceso de decisión sobre aquello que se considera importante. Es la ciencia del Estado, en la medida en que permite ver cómo se desarrolla la acción frente a esos problemas sociales, cómo se construyen las relaciones y cómo se transforma la realidad a partir de esos procesos decisionales.

De este modo, iniciar con el análisis de sus componentes, actores y tipología implica la revisión individual de los cerca de 6.823[1] documentos disponibles en el portal, lo que lo convierte en un proceso manual y costoso por la inversión de tiempo y personal requerido. Desde la bibliotecología, la clasificación, el análisis y el tratamiento de la información disponible en la web supone también un reto para el profesional de la información ya que el crecimiento exponencial de la información hace difícil su tratamiento con métodos y técnicas manuales. Es por ello, que luego de una revisión inicial a los documentos disponibles y al proceso de captura, organización y revisión de la información, se propone el diseño de un modelo automatizado de clusterización de la información disponible en los documentos CONPES y así brindar mayor cobertura en los documentos que se analizan además de mejorar los procesos de toma de decisiones.

Para dar atención a esta necesidad, se plantea este proyecto de investigación que propone la automatización del proceso de revisión a través de la construcción de un modelo de análisis computacional para para la clusterización automática de los documentos de políticas nacionales de 2010 a 2020 disponibles en el CONPES que permita el análisis completo e intencionado de las políticas públicas y el desarrollo de herramientas de análisis de corpus de texto, lo que permitiría una optimización de recursos y una mayor cobertura y alcance de los documentos revisados. Para la comunidad académica los resultados de este proyecto de investigación serán un aporte a los desarrollos de procesamiento de lenguaje natural que se proponen desde diversas disciplinas,

1 Cifra aproximada https://www.dnp.gov.co/CONPES/documentos-conpes Consultada el 25 de agosto de 2022

además de ser un modelo que puede ser aplicado en la revisión de información disponible en la web y puede derivar en la aplicación de ontologías para la marcación automática de textos, el análisis de sentimientos para identificar las palabras positivas y negativas de un texto, el uso de algoritmos para la clasificación de textos, así como la extracción de entidades, el análisis de frecuencias y la clasificación con vocabularios controlados.

2 Pregunta de Investigación

¿Cómo diseñar un modelo de análisis computacional para la clusterización automática de los documentos de políticas nacionales CONPES de 2010 a 2020?

3 Objetivos

3.1 Objetivo general

Diseñar un modelo de análisis computacional para la clusterización automática de los documentos de políticas nacionales CONPES 2010-2020 para reconocer los patrones, estructuras y modelos de articulación en torno a la política pública.

3.2 Objetivos específicos

- Identificar modelos y requerimientos para el Procesamiento de Lenguaje Natural para saber si pueden ser adaptados en una propuesta de análisis computacional de los documentos CONPES.
- Analizar los componentes y requerimientos del procesamiento de lenguaje natural para establecer aquellos que van a integrar el modelo de análisis computacional de los documentos CONPES.
- Desarrollar el algoritmo de análisis computacional en Python para el diseño del modelo de Procesamiento de Lenguaje Natural.
- Ejecutar el modelo de análisis computacional en prueba sobre los documentos de políticas nacionales CONPES para validar su funcionalidad.

4 Marco Contextual

El desarrollo de políticas públicas en Colombia que este modelo pretende analizar está impulsado por el ***Departamento Nacional de planeación (DNP)*** institución administrativa, que hace parte de la rama ejecutiva y que como lo presenta en su página web oficial:

> impulsa la implantación de una visión estratégica del país en los campos social, económico y ambiental, a través del diseño, la orientación y evaluación de las políticas públicas colombianas, el manejo y asignación de la inversión pública y la concreción de las mismas en planes, programas y proyectos del Gobierno. (Departamento Nacional de Planeación).

El DNP se origina en 1958 como comité asesor del presidente de la república para orientar el ciclo de las políticas públicas, la priorización de los recursos de inversión y la coordinación del Plan Nacional de Desarrollo con los ministerios, departamentos administrativos y entidades territoriales. El DNP se encarga de presentar el diseño, hacer seguimiento y evaluación de los proyectos financiados con recursos nacionales a través de la aprobación y revisión de los documentos CONPES que plasman "las decisiones aprobadas entre diferentes entidades e instituciones del Gobierno nacional, donde se establecen acciones específicas para alcanzar los objetivos propuestos"(DNP, 2022) y que son gestionados por el ***Consejo Nacional de Política Económica y Social (CONPES)*** "máxima autoridad nacional de planeación y se desempeña como organismo asesor del Gobierno en todos los aspectos relacionados con el desarrollo económico y social del país", para lograrlo, orienta a los organismos nacionales encargados de la gestión económica y social a través de los documentos CONPES.

Teniendo en cuenta la importancia de estos documentos y en el ejercicio de analizar el potencial presente en las herramientas de planeación en Colombia y dispuestas en las políticas públicas, el grupo ***Colav***: Colaboratorio de Vinculación para las Ciencias Sociales Computacionales y las Humanidades Digitales el cual está integrado por investigadores de varias facultades y dependencias de la Universidad de Antioquia, en su línea de trabajo de Gobernanza y

políticas públicas (Golab) desarrolla una iniciativa conjunta entre la Universidad de Antioquia y la Universidad Externado de Colombia orientada “la generación de experiencias de innovación pública, gobierno digital, participación ciudadana y utilización de tecnologías de la información e implementación de herramientas computacionales”(Colav, 2022) línea de trabajo que enmarca la formulación de este proyecto de investigación desde el análisis computacional aplicada a los documentos CONPES.

5 Marco Conceptual

Para la realización y desarrollo de este proyecto de investigación, es indispensable definir los conceptos claves, con el fin de obtener una buena comprensión conceptual y fundamento teórico en el contexto de aplicación de la investigación. Como concepto más amplio del tema, el ***Análisis Lingüístico del texto*** hace referencia a la lingüística cuyo objeto de estudio como lo presenta el Centro Virtual Cervantes (2022) es el texto y se convierte en insumo para comprender la estructura de los documentos CONPES que se desean analizar. Crystal (2008) define el análisis lingüístico del texto como "una rama de la lingüística que se ocupa de la descripción y el análisis de textos extensos (hablados o escritos) en contextos comunicativos" (p.3), es decir que el análisis de los documentos implica más que el estudio de las oraciones como lo propuso Chomsky y se analice también el discurso y la intensión comunicativa. Aplicar este concepto en el desarrollo del proyecto permitirá comprender las estructuras lingüísticas como la macroestructura, superestructura e identificar otras unidades de análisis dentro de los corpus textuales (Bernárdez,1982).

> La aplicación del análisis lingüístico desde la perspectiva computacional que se espera implementar desde este proyecto implica el desarrollo de algoritmos de Análisis de texto computacional para automatizar tareas lingüísticas que se hacen de forma manual, tales como la captura y el marcado de texto.

El Análisis de texto computacional hace referencia a las herramientas diseñadas para el análisis de textos digitales y la transformación de textos en datos cuantitativos [2] procesables. O´Connor & et al (2011, p.1) exponen que "el análisis computacional se basa en técnicas de procesamiento de lenguaje natural, recuperación de información, minería de texto y aprendizaje automático y debe entenderse correctamente como una clase de metodologías cuantitativas de las ciencias sociales". Término que resulta pertinente de acuerdo con el tipo de investigación de enfoque social que se pretende realizar y que aporta técnicas y herramientas para el análisis masivo de corpus textuales lo que representa una ventaja para abordar el volumen de documentos de

2 Justin Chun-ting HO

políticas públicas dispuestos para revisar. La Biblioteca Schaffer (2019) refiere que puede emplearse el análisis de texto computacional para identificar patrones y tendencias a gran escala y que sus técnicas incluyen el análisis de palabras claves, el reconocimiento de entidades nombradas, el análisis de sentimientos, la estilometría, el modelado de temas y el modelado.

El análisis de texto computacional facilitará la transformación de los documentos de políticas nacionales en datos procesables, lo que permitirá el procesamiento de un mayor volumen de documentos en menor tiempo vinculando a esta estrategia modelos de procesamiento de lenguaje natural. Andrea, P., & Cañón, B (2007) presentan el procesamiento de lenguaje natural, como una subdisciplina de la inteligencia artificial y rama de la ingeniería lingüística computacional que busca crear programas que puedan analizar, entender y generar lenguajes de comunicación entre los humanos y las maquinas.

Para aplicar el análisis lingüístico, computacional y las herramientas asociadas a la revisión y preparación de los corpus textuales, es necesario entender el ***Procesamiento de Lenguaje Natural (PLN)*** como disciplina de la inteligencia artificial y la lingüística computacional Según (Zeroual & Lakhouaja, 2018) y que tiene como objetivo "aprender, comprender, reconocer y producir contenido de lenguaje humano". Es decir, trata de lograr una comunicación eficiente entre el humano y la máquina. La digitalización ha brindado a los profesionales de la información la posibilidad de recopilar una gran cantidad de datos textuales para investigar fenómenos contemporáneos. El Procesamiento del lenguaje natural, permite a los académicos extraer fácilmente información valiosa contenida en conjuntos de datos textuales y evitar el trabajo de análisis documental individual.

Esta técnica tiene sus orígenes en los años cincuenta de la mano con el surgimiento de la inteligencia artificial y pueden identificarse tres tipos de procesamiento diferentes:

Interfaces en lenguaje natural. Cuyo objetivo es el desarrollo de un interfaz de comunicación que permita a la máquina interpretar y ejecutar órdenes que se expresan en lenguaje natural. Su desarrollo podría facilitar a los usuarios la localización de recursos, la comunicación entre aplicaciones informáticas, la búsqueda de información, entre otras actividades realizadas en la Web. (Solís Sánchez, Florencia Juárez, & Acosta Guadarrama, 2018)

Traducción automática. Es este elemento uno de los objetivos principales del Procesamiento de Lenguaje Natural y permite obtener corpus anotados a partir de corpus provenientes de otros idiomas, los cuales pueden ser aplicables a diferentes tareas de procesamiento de lenguaje natural (PLN). (Peña, J. Bucheli, V. & Gutiérrez, R, J., 2022)

Procesamiento de textos. Ante el crecimiento exponencial de la información textual es necesario el desarrollo de técnicas y algoritmos para el procesamiento de textos que no solo trabajen en la extracción automática de textos si no en el análisis de la información dispuesta en ellos. El procesamiento de lenguaje natural es un área central en cualquier aplicación de análisis de datos textuales. (Talamé, L., Cardoso, A., & Amor, M,2019)

Algunos enfoques de PLN se basan en técnicas de aprendizaje automático para el análisis de textos, pero la principal dificultad radica en la identificación de los textos ya que requiere un análisis exhaustivo desde el nivel referencial, estructural, léxico y pragmático de los documentos. Aunque las herramientas computacionales para el análisis lingüístico de texto son de un enfoque más cuantitativo, en este caso se aplicaran a una propuesta social desde la ***Bibliotecología*** como herramientas que permitan evidenciar la ejecución de los planes de gobierno a través de las políticas públicas.

Para la ejecución de este proyecto, se desarrolla un algoritmo apoyado en una herramienta fundamental para el Procesamiento de Lenguaje Natural y es la ***Clusterización***, que tiene como propósito segmentar la información de acuerdo con unos criterios definidos de similitud, de cumplimiento de características o patrones (Sanabria, 2004). por medio de esta técnica, se consigue la agrupación de corpus textuales en este caso los documentos CONPES a base de características comunes.

Las técnicas de clusterización son técnicas de clasificación no supervisada de patrones en conjuntos denominados clústers previamente determinados, de manera que los miembros de cada grupo estén lo más cerca posible de sus centroides. El algoritmo funciona de forma iterativa, actualizando el centro de los clústeres de manera de ir reduciendo las distancias entre los miembros de cada cluster y su centro Lee t al. (2012). La mayoría de los algoritmos de agrupamiento se basan en dos técnicas denominadas agrupamiento jerárquico y agrupamiento de partición. El primer

método es dividir la matriz grande (que contiene todos los objetos) en matrices más pequeñas hasta obtener una matriz de un solo elemento. La técnica de unión, por otro lado, es lo opuesto a la técnica de división. Su construcción comienza con los elementos de la matriz, que se almacenan en la página y se interpretan como matrices únicas. Luego hay un paso donde dos subconjuntos similares se combinan y almacenan en nodos. Este proceso se repite hasta llegar a la raíz del árbol que contiene todos los elementos del arreglo (Nielsen, 2016).

Para el caso de los algoritmos de agrupamiento particionado Nada & Panda (2014) mencionan, que consiste en dividir el conjunto de datos en varios grupos de acuerdo con un criterio seleccionado, llamado medida de aptitud, que afecta directamente la naturaleza de la clusterización. Las técnicas de clusterización han sido ampliamente utilizadas en múltiples aplicaciones tales como reconocimiento de patrones, análisis de datos, procesado de imágenes o estudios de mercado. Gracias a la clusterización, se pueden identificar regiones tanto pobladas como dispersas y, por consiguiente, descubrir patrones de distribución general y correlaciones interesantes entre los atributos de los datos y para este caso puntual, el análisis de los documentos CONPES que contienen las políticas públicas.

Torres-Melo & Santander (2013) definen las P***olíticas Públicas*** como "una construcción social donde el gobierno, como el orientador de la acción colectiva, interactúa con múltiples y diversos actores sociales y políticos" que permiten legitimar al estado como proveedor de servicios que supla las necesidades y solucione problemas públicos. Las políticas son determinantes para la ejecución de la gestión pública y deben ser constantemente medidas y evaluadas, por esta razón resulta pertinente la iniciativa de diseñar un modelo de análisis computacional para el para la clusterización automática de las políticas nacionales que permita hacer veeduría y seguimiento a la administración pública.

Las políticas públicas como resultado de la construcción colectiva deben disponerse a través de herramientas que proporcione el gobierno, para que la ciudadanía se informe y participe. En Colombia, se ha adoptado el modelo de ***E-gobierno*** que como lo define Organización de Estados Americanos (OEA), es una estrategia de gestión gubernamental que "utiliza las tecnologías de información y comunicación para ayudar a los gobiernos a ser más accesibles a los electores, mejorar los servicios y a ser más eficientes, y a estar cada vez más conectados con otras partes de la sociedad" (2022), un mecanismo innovador que le permite al gobierno nacional garantizar la

transparencia, la participación y la colaboración de las políticas públicas optimizando el uso de los recursos públicos.

Esta estrategia de gobierno abierto es la que permite acceder a la información de políticas públicas dispuesta en los documentos CONPES, para ser recolectada, estructurada y posteriormente analizada a través del modelo de análisis que este proyecto propone.

6 Metodología

6.1 Línea de Investigación

El diseño del modelo de análisis computacional para para la clusterización automática de las políticas nacionales CONPES 2010-2020, se enmarca en la línea de investigación **Estudios interdisciplinarios de la gestión de la información y el conocimiento** de la Escuela Interamericana de Bibliotecología ya que como presenta en su página web, "esta línea aborda temas interdisciplinarios, nuevas tendencias en el uso de la información y tecnologías aplicadas en la producción, organización, transferencia, comunicación, uso y apropiación de la información"(Escuela Interamericana de Bibliotecología, 2022).

6.2 Enfoque de Investigación

Este proyecto es una investigación de enfoque mixto que Johnson y Onwuegbuzie (2004) definen como "(…) el tipo de estudio donde el investigador mezcla o combinas técnicas de investigación, métodos, enfoques, conceptos o lenguaje cuantitativo o cualitativo en un solo estudio" (p. 17). Ya que, para abordar la pregunta de investigación, es necesario utilizar técnicas de enfoque cualitativo como el análisis de contenido y el uso de técnicas y herramientas de enfoque cuantitativo como la revisión de documentos.

6.3 Nivel de Investigación

En cuanto al nivel de esta investigación según Hurtado de Barrera (2000) y su metodología de la investigación holística se ubica en el nivel ***Integrativo Interactivo*** que implica:

> "La realización de acciones por parte del investigador, ya sea solo o conjuntamente con algún grupo o comunidad, con el propósito de modificar una situación o evento (…) y ejecutar acciones para modificar un evento, y recoge información durante el proceso con el fin de reorientar la actividad." (Hurtado de Barrera, 2000)

Lo que se ajusta al objetivo de este proyecto que es diseñar un modelo de análisis computacional que permita la clusterización automática de los documentos de políticas nacionales, el cual será revisado en un proceso de aseguramiento de la calidad del algoritmo por un equipo de trabajo especializado en políticas públicas específicamente en documentos CONPES y se ajustará de acuerdo con la información recolectada en la revisión.

6.4 Técnicas e Instrumentos

Teniendo en cuenta el enfoque metodológico del proyecto y para dar cumplimiento a los objetivos propuestos para este trabajo, se proponen las siguientes técnicas e instrumentos:

Para abordar el primer objetivo que pretende Identificar modelos y requerimientos para el Procesamiento de Lenguaje Natural para saber si pueden ser adaptados en una propuesta de análisis computacional de los documentos CONPES , se realizó una *Revisión Documental* que como plantea Hernández Sampieri (2014) "implica detectar, consultar y obtener la bibliografía (referencias) y otros materiales que sean útiles para los propósitos del estudio, de donde se tiene que extraer y recopilar la información relevante y necesaria para enmarcar nuestro problema de investigación". (p.63). Silamani Guirao (2015) ha definido ocho tipos de revisión documental "narrativa, integradora, panorámica, análisis conceptual, sistemática, sistematizada, revisión de revisiones, realista" y para este proyecto se estableció una revisión documental panorámica, que "tienen por objeto identificar rápidamente los conceptos clave que sustentan un área de investigación y las principales fuentes y tipos de evidencias disponibles."

Para la revisión documental, se establecieron inicialmente las palabras claves asociadas al tema de la investigación y sus equivalentes en inglés y portugués, y adicionalmente algunos términos asociados que aportan en la obtención de resultados más idóneos (Tabla 1.)

Tabla1

Términos de búsqueda.

Palabras Claves	Keywords	Palavras Chaves
Español	**Inglés**	**Portugués**
Políticas Públicas	**Public politics**	**Políticas públicas**
Políticas Nacionales	National Policies	Políticas Nacionais
Política Interna	Internal Policy	Política Interna
Política Interior	Domestic policy	Politica domestica
Políticas interiores	internal policies	
Análisis computacional	**Computational analysis**	**Análise computacional**
Análisis de algoritmos	Algorithm analysis	Análise de algoritmo
Procesamiento de Lenguaje Natural	**Natural Language Processing**	**Processamento de linguagem natural**
PLN	NLP	NLP
Análisis del lenguaje	language analysis	análise de linguagem
CONPES	**CONPES**	**CONPES**
Consejo Nacional de política y economía social	Consejo Nacional de política y economía social	Consejo Nacional de política y economía social
Modelos	**Models**	**Modelos**
Prototipo	Prototype	Protótipo
Análisis Lingüístico	**Linguistic Analysis**	**Análise Linguística**
Análisis lingüístico de texto	Linguistic analysis of text	Análise linguística do texto
Vectorización	**Vectorization**	**Vetorização**
Análisis Vectorial	**Vector analysis**	**Análise vetorial**
Clustering	**Clustering**	**Agrupamento**
Clusterización		Clustering
Análisis Multivariado	**Multivariate analysis**	**Análise multivariada**

Estos términos de búsqueda se organizaron con operadores booleanos y comillas a través de ecuaciones de búsqueda para facilitar la recuperación de la información, las cuales se aplicaron en fuentes de búsqueda nacionales e internacionales que incluyen repositorios, buscadores, directorios de revistas, revistas como **E-*lis, Git Hub, DOAJ, Google, Google Scholar,***

Códice, El Repositorio de la Universidad Javeriana y el Repositorio de la Universidad de Antioquia, los resultados obtenidos se dispusieron en una bitácora (**Tabla 2**) que registra la fecha de búsqueda, la ecuación de búsqueda utilizada, los resultados obtenidos con su aplicación y la cantidad de artículos con mayor relevancia para el desarrollo del trabajo y cuyo periodo de publicación fuera entre 2018 y 2022.

Tabla 2
Bitácora de búsqueda

Construcción de un modelo de análisis computacional para la clusterización automática de los documentos de políticas nacionales CONPES.					
Fecha de Inicio de la búsqueda:		Fecha final de la búsqueda:			
FECHA	RECURSO	ECUACIONES DE BÚSQUEDA	RESULTADOS	PERTINENCIA	OBSERVACIONES

Como instrumento para este objetivo se diseñó una ficha ***de Contenido***, que como lo define Carreras Panchón son “las que realizamos para anotar el contenido de los trabajos leídos. Estas fichas son indispensables para la elaboración de la introducción y discusión de los trabajos”. (p.10) lo que permitirá organizar la información recopilada de manera ordenada y metódica para su posterior análisis. **Ver anexo 1**.

Para el desarrollo del segundo objetivo que consiste en Analizar los componentes y requerimientos del procesamiento de lenguaje natural para establecer aquellos que van a integrar el modelo de análisis computacional de los documentos CONPES, se realizó un *Análisis comparativo* entendiendo que *“el método comparativo tiene como objetivo la búsqueda de similitudes y disimilitudes” (Sartoti, 1984 como se citó en Tonon, 2011).* Este ejercicio pretende identificar los elementos, flujos de trabajo, librerías de procesamiento de lenguaje natural para el lenguaje de programación Python, modelos de preprocesamiento de texto e incluir otros elementos que sean fundamentales para el diseño del modelo que se pretende construir.

Y para ello se empleó una *matriz comparativa* que como lo presenta la Universidad de Extremadura en su portal del Sistema de Bibliotecas (2021):

> "permite organizar la información de acuerdo con unos criterios previamente establecidos. La finalidad principal es establecer las diferencias entre los conceptos que se tratan. Está formado por un número determinado de columnas, permite identificar los elementos que se desea comparar y permite escribir las características de cada objeto o evento".

Esta matriz, analiza los siguientes aspectos:

Tipo de modelo: hace referencia si el modelo que presenta el recurso corresponde a un algoritmo de programación supervisado o no supervisado.

Tipo de tecnología: permite identificar si en los estudios se aplican elementos de Machine Learning, Deep Learning y las librerías empleadas para el trabajo en Python.

Requerimientos: hace referencia a la función del algoritmo y las necesidades de los usuarios del modelo.

Modelo gráfico: si los recursos analizados presentan flujos de trabajo, flujos de información y diagramas.

Otros elementos de análisis: Debilidades y fortalezas presentes en los documentos a partir de la experiencia del autor o que se evidencian a través de la lectura.

Observaciones: Un espacio para comentarios, dudas, nuevos aspectos a considerar, entre otros.

Lo que facilitará la identificación de los componentes y requerimientos para el procesamiento de lenguaje natural. **Ver anexo 2.**

Para lograr el tercer objetivo del proyecto que consiste en Desarrollar el algoritmo de análisis computacional en Python para el diseño del modelo de Procesamiento de Lenguaje Natural, se plantea un *modelo de Arquitectura de Software* que García-Holgado y García-Peñalvo (2014) definen como:

> "Los modelos construidos a partir de las diferentes librerías, (…) En dichos patrones se expresa una descripción de los subsistemas y componentes de un sistema de software que establece las relaciones entre ellos. Dichos subsistemas y componentes generalmente se especifican en diferentes vistas para mostrar las propiedades funcionales y no funcionales relevantes de un sistema de software, la definición de un patrón de arquitectura es el resultado principal de la actividad de diseño de software". (p.94)

El uso de este modelo permitirá la aplicación de los elementos identificados en la matriz comparativa y hacer el diseño inicial del algoritmo. Se proponen dos instrumentos para abordar este objetivo, inicialmente realizar una ***matriz de requerimientos técnicos*** que Madrid, F., & Fernanda, L. (2020) define como "los requerimientos de operación y funcionamiento que debe tener el algoritmo para que este pudiera ser implementable" (p.12) para establecer los parámetros de trabajo del algoritmo. (***Tabla 3***)

Tabla3

Requerimientos Técnicos.

Construcción de un modelo de análisis computacional para la clusterización automática de los documentos de políticas nacionales CONPES 2010-2020
Requerimientos técnicos
R 001 - Obtención de información relevante y específica
Para el diseño y construcción del modelo.
R 002 - Diseño del modelo de análisis computacional para la clusterización automática de los documentos de políticas nacionales
Esta función debe generar un modelo de captura de documentos CONPES y hacer un preprocesamiento del corpus textual para su posterior clusterización.
R 003 - Creación de un servicio con potencial de escalabilidad e interoperabilidad entre otros servicios
Que el modelo pueda ser ajustado y aplicado en otros documentos o periodo de tiempo.
R 004 - Creación de un servicio que sea consumible desde diferentes usuarios
El servicio debe poderse visualizar en cualquier equipo mientras este tenga un navegador y acceso a internet.
R 005 - Calidad
La funcionalidad del algoritmo debe ser de buena calidad para generar confiabilidad a los usuarios.
R 006 - Rendimiento
El servicio debe tener una adecuada velocidad de respuesta y el menor consumo de recursos posible
R 007 - Mantenimiento
El sistema estará en constante mantenimiento, verificación y actualización.
R 008 - Restricciones
Para el acceso y uso del modelo.

Posteriormente, se espera realizar el ***Diseño del Algoritmo*** para el análisis computacional que Gallardo y García definen como "una secuencia ordenada de pasos que conduce a la solución de un problema concreto, sin ambigüedad alguna, en un tiempo finito". (p.3) Los algoritmos son independientes tanto del lenguaje de programación como del ordenador que los ejecuta y deben ser precisos, definidos y finitos. El diseño se elaborará en lenguaje de programación Python en el entorno de ***Google Colaboratory*** que Baume (2021) define como una herramienta para escribir y ejecutar código Python en la nube de Google para fomentar la investigación sobre Aprendizaje de Máquina e inteligencia Artificial. (p.1), este instrumento permitirá el diseño y verificar la viabilidad de su implementación. **Ver anexo 3.**

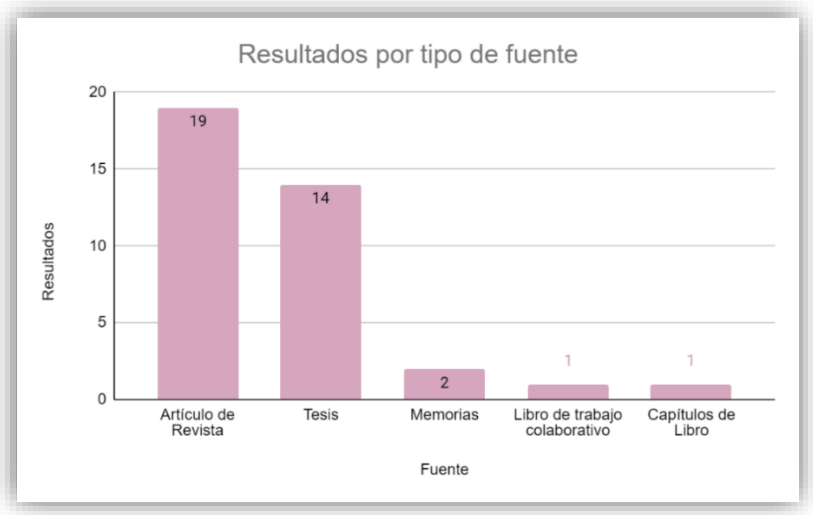
7 Resultados

7.1 Revisión de Literatura

Para iniciar la búsqueda de información, se realizó una selección de palabras claves y términos asociados al tema y título propuesto para el proyecto *Construcción de un modelo de análisis computacional para la clusterización automática de los documentos de políticas nacionales CONPES 2010-2020* las cuales se organizaron en ecuaciones de búsqueda acompañadas de comillas y operadores boleanos para facilitar la recuperación de información y así iniciar una búsqueda en los repositorios y revistas propuestos. En una bitácora se registra la fecha de búsqueda, el recurso empleado con enlace vinculado a los resultados, la ecuación de búsqueda utilizada, los resultados obtenidos con su aplicación y la cantidad de artículos con mayor relevancia para el posterior desarrollo del trabajo.

Como lo demuestra la bitácora, con la aplicación de las ecuaciones de búsqueda, se obtuvieron se establecieron cuarenta y seis ecuaciones de búsqueda con las palabras claves establecidas asociadas al proyecto las cuales se aplicaron en las fuentes de información establecidas previamente y se obtuvieron 699.998 resultados de los cuales se seleccionaron 37 documentos de diversas fuentes por su pertinencia para para la fundamentación teórica del proyecto, la conceptualización, por presentar experiencias de aplicación de modelos computacionales de procesamiento de lenguaje natural y librerías y comandos de Python que puedan aportar al diseño del modelo este proyecto (**¡Error! No se encuentra el origen de la referencia.**)

Figura 1
Resultados de búsqueda por fuente



Esta revisión permitió reconocer el desarrollo del procesamiento de lenguaje natural a través de modelos computacionales, las condiciones y limitaciones de su aplicación, además de identificar y conceptualizar los elementos necesarios para la construcción de un algoritmo de análisis. De esta revisión inicial se pudo establecer que la producción de literatura asociada al procesamiento del lenguaje natural de corpus textuales en español es limitada. Además, que no existe un modelo de análisis computacional que pueda ser aplicado o adaptado al procesamiento de los documentos CONPES lo que refuerza la pertinencia de la construcción del modelo que se pretende desarrollar en este proyecto de investigación.

El repositorio *Git*Hub es un servicio en línea para el alojamiento de código fuente y la colaboración en proyectos de software, Dabbish et al. (2012) mencionan que en esta herramienta se almacenan, evalúan, se documentan y organizan sistemáticamente los proyectos de código en acceso abierto para ser descargado y revisado por cualquier usuario por lo que era fundamental realizar una revisión de esta fuente. El interés inicial era recuperar modelos, librerías, o código disponible para la revisión de políticas públicas, PLN y clusterización y de esta búsqueda e encontraron 2.189 resultados asociados al *Procesamiento de Lenguaje* Natural, 5.485 asociados a la *clusterización* y de ellos 1.515 pertenecen al lenguaje de programación Python que es lenguaje elegido para desarrollar el modelo. Por tratarse de un repositorio de acceso libre, algunos de los resultados obtenidos corresponden a códigos inconclusos o ejercicios prácticos para el aprendizaje

sobre PLN y clusterización mas no de modelos establecidos para el análisis textual de políticas públicas, pero si bien esta revisión no evidenció un modelo aplicable a los documentos de políticas nacionales CONPES, si existe una librería desarrollada por el DNP denominada *ConTexto* que proporciona herramientas para realizar análisis de texto usando el lenguaje de programación Python y que surge como solución a tres principales aspectos:

> "primero, la necesidad de integrar todos los esfuerzos y desarrollos que ha hecho la Unidad de Científicos de Datos (UCD) del DNP, en proyectos relacionados con la analítica de texto, segundo, evitar reprocesos en la construcción de scripts para estas tareas, y finalmente, aumentar la cantidad de librerías enfocadas en el análisis de texto en español que existen actualmente". (DNP, 2021)

Esta librería presenta información relevante para el procesamiento de texto y al ser un algoritmo adaptado a los documentos CONPES tiene un desarrollo el análisis y preprocesamiento de los mismos, pero no aborda los elementos de limpieza de texto que garantizan la depuración de otros elementos que pueden generar ruido en el análisis computacional, además no contemplan la clusterización de corpus elemento fundamental de esta iniciativa.

7.2 Análisis Comparativo

Posterior a la revisión de literatura se analizan los documentos recuperados que contiene información que se considera relevante en primera instancia y por lo tanto representan la base para la construcción del modelo computacional, y se pudo establecer que los modelos de análisis computacional aplicados en los estudios reseñados utilizan elementos de la inteligencia artificial (IA) como el Deep Learning que como mencionan LeCun & Hinton (2015) "permite que los modelos computacionales que se componen de múltiples capas de procesamiento aprendan representaciones de datos con múltiples niveles de abstracción" (p.436) lo que facilita que el algoritmo diseñado descubrir patrones y estructuras en grandes conjuntos de datos y Machine

Learning que busca aprender automáticamente relaciones y patrones significativos a partir de ejemplos y observaciones (Bishop,2006) como es el caso de la clusterización elemento fundamental de este proyecto.

Igualmente, los autores revisados coinciden que para el desarrollo de modelos automatizados existen dos tipos de algoritmos: supervisados y no supervisados, al tratarse este proyecto de la construcción de un modelo de clusterización García & Bellas (2019) plantean que este método de agrupamiento es no supervisado, ya que a diferencia de los métodos de aprendizaje supervisados no requieren un valor o etiqueta de referencia en la salida para realizar su aprendizaje, sino que trata al conjunto de entradas como una serie de variables aleatorias y construye un modelo a partir de ellas, lo que permitirá la agrupación de los corpus textuales basado en la detección de características y correlaciones entre los datos.

La revisión de literatura también demuestra que, para la construcción del algoritmo de captura, preprocesamiento y clusterización existen diversas librerías para trabajar con los corpus textuales y su aplicación va de acuerdo con el tipo de página web que se pretende acceder, el tipo de documentos que se encuentran disponibles, el etiquetado de la página, entre otros. Los hallazgos obtenidos en la revisión de literatura sobre librerías de procesamiento de corpus textuales usadas en Python como Beautifulsoup, Pandas, NLTK fueron verificados al realizar la revisión de los algoritmos alojados en el repositorio de trabajo GitHub

Posterior a la captura de los documentos es necesario realizar un preprocesamiento del corpus para obtener un texto más depurado, ya que al migrar los datos al entorno de trabajo en Google Colab este contiene cadenas de texto sin procesar, y con términos o caracteres que pueden entorpecer el análisis de la información. En la literatura revisada, se encontraron algunos softwars y herramientas que hacen preprocesamiento de texto, como es el caso de ***Freeling,***[3] ***Treetagger***[4] y ***LT TTT***[5] que hacen tokenización automática de textos, ***UAM Corpus Tools*** que es una herramienta

[3] http://garraf.epsevg.upc.es/freeling/
[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
[5] https://www.ltg.ed.ac.uk/

gratuita para la anotación y recuperación de información estadística sobre el texto como la cantidad de oraciones y la longitud de las palabras. La dificultad al emplear este tipo de herramientas es que no pueden adaptarse y no cumplen con los criterios de limpieza que se esperan en el proyecto por lo que se requiere el diseño de un modelo independiente que cumpla con las necesidades específicas que requiere el análisis de los documentos CONPES.

Este modelo tendrá librerías y funciones que se emplean en Python en algoritmos de procesamiento de lenguaje natural y que debido a la eficiencia en la aplicación son reiterativas en la literatura como es ***Natural Language Toolkit*** que permite realizar tareas de preprocesado de texto, entre las que encontramos, la clasificación y tokenización, ***Word tokenize*** que divide el texto en palabras para facilitar su análisis, **Stopwords** que elimina palabras como contracciones y artículos que no son necesarios para el corpus, entre otros.

Para la parte final del modelo, que espera realizar una agrupación de los corpus textuales a través de algoritmos de clusterización la revisión documental permite comprender que hay gran variedad de algoritmos igual de eficaces en el análisis de datos y si bien cada método queda a elección del desarrollador Andritsos (2002) menciona que si existen ciertas características que todo algoritmo de clusterización debería intentar satisfacer y ellas son:

- Escalabilidad: La capacidad del algoritmo de trabajar correctamente con conjuntos de datos de gran tamaño.
- Análisis de mezclas de atributos: capacidad de analizar atributos individuales, así como mezclas de atributos.
- Encontrar grupos con forma arbitraria: lo ideal es que un algoritmo sea lo más genérico posible, siendo capaz de detectar clústeres de cualquier forma.
- Mínimo número de parámetros a introducir: muchos algoritmos requieren que el usuario introduzca parámetros predefinidos para el funcionamiento del algoritmo como, por ejemplo, el número de grupos. Sin embargo, es deseable en grandes conjuntos de datos requieran el mínimo número de inputs de cara a evitar que se encuentren resultados sesgados.

- Soportar ruido: los algoritmos de clusterización deben ser capaces de mostrarse insensibles ante la presencia de desviaciones, para ofrecer resultados de calidad. La detección de anomalías es un problema diferente.
- Insensibilidad al orden de los inputs: hay muchos algoritmos que ofrecen resultados totalmente diferentes cuando el dataset es introducido en órdenes distintos. Esto es indeseable, por lo que un buen algoritmo debe ser insensible al orden en el que los datos se le presentan.

Por tanto, a continuación, se presentan los algoritmos de clusterización identificados en la revisión de literatura con sus características para facilitar la comparación:

Tabla 3

Algoritmos de clusterización.

Algoritmo	Inputs	Optimizados para	Estructura del Cluster
Particionales			
K-means	N° Clústeres	Grupos separados	Esférica
PAM	N° Clústeres	Grupos separados, set de datos pequeños	Esférica
Clara	N° Clústeres	Set de datos grandes	Esférica
Basados en Densidad			
DBSCAN	minPTS	Grupos arbitrarios, set de datos grandes	Arbitratia
HDBSCAN	min_Cluster_Size	Grupos arbitrarios, set de datos grandes	Arbitratia
OPTICS	minPTS	Grupos arbitrarios, set de datos grandes	Arbitratia
DENCLUE	minPTS	Grupos arbitrarios, set de datos grandes	Arbitraria
Basados en cuadrículas			
STING	N° de celdas en nivel inferior, N° de objetos en celda	Set de datos Grandes	Límites verticales y horizontales
CLIQUE	Tamaño de rejilla, N° minimo de pu	Dataset de alto numero de dimensiones	Arbitraria

En el desarrollo del modelo de análisis computacional para la clusterización automática de los documentos de políticas nacionales CONPES 2010-2020 se elige el método K-means que fue propuesto por Stuart Lloyd en 1957 y es uno de los algoritmos más usados por su eficacia y de

acuerdo con la revisión de los diferentes algoritmos de agrupación puede aplicarse cuando se tienen datos no etiquetados, es decir, datos sin categorías o grupos definidos como es el caso de los datos disponibles para trabajar el modelo, además de ser un modelo de fácil aplicación para personas que no son expertas en el uso del lenguaje computacional.

Anil K. Jain (2008) en su conferencia del XIX Congreso Internacional sobre reconocimiento de patrones (ICPR) explica el modelo así:

1. Se escogen K elementos al azar que forman los grupos iniciales, siendo ellos los patrones o centroides (ci) de los mismos.
2. Se asignan los objetos restantes al grupo cuya distancia euclídea sea menor.
3. Se recalcula el centroide de cada grupo como la media de los elementos que componen el mismo, buscando minimizar el índice:

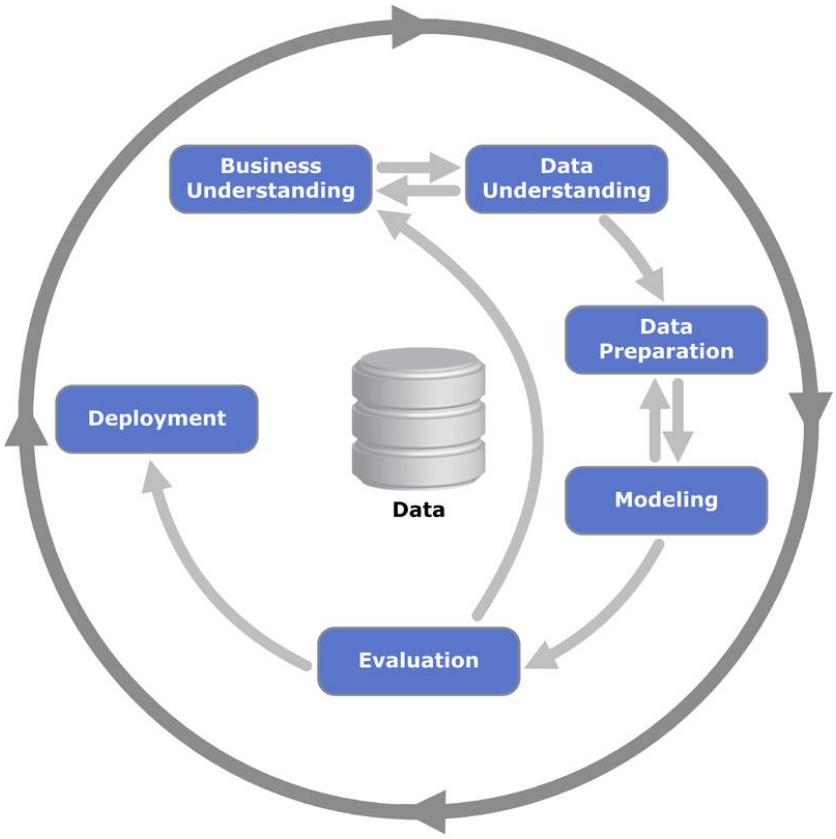
$$J(M) = \sum_{i=1}^{k} \sum_{j=1}^{c_i} \|x_i - m_j\|^2$$ (Ecuación 1)

4. Los dos pasos anteriores se repiten cíclicamente hasta que todos los centroides permanezcan constantes o se cumpla cualquier otra condición de finalización. Aunque no se garantiza una solución óptima, el algoritmo siempre finaliza (converge).

7.3 Diseño del Algoritmo

Para el diseño del modelo de análisis computacional para la clusterización automática de los documentos de políticas nacionales CONPES 2010-2020 se toman elementos de la metodología para análisis de datos CRISP-DM (*Cross Industry Standard Process for Data Mining*), una de las más empleadas actualmente para el desarrollo de proyectos de minería de datos que se presentó en Bruselas en 1999 como una guía de análisis de datos que obedece a un modelo jerárquico que va de lo general a lo específico (Gallardo, 2018).

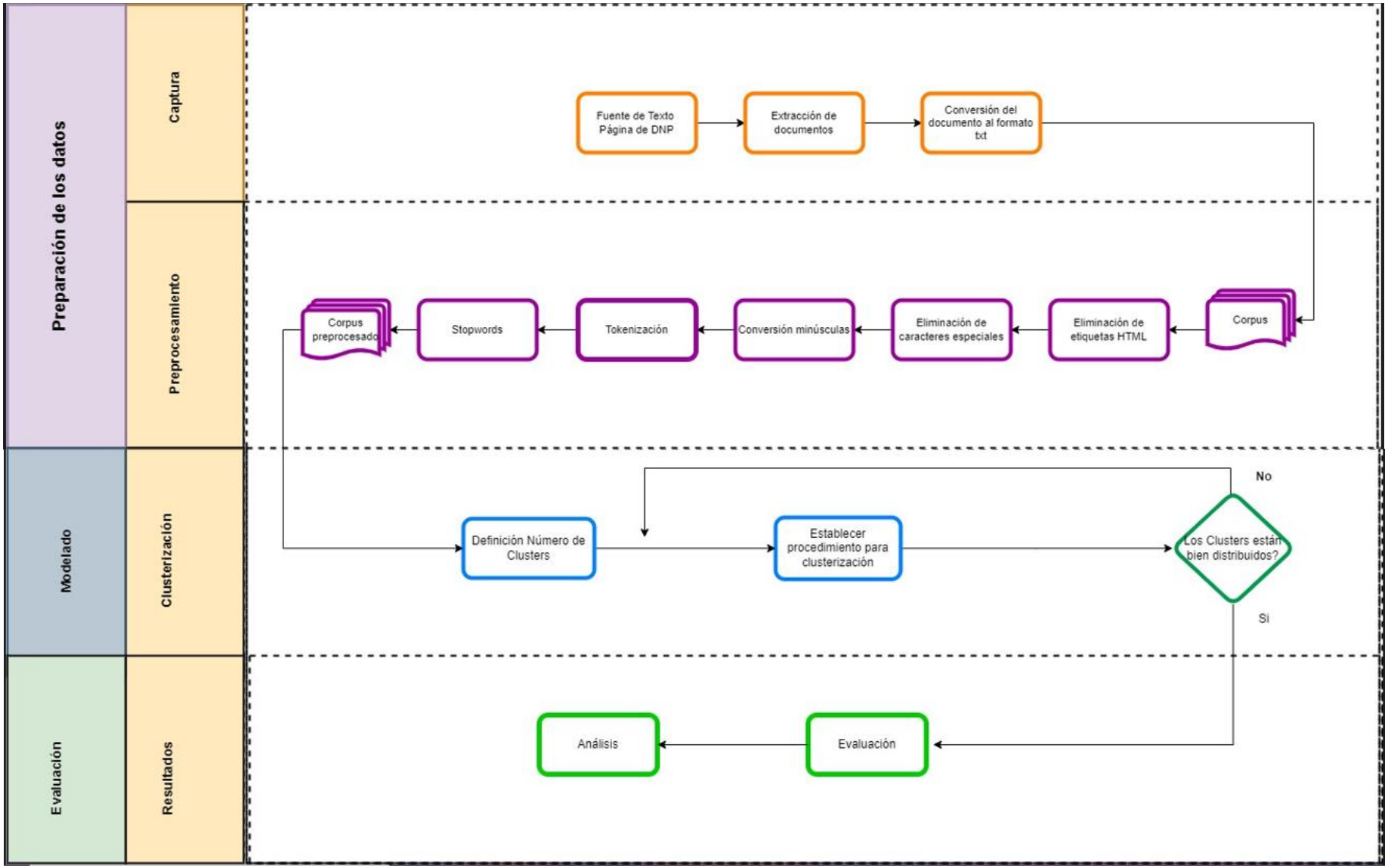
Figura 2
Modelo CRISP



Nota. Tomado de Health Data Miner https://bit.ly/3TMg1Z8

De esta metodología se trabaja específicamente las fases de preparación de los datos, el modelamiento y la evaluación para la creación del modelo de análisis computacional para la clusterización de los documentos de políticas públicas.

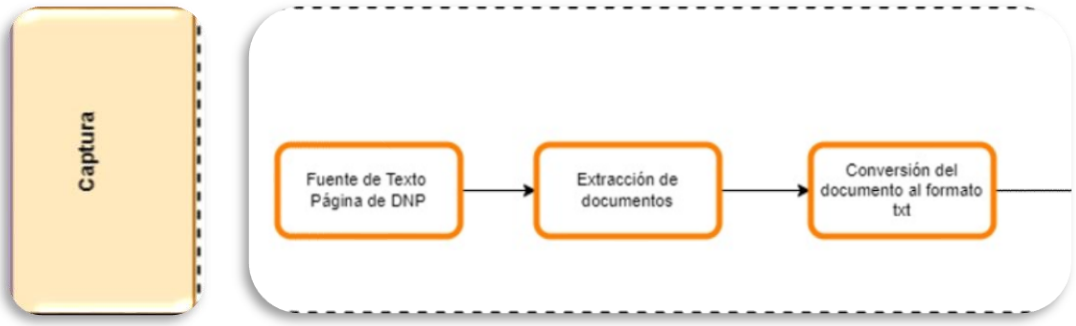
Figura 3
Modelo de análisis computacional para la clusterización automática de los documentos de políticas nacionales CONPES 2010-2020.


Preparación de los datos
Modelado
Evaluación
Captura
Preprocesamiento
Clusterización
Resultados
Fuente de Texto Página de DNP
Extracción de documentos
Conversión del documento al formato txt
Corpus
Eliminación de etiquetas HTML
Eliminación de caracteres especiales
Conversión minúsculas
Tokenización
Stopwords
Corpus preprocesado
Definición Número de Clusters
Establecer procedimiento para clusterización
Los Clusters están bien distribuidos?
No
Si
Evaluación
Análisis

7.3.1 Preparación de los datos.

Esta fase efectúa la recolección inicial de los datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se van a utilizar posteriormente (Rodríguez,2010). Estas pueden ser técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para explotación de los datos y que para este caso en particular es de *Clusterización*. La preparación de los datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. En este modelo esta fase esta compuesta por dos momentos *captura y preprocesamiento*.

7.3.1.1 Captura



El procedimiento de captura inicia con la revisión de la Fuente de Texto que es la página del Departamento Nacional de Planeación donde se encuentran alojados los documentos CONPES https://www.dnp.gov.co/ este portal se encuentra indexado en la página del Gobierno digital de Colombia en la que se da acceso a los datos, trámites e información pública del país. En la parte superior se encuentra un menú desplegable con la información asociada al DNP sobre acceso a la información, atención y servicios a la ciudadanía, estructura organizativa de la entidad, Normativa que orienta la emisión de los documentos de políticas públicas y novedades de prensa asociadas al

Departamento. Se encuentra también un botón de búsqueda que permite la búsqueda de términos sencillos, compuestos, el uso de comillas y operadores boleanos y permite filtrar los resultados por documentos o imágenes asociadas con la búsqueda.

En la franja izquierda del portal, se encuentra un menú vertical que corresponde a la colección de los documentos dispuestos por el DNP para la consulta como los archivos de economía, estudios DNP, la Revista de planeación y desarrollo, libros DNP, el boletín del sistema de indicadores sociodemográficos de Colombia, los boletines de divulgación económica y los documentos CONPES sociales y económicos que son el interés de este proyecto. En este mismo segmento se encuentran filtros para la selección de la información por año de publicación, autor corporativo, autor secundario y autor personal.

Y en la franja central del documento se encuentran listados los documentos de acuerdo con el tipo de documentos seleccionados en la colección. Para este caso de los documentos CONPES, el portal presenta el título del documento, el número del CONPES, el autor, los datos de edición, los descriptores temáticos, el idioma, los descriptores propuestos, los descriptores geográficos, la cantidad de páginas del documento y el acceso al documento en pdf.

Para realizar el diseño del modelo, es preciso identificar la estructura, clasificación y elementos que componen los documentos CONPES por lo que se realizó una revisión de los documentos de políticas públicas de 2010 a 2020 que es el periodo seleccionado para el análisis. En esta revisión se hallaron 1.223 documentos, 204 correspondientes a Conpes Sociales y 1.019 que pertenecen a Conpes Económicos los cuales se encuentran clasificados en cuatro tipos: Concepto favorable, Declaración de importancia, Plan Operativo y Política Nacional y cuya estructura es la siguiente:

- *Título y número de Conpes*: que hace referencia al título del documento y al número asignado por el DNP.
- *Ministerios vinculados en el Conpes:* registra los ministerios que están vinculados en la construcción y publicación del documento.
- *Fecha de publicación:* Fecha de publicación del documento.

- *Firmas:* de los ministerios y los ministros que se encuentran nombrados al momento de la publicación del documento, del director (a) Directora del Departamento Nacional de Planeación, del Subdirector general sectorial y el Subdirector general territorial.
- *Resumen Ejecutivo:* que presenta un resumen o abstract del documento.
- *Clasificación:* Clave alfabética o alfanumérica asignada por el CONPES.
- *Palabras claves:* Términos o materias asociadas al documento.
- *Introducción:* presentación del documento y sus componentes
- *Antecedentes y justificación:* Documentos asociados al documento y la justificación de su elaboración.
- *Recomendaciones:* que van asociadas a la aplicación del documento o la continuidad del proyecto.
- *Anexos:* Los datos que apoyan lo descrito dentro del documento CONPES.
- *Bibliografía:* Las referencias bibliográficas citadas dentro del documento.

Una vez se reconoce la página las características de la página web y los documentos se procede a diseñar el algoritmo para obtener los documentos que se desean analizar. Se establecen el rango de los números de los documentos CONPES que se van a obtener de acuerdo con las categorías social y económico y la url del catálogo de los documentos para acceder a él. (ver figura 4)

Figura 4
Rango de documentos a trabajar

```
# Preparación de los datos
# rango documentos categorias
doc_eco_target = [3636,4021]
doc_soc_target = [131,181]
url_template = "https://colaboracion.dnp.gov.co/CDT/Conpes/%s/%d.pdf"
```

Se construye un ciclo para capturar los documentos y descargarlos en la carpeta seleccionada para ello con la herramienta *requests*[6] que permite conectarse con los servidores que contienen los archivos y realizar solicitudes HTTP a cualquier API, *Os*[7] que permite acceder a funcionalidades dependientes del Sistema Operativo. Sobre todo, aquellas que refieren información sobre el entorno de este y permiten manipular la estructura de directorios, *Re*[8] que define patrones de caracteres a emparejar y extraer de una cadena de texto y *Pdf Plumber* [9]que es una librería de extracción de texto de PDFs, dispone de distintos modos de extracción con los que obtener información carácter a carácter incluso tablas. (ver figura 5)

Figura 5
Captura de documentos

```
# Herramientas descargas
def downloader(url,file_name): # verify: ssl certificates
    '''docs'''
    # descargar pdfs
    response = requests.get(url, verify=False)
    # save pdf
    with open('./'+file_name, 'wb') as f:
        f.write(response.content)

# collector
def collector(categoria,id_doc,url_template,path='./'):
    '''docs here'''
    #
    url = url_template % (categoria,id_doc)
    downloader(url,str(id_doc)+'.pdf') # here!

# collect by ranges
def collector_by_ranges(ranges,categoria,url_template):
    for id_doc in range(ranges[0],ranges[1]):
        collector(categoria,id_doc, url_template)
```

[6] https://datagy.io/python-requests-get-request/
[7] https://uniwebsidad.com/libros
[8] https://docs.python.org/es/3/library/re.html
[9] https://computersciencehub.io/python/pdfplumber-extracting-text-out-of-pdf/

```
all_files = [i for i in os.listdir() if '.pdf' in i]
all_files

def doc_pages_scroller(name_file_input,name_file_output):
    with pdfplumber.open("./"+name_file_input) as pdf:
        pages = pdf.pages
```

Se deben crear dos líneas de código para acceder a los documentos CONPES, de acuerdo con la tipología de documentos Social y económico. (ver figura 6)

Figura 6
Acceso a documentos por tipología

```
collector_by_ranges(doc_soc_target,'Social',url_template)
```

```
collector_by_ranges(doc_eco_target,'Económicos',url_template)
```

De esta captura se obtuvieron 434 documentos en pdf 384 de ellos corresponden a CONPES económicos y 50 a CONPES sociales lo que difiere de los 1.223 que registran en el catálogo del DNP. Se hizo necesario una revisión de los documentos disponibles para verificar las causas de la diferencia o si se debe a un error en la construcción del algoritmo y se evidenció que en el periodo 2010-2020 los documentos CONPES y sus anexos en su mayoría tienen registros duplicados y en algunos casos triplicados. (ver figura 7)

Figura 7
Evidencia de registros duplicados

3708	Concepto favorable a la nación para contratar un empréstito externo con la banca multilateral hasta por la suma de US$ 46.000.000 o su equivalente en otras monedas destinado a financiar el programa de apoyo en gestión al plan de educación de calidad para	4 de noviembre de 2011
3708	3708-Anexo	4 de noviembre de 2011
3708	Concepto favorable a la Nación para contratar un empréstito externo con la banca multilateral hasta por la suma de US$ 46.000.000 destinado a financiar el programa de apoyo en gestión al plan de educación de calidad para la prosperidad	4 de noviembre de 2011
3708	Concepto favorable a la nación para contratar un empréstito externo con la banca multilateral hasta por la suma de US$ 46.000.000 o su equivalente en otras monedas destinado a financiar el programa de apoyo en gestión al plan de educación de calidad para	4 de noviembre de 2011
3708	3708-Anexo	4 de noviembre de 2011
3708	Concepto favorable a la Nación para contratar un empréstito externo con la banca multilateral hasta por la suma de US$ 46.000.000 destinado a financiar el programa de apoyo en gestión al plan de educación de calidad para la prosperidad	4 de noviembre de 2011

Fuente: Dnp

Una vez realizada esta revisión y establecer la cantidad exacta de los documentos, se procede a la transformación de los documentos de pdf a txt, ya que este formato facilita acceder al contenido de los documentos para realizar el procesamiento y posterior clusterización; pero al revisar la migración de los documentos se encontraron 29 archivos que no descargaron correctamente lo que requirió una revisión de los mismos y del acceso a ellos a través del portal del DNP y se pudo identificar que los documentos existen y son de validez, pero no están alojados en el catálogo de documentos y que tienen una variación en identificación del documento y en la url de acceso lo que implicó una labor manual de búsqueda, descargue y almacenamiento en la carpeta de documentos de cada uno de estos elementos y así no dejar documentos sin analizar del periodo establecido para la revisión. (ver figura 8)

Figura 8
Error de acceso al documento

```
<html>
<head>
<meta http-equiv="Content-Type" content="text/html;charset=utf-8" /><meta name="SharePointError" content=""/>
<meta name="Description" content=""/>
<meta name="Keywords" content=""/>
<meta name="Robots" content="NOINDEX"/>
<script type="text/javascript">
<!--
window.location.replace("https:\u002f\u002fcolaboracion.dnp.gov.co\u002fPages\u002fPageNotFoundError.aspx?
requestUrl=https:\u002f\u002fcolaboracion.dnp.gov.co\u002fCDT\u002fConpes\u002fEcon\u00f3micos\u002f3855.pdf")
-->
</script>
</head>
<body>
<!------------------------------------------------------------------------>
File Not Found
<!------------------------------------------------------------------------>
</body>
</html>
```

Para esta conversión se usó un condicional con el módulo de os *Path.exists* [10]que se usa para verificar si la ruta especificada para ubicar el documento existe, el módulo *Mkdir* para crear una nueva carpeta para almacenar los documentos transformados en txt y *Range* para crear una secuencia que recorra los documentos en los parámetros establecidos. (ver figura 9)

Figura 9
Conversión de los documentos en txt

```
# crear directorio txts
if os.path.exists('./txts'):
    pass
else:
    os.mkdir('./txts')

for i in range(1,len(pages)):
    with open('./txts/' + name_file_output + '.txt',"a") as f:
        f.write(pdf.pages[i].extract_text()+'\n')
#print(print(pdf.pages[i].extract_text()))
```

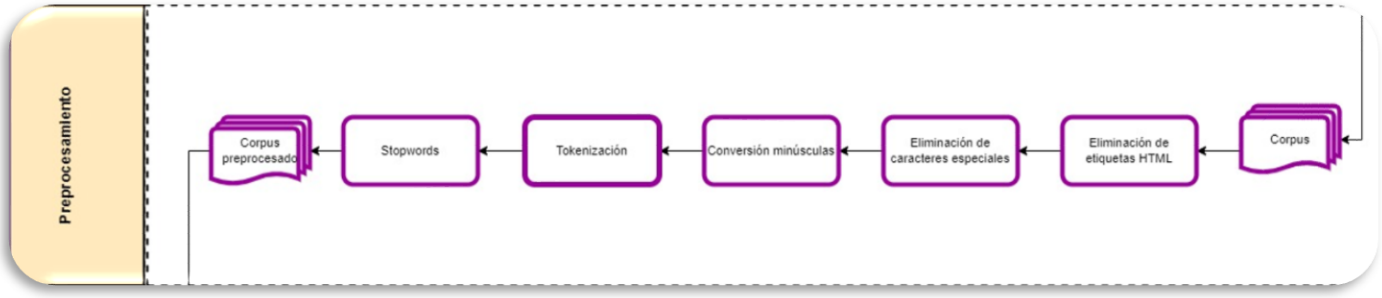
[10] https://www.geeksforgeeks.org/python-os-path-exists-method/

Se escribe un ciclo para el manejo de errores que indique al algoritmo como ejecutarse cuando se ingresa el nombre del archivo no está dentro del directorio y a su vez registre en una lista con los archivos con error y de esta manera finalizar la primera parte de captura. (ver figura 10)

Figura 10
Ciclo para listar errores

```
#archivos txt
errors=[]
for name_file_input in all_files:
    print(name_file_input)
    try:
        doc_pages_scroller(name_file_input,name_file_input)
    except:
        errors.append(name_file_input)
```

7.3.1.1 Preprocesamiento



Pyle (1999) afirma que el preprocesamiento de información tiene como propósito fundamental la “manipulación y transformación de los datos en bruto de manera que permita exponer o al menos facilitar la exposición de la información contenida en el arreglo de datos” (p.32). Esta tarea complementa a la anterior y pueden aplicarse diversidad de técnicas que pueden para optimizar la calidad de los datos disponibles.

Se instala la librería *NLTK*[11] *(Natural Language Toolkit)* que es un conjunto de librerías y programas para Python que permiten llevar a cabo tareas relacionadas con el Procesamiento del Lenguaje Natural de manera eficiente, de esta librería se usa *Word Tokenize* [12]que permite dividir las cadenas de texto en oraciones o palabras para facilitar su análisis, para este modelo se realizará la tokenización por palabras. (ver figura 11)

Figura 11
Ciclo para listar errores

```
#!pip install nltk
```

```
# librerias
#procesamiento de texto
import nltk
from nltk.tokenize import word_tokenize
```

Se emplea de esta librería el módulo *Stopwords*[13] que permite identificar en la biblioteca de PLN un listado de palabras que no aportan información valiosa al texto, palabras como artículos o preposiciones que por el contrario pueden generar distorsión al momento de analizarlo. El listado de stopwords para el idioma español aún es limitada pero este tipo de proyectos contribuyen a su construcción. (ver figura 12)

Figura 12
Descarga de librería NLTK

[11] https://www.nltk.org/
[12] https://www.nltk.org/api/nltk.tokenize.html
[13] https://pythonspot.com/nltk-stop-words/

```
nltk.download('punkt')
nltk.download('stopwords')
```

```
from nltk.corpus import stopwords
```

Se utiliza el método *Text.lower* [14]método convierte todos los caracteres en mayúsculas de una cadena en caracteres en minúsculas y los devuelve. (ver figura 13)

Figura 13
Conversión texto en minúsculas

```
text_min= input (var4)
text_min.lower()
```

Se crea un ciclo que ubique, abra y obtenga de la carpeta creada en el algoritmo anterior con los documentos en formato txt y los almacene en una lista. (ver figura 14)

Figura 14
Almacenamiento de textos en lista

```
listatxt = [s for s in (os.listdir('/content/gdrive/MyDrive/UdeA/Trabajodegrado/Extractos/txts')) if (".txt" in s)]
y.append(s)
y
```

[14] https://www.programiz.com/python-programming/methods/string/lower

```
var3 = ""
var5 = ""
var4 = []

for x in y:
  var3 = open("/content/gdrive/MyDrive/UdeA/Trabajodegrado/Extractos/txts/" + x, 'r')
  var5 = var3.readlines()
  var4.append(var5)
  var3.close()
var4
```

Se aplican las librerías y módulos antes mencionadas a la lista de documentos CONPES y así obtener el corpus textual preprocesado y listo para la siguiente etapa del proceso.

7.3.1 Modelado

En esta fase se seleccionan las técnicas de modelado más apropiadas para el proyecto. Para esta selección, se consideró el objetivo principal del proyecto y la relación con las herramientas disponibles para el análisis de datos a través de algoritmos no supervisados y se ejecuta la técnica seleccionada sobre los datos previamente preparados. (ver figura 15)

Figura 15
Clusterización

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
from sklearn.metrics import adjusted_rand_score

documents = txt_lst

vectorizer = TfidfVectorizer(stop_words=stoplist)
X = vectorizer.fit_transform(documents)

true_k = 45
model = KMeans(n_clusters=true_k, init='k-means++', max_iter=100, n_init=1)
model.fit(X)

print("Top terms per cluster:")
order_centroids = model.cluster_centers_.argsort()[:, ::-1]
terms = vectorizer.get_feature_names()
for i in range(true_k):
    print("Cluster %d:" % i),
    for ind in order_centroids[i, :10]:
        print(' %s' % terms[ind]),
    print

print("\n")
```

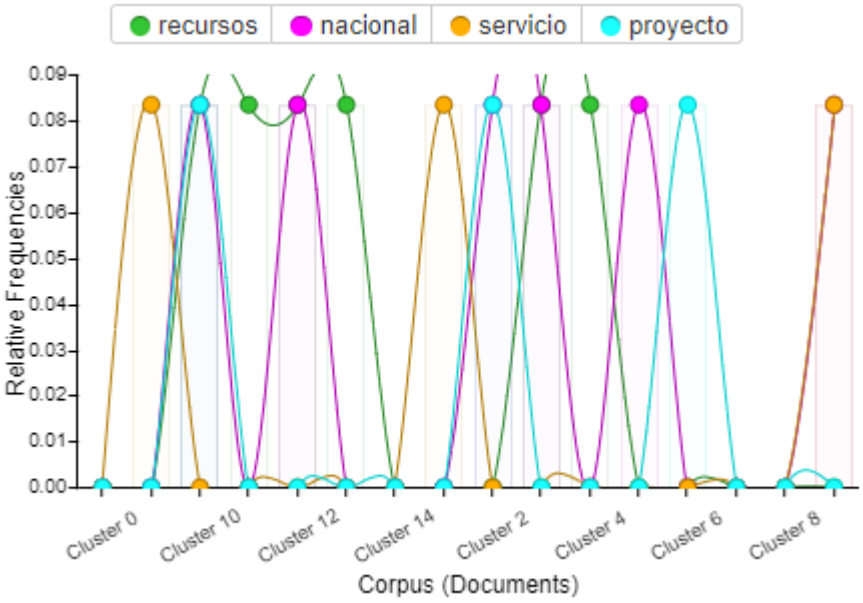
Se obtuvieron 45 clusters agrupados por similaridad textual de los documentos, organizados en grupos de 10 términos de mayor relevancia por documento. (ver figura 16)

Figura 16
Ejemplo de Cluster

```
Cluster 25:
 ordenamiento
 pot
 territorial
 pod
 pemot
 suelo
 territoriales
 planes
 entidades
 desarrollo
```

De acuerdo con la agrupación obtenida se obtuvo una gráfica de tendencia de los documentos. (ver figura 17)

Figura 17
Gráfica de tendencias en los documentos



Además de establecer una gráfica de relaciones entre los términos más usados en el corpus textual. (ver figura 18)

Figura 18
Gráfica de relaciones entre términos

7.3.1 Evaluación

Se analizaron 523 documentos con un promedio de 74 páginas y el tiempo de ejecución del algoritmo es de 19 minutos aproximadamente, lo que presenta una optimización del tiempo y recursos invertidos en el análisis inicial frente a una lectura y selección de términos manual que tarda aproximadamente 45 minutos sin establecer relaciones entre los mismos.

Este modelo, permite al profesional de la información ocuparse de tareas enfocadas al análisis propio de la información.

8 Cronograma

Actividad	Meses/Semanas																	
	1				2				3				4				5	
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2
Revisión del anteproyecto con el Asesor	X	X																
Revisión y ajuste del anteproyecto y de los instrumentos para la recolección de información			X															
Revisión de literatura para identificar modelos y requerimientos del procesamiento de lenguaje natural				X	X													
Analizar los componentes y requerimientos del procesamiento de lenguaje natural a través de una matriz						X	X											
Desarrollar el algoritmo de análisis computacional en Python								X	X	X								
Ejecutar el modelo de análisis computacional en prueba sobre los documentos											X	X	X					
Modelo de Validación														X				
Ajuste del modelo															X			
Desarrollo de documento de consolidación del informe final																X	X	
Socialización de resultados																		X

9 Conclusiones

A continuación, se presentan las conclusiones que reúnen los elementos más importantes y expuestos a lo largo de los resultados.

- No se evidenciaron desarrollos en modelos de procesamiento de lenguaje natural asociados a políticas públicas que facilite el análisis intencionado de las mismas.
- Hay ausencia de modelos, códigos y experiencias en procesamiento de lenguaje natural disponibles en la web lo que hace complejo la creación de un modelo ya que no hay referencias que orienten su desarrollo.
- Este proyecto evidencia la importancia del trabajo interdisciplinar como mecanismo de enriquecimiento de las prácticas bibliotecológicas y en el desarrollo de herramientas que permitan nuevas formas de organización y tratamiento de la información digital.
- El modelo es aplicable a documentos CONPES de otros periodos de tiempo lo que permitiría un análisis conjunto de políticas públicas.
- Si bien los documentos CONPES están dispuestos para consulta pública hay una deficiente disposición de los registros de los documentos CONPES en el archivo de DNP lo que dificulta el desarrollo de modelos automatizados.
- No existe una normalización en las URL de acceso a los documentos CONPES por parte del DNP.

10 Recomendaciones

Los resultados de este trabajo de grado permiten identificar algunos aspectos a mejorar en el proceso de captura y análisis de la información.

- Se sugiere para dar continuidad a la aplicación del modelo en otros periodos de tiempo y así realizar un análisis completo de las políticas públicas colombianas disponibles.
- Se recomienda dar continuidad a esta iniciativa con el desarrollo de nuevos modelos que no solo permitan la clusterización sino la categorización de los documentos, la identificación de actores, la verificación de cumplimiento de las políticas, entre otros elementos de análisis.
- Por otro lado, se sugiere al DNP la depuración de los registros y así evitar la duplicidad de estos.
- La normalización de la dirección URL de acceso a los documentos CONPES por parte del DNP para facilitar el acceso estandarizado a la información.

Referencias

Andritsos, Periklis. (2002). Data Clustering Techniques.

Andrea, P., & Cañón, B. (s/f). *pln procesamiento del lenguaje natural en la recuperación de información*. Rclis.org. Recuperado el 11 de julio de 2022, de http://eprints.rclis.org/9598/1/PROCESAMIENTO_DEL_LENGUAJE_NATURAL_EN_LA_RECUPERACI_N_DE_INFORMACI_N.pdf

Arboleda, S., Sánchez, F., Liendo, N., Ángel, S., Losada, R., Rivas, J., Martínez, D., Muñoz, P., Valencia, M., Acosta, C., Ortíz, C. (2020). *Manual de Ciencia Política y Relaciones Internacionales.* https://repository.usergioarboleda.edu.co/bitstream/handle/11232/1458/El%20ana%CC%81lisis%20de%20poli%CC%81ticas%20pu%CC%81blicas.pdf?sequence=1&isAllowed=y

Baume, G. (2021) Edu.ar. Recuperado el 28 de junio de 2022, de http://fcaglp.unlp.edu.ar/~gbaume/grupo/Publicaciones/Apuntes/GoogleColab.pdf

Bishop, C. M. (2006). Pattern recognition and machine learning (Information science and statistics). Springer-Verlag New York, Inc.

Bernárdez, E. (1982). Introducción a la lingüística del texto. Madrid: Espasa-Calpe

Cervantes, C. C. V. (s/f). *CVC. Diccionario de términos clave de ELE. Lingüística del texto*. Recuperado el 13 de julio de 2022, de https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/linguisticatextual.htm

Consejo Nacional de Política Económica y Social - CONPES. (2021, julio 7). Ministerio de Ambiente y Desarrollo Sostenible. https://www.minambiente.gov.co/planeacion-y-seguimiento/consejo-nacional-de-politica-economica-y-social-conpes/

Contreras, H. Z. (2001). Procesamiento del Lenguaje Natural basado en una "gramática de estilos" para el idioma español. Facultad de Ingeniería, Colombia, Universidad de Los Andes., 54. http://www.saber.ula.ve/bitstream/123456789/13157/1/hc_propuestatesis.pdf

Consejo Nacional de Política Económica y Social, CONPES. (s/f). Gov.co. Recuperado el 19 de julio de 2022, de https://www.dnp.gov.co/CONPES

Crystal,D. Dictionary of Linguistics and Phonetics , 6th ed. Blackwell, 2008

E. Cambria, P. Gastaldo, F. Bisio, and R. Zunino, "An ELM-based model for affective analogical reasoning," Neurocomputing, vol. 149, no. Part A, pp. 443–455, 2015.

Departamento Nacional de Planeación. (s/f). Gov.co. Recuperado el 18 de julio de 2022, de https://www.dnp.gov.co/

Escobar, J., Francy, Y., Bonilla-Jimenez, I., El, U., & Resumen, B. (s/f). Cuadernos Hispanoamericanos de Psicología. Sacopsi.com. Recuperado el 28 de junio de 2022, de http://sacopsi.com/articulos/Grupo%20focal%20(2).pdf

Escuela Interamericana de Bibliotecología. (2022). Líneas de Investigación. Recuperado de LineasinvestigaciónEib

Gallardo, J. (s.f.) Facultad De Ciencias, I. (s/f). *Tema 2. Diseño de algoritmos y programas 1*. Uma.es. Recuperado el 28 de junio de 2022, de http://www.lcc.uma.es/~pepeg/modula/temas/tema2.pdf

Gallardo, J.A. (2018). Metodología para el desarrollo de proyectos en Minería de Datos CRIPS-DM.http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037

García-Holgado, A., & García-Peñalvo, F.J. (2014). Architectural pattern for the definition of eLearning ecosystems based on Open Source developments. *2014 International Symposium on Computers in Education (SIIE)*, 93-98.

Guirao Goris, Silamani J. Adolf. (2015). Utilidad y tipos de revisión de literatura. Ene, 9(2) https://dx.doi.org/10.4321/S1988-348X2015000200002

Herrera-Kit, P., Balanzó Guzmán, A., Parra Moreno, J., & Rivera, M. (2021). Mecanismos de colaboración interinstitucional: prácticas típicas. Innovar, 31(79),145-159 https://doi.org/10.15446/innovar.v31n79.91888

Hurtado, J. (2010). Metodología de la investigación: Guía para la comprensión holística de la ciencia. Quirón Ediciones.

Johnson, B. y Onwuegbuzie, A. (2004, October). Mixed Methods Research: A Research Paradigm Whose Time Has Come [Los métodos de investigación mixtos: un paradigma de investigación cuyo tiempo ha llegado]. Educational Researcher, 33(7), 14-26. Recuperado de http://edr.sagepub.com/cgi/content/abstract/33/7/14

Kontio, J., Lehtola, L. and Bragge, J., Using the Focus Group Method in Software Engineering: Obtaining Practitioner and User Experiences, pp. 271-280, 2004.

LeCun, Y., Bengio, Y. & Hinton, G. Aprendizaje profundo. Naturaleza 521, 436–444 (2015). https://doi.org/10.1038/nature14539

Lee, CY, Antonsson, E., 2000. Agrupamiento particional dinámico utilizando estrategias de evolución, en: 2000 26th Annual Conference of the IEEE Industrial Electronics Society. IECON 2000. 2000 Conferencia Internacional IEEE sobre Electrónica Industrial, Control e Instrumentación. Tecnologías del siglo XXI, IEEE. págs. 2716–2721.

Madrid, F., & Fernanda, L. (2020). Diseño e implementación de algoritmos de procesamiento de lenguaje natural para automatizar la construcción de resúmenes de consultas médicas en la IPS Neumomed S.A.S. mediante técnicas de Machine Learning. Medellín, Colombia.

Colombia. Ministerio de Ciencia, Tecnología e Innovación [Minciencias]. (2021). *Documentos Conpes*. https://bit.ly/3Ed9a5v

Nanda, S.J., Panda, G., 2014. A survey on nature inspired metaheuristic algorithms for partitional clustering. Swarm and Evolutionary computation 16, 1–18

Nielsen, F., 2016. Introduction to HPC with MPI for Data Science. Springer

O'connor, B., Bamman, D., & Smith, N. A. (s/f). *Computational text analysis for social science: Model assumptions and complexity*. Psu.edu. Recuperado el 19 de julio de 2022, de https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.228.3731&rep=rep1&type=pdf

OEA. (2009). *OEA - Organización de los Estados Americanos: Democracia para la paz, la seguridad y el desarrollo*. https://www.oas.org/es/temas/egovt.asp

Peña, J. Bucheli, V. & Gutiérrez, R, J., (2022) View of digital texts as a teaching alternative of the mother language. (s/f). Edu.Co. Recuperado el 5 de noviembre de 2022, de https://revistas.uptc.edu.co/index.php/linguistica_hispanica/article/view/13436/11480

Prieto Rodríguez, M., & March Cerdá, J. (2002). Paso a paso en el diseño de un estudio mediante grupos focales. *Atención Primaria*, *29*(6), 366-373.

Rodríguez, (2010). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM*.

Rubén, E., García, A., & Bellas Bouza, F. J. (2019). *Desarrollo de algoritmos no supervisados para información sensorial en robótica autónoma*. Udc.es. Recuperado el 28 de septiembre de 2022, de https://ruc.udc.es/dspace/handle/2183/25120

Sampieri, R. H., Collado, C. F., Lucio, P. B., Valencia, S. M., & Torres, C. P. M. (2014). *Metodología de la investigación* (6.ª ed.). McGraw-Hill Education.

Sanabria Garzón, J. (2004). Herramienta software para implementar minería de datos: clusterización utilizando lógica difusa. Orinoquia, 8(1), 15-23. Obtenido de https://www.redalyc.org/pdf/896/89680103.pdf

Schaffer Library. (2020). *Subject Research, Course Guides, Documentation: Digital scholarship: Computational text analysis*. https://libguides.union.edu/digital-scholarship/cta

Solís Sánchez, A., Florencia Juárez, R., Acosta Guadarrama, J. C., & López Orozco, F. (2018). Interfaz de lenguaje natural para deducción de información almacenada en ontologías. Research in Computing Science, 147(6), 189–205. https://doi.org/10.13053/rcs-147-6-15

Torres-Melo, J., & Santander, J. (2013). Introducción a las políticas públicas. Recuperado el 25 de julio de 2022, de https://www.funcionpublica.gov.co/eva/admon/files/empresas/ZW1wcmVzYV83Ng==/imgproductos/1450056996_ce38e6d218235ac89d6c8a14907a5a9c.pdf

Tonon, G. (2011). La utilización del método comparativo en estudios cualitativos en ciencias políticas y ciencias sociales: diseño y desarrollo de una tesis doctoral. KAIROS. Año 15. N.º 27. Mayo de 2011. https://dialnet.unirioja.es/servlet/articulo?codigo=3702607

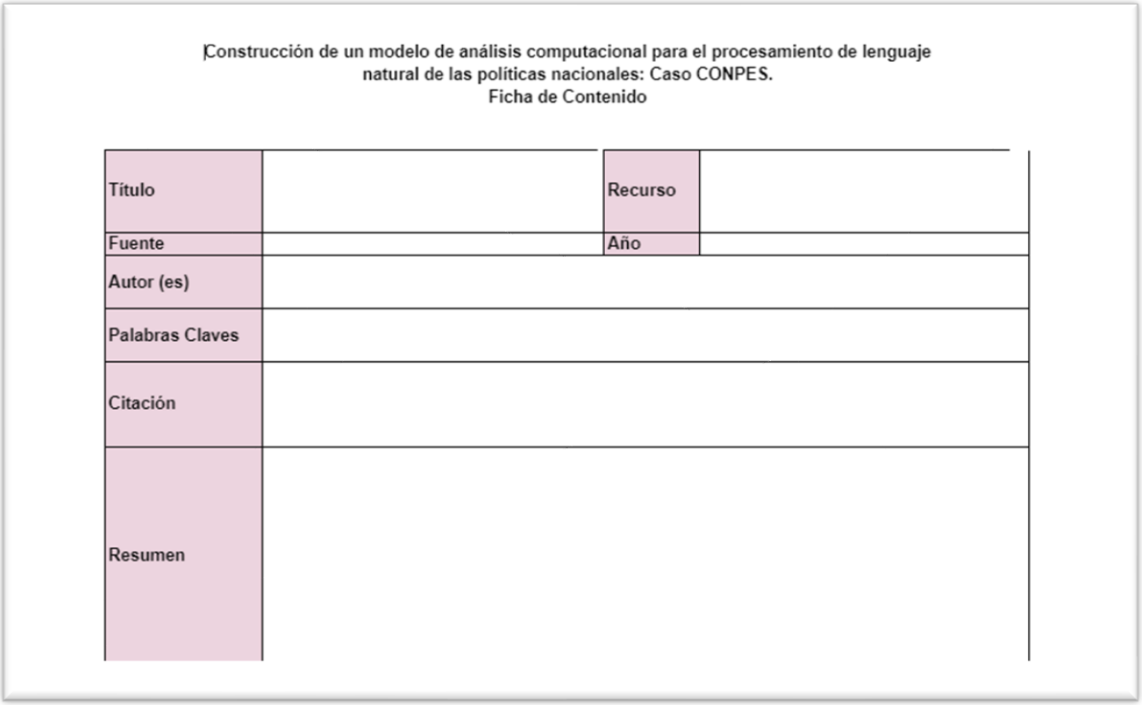
Universidad de Extremadura. Biblioguías: Técnicas de estudio: Cuadro comparativo. (2016). https://biblioguias.unex.es/c.php?g=572102&p=3944896

Zeroual, I., & Lakhouaja, A. (2018). Data science in light of natural language processing: An overview. Procedia Computer Science, 127, 82–91. https://doi.org/10.1016/J.PROCS.2018.01.101

Anexos

Anexo 1. Ficha de Contenido

https://docs.google.com/document/d/1eoL_E35ZUpY2obd-8o3xL0x3E3wUAM8Y/edit?usp=sharing&ouid=110821793224683673303&rtpof=true&sd=true

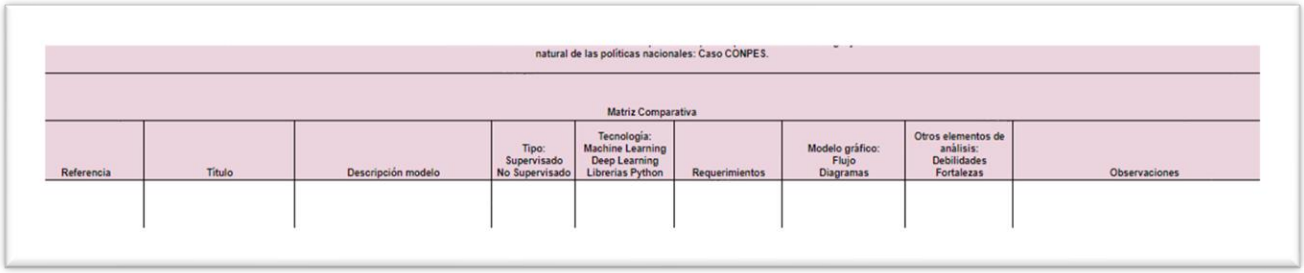


Construcción de un modelo de análisis computacional para el procesamiento de lenguaje natural de las políticas nacionales: Caso CONPES.
Ficha de Contenido

Título		Recurso	
Fuente		Año	
Autor (es)			
Palabras Claves			
Citación			
Resumen			

Anexo 2. Matriz Comparativa

https://docs.google.com/spreadsheets/d/1nZIb-VkM9tsb0XGYqhbTUT3lqYn_DjP1Tvn9rs_m4cI/edit?usp=sharing



natural de las políticas nacionales: Caso CONPES.

Matriz Comparativa

Referencia	Título	Descripción modelo	Tipo: Supervisado No Supervisado	Tecnología: Machine Learning Deep Learning Librerías Python	Requerimientos	Modelo gráfico: Flujo Diagramas	Otros elementos de análisis: Debilidades Fortalezas	Observaciones

Anexo 3. Entorno para diseño de Algoritmo

https://colab.research.google.com/drive/1OPdAYITH6LnErhEoLvrp3w2XqVPWsvxB?usp=sharing