

**IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN
SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES**



FUNDACIÓN
**UNIVERSITARIA
DE POPAYÁN**
35 ANIVERSARIO

CAROL ELIZABETH VASQUEZ BURBANO
JUAN MANUEL BOLAÑOS BOLAÑOS

FUNDACIÓN UNIVERSITARIA DE POPAYÁN
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS
GRUPO DE INVESTIGACIÓN IMS
Popayán, junio de 2019

**IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN
SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES**



FUNDACIÓN
**UNIVERSITARIA
DE POPAYÁN**
35 ANIVERSARIO

CAROL ELIZABETH VASQUEZ BURBANO
JUAN MANUEL BOLAÑOS BOLAÑOS

TRABAJO DE GRADO
INGENIERÍA DE SISTEMAS

DIRECTOR: CRISTIAN CAMILO ORDOÑEZ QUINTERO
CO. DIRECTOR: JOSE ARMANDO ORDOÑEZ

FUNDACIÓN UNIVERSITARIA DE POPAYÁN
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS
GRUPO DE INVESTIGACIÓN IMS
Popayán, junio de 2019

DEDICATORIA

A mis padres por ser un referente de lucha, de amor, de familia, de humildad, de liderazgo, de vida.

Juan Manuel Bolaños Bolaños

En primer lugar a Dios por darme la oportunidad de llegar hasta este momento tan importante de mi formación profesional, quien me ha dado fortaleza para continuar cuando he sentido que mis fuerzas se agotan. A mis padres que han sido mi apoyo incondicional, mi modelo a seguir, mi compañía durante todo mi trayecto estudiantil y de vida. A mi hermano Edison, mi confidente que a pesar de que solo en los últimos años nos hemos acercado, es mi consejero, quien me escucha sin cuestionarme y sin intentar cambiar mi forma de ser. A mis amigos y compañeros los cuales me han acompañado durante todos estos años con su alegría sacando en mí muchas sonrisas.

Carol Elizabeth Vasquez Burbano

AGRADECIMIENTOS

A toda mi familia por ser quienes me han guiado y han estado a mi lado en cada uno de los proyectos que he emprendido en lo que va de mi vida.

A mis amigos por estar ahí en las buenas en las malas, en las alegrías y tristezas, por compartir conmigo tantas experiencias.

A mis maestros que con mucha paciencia ayudaron a que entendiera un poco más lo que pasa con este loco mundo.

A mis compañeros de estudio, un gran grupo lleno de humildad, de compromiso, de trabajo duro, de quienes he aprendido muchísimo y con quienes se han pasado gratos momentos, cinco años, son cinco años.

A nuestro director de trabajo de grado Cristian Ordoñez por ser un gran maestro y su compromiso con algo tan importante como lo es la educación.

A mi compañera de estudio y de tesis Carol Vasquez por ser un referente a lo largo de todos estos cinco años. A TODOS USTEDES GRACIAS

Juan Manuel Bolaños Bolaños

A Dios por tantas bendiciones, por protegerme, darme salud, inteligencia y fuerzas para superar los obstáculos que se me presentan a lo largo de la vida.

A mi madre por confiar en mí, en que lo lograría, que sin duda me ha demostrado su amor incondicional, me ha enseñado el camino del bien y ha celebrado conmigo cada uno de mis triunfos, juntas lo logramos.

A mi padre, que gracias a su temperamento me ha logrado encaminar hacia lo correcto, me ha convertido en una persona valiente y luchadora, que consigue lo que se propone.

A mi hermano, por cuidarme siempre y por aconsejarme cuando lo he necesitado, por su disposición de ayuda durante mi carrera, ya somos colegas.

A Felipe Bedoya, por apoyarme en estos últimos años, por tantas alegrías y por motivarme cuando siento que mis fuerzas se acaban.

A Cristian Ordoñez, por el acompañamiento durante la elaboración del Trabajo de Grado, por ser un gran docente y por su valiosa guía y asesoramiento.

A Juan Manuel Bolaños, mi compañero de tesis, porque juntos lo conseguimos, una gran elección al apoyarnos juntos en este camino.

A mis docentes por transmitirme su conocimiento, ya que gracias a ellos he logrado llegar al final de mi carrera profesional.

Carol Elizabeth Vasquez Burbano

TABLA DE CONTENIDO

CAPÍTULO I.....	1
1. ASPECTOS GENERALES DE LA INVESTIGACIÓN	1
1.1 PLANTEAMIENTO DEL PROBLEMA	1
1.2 JUSTIFICACIÓN.....	2
1.3 APORTES.....	3
1.4 OBJETIVOS.....	3
1.4.1 OBJETIVO GENERAL.....	3
1.4.2 OBJETIVOS ESPECÍFICOS.....	4
1.5 ORGANIZACIÓN DEL DOCUMENTO	4
CAPÍTULO II.....	5
2. MARCO REFERENCIAL	5
2.1 INTRODUCCIÓN	5
2.2 MARCO CONCEPTUAL	5
2.2.1 Búsqueda y Recuperación de Información (ISR)	6
2.2.2 Buscador ISGR.....	6
2.2.3 Clustering	6
2.2.4 KMEANS	7
2.2.5 DBSCAN	7
2.3 ESTADO DEL ARTE.....	8
CAPÍTULO III.....	14
3. CARACTERÍSTICAS DE CLUSTERING BASADAS EN DISEÑO CENTRADO EN EL USUARIO PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES.....	14
3.1 INTRODUCCIÓN	14
3.2 CARACTERITICAS DEL CLUSTERING	14
3.3 PRUEBAS DE DISEÑO CENTRADO EN EL USUARIO	15
3.4 EJECUCIÓN DE PRUEBAS	17
3.5 MÉTODOS DE CLUSTERING	24
3.6 ETIQUETAS DE CLUSTERING	25
3.7 ALGORITMOS DE CLUSTERING A UTILIZAR	25
3.7.1 KMEANS	26
3.7.2 DBSCAN	26

3.8	DIFERENCIAS ENTRE DBSCAN Y KMEANS	27
3.8.1	DBSCAN	27
3.8.2	KMEANS	27
CAPÍTULO IV.....		28
4.	IMPLEMENTACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES.	28
4.1	INTRODUCCIÓN	28
4.2	ARQUITECTURA DEL BUSCADOR	28
4.3	IMPLEMENTACIÓN DEL BUSCADOR WEB.....	28
4.3.1	Motor de Base de Datos	28
4.3.2	Página web.....	29
4.3.3	Algoritmos de Agrupamiento.....	29
4.3.4	Implementación del algoritmo KMEANS	32
4.3.5	Implementación del algoritmo DBSCAN	33
4.3	INDEXACIÓN DE DOCUMENTOS POR MEDIO DE CLUSTERING	34
CAPÍTULO V.....		39
5.	EVALUACIÓN POR EXPERTOS DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES.....	39
5.1	INTRODUCCIÓN	39
5.2	INSTRUMENTO DE VALIDACIÓN	39
5.3	CASOS DE SENTENCIAS.....	40
5.4	RESULTADOS	42
5.4.1	Pregunta Q1.....	43
5.4.2	Pregunta Q2.....	44
5.4.3	Pregunta Q3.....	46
5.4.4	Pregunta Q4	47
CAPÍTULO VI.....		49
6.	CONCLUSIONES Y RECOMENDACIONES	49
6.1	CONCLUSIONES	49
6.2	SUGERENCIAS PARA FUTURAS INVESTIGACIONES	51

TABLA DE TABLAS

Tabla 1. Prueba 1	19
Tabla 2. Prueba 2	20
Tabla 3. Prueba 3	21
Tabla 4. Prueba 4	22
Tabla 5. Prueba 5	23
Tabla 6. Algoritmo KMEANS.....	32
Tabla 7. Algoritmo DBSCAN.....	33
Tabla 8. Pasos para la indexación de documentos	35
Tabla 9. Funcionamiento buscador Jurisprudencia	38
Tabla 10. Tabla de Preguntas	40
Tabla 11. Solución de cada evaluador por caso.....	42
Tabla 12. Resultados pregunta Q1	43
Tabla 13. Resultados pregunta Q2.....	44
Tabla 14. Resultados pregunta Q3.....	46
Tabla 15. Resultados pregunta Q4.....	47

TABLA DE ECUACIONES

Ecuación 1. Distancia Euclidiana	27
--	----

TABLA DE FIGURAS

Figura 1. Prueba presentada Usuario 1	19
Figura 2. Prueba presentada Usuario 2	20
Figura 3. Prueba presentada Usuario 3.....	21
Figura 4. Prueba presentada Usuario 4.....	22
Figura 5. Prueba presentada Usuario 5.....	23
Figura 6. Arquitectura Buscador.....	28
Figura 7. Diagrama de flujo funcionamiento del código	30
Figura 8. Indexación de Documentos.....	34
Figura 9. Buscador de Documentos Jurisprudenciales	36
Figura 10. Campo de Ingreso Texto	36
Figura 11. Elección método agrupamiento.....	36
Figura 12. Botón búsqueda - Año	37
Figura 13. Resultados y resumen	37
Figura 14. Visualización de la Sentencia.....	38

TABLA GRÁFICOS

Gráfico 1. Resultados pregunta Q1	44
Gráfico 2. Resultados pregunta Q2	45
Gráfico 3. Resultados pregunta Q3	46
Gráfico 4. Resultados pregunta Q4	48

RESUMEN

En esta investigación se pretende contribuir en el mejoramiento de los procesos jurisprudenciales en cuanto a la búsqueda de sentencias de tal manera que tenga un alto porcentaje de exactitud y arroje rápidamente el resultado de esta. Para ello se utilizaron dos algoritmos de clustering (KMEANS y DBSCAN) los cuales fueron evaluados y probados en cuanto a efectividad y precisión; a partir de los resultados obtenidos se determinó el mejor algoritmo el cual va a ser utilizado en la herramienta de búsqueda de sentencias.

Palabras clave: agrupamiento, jurisprudencia, tutela, aprendizaje no supervisado.

ABSTRACT

This research pretends to contribute to jurisprudential process improvement as to judgements search in the way to get high precision percentage and throw a fast result about it. Two clustering algorithms (KMEANS & DBSCAN) were used, they were evaluated and tested in terms of effectiveness and accuracy; from obtained results was posible to determine the best algorithm which is going to be used for the judgement search tool.

Key Words: clustering, jurisprudence, tutela, unsupervised learning.

INTRODUCCIÓN

En Colombia el precedente judicial facilita la toma de decisiones por parte de los jueces basándose en sentencias anteriores. Sin embargo, los profesionales del derecho deben buscar entre un gran número de documentos las sentencias que sirvan como soporte para sus casos en curso. Frente a las necesidades del derecho, es común acudir a la denominada informática jurídica como la técnica interdisciplinaria que tienen por propósito la aplicación de la informática a la recuperación de información jurídica, así como la elaboración y aprovechamiento de instrumentos de análisis y tratamiento de dicha información, necesarios para una toma de decisión con repercusiones jurídicas.

La agrupación (Clustering) de documentos web, ha tomado gran interés por parte de investigadores académicos y científicos, los cuales han involucrado modelos de recuperación de información complementados con algoritmos de clustering, buscando aumentar la cobertura de los documentos presentados para que el usuario los revise, estos métodos se han estudiado recientemente debido a la aplicabilidad en áreas tales como, motores de búsqueda, recuperación de información, minería web. Agrupar documentos y organizarlos en diferentes **grupos** de texto para poder realizar una búsqueda eficiente por ello existe una variedad de algoritmos que realizan Clustering entre ellos KMEANS [1] DBSCAN [2][3] etc. Estos métodos son evaluados para buscar como resultado un agrupamiento aceptable y eficiente[4], por esta razón se dará un aporte para contribuir en el mejoramiento de los procesos de búsqueda de documentos jurisprudenciales en Colombia, buscando obtener un porcentaje elevado en de exactitud y agilidad (Tiempo de respuesta) . Para ello se utilizaron dos algoritmos de clustering reconocidos en el estado del arte denominados KMEANS y DBSCAN estos algoritmos fueron evaluados cualitativamente por parte de usuarios expertos en el área judicial utilizando variables como tiempo y precisión. A partir de los resultados obtenidos se determinó el mejor algoritmo el cual se utilizará en una herramienta de búsqueda de sentencias jurisprudenciales para la legislación colombiana.

CAPÍTULO I

1. ASPECTOS GENERALES DE LA INVESTIGACIÓN

1.1 PLANTEAMIENTO DEL PROBLEMA

En Colombia y en muchos otros países, la capacidad que tiene el juez para atemperar sus decisiones y la interpretación de la norma a las circunstancias propias del caso, ocupa en este momento un lugar fundamental en la función judicial, por lo tanto, un juez no sólo se limita a la mera aplicación de normas vigentes, sino que, por el contrario, acude a justificaciones de su propio razonamiento. Así las cosas, el derecho observa necesario imponer criterios, límites y técnicas que garanticen la seguridad jurídica y la igualdad en el acceso a la administración de justicia, para que los individuos no estemos indefensos frente a la subjetividad de los jueces, siendo una de estas técnicas el precedente judicial [1]. En los conflictos sometidos al conocimiento de una autoridad jurisdiccional o administrativa, en los que se pretenda acudir al precedente, es necesario identificar decisiones judiciales anteriores o precedentes que versen sobre hechos similares a los que se estén resolviendo actualmente, así como debe distinguirse con claridad en cada una de ellas, los argumentos jurídicos que fundamentaron la decisión que puedan aplicarse en el caso en cuestión. Es necesario tener en cuenta que las decisiones judiciales tienen un efecto diferente dependiendo de la autoridad que las profiera, por lo cual dichos fallos habrán de discriminarse por jurisdicción, especialidad y jerarquía del fallador.

Es evidente que determinar el precedente judicial para un caso concreto, es una ardua tarea que implica estudio a profundidad de gran cantidad de textos jurisprudenciales, tanto para los funcionarios judiciales y administrativos en la resolución de conflictos, como para los abogados y usuarios de la administración de justicia que constantemente buscan argumentos en pro de sus intereses; es así como se reconoce la necesidad de procesar este cúmulo de información de una forma sencilla y ágil; que aunque ya cuentan con algunas herramientas puestas a su disposición como LEGIS¹ el cual es un software por donde se pueden realizar búsquedas filtradas sobre casos jurisprudenciales, o la página oficial de búsqueda de la jurisprudencia colombiana², no ofrecen una solución real al problema debido a que su funcionamiento se limita a operaciones básicas de búsqueda las cuales no son las más óptimas y pueden ser mejoradas por medio de la implementación

¹ Esta herramienta se puede ubicar en www.legis.com.co

² Página oficial <http://jurisprudencia.ramajudicial.gov.co/WebRelatoria/consulta/index.xhtml>

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

de métodos de Clustering, debido a esto se piensan realizar pruebas con una herramienta capaz de agilizar estos procesos basados en la fluidez y la precisión comparando elementos clave de cada sentencia sin olvidar su contexto y las cuales servirán para categorizar cada búsqueda, ahorrando tiempo en los procesos de obtención de sentencias dadas con anterioridad y agilizando la ejecución de actividades tanto de funcionarios como de los ciudadanos. Algunos mecanismos de inteligencia artificial han sido usados en diversos campos para el análisis de la información, particularmente las áreas de procesamiento de lenguaje natural y la minería de texto como se puede observar en las distintas investigaciones que se referencian en el estado del arte. Sin embargo, estas aproximaciones no han abordado el dominio específico del precedente judicial en Colombia; la ventaja de realizar pruebas con diferentes técnicas de Clustering es que, al realizar un análisis minucioso de cada una, se logró elegir la más adecuada para solucionar la problemática mencionada con el fin de que la búsqueda sea precisa y tenga un alto rendimiento.

El presente trabajo se enmarcó dentro del proyecto “Plataforma para la identificación y búsqueda de precedentes judiciales soportada en inteligencia artificial” financiado por el Sistema General de Regalías a través del proyecto ID-3848 denominado “Red de Formación del Talento Humano para la Innovación Social y Productiva en el departamento del Cauca” (Convenio específico No. 6-81.2/022 de 2017 VRI celebrado entre la Universidad del Cauca y La Fundación Universitaria de Popayán).

1.2 JUSTIFICACIÓN

Frente a las necesidades del derecho, es común acudir a la denominada informática jurídica, definida en [2] “como la técnica interdisciplinaria que tienen por propósito la aplicación de la informática a la recuperación de información jurídica, así como la elaboración y aprovechamiento de instrumentos de análisis y tratamiento de dicha información, necesarios para una toma de decisión con repercusiones jurídicas”, por eso se pretende implementar una técnica de Clustering para los procesos de recuperación de datos de forma eficaz y con un porcentaje de exactitud bastante elevado debido a que en Colombia no se ha evidenciado la existencia de una herramienta con éstas características y que permita un uso eficiente en la recuperación de sentencias dadas por la jurisprudencia Colombiana. Por eso es importante disponer de mecanismos que aporten en el desarrollo del proyecto por lo cual el estado colombiano le ha

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

otorgado alta prioridad a este tema; esto se logra evidenciar en la convocatoria estratégica de innovación para la justicia de parte del ministerio colombiano de las telecomunicaciones y Colciencias [3], en donde se resalta la importancia del precedente judicial para la seguridad jurídica de los administrativos y de los ciudadanos plantea las siguientes líneas temáticas:

- Generar las herramientas y aplicaciones que permitan la identificación y divulgación de Referentes de Unificación Jurisprudencial y Líneas jurisprudenciales.
- Generar las herramientas y aplicaciones que permitan la identificación y divulgación de doctrinas probables, esto es, tres decisiones de la Corte Suprema como Tribunal de Casación sobre un mismo punto de derecho, según el artículo 7 del Código General del Proceso y el artículo 40 de la ley 153 de 1887.
- La herramienta implementada permite agilizar los procesos jurisprudenciales en cuanto a dar una respuesta más oportuna y rápida a una sentencia, ya que el profesional en el área de derecho por medio del buscador el cual contiene un compilado de tutelas de la corte constitucional encontrará una línea base para generar un documento con el fin de dar solución a un usuario que requiere de su servicio.

1.3 APORTES

- ✓ Registro de Software

1.4 OBJETIVOS

1.4.1 OBJETIVO GENERAL

Implementar técnicas de Clustering para un sistema de recuperación de documentos jurisprudenciales

1.4.2 OBJETIVOS ESPECÍFICOS

- ✓ Determinar las características de Clustering basadas en diseño centrado en el usuario para un sistema recuperación de documentos jurisprudenciales.
- ✓ Implementar técnicas de Clustering del estado del arte para un sistema recuperación de documentos jurisprudenciales.
- ✓ Evaluar por expertos las técnicas de Clustering implementados en el sistema recuperación de documentos jurisprudenciales.

1.5 ORGANIZACIÓN DEL DOCUMENTO

CAPITULO 1 - ASPECTOS GENERALES: se presentan aspectos generales como el planteamiento del problema, justificación, aportes que realizará el proyecto, objetivos generales y específicos que se cumplen durante la ejecución del mismo.

CAPITULO 2 - MARCO DE REFERENCIA: se presenta el marco conceptual el cual describe los conceptos más relevantes del proyecto Y el estado del arte que contiene los resultados de otras investigaciones similares al tema de investigación del proyecto actual.

CAPITULO 3 - CARACTERÍSTICA DE LOS ALGORITMOS DE CLUSTERING IMPLEMENTADOS: se presentan las características de los algoritmos implementados para el sistema de recuperación de documentos jurisprudenciales, los métodos utilizados y las etiquetas.

CAPITULO 4 - IMPLEMENTACIÓN DE LAS TÉCNICAS DE CLUSTERING EN EL BUSCADOR DE DOCUMENTOS JURISPRUDENCIALES: se presenta el proceso de implementación de los algoritmos de clustering y la explicación detallada de las librerías utilizadas para cumplir con los objetivos propuestos.

CAPITULO 5 - EVALUACIÓN DEL SISTEMA POR EXPERTOS: se presenta el proceso de evaluación y pruebas realizadas por expertos en el ámbito judicial de tal manera que la búsqueda logre cumplir los objetivos propuestos.

CAPITULO 6 – CONCLUSIONES Y RECOMENDACIONES: por último, se presentan conclusiones, recomendaciones y futuras propuestas de investigación.

CAPÍTULO II

2. MARCO REFERENCIAL

2.1 INTRODUCCIÓN

En el presente capítulo se indica el marco conceptual, el cual contiene las recopilaciones teóricas, definición de conceptos utilizados y posteriormente un estado del arte, presentando resultados de investigaciones previas relacionadas con el tema del proyecto de investigación.

2.2 MARCO CONCEPTUAL

En esta investigación se implementó una herramienta tecnológica para agilizar los resultados al momento de realizar una búsqueda utilizando un buscador que está conformado por un campo en el cual se digita un texto y posteriormente se selecciona el tipo de algoritmo con el que desea buscar para finalmente obtener los resultados que se relacionan con el texto ingresado garantizando una mayor precisión. Para poder realizar este proceso se categorizaron los documentos usados por los jueces, en este caso las sentencias las cuales traen ciertos patrones que ayudan a diferenciarse unas de otras por lo que encontrar estos patrones y categorizarlos entre cientos de documentos es un proceso complejo para un ser humano, pero no para una máquina. En este apartado se hará uso de un método llamado Clustering [4] el cual implementa una serie de algoritmos o pasos lógicos para realizar una tarea la cual consiste en categorizar estos documentos según los criterios de búsqueda dados por el usuario y reunirlos en un mismo sitio, esto se traduce en búsquedas más precisas sobre un tema ya que los elementos encontrados pertenecerían todos a un mismo criterio. En este punto se tuvo en cuenta el uso de ISR (Information Search and Retrieval) o en español Búsqueda y Recuperación de Información [5], la cual es la parte de la informática que se encarga de la búsqueda de información en documentos electrónicos o cualquier medio digital, video, imágenes, audios y su objetivo es la recuperación de este material y mostrar su información de forma escrita y muy relevante, este proceso se ve reflejado en diferentes herramientas de uso cotidiano como son los buscadores [6] que normalmente son usados por diferentes tipos de usuarios. Se hizo uso de un diseño centrado en el usuario (DCU) [7], con lo que se determinan los elementos en los cuales los usuarios podrán influir en el diseño de la herramienta para obtener un resultado final acorde a sus especificaciones,

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

garantizando que se cumpla el concepto de usabilidad[8] el cual consiste en mejorar la experiencia del usuario al interactuar con el producto.

Según el criterio de la doctrina, se puede establecer que la jurisprudencia es un criterio auxiliar, en el sentido que, cuando las disposiciones de la Constitución y de las demás fuentes formales del derecho no tienen un sentido unívoco, que sea capaz de eliminar toda indeterminación, la jurisprudencia auxilia al entendimiento pleno del sentido de dichas fuentes formales, pues en ella se encuentran las normas adscritas que expresan su significado en sentido prescriptivo.

A continuación, se presentan conceptos clave de este trabajo.

2.2.1 Búsqueda y Recuperación de Información (ISR)

Es un proceso articulado y, en muchas ocasiones, retroalimentado que se inicia cuando una persona tiene un problema que quiere resolver mediante la obtención de cierta información que termina cuando la persona resuelve este problema con la información obtenida y que se implementa a través de la identificación y localización de los documentos que contienen esta información que es pertinente para satisfacer la necesidad de información de la persona[5].

2.2.2 Buscador ISGR

Un buscador es un sistema informático que permite encontrar páginas web o resultados en base a la frase o palabra que hayamos ingresado y estemos buscando [6]. Debido a que en el proceso de la implementación se trabajó una metodología de diseño centrado en el usuario el cual es un término general que se utiliza para describir el diseño en el que el usuario influye en el resultado final [7] pero tratando que tenga la mejor usabilidad posible, la cual es definida como la medida de la calidad de la experiencia que tiene un usuario cuando interactúa con un producto o sistema [8], con el fin de desarrollar un software que permita darle una experiencia de calidad y de uso al usuario que va a interactuar con el mismo.

2.2.3 Clustering

Permite dividir los datos en grupos (clusters), de tal forma que los grupos capturan la estructura natural de los datos y también se pueden dividir datos sin etiqueta en

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

grupos (clusters) de tal forma que datos que pertenecen al mismo grupo son similares, y datos que pertenecen a diferentes grupos son diferentes [4].

2.2.4 KMEANS

Este algoritmo creado por (MacQueen, 1967) [1] es un método de agrupación eficiente con un único parámetro de entrada el cual es el número k de agrupaciones a generar, donde cada clúster tiene asociado un centroide (centro geométrico del clúster), donde sus puntos se asignan al cluster cuyo centroide esté más cerca (utilizando métrica de distancia euclidiana en su gran mayoría), en su iteración, los centroides se van actualizando en función de las asignaciones de puntos a clusters, hasta que los centroides dejen de cambiar y los criterios de parada se cumplan, este método se implementa en el presente proyecto con la ayuda de la librería scikit-learn de Python.

2.2.5 DBSCAN

Este algoritmo publicado por (Ester et. Alabama, 1996)[3] es un método utilizado recursivamente en minería de datos dada su capacidad para encontrar grandes cantidades de datos de forma aleatoria, este utiliza dos parámetros principales como son ' ϵ ' y 'minPoints' donde ϵ define el radio de "Región del vecindario" y "minPuntos" define el número mínimo de puntos que genera para encontrar una agrupación, este contiene una característica única dado que funciona bien incluso con conjuntos de datos ruidosos o datasets dañados.

Dentro de la herramienta de software implementada se utiliza para ser evaluado junto a K-MEANS, su implementación se realiza mediante la librería scikit-learn de Python.

En el proyecto de investigación se eligieron los algoritmos K-MEANS y DBSCAN ya que por medio del estado del arte se logró identificar que son dos de los métodos más utilizados para realizar agrupamiento de documentos y en el caso de Python existe la librería scikit-learn la cual permite realizar con más agilidad el desarrollo de la herramienta de software.

2.3 ESTADO DEL ARTE

Día a día la información crece de forma exponencial, millones de datos son enviados a redes locales, en nuestros ordenadores e internet, esto hace que encontrarla de manera oportuna y ágil sea tedioso por la cantidad que existe a disposición de consultar, por lo tanto, se genera una necesidad de recurrir a elementos avanzados de informática para darle un uso de forma eficiente, es decir, una buena organización, búsqueda y manipulación. En este caso se dio énfasis a uno de los métodos de minería de datos llamado Clustering el cual se encarga de entender la estructura macroscópica y relaciones entre documentos, considerando las formas en las que estos son similares y diferentes, es decir que su enfoque va dado a segmentar el conjunto completo de datos en subgrupos homogéneos y estos se agrupan en clusters [4].

Al implementar estas técnicas de clustering se evita que ciertos elementos de características particulares se mezclen con otros totalmente diferentes lo necesario de forma ordenada y agrupada. Este proceso de separación sistemática de archivos por características dadas se denomina Clustering el cual puede ser aprovechado en el sistema jurisprudencial colombiano utilizando sentencias dadas con anterioridad y a las cuales se les consigue dar un uso posterior para tomar una decisión respecto a una nueva sentencia, debido a que la ley permite que sean tomadas en cuenta siempre y cuando la información de la sentencia antigua con la sentencia a dictaminar sea similar. Buscar estas similitudes de forma precisa y rápida en las sentencias pasadas ayuda a que el sistema judicial colombiano se descongestione ya que proporcionaría de forma eficiente documentos que le sean de gran apoyo a los jueces para su posterior análisis y así dictar un veredicto rápidamente de la sentencia actual, todo esto utilizando diferentes métodos de Clustering con el fin de encontrar la solución más óptima y con mayor exactitud tanto en búsqueda de resultados como en eficiencia, ya que al probar distintos métodos y realizar el respectivo análisis a cada uno de ellos, finalmente se puede conseguir utilizar el más apto para la problemática planteada. Por ello a continuación se realiza un reporte detallado de las investigaciones previas relacionadas con el anterior problema:

En 2016 la base de datos oficial de la corte en China registro 19.706.614 documentos judiciales con un crecimiento diario, llevar a cabo un herramienta que gestionara la extracción de información sobre estos documentos y su uso por parte de jueces y personas llegadas al tema del derecho penal era necesario para esto por eso en [9] se propuso un sistema de consulta legal para China el cual se basó en algoritmos genéticos y aproximaciones de tipo KNN dando gran importancia al contexto de la región. Además de las características legales de este

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

país, la importancia de la semántica fue esencial para el proceso de creación, esto significa que la más usada era tomada mientras la más ambigua era desechada, lo cual facilitó el mejor uso de los recursos y mayor velocidad en las respuestas sin perder la forma. Con este estudio se demostró la eficiencia de los Algoritmos Genéticos (GA) con KNN y que la investigación sobre este campo es necesario debido a la creciente cantidad de datos, como se puede ver también en [10] el uso de la semántica y la estructura de la documentación fue de gran importancia y sobre este para poder dar solución al problema de recuperación de datos y dejando a cargo una sola herramienta de categorización y agrupamiento KNN, la recurrencia de los escritos sobre como la forma de las leyes cambian y su debida interpretación deja ver la importancia sobre la semántica, el contexto, y demás temas relacionados con las culturas como se puede ver en [11] cuyo software EUNOMOS, trabajo sobre ontología jurídicas para poder dar solución a la gran cantidad de temas jurídicos a nivel internacional, el trabajo de esta empresa se basó en identificar diferentes formas de cómo se dan las leyes, como se interpretan y como se toman a nivel jurídico y su contexto, para este software se hace uso de la categorización según el país, se usó esto para poder categorizar cada una de las leyes según su contenido en una base de datos por medio de etiquetado y así poder ubicarlos posteriormente, en [12] se planteó una extracción de palabras, frases por medio de la construcción de un repositorio de documentos sin estructura es decir sin formato, esta información se guarda en bloques de información creado bajo herramientas XML con etiquetas más elaboradas dirigidos de procesamiento de lenguaje natural PLN, esto para poder tener documentos más fieles de acuerdo a la semántica que se maneja, identificando palabras o frases importantes y transformando los menos importantes, en [14] se trabajó una metodología de clasificación, agrupación y búsqueda de documentos la cual se encuentra basada en redes neuronales, esto ayuda todos aquellos que necesiten una herramienta para hacer uso de estos documentos en diferentes casos para que haya una aplicación de la ley y administrar escritos de juicios criminales de manera más eficiente. También se hizo un tratamiento sobre las palabras chinas en este caso para darle efectividad a al proceso de agrupamiento de textos mediante el esquema de extracción de términos para seleccionar las palabras clave con la frecuencia más alta como entradas de la Red de Propagación. Se seleccionamos siete categorías criminales como objetivo de salidas presentando resultados muy altos al momento de encontrar casos criminales para uso del usuario.

En [15] se propone un método para realizar resúmenes de multi-documentos altamente informativos, en el cual se utiliza Clustering con un enfoque híbrido basado en estadísticas que calculan la importancia de las oraciones en función de elementos de texto, además las simplifica por medio de lenguaje natural para

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

mostrar solo la información necesaria, con el objetivo de generar resúmenes más útiles. Utilizan una versión adaptada del algoritmo K-Means para agrupar las oraciones por palabras claves. Al realizar el agrupamiento por similitud y por palabras claves se obtiene un agrupamiento doble, con el objetivo de elegir las oraciones que contienen la información más importante y además que no redunde con otras oraciones para que finalmente se conforme el documento resumido. En esta investigación se plantea que se debe realizar primero el agrupamiento doble y seguido de ello la simplificación ya que se tiene la información completa y es más preciso realizar el Clustering antes de eliminar información que puede ser importante por otra parte en [16] se habla de la importancia de los algoritmos para agrupar documentos los cuales pueden facilitar el descubrimiento de nuevos y útiles conocimientos analizando los mismos. Se presenta un enfoque que aplica algoritmos de agrupamiento para el análisis forense de computadores incautados en investigación policial. Se realizó una experimentación con seis algoritmos de agrupamiento con el fin de mostrar el potencial para realizar la tarea planteada, K-means partitional, K-medoids, Single Link, Complete Link, Average Link y el algoritmo de conjunto cúmulo CSPA; se aplicaron a cinco conjuntos de datos del mundo real obtenidos de computadores incautados en el área de la investigación. Estos algoritmos se llevaron a cabo con diferentes combinaciones en sus parámetros y resultaron dieciséis instancias de algoritmos diferentes y a partir de ello se compararon sus resultados relativos. Finalmente obtuvieron como resultado que los mejores algoritmos al realizar el experimento fueron Complete Link y Average Link, aunque estos tuvieron altos costos computacionales, demostraron que son los adecuados para realizar agrupación de documentos y generar resúmenes más completos.

En [17] se propuso un algoritmo de clasificación reforzado y de agrupación mediante Clustering, el cual tiene un enfoque para el resumen de múltiples documentos, manteniendo las principales características del conjunto de documentos originales, la centralización de la información se da en la agrupación de oraciones relacionadas para proporcionar resúmenes más informativos. Se definen tres funciones de clasificación diferentes Global Ranking (Sin agrupamiento), Local Ranking (Dentro de los clusters), Condittional Ranking (Por medio de los clusters) las cuales componen un gráfico construido a partir del conjunto de documentos dado, es decir, clasificaciones globales dentro del clúster y condicionales respectivamente basado en K-Clusters inicialmente y convirtiendo cada oración en un vector k-dimensional donde cada dimensión es un componente con respecto a un clúster que se mide por la distribución del rango, y con ello se asigna nuevamente a la agrupación que se asemeje a la nueva medida. Como resultado se obtuvo que la calidad de agrupación y clasificación están reforzadas, y las pruebas demuestran la eficacia del enfoque propuesto el cual es basado en

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

clusters. En [18] se considera el agrupamiento de documentos como un análisis de datos, este proceso de obtención de datos se basa en la clasificación de los documentos más acertado respecto a nuestra búsqueda por medio de procesos usados en la minería de datos, y aunque bien se sabe la minería de datos se usa más que todo para el análisis numérico, en este artículo se da a conocer una técnica donde se transforman los datos de un documento en una serie de datos estructurados document-term Matrix (DTM) usando estas técnicas se permite obtener dichos resultados como documentos que están conformados en un arreglo de filas y columnas de la DTM respectivamente permitiendo ver la frecuencia de los términos que hay en cada documento. Estos resultados bajo este algoritmo han sido probados con varios datasets variando el número de documentos con una buena diferencia de categorías que había en el conjunto, estos resultados fueron más satisfactorios que los resultados dados por algoritmos de agrupamiento tradicionales. En [19] se aborda el problema de extracción de palabras claves de conversaciones en audio, se propone un algoritmo voraz, un método para encontrar múltiples consultas como resultado del conjunto de palabras claves, con el fin de maximizar las posibilidades de hacer al menos una recomendación relevante al usar estas consultas para buscar en Wikipedia en idioma inglés. En el documento se adopta la perspectiva de la recuperación just-in-time y agrupamiento en clusters, en el cual se propone un enfoque de dos etapas para la formulación de preguntas implícitas, el primero es la extracción de palabras claves de un fragmento de conversación y la segunda es la agrupación de la palabra clave para establecer varias consultas. Ya que los motores de búsqueda existentes en la web no son tan eficaces ni proporcionan rápidamente lo que el usuario desea buscar, el método utilizado recompensa esta deficiencia al recomendar espontáneamente con el reconocimiento de voz los resultados que están relacionados con la búsqueda que desea realizar. Por ejemplo, cuando los usuarios participan en una reunión, sus necesidades de información pueden ser rápidas como consultas implícitas que se construyen en el grupo de las palabras pronunciadas, obtenidas a través de auto-reconocimiento de voz matemático (ASR). Estas consultas implícitas se usan para recuperar y recomendar documentos de la Web o un repositorio local. El sistema planteado just-in-time permite realizar consultas implícitas a partir de la entrada de conversaciones.

En [20] se realizó un estudio se muestra cómo se agruparon dos elementos la Ontología y el Clustering Jerárquico pueden unirse para darle eficiencia a la recuperación de datos por medio de una recolección de data sets lo cual son una serie de documentos candidatos que se van recopilando, estos elementos entran a un pre procesamiento lo cual se caracteriza por remover el ruido o elementos basura, cada uno de los elementos se analizan bajo ciertos criterios como su estándar, luego se pasa a la extracción donde por medio de un algoritmo se

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

seleccionan los mejores candidatos para la búsqueda y con estos datos prepararlos para su respectiva agrupación (Clustering), en este punto se tienen los elementos principales para ser catalogados, en este caso se crearan clusters planos, los cuales no tienen una estructura relacionada con otro, cada uno posee su propia estructura, esto hace más eficiente para un primer Clustering, con cada uno de estos clúster se dividen en tipo de documentos formando un cluster jerárquico el cual consiste en darle jerarquía a lo ya organizado, los documentos se dividen en formatos PDF o Word, y estos a su vez se pueden dividir en Papers, libros o reportes, lo cual se traduce en un sistema eficiente para ejecutar la búsqueda del documento en cuestión y por ultimo ubicar el archivo y recuperar la información que se buscaba en él. Con este método se clasifica de forma eficiente un gran dataset, reduciendo el nivel de ruido llevando a una mayor exactitud incrementando la velocidad y recuperación de la información en un poco tiempo.

En [21] se realizó una investigación que se enfoca en aplicación de Clustering en data mining para abarcar la problemática de la cantidad de documentos de texto electrónicos disponibles. El método propuesto de agrupamiento de documentos se basa en la metodología MapReduce que mejora el rendimiento de agrupamiento de documentos y permite la manipulación de una gran base de datos, este método es basado en el algoritmo KMEANS. El algoritmo que se propone en el documento utiliza centroides óptimos para la agrupación en el algoritmo KMEANS basado en la Optimización de Enjambre de Partículas (PSO). PSO se utiliza para aprovechar su ventaja global y capacidad de búsqueda para proporcionar centroides óptimos que ayuda a generar clústeres con una precisión mejorada. Inicialmente se eligió un conjunto de datos de documentos para el análisis, con estos documentos se realiza un pre procesamiento para reducir el número de atributos que se convierten en un conjunto de términos que se pueden incluir en el modelo vectorial, esto se realiza para representar los documentos en forma vectorial con el fin de determinar el peso, por último los vectores de los documentos se agrupan utilizando el algoritmo híbrido de agrupación PSO KMEANS basado en MapReduce (MR-KMEANS).

En [23] se propone un nuevo enfoque al trabajo llamado Maletín de conceptos el cual es un método con el que se pueden extraer palabras de un documentos y poder trabajar con estas pero este tipo de elementos normales poseen un problema el cual se basa en el manejo de la palabra la cual la toma como elemento único sin tener en cuenta lo que hay a su alrededor, o sea queda fuera del contexto, a partir de esto se da origen a una variación del trabajo original bolsa de conceptos como un método alternativo de representación de documentos que supera las debilidades de estos dos métodos. Este método propuesto crea conceptos mediante Clustering de vectores de palabras generados a partir de

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

word2vec, y usa las frecuencias de estos clusters de conceptos para representar vectores de documentos. A través de estos conceptos basados en datos, el método propuesto incorpora el impacto de palabras semánticamente similares en la preservación de la proximidad de documentos de manera efectiva. Con un esquema de ponderación adecuado, como la frecuencia de documentos inversos en frecuencia conceptual, el método propuesto proporciona una mejor representación de documentos que los métodos sugeridos anteriormente, y también ofrece una interpretación intuitiva detrás de los vectores de documentos generados. Con base en el método propuesto, los modelos de minería de textos construidos posteriormente, como el árbol de decisiones, también pueden proporcionar razones interpretables e intuitivas sobre por qué ciertas colecciones de documentos son diferentes de otras.

En [24] se hace un estudio sobre el agrupamiento y la clasificación que son dos de los elementos más importantes para la obtención de resultados en la recuperación de información, por lo que en este artículo se trata la forma de cómo dar mejores resultados por medio de una técnica la cual aproxima la semántica, Para esto se hace uso latent semantic indexing (LSI) y un algoritmo min-cut, el cual se mejora por medio de nuevas herramientas de selección que han sido enfocadas en encontrar conjunto de características que reúne el punto crucial de los documentos en el corpus sin deteriorar el resultado del proceso de construcción. Mientras el LSI adquiere los sinónimos el algoritmo min-cut ayuda a generar grupos (cluster) eficientes en cada proceso de agrupamiento (Clustering), para dar validación a estos datos y dar el número de clusters se usa el coeficiente de solhouette que une dos elementos cohesión y separación de clusters después de todo el proceso de agrupamiento (Clustering), esto da una serie de elementos que deben tener un ranking el cual será trabajado por medio de técnicas y elementos como Dirichlet el cual es un método de interpretación y consistencia de validación. Este estudio combinó un elemento de aproximación de semántica, agrupamiento y fueron eficientes mostrando un comportamiento satisfactorio pero enmarcado en dentro de lo normal.

En el presente proyecto de investigación se realizó un aporte utilizando los métodos de Clustering K-MEANS y DBSCAN para solucionar una problemática en Jurisprudencia, de tal manera que se consiga dar un apoyo en la rama judicial para que los jueces logren dar respuesta a sentencias de manera más rápida y eficaz proporcionándoles documentos (sentencias) relevantes es decir que le sean de gran ayuda ya que ellos tienen la posibilidad de dar un veredicto final en base a sentencias pasadas con la condición de que éstas tengan similitud con la actual, en los trabajos relacionados se utiliza este método pero orientado a otros ámbitos, teniendo en cuenta que en la mayoría de éstos su aporte es en otro idioma.

CAPÍTULO III

3. CARACTERÍSTICAS DE CLUSTERING BASADAS EN DISEÑO CENTRADO EN EL USUARIO PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

3.1 INTRODUCCIÓN

Dentro del marco del desarrollo del proyecto, las actividades que se realizaron en función del cumplimiento de los objetivos propuestos, en la primera fase se realizaron las respectivas investigaciones y consultas del estado actual del conocimiento en las áreas y tecnologías pertinentes en la elaboración del proyecto, incluyendo el diseño centrado en el usuario; en la segunda fase se realizó la implementación del software basado en los requerimientos obtenidos y las respectivas pruebas; en la tercera fase se realizó la publicación de los resultados con la elaboración y presentación de un artículo de investigación.

3.2 CARACTERÍSTICAS DEL CLUSTERING

En la actualidad, se han desarrollado algunas herramientas tecnológicas como solución a la problemática mencionada estas incorporan en su gran mayoría modelos de espacio vectorial (SVM) cuya efectividad se va deteriorando frente las continuas consultas de los usuarios de igual manera existen herramientas del ámbito judicial, orientadas al uso del procesamiento de lenguaje natural[9]. En este escenario, las tecnologías de recuperación automática de información representan una buena aproximación inicial a la solución del problema. Los sistemas de recuperación de documentos jurisprudenciales se componen por tareas de búsqueda, consulta de información, clasificación de resultados, almacenamiento de la información, clasificación de documentos en grupos pre definidos y agrupamiento de documentos en conjuntos definidos a partir del análisis automático del contenido del documento (clustering) [12].

Igualmente se realizó una exploración de las necesidades específicas de los potenciales usuarios del sistema. Para esta fase se utilizará la metodología Diseño Centrado en el Usuario [25] para la fase de exploración, en la identificación de necesidades del usuario y en la problemática que él presenta, a partir del análisis de sus requerimientos se tomó la decisión de cuál es la mejor técnica de Clustering a utilizar, es decir la que sea capaz de solucionar el problema de

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

manera ágil, óptima y precisa, además con ello se encontró la mejor manera de agrupar la información según convenga.

La cual se realizó con las siguientes actividades:

- ✓ Identificar las personas que utilizarán el software a desarrollar.
- ✓ Realizar el levantamiento de los requerimientos tanto funcionales como de usabilidades, es decir necesidades y objetivos de los usuarios.
- ✓ A partir de la información recolectada se debe realizar el diseño de la solución.
- ✓ Partiendo de diferentes posibles soluciones se evaluó por medio de las personas seleccionadas cual es la mejor y con ello se definió la más acorde al usuario.

3.3 PRUEBAS DE DISEÑO CENTRADO EN EL USUARIO

Se detectaron elementos de búsqueda por los cuales se guía el usuario y los cuales son fundamentales para comenzar a diseñar la aplicación, para esta parte se trabajó con elementos basados en la experiencia centrada al usuario para así poder detectar cuales elementos son los principales para comenzar a desarrollar la aplicación, todo esto gracias a una serie de actividades donde participan personas que trabajan con la rama judicial y personas que no han tenido trato con esta.

Para establecer estos parámetros de búsqueda por parte de los usuarios se realizaron una serie de pruebas en ambientes controlados que cumplieran con las especificaciones detalladas en el Protocolo Pruebas de Uso.

Mientras se realizaron las pruebas se pudieron identificar las etapas del proceso de uso, las herramientas utilizadas por los usuarios y las dolencias de los usuarios a la hora de realizar las consultas para esto se contó con una serie de herramientas listadas a continuación:

1. Computador Portátil con el software Morae previamente instalado y verificado.
2. Conexión Eléctrica
3. Acceso a Internet
4. Una Mesa con al menos dos sillas
5. Hojas en blanco
6. Lapiceros

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

7. Protocolos de uso
8. Grabador de Sonido.
9. Cámara de Vídeo (no es obligatoria)

Las actividades comenzaron dando la bienvenida a cada uno de los participantes a la prueba de uso, en esta parte se explicó en detalle el objetivo de la prueba y como se desarrollarían cada una de las actividades, todo esto supervisado por cada uno de los compañeros, ayudantes y demás personas presentes en la prueba. Terminado el acto de bienvenida se da paso al **cuestionario inicial**, en este apartado se hicieron una serie de preguntas con las cuales se busca conocer aspectos correspondientes a la actividad y el uso de las diferentes herramientas informáticas para el desarrollo de la actividad de cada cliente.

1. ¿Cuál es su Nombre?
2. ¿Cuáles son las funciones generales del cargo que desempeña y/o cuáles son las actividades propias de la profesión que ejerce en la actualidad?
3. ¿Ha consultado jurisprudencia en la última semana? ¿con qué frecuencia lo hace?
4. ¿Para qué utiliza las sentencias consultadas?
5. Normalmente ¿cómo realiza usted la consulta de sentencias de jurisprudencia?, ¿qué herramientas o métodos utiliza?

Finalizando la primera fase se comienza la ejecución de la Prueba de Uso, se le explica a los usuarios los pasos para comenzar la prueba los cuales son:

1. Asignar un Tema del Banco de Temas.
2. Indicar que para hacer la consulta dispone del computador portátil suministrado por el colaborador técnico, donde quedará registrado todo lo que él haga mientras lo esté usando.
3. El usuario dispondrá de 5 minutos en los cuales el podrá hacer uso de cualquier programa instalado para la solución de este problema y dejando claro en el caso de que no se dé solución lo que interesa son los datos recopilados.
4. El usuario podrá realizar en cualquier momento una pregunta por alguna inquietud que se presente mientras se realiza la prueba.
5. Iniciar con la prueba.
6. Si el usuario finaliza antes del tiempo, se procede al cuestionario final.

Dados los pasos, se comenzó la respectiva prueba, se dispuso un equipo para cada uno de los usuarios citados, estos equipos tenían una configuración inicial

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

que aseguraron unos resultados precisos, con todo en orden se dio la orden para comenzar y desde ahí los cinco minutos de tiempo por prueba.

3.4 EJECUCIÓN DE PRUEBAS

Banco de Temas:

1. Menor de edad a la cual el médico adscrito a la EPS le prescribe un medicamento no incluido en el POS (PBS) y la EPS no le autoriza la entrega.
2. Profesional que presenta la documentación para participar en concurso de méritos ante una Entidad Pública, la cual es rechazada por falta de requisitos por parte de la Universidad contratada para adelantar el concurso. La resolución que decidió el rechazo fue publicada en la página del Ministerio de Educación, pero no en la página del concurso, por lo cual el docente no pudo presentar recursos en contra de dicha decisión.
3. Hombre soltero que acudió a Bienestar Familiar para adoptar un menor de edad, pero su solicitud fue rechazada por su orientación homosexual.
4. Interno carcelario que duerme en una colchoneta en mal estado y a pesar de diferentes solicitudes la dirección del centro carcelario no le ha cambiado la colchoneta, argumentando que la USPEC, encargada de proveer los recursos para dotación, no ha girado los recursos para adquirir elementos que permitan el cambio de aquellos en malas condiciones.
5. Trabajadora que recibe un salario mínimo e incapacitado por accidente laboral, a quien la ARL a la que se encuentra afiliada no le ha pagado las incapacidades correspondientes a 3 meses.
6. Víctima del conflicto interno que después de haber radicado hace más de 1 año la solicitud para su inclusión en el Registro Único de Víctimas, la DIRECCIÓN NACIONAL PARA LA ATENCIÓN Y REPARACIÓN INTEGRAL A LAS VÍCTIMAS no le ha dado respuesta a la petición. Solicita además que le sea reconocida la indemnización respectiva.
7. Se abre una licitación para una obra pública, en cuyo pliego de peticiones aparecen 20 ítems, cada uno con unas descripciones específicas para los proponentes. Después de publicados y 1 día antes del cierre de convocatoria,

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

la entidad agrega un ítem más. Uno de los proponentes que no pudo presentar el ítem 21 solicita que se ordene la ampliación de plazo o la nulidad del ítem 21.

8. A un paciente afiliado a una EPS, que devenga \$1'000.000 mensual, se le remite a una cita ambulatoria con especialista en la ciudad de Cali. Solicitó el pago de transporte y viáticos alegando falta de recursos para cubrir los gastos que implica el desplazamiento a la cita médica en otra ciudad. La EPS niega el pago de dichos costos por no tratarse de una persona que devenga igual o menor que un salario mínimo mensual.
9. Una persona en un proceso judicial considera que el juez le ha vulnerado sus derechos por no haber valorado una prueba testimonial que pudo haber modificado la sentencia en favor suyo. Interpone tutela en contra de dicha decisión judicial.
10. Una persona adulta solicita a COLPENSIONES el reconocimiento de la pensión de sobrevivientes y afiliación a una EPS en favor de su hermano mayor de edad con discapacidad mental, para que reciba la sustitución pensional de la mesada que recibía su padre quien recientemente falleció. COLPENSIONES niega el reconocimiento por considerar que no se adelantó el trámite para el reconocimiento de la discapacidad del hijo del pensionado estando éste último en vida.

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

Prueba 1.

Usuario:	FERNANDO CHACON	Fecha:	31 Agosto 2018
Profesión:	INGENIERO CIVIL	Tema:	7

Proceso y Sugerencias:

El usuario no obtiene un resultado satisfactorio para poder resolver el caso en cuestión, y aunque se apoyó en el navegador Google este fue ineficiente para él no importaba el número de pestañas que fuera. Al responder el siguiente cuestionario una de las inquietudes que mostro ante el navegador fue que no tenía una información lo suficientemente precisa para poder resolver ningún caso y un resumen algo corto para poder basarse en dichos resultados, también cuestiono el poco uso de un elemento temporal que determinara cuál de los resultados es el más reciente dando a entender la importancia de poder reducir los resultados de éxito a través de un filtro de tiempo, para este caso año.

Evidencia:

The screenshot shows a recording session in 'Jurisprudencia - Morae Manager'. The main window displays a Google search for 'tiempo minimo o maximo para solicitar un adendo a una licitacion'. The search results are visible, showing a list of documents with titles like 'EXPEDICIÓN DE ADENDOS' and 'Una entidad estatal expidió una adenda en un proceso de...'. The interface includes a file explorer on the left, a search bar at the top, and a video player at the bottom. The recording software interface is overlaid on the browser window, showing various controls and a timeline at the bottom.

Figura 1. Prueba presentada Usuario 1

Tabla 1. Prueba 1

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

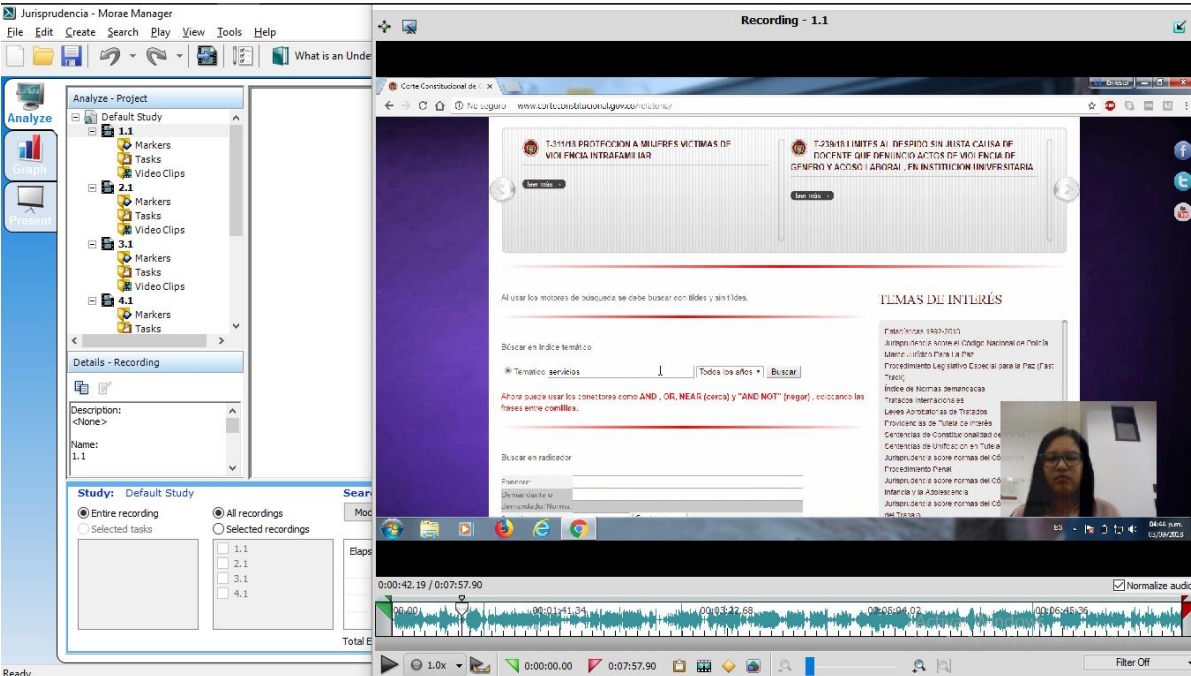
Prueba 2.			
Usuario:	Usuario 2	Fecha:	03 septiembre 2018
Profesión:	Abogada	Tema:	1
Proceso y Sugerencias:			
<p>El usuario se encarga de realizar la búsqueda por medio de la página de la Relatoría de la Corte Constitucional, para este caso sus resultados son efectivos dando con ella en la primeras búsqueda, para llegar a esta sentencia la usuaria hizo uso de palabras muy precisas en el ámbito del derecho por lo que deja clara que la sintaxis es un elemento muy importante a tener en cuenta, además de tomar como referencia la fecha que se presenta para poder identificar la resolución dando importancia a su fecha, dando como sugerencia dos elementos claves como lo son las palabras usadas para la búsqueda y su respectiva fecha.</p>			
Evidencia:			
			
Figura 2. Prueba presentada Usuario 2			

Tabla 2. Prueba 2

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

Prueba 3.

Usuario: Usuario 3 **Fecha:** 03 septiembre 2018

Profesión: Abogado **Tema:** 2

Proceso y Sugerencias:

El usuario de la prueba 3 no tuvo éxito al encontrar el resultado esperado debido a que tomo mal la forma en como ingresar las palabras clave necesarias para poder llegar a ella, además para poder validar sus resultados debe ingresar a dicha sentencia perdiendo de vista los diferentes resultados y cada uno de resúmenes gastando tiempo necesario para poder seguir con eficiencia su búsqueda, el usuario no deajo sugerencias y deajo en claro la satisfacción sobre el buscador Google pero la actividad deajo claro ciertos temas como unos resúmenes mejor elaborados y más extensos sobre cada una de las sentencias además el poder visualizar cada una de estas sentencias sin perder de vista los resultados antes dados.

Evidencia:

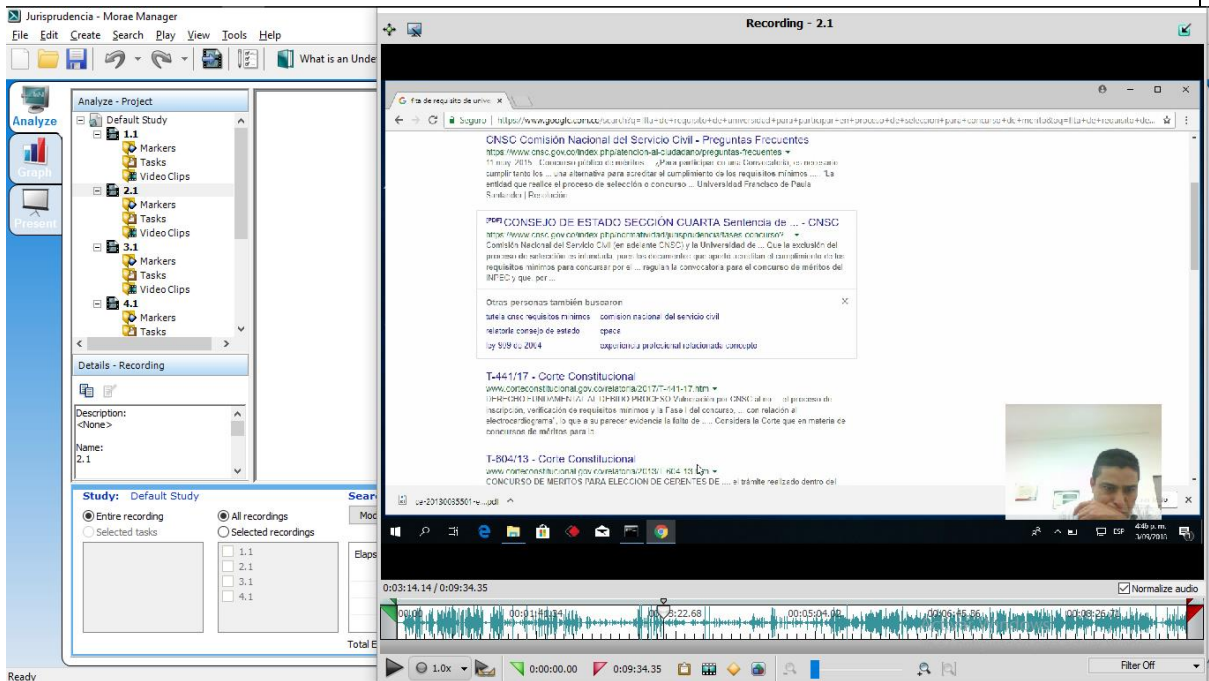


Figura 3. Prueba presentada Usuario 3

Tabla 3. Prueba 3

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

Prueba 4.			
Usuario:	Usuario 4	Fecha:	03 septiembre 2018
Profesión:	Abogada	Tema:	2
Proceso y Sugerencias:			
<p>En esta prueba el usuario paso a realizar la búsqueda directamente por google, en su primer intento no encontró nada, entonces decidió añadir ciertos parámetros extras en una nueva búsqueda para poder darle solución al tema dado, en esta nueva búsqueda concateno cada una de las palabras por medio del símbolo “+” pero agrego un término extra a su anterior referencia de búsqueda la cual era “corte” esto trajo consigo resultados diferentes y satisfactorios dando con una sentencia del año 2011 la cual procedió a revisar y a descargar para su posterior uso. Con el tema solucionado el usuario sugirió que las búsquedas fueran efectivas a partir de las palabras clave ingresadas en la búsqueda, así como lo hace Google, que se haga énfasis en las fechas para poder elegir cada los resultados y mantener la búsqueda sencilla, sin que se tengan que agregar filtros.</p>			

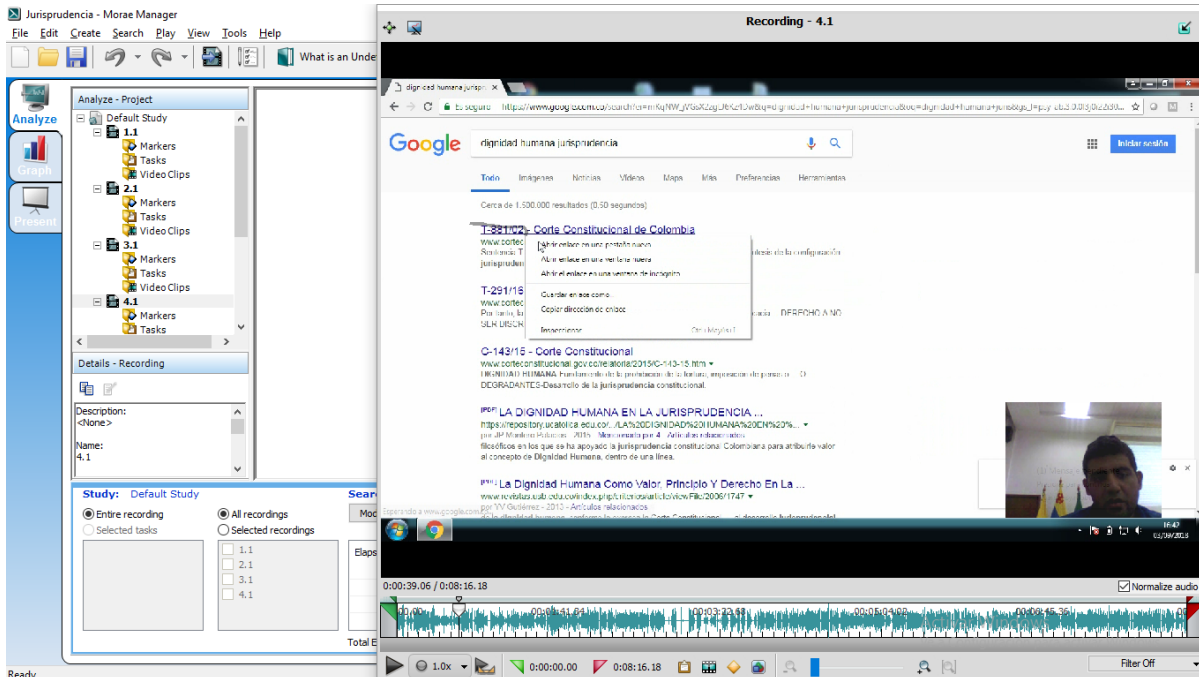
Evidencia:

The screenshot displays two overlapping windows. On the left is the 'Jurisprudencia - Morae Manager' application, showing a project analysis interface with a tree view of tasks and markers. On the right is a web browser window titled 'Recording - 3.1' showing search results for 'adoption of children by same-sex couples'. The search results include articles from 'La batalla que libró la primera pareja gay por adoptar', 'ICCF adopción de niños por parejas del mismo sexo - Universidad Libre', 'Parejas del mismo sexo podrán adoptar: Corte Constitucional', 'Corte Constitucional de Colombia', and 'Requisitos del ICCF para la adopción por parte de parejas - Abogados'. A video player is visible in the bottom right corner of the browser window, showing a woman's face. The software interface on the left has a 'Details - Recording' section with a 'Description: <None>' and a 'Study: Default Study' section with radio buttons for 'Entre recording' and 'All recordings', and a list of tasks (1.1, 2.1, 3.1, 4.1) with checkboxes.

Figura 4. Prueba presentada Usuario 4

Tabla 4. Prueba 4

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

Prueba 5.			
Usuario:	Usuario 5	Fecha:	03 septiembre 2018
Profesión:	Abogado	Tema:	2
<p>Proceso y Sugerencias:</p> <p>El usuario hizo un recorrido algo más largo para llegar a su posible solución comenzando por google y colocando una serie de palabras de carácter no tan elaboradas fue una búsqueda sencilla posteriormente abrió el resultado que mejor identificó como sentencia en una pestaña nueva abriendo cada sentencia en una pestaña nueva, al no encontrar algo que cumpliera con su búsqueda procedió a colocar palabras más a fines con el derecho elaborando una consulta más robusta hacia el tema, teniendo esto se dedicó a leer los resúmenes presentados por el buscador y comenzó a abrir las sentencias que mejor definían su búsqueda, finalmente revisó la primera sentencia buscando que coincidiera con sus parámetros de búsqueda y efectivamente coincidió, en ese momento manifestó que ya con lo que tenía podría elaborar una buena defensa. Como sugerencia el usuario considera fundamental enfocar la búsqueda hacia los derechos de las personas (puntal al caso) para poder hacerla más fácilmente. También considera que leyendo el resumen de cada sentencia o sus encabezados tienen suficiente información para saber si le sirve o no dicha sentencia. Confía mucho en el título y descripción del resultado de la búsqueda de Google</p>			
<p>Evidencia:</p>  <p>The screenshot shows a Google search results page for the query 'dignidad humana jurisprudencia'. The search results include several links to legal documents, such as 'Corte Constitucional de Colombia' and 'LA DIGNIDAD HUMANA EN LA JURISPRUDENCIA...'. Overlaid on the bottom left is the 'Morae Manager' software interface, which displays a project named 'Analyze - Project' with a list of tasks and markers. A 'Recording - 4.1' window is also visible, showing a video recording of the user's screen and a microphone icon, indicating that the user's actions were being recorded for evidence.</p>			
Figura 5. Prueba presentada Usuario 5			
Tabla 5. Prueba 5			

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

Características:

Con base en las entrevistas realizadas se logró identificar que para la presente investigación se deben tener en cuenta las siguientes consideraciones al realizar el Clustering:

- ✓ El buscador debe ser sencillo y fácil de usar por parte del usuario que va a utilizar el software.
- ✓ Para realizar la búsqueda es importante el diligenciamiento de dos parámetros los cuales son el año y una descripción de la información que se desea encontrar.
- ✓ El Clustering se debe realizar partiendo del año de la sentencia y el resumen de la misma, tomando el resumen como el escrito que contiene lo más importante.
- ✓ Los resultados que arroje la búsqueda se deben mostrar agrupados por el año ingresado por el usuario y ordenados de mayor a menor en cuanto al porcentaje de similitud del resumen con el texto ingresado en el buscador.
- ✓ El software debe permitir descargar la sentencia.

3.5 MÉTODOS DE CLUSTERING

Se realizaron las actividades relacionadas con el desarrollo del prototipo experimental, sus diferentes módulos y otros componentes necesarios para el funcionamiento del prototipo, como por ejemplo índices, algoritmos de agrupación.

- ✓ Diseño del buscador (prototipo).
- ✓ Desarrollo de clúster para la abstracción, documentos.
- ✓ Mecanismo para la indexación y búsqueda de documentos jurisprudenciales.
- ✓ Implementación de algoritmos para la agrupación de documentos legales, resultados de búsquedas realizadas por los usuarios.

3.6 ETIQUETAS DE CLUSTERING

- ✓ Texto ingresado en el buscador: espacio donde el usuario ingresara las palabras para la búsqueda de su sentencia.
- ✓ Año de la sentencia: la importancia de este espacio se da luego de que gran parte de los entrevistados en las pruebas a usuarios lo declararan como algo importante.
- ✓ KMEANS y DBSCAN: métodos a comparar en cada una de las sentencias y poder generar los resultados respectivos.
- ✓ Resumen de Sentencias: Espacio donde se muestran los resultados de la búsqueda y sus respectivos resúmenes.
- ✓ Visualizador de sentencias: Espacio donde se visualiza el documento completo de la sentencia elegida a partir del resultado elegido.

3.7 ALGORITMOS DE CLUSTERING A UTILIZAR

Los dos algoritmos utilizados hicieron el clustering con registros de un dataset, el cual contiene el texto de las sentencias extraídas de la página web de la Corte Constitucional.

Con el análisis realizado y plasmado en el estado del arte se destacó el uso de dos métodos de agrupamiento, KMEANS y DBSCAN esto debido a su uso variado en el área del análisis de datos donde se obtienen excelentes resultados en la búsqueda y agrupación de documentos, datos, etc. Además de contar con una comunidad que sigue desarrollando herramientas que complementan cada uno de estos métodos de agrupamiento.

En el estado del arte se evidenciaron métodos como K-medoid, Algoritmos Genéticos, KNN, entre otros pero estos son aplicados en otro tipo de ámbitos ya que para el tema de agrupación de documentos no son lo suficientemente óptimos.

Al ingresar un texto de búsqueda y el año de la sentencia, el sistema consulta en la base de datos los registros que tengan similitud con la frase ingresada en el

buscador, posteriormente se convierten en una matriz de datos (dataset) numéricos para su respectivo análisis. Tanto en KMEANS como en DBSCAN el agrupamiento se basa en la información contenida en la columna “descripción”, estos valores serán evaluados en cada una de las distancias k que corresponden entre centroides-muestras para KMEANS y nucleos-muestras en el caso de DBSCAN.

3.7.1 KMEANS

Para este algoritmo se usó un $K=8$ lo cual se traduce en tener 8 agrupaciones del dataset ya que al elegir menor cantidad de clusters la cantidad de sentencias agrupadas eran de gran cantidad pero generaban resultados poco precisos al momento de buscar y al utilizar más de 8 clusters para el agrupamiento se encontraban muy pocos por grupo disminuyendo su precisión; por lo tanto se decidió utilizar un valor de $K=8$ la cual es una medida justa agrupando los mejores resultados preparado con anterioridad a continuación se observa el funcionamiento general del algoritmo.

Después de haber terminado el proceso se pueden determinar los K clusters desde los de mayor población hasta los de menor población.

NOTA: Estos grupos se consideran etiquetas originales que se utilizarán en el siguiente paso.

3.7.2 DBSCAN

Para este algoritmo se trabajó con un $\epsilon = 1$ y un $\text{minPuntos} = 2$ los cuales garantizaran debido a que el algoritmo DBSCAN trabaja basado en densidad, a mayor cantidad de sentencias presenta mejor precisión al arrojar los resultados de las búsquedas. Este algoritmo fue creado para comprender datos numéricos debido a su proceso basado en densidad cambiando de forma drástica cuando se trabaja con datos convertidos de texto a valores de vector, donde cada una de las palabras crea un patrón numérico según su uso y su contexto. un trabajo óptimo y con poco ruido, a continuación, se presenta el funcionamiento general del algoritmo.

3.8 DIFERENCIAS ENTRE DBSCAN Y KMEANS

3.8.1 DBSCAN

- ✓ Necesita escoger una distancia prudente en términos de los datos a ser agrupados, esta distancia se da por el radio que el usuario introduce en el parámetro r del algoritmo DBSCAN.
- ✓ Necesita que el usuario ingrese 2 parámetros, uno es el radio bajo el cual se dará el barrido para tomar los datos y el otro término es el eps que representa el número mínimo de muestras que se deben tener en dicho radio para poder hacer de este un cluster.
- ✓ La distancia entre dos puntos representará la densidad de puntos y mostrada, si dos puntos de la misma clase se encuentran bajo el mismo barrido del radio antes introducido entre punto y punto pueden ser agrupados en la misma clase.
- ✓ En el algoritmo DBSCAN no se necesita saber el número de clusters para poder crear las agrupaciones ya que su forma de Clustering se da gracias a los parámetros introducidos con anterioridad.

3.8.2 KMEANS

- ✓ La agrupación se da mediante distancia euclidiana entre el centroide escogido al azar y el punto de la muestra.

$$d_{xy} = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$$

Ecuación 1. Distancia Euclidiana

- ✓ Los puntos se agrupan según la distancia euclidiana óptima debido a que el proceso de medición entre centroide y cada muestra se realiza hasta encontrar las distancias más adecuadas.
- ✓ El usuario debe ingresar un parámetro K el cual representa la cantidad de clusters determina la semejanza de los centroides y cada una de las muestras del dataset dado con anterioridad.

CAPÍTULO IV

4. IMPLEMENTACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES.

4.1 INTRODUCCIÓN

En el capítulo de implementación de técnicas de clustering se indica el proceso de desarrollo de los algoritmos de clustering K-MEANS y DBSCAN, así como la explicación detallada de las librerías utilizadas para cumplir con los objetivos propuestos.

4.2 ARQUITECTURA DEL BUSCADOR

Se presenta la arquitectura soportada en algoritmos de clustering para extraer e identificar características, que permitan optimizar los procesos de búsqueda de jurisprudencia en el ámbito judicial colombiano. A continuación en la ilustración 1, se presentan los elementos que fueron utilizados para el diseño de la arquitectura.

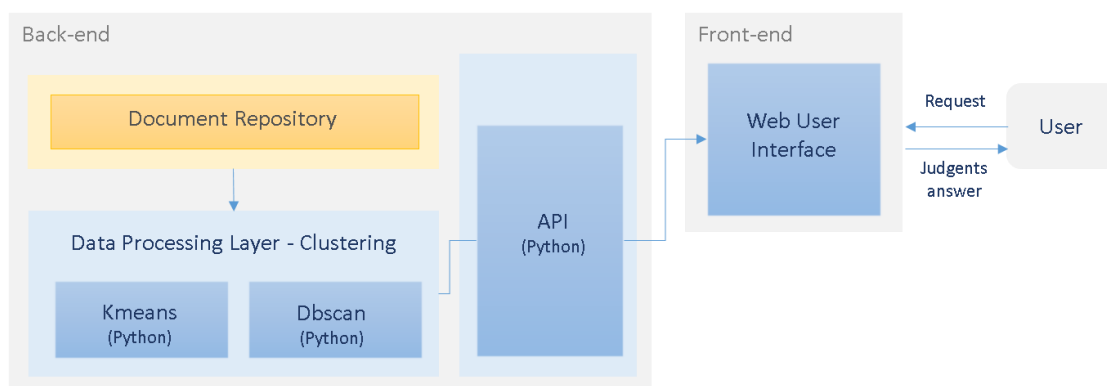


Figura 6. Arquitectura Buscador

4.3 IMPLEMENTACIÓN DEL BUSCADOR WEB

4.3.1 Motor de Base de Datos

Se utilizó el gestor de base de datos MySQL debido a la adaptabilidad que posee con el sistema de recuperación de documentos de jurisprudencia llamado Scrappy el cual se encarga de buscar documentación de forma automática para luego almacenarla en una base de datos.

4.3.2 Página web

Es la encargada de mostrar el buscador, dicha página fue implementada en php debido a su amplia documentación para su construcción como para su implementación.

4.3.3 Algoritmos de Agrupamiento

Implementado en Python debido a la capacidad de dicho lenguaje para poder manejar dichos algoritmos y la gran complejidad que existe en ellos, esto unido a la documentación hace de este la mejor opción para la implementación de estos.

Librerías usadas en la implementación de los algoritmos de agrupamiento:

- ✓ Sklearn.feature_extraction.text³: procesamiento de texto como si fuera una serie de elementos numéricos convirtiendo cada uno de estos elementos en matrices y vectores con los cuales podremos relacionar valores, con esta parte de la librería implementada se comienza a trabajar sobre estos datos para poder llevar a cabo el clustering.
- ✓ Sklearn.metrics import silhouette_score: o valor de silhouette es una medida que se rige a que tan similares son cada uno de los elementos al grupo al cual pertenece en relación a otros grupos. Esta puede variar entre -1 y 1 donde el valor alto es el mejor e indica que cada elemento del grupo es similar y que no corresponde a otros elementos de otro grupo por el contrario si es bajo esto podría dar valor a muy pocos clusters debido a la baja cohesión entre elementos sus elementos.
- ✓ Sklearn.cluster import KMEANS: Encargado de implementar todo lo relacionado con los procesos de clustering (agrupamiento) en este caso específico se hace KMEANS, esta parte de la librería sklearn es la más importante en el momento de la implementación.
- ✓ Pandas: potente librería para el análisis de datos con la cual se pueden obtener mejores resultados gracias al uso que este le da a su estructura básica llamada dataframe el cual es la recopilación de todos los elementos necesarios para el estudio de determinado caso.
- ✓ Mysql.connector: librería encargada de gestionar las conexiones a la base de datos sql sobre la cual se trabaja.

³ Herramientas disponibles en <https://scikit-learn.org/stable/>.

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

- ✓ Numpy: librería utilizada para gestionar las operaciones complejas entre diferentes arreglos, complemento ideal para la gestión de estos datos con Pandas.
- ✓ Json: el cual es un formato de texto sencillo con el cual se puede realizar el intercambio de datos entre el programa hecho en python y la página web en php.
- ✓ Nltk: [26] Abreviatura de NATURAL LENGUAJE TOOL KIT, librería que contiene diferentes recursos enfocados en la interpretación e implementación de datos inspirados en lenguaje natural humano; en su mayoría están dedicadas al procesamiento de texto para este caso se hará uso de la herramienta stopwords que se explicará más adelante.

Teniendo en cuenta la explicación de cada una de las librerías se inicia con la implementación del código.

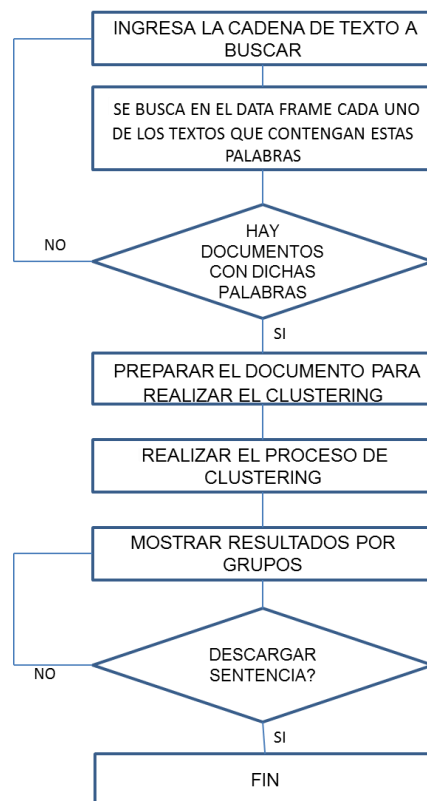


Figura 7. Diagrama de flujo funcionamiento del código

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

El primer elemento es la declaración de la función stopwords de la librería NLTK, la cual se encarga de analizar las palabras vacías en pocas palabras sin significados como lo son artículos, pronombres, preposiciones, etc. Muchos de los algoritmos utilizados para esta práctica vienen dados por un diccionario antes cargado y otros funcionan de forma natural evitando la limitación de estos, para este caso se utiliza un algoritmo basado en listas para lo cual se cargará con un diccionario basado en el ESPAÑOL.

La clase ClassControl declarando así el inicio de todo el proceso de conexión a la respectiva base de datos, la recolección de las palabras en la página web para luego ser tratadas y pasar a su clusterización.

Conexión a base de datos con los parámetros necesarios para uso de los documentos que se encuentran en esta (DATAFRAME).

Declaración del método run_query_KMEANS (correr_busqueda), la cual recibirá dos parámetros directamente del buscador, y estos son los datos que se tomaron como elementos de importancia para la búsqueda y encontrados anteriormente con la experiencia de usuarios, estos parámetros son el texto que se quiere encontrar y el año en el cual fue publicado.

Se pasa la palabra a un solo formato de minúsculas esto con el fin de que todo quede en un solo formato para poder categorizar cada una de estas y poder separar una a una por el reconocimiento del espacio en la función split, este luego se comienza la búsqueda de cada una de estas hasta encontrar la última y poder llevarla al query que ira a la base de datos.

Se crea la consulta dinámicamente con el texto de búsqueda y elimina palabras vacías al implementar el método stopwords con la consulta anteriormente creada para poder ejecutar la consulta en la base de datos, este retornara cada uno de los elementos encontrados con cada una de las palabras.

Continuadamente se pasa a la indexación de cada uno de los id de cada una de las tutelas que cumplen con la búsqueda y con estos elementos comenzar el tratamiento para comenzar el proceso de clusterización.

Al inicio del código se habló de la librería sklearn.feature_extraction.text cuya principal función es poder extraer caracteres y poderlos ver en un formato compatible con algoritmos de aprendizaje numéricos automáticos.

Esta se complementa con TfidfVectorizer que convierte una serie de documentos en una matriz de características TF-IDF donde cada uno de estos donde TF significa frecuencia de término eso quiere decir que los entre más elementos de un documento sea igual y en valores más altos de este simplemente significarían una

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

mayor importancia en el texto dado pero en ese caso, los documentos de mayor tamaño tendrían más apariciones de palabras que los documentos más pequeños. Por lo tanto, una mejor representación sería normalizar la aparición de la palabra con el tamaño del cuerpo y se denomina término-frecuencia, en pocas palabras no importa el número de palabras según el texto, si no el valor de igualdad de estas palabras respecto al tamaño del documento, seguidamente estos valores deben ser los óptimos por lo que se da el uso de fit_transformer, con un elemento numérico optimizado llamado X se procede a realizar el clustering.

Ambos códigos son similares al tratamiento de datos para poder trabajarlos bajo los 2 algoritmos de clustering y en esta parte de la implementación del código radica su diferencia ya que aquí es donde se realiza el KMEANS o el DBSCAN y a continuación se explicarán los elementos que cada uno de estos necesitan para poder lograr su objetivo.

4.3.4 Implementación del algoritmo KMEANS

Para este modelo se definieron la cantidad de Clusters los cuales fueron 8; posteriormente se declaró la variable donde se guardan los resultados del algoritmo KMEANS los cuales fueron optimizados y ubicados en doc. Mediante el id de la resolución y al cluster al que pertenece.

Algoritmo KMEANS	
Paso 1.	Selecciona $K=8$ del dataset X y se eligen los centroides iniciales al azar.
Paso 2.	Se calcula la distancia euclidiana entre cada uno de los datos y los K clusters mediante la formula $dxy = \sqrt{\sum_{i=1}^n (Xi - Yi)^2}$ y se dividen cada uno de estos datos según la cercanía del clúster con cada centro.
Paso 3.	Recalcula el centro de estos datos.
Paso 4.	Calcula la medida estándar de la función.
Paso 5.	Si ha llegado al número máximo de iteraciones terminar, si no volver al Paso 2.

Tabla 6. Algoritmo KMEANS

4.3.5 Implementación del algoritmo DBSCAN

Para el modelo de DBSCAN se indicaron dos variables las cuales son: eps de tipo float que representa la máxima distancia (radio) entre cada una de las muestras para así formar su vecindario, y por último la variable min_samples la cual determina la cantidad de muestras en un vecindario para que un punto se considere como un punto central, esto incluye el punto en sí.

Una predicción de clase es: dado el modelo finalizado y una o más instancias de datos, predice la clase para las instancias de datos no sabemos las clases de resultados para los nuevos datos, es por eso que necesitamos el modelo en primer lugar. Podemos predecir la clase para nuevas instancias de datos usando nuestro modelo de clasificación finalizado en scikit-learn usando la función predict.

En esta parte del código se recorre cada uno de los arreglos se dan los primeros resultados con mejor rendimiento y estos son guardados en los distintos arreglos y sus resultados son enviados al buscador principal.

Algoritmo DBSCAN	
Paso 1.	Selecciona un punto P del dataset al azar, este punto P es el valor sobre la descripción del documento la cual fue dada al transformar el arreglo inicial en un dataset.
Paso 2.	Recupera todos los puntos de densidad, en pocas palabras descripciones con valor cercano y alcanzables desde P y que se encuentran en un EPS(radio).
Paso 3.	Si P es un núcleo y el minPuntos que es la cantidad de vecinos necesarios para formar un núcleo P se cumple, un clúster es formado, esto indica que las descripciones entre cada documento son muy similares a la de núcleo garantizando los resultados sobre este cluster.
Paso 4.	Si P es un borde, no hay puntos de densidad alcanzables de P DBSCAN visita el próximo punto del dataset.
Paso 5.	Continúa el proceso hasta que todos los puntos hayan sido procesados.

Tabla 7. Algoritmo DBSCAN

4.3 INDEXACIÓN DE DOCUMENTOS POR MEDIO DE CLUSTERING

Estos métodos son utilizados para generar un conjunto de agrupaciones basado en las búsquedas de los usuarios para ilustrar la implementación y la búsqueda de la propuesta se presenta en la siguiente, donde el proceso se determina bajo la búsqueda del documento por parte del usuario en el sistema en este paso se utiliza dos técnicas, primero convertir el texto basado en el campo descripción de la base de datos, el proceso de búsqueda se describe a continuación.

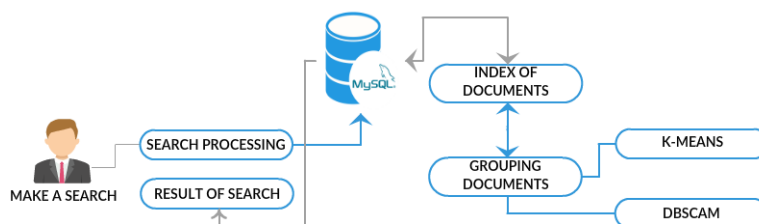


Figura 8. Indexación de Documentos

INDEXACIÓN DE DOCUMENTOS	
1	El usuario ingresa el texto de búsqueda y selecciona el año en el cual desea encontrar la sentencia.
2	El texto ingresado en el campo de búsqueda es procesado por la librería Stopwords de tal manera que quede limpio de palabras vacías.
3	Se realiza la consulta en base de datos de los documentos que tienen una similitud teniendo en cuenta el texto de búsqueda con palabras vacías eliminadas y el año seleccionado en la interfaz.
4	El resultado obtenido de la consulta realizada en base de datos que arroja los documentos, es almacenado en un DataFrame que contiene los campos: Identificador de la Tutela, Título, Texto Tutela, Link, y Año.
5	Para realizar el clustering se tiene en cuenta el texto de la tutela; debido a que los algoritmos utilizados (KMEANS y DBSCAN) necesitan información transformada en matrices de datos para realizar el agrupamiento y no lo hacen a partir de palabras directamente, el contenido de cada documento debe ser tratado; para ello se utilizan las librerías de sklearn (TfidfVectorizer y TfidfTransformer), TfidfVectorizer convierte en tokens cada palabra de cada documento y realiza el conteo de las veces que se repite; TfidfTransformer transforma los recuentos en bruto a una matriz de valores de TF / IDF (Frecuencia de ocurrencia de términos en los documentos obtenidos en la consulta).

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

6	Posteriormente con la información transformada se realiza el clustering con los algoritmos (KMEANS y DBSCAN) utilizando las librerías de sklearn; cada algoritmo realiza el agrupamiento y como resultado se obtienen los grupos, cada grupo contiene los identificadores de las tutelas que tienen similitudes entre ellas.
7	Finalmente con el texto ingresado en el buscador se realiza la predicción por medio de la librería predict para KMEANS y fit_predict para DBSCAN, el resultado que arroja la predicción indica a que grupo de documentos pertenece la búsqueda realizada y estos son mostrados en la interfaz gráfica.

Tabla 8. Pasos para la indexación de documentos

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

4.3 DESARROLLO DEL SISTEMA

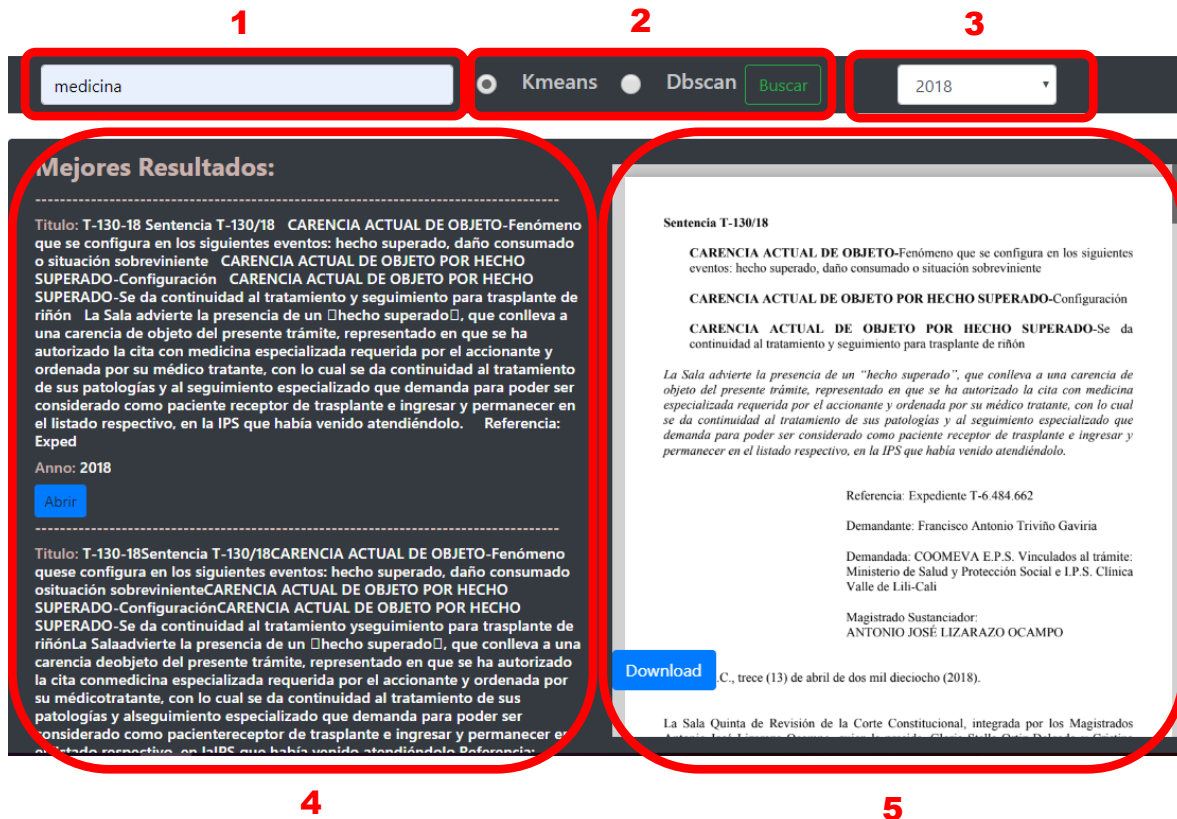

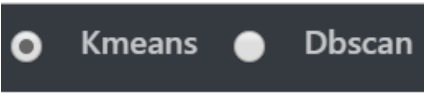


Figura 9. Buscador de Documentos Jurisprudenciales

<p>Campo de ingreso de búsqueda</p>	 <p>Figura 10. Campo de Ingreso Texto</p>
<p>En este campo el usuario ingresara las palabras de búsqueda requeridas para ubicar las diferentes sentencias.</p>	
<p>Elección de método de agrupamiento.</p>	 <p>Figura 11. Elección método agrupamiento</p>
	<p>En cada caso a ser evaluado el usuario debe seleccionar un método de agrupamiento con el cual se realizará la indexación de los documentos.</p>

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

<p>Selección de año en el cual se quiere ubicar la sentencia y botón buscar.</p>	<div data-bbox="769 226 1240 310" data-label="Image"> </div> <p align="center">Figura 12. Botón búsqueda - Año</p> <p>El usuario podrá elegir el año en el cual ubicar la sentencia para pasar luego a la búsqueda de dicho documento mediante el botón Buscar ubicado al lado izquierda de este.</p>
<p>Resultados y Resumen de Sentencias</p>	<div data-bbox="721 630 1299 1239" data-label="Image"> </div> <p align="center">Figura 13. Resultados y resumen</p> <p>En la presente sección se visualizan los resúmenes de cada una de las sentencias ubicadas por año, desde el mejor resultado hasta el peor, debajo se presenta el botón abrir el cual nos permite visualizar el documento en el visor ubicado en el lado derecho de la ventana.</p>

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

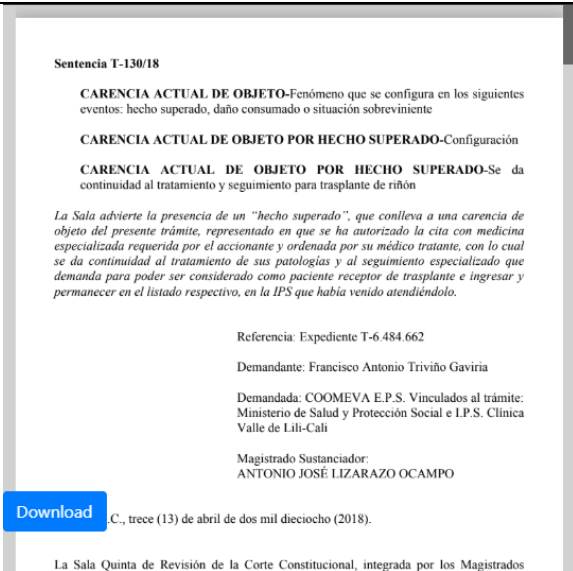
<p>Visualización de la sentencia que se escogió con su respectivo botón para su descarga.</p>	 <p>The screenshot shows a legal document titled 'Sentencia T-130/18'. It contains several sections: 'CARENCIA ACTUAL DE OBJETO-Fenómeno que se configura en los siguientes eventos: hecho superado, daño consumado o situación sobreviniente', 'CARENCIA ACTUAL DE OBJETO POR HECHO SUPERADO-Configuración', and 'CARENCIA ACTUAL DE OBJETO POR HECHO SUPERADO-Se da continuidad al tratamiento y seguimiento para trasplante de riñón'. There is a paragraph starting with 'La Sala advierte la presencia de un "hecho superado"...'. Below this, there are fields for 'Referencia: Expediente T-6 484 662', 'Demandante: Francisco Antonio Triviño Gaviria', 'Demandada: COOMEVA E.P.S. Vinculados al trámite: Ministerio de Salud y Protección Social e I.P.S. Clínica Valle de Lili-Cali', and 'Magistrado Sustanciador: ANTONIO JOSÉ LIZARAZO OCAMPO'. A blue 'Download' button is visible over the text. At the bottom, it says 'C., trece (13) de abril de dos mil dieciocho (2018). La Sala Quinta de Revisión de la Corte Constitucional, integrada por los Magistrados'.</p>
	<p>Al darle abrir en el botón de cada resumen en la ventana de resultados como se muestra en la Ilustración 13 se invocara el documento completo como se observa en la Ilustración 14, si el usuario está satisfecho con dicho documento puede usar el botón download para poder descargarlo a su dispositivo y darle su respectivo uso.</p>

Figura 14. Visualización de la Sentencia

Tabla 9. Funcionamiento buscador Jurisprudencia

CAPÍTULO V

5. EVALUACIÓN POR EXPERTOS DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES.

5.1 INTRODUCCIÓN

En este capítulo se presenta el proceso de evaluación y pruebas realizadas por expertos en el ámbito judicial de tal manera que la búsqueda logre cumplir los objetivos propuestos.

5.2 INSTRUMENTO DE VALIDACIÓN

Esta sección describe los experimentos desarrollados para evaluar la satisfacción del usuario con el sistema. Estos usuarios fueron escogidos en base a su oficio en el sector de derecho y la jurisprudencia por lo que se buscaron abogados cuya carrera estuviera finalizada o estudiantes que estuvieran prontos a finalizar su estudio en derecho, también se dio como un filtro el cual garantizara que cada uno de estos este ejerciendo su labor de tiempo completo en este oficio. Cada uno de ellos interactuó con el sistema y luego proporcionaron sus reflexiones sobre su grado de satisfacción alcanzado en diferentes aspectos. Por otro lado, se integra un cuestionario que permita evaluar a los usuarios, el sistema este se basa en el modelo propuesto en [16] para evaluar el éxito del Sistema de recuperación de documentos basado en Clustering aplicado en el ámbito judicial se adoptado este modelo debido a las numerosas similitudes con el caso en mano, por lo tanto, debido a su facilidad de aplicación con una baja necesidad de adaptación. Además, desde el punto de vista del usuario es más fácil para ellos proporcionar información si hay pocas preguntas, considerando que los usuarios por lo general son reacios a responder encuestas. En particular, se plantearon 4 preguntas con una escala de respuesta por el usuario de 1 a 5 donde 1 es deficiente, 2 Insuficiente, 3 aceptable, 4 bueno, 5 aceptable.

TABLA DE PREGUNTAS	
PREGUNTA Q1	¿Cuál de los dos algoritmos genera mejores agrupaciones en la búsqueda KMEANS o DBSCAN? En este caso nos centramos en observar si los algoritmos les generan confianza en agrupación de las búsquedas.
PREGUNTA Q2	¿Qué algoritmo consideras que tiene mejor precisión a la hora de la búsqueda KMEANS o DBSCAN? En este caso nos centramos en observar que algoritmo les genera confianza en precisión de las búsquedas.
PREGUNTA Q3	¿Cuál algoritmo genera una búsqueda más ágil en tiempo de respuesta dentro del sistema KMEANS o DBSCAN? La idea es medir la reacción de los usuarios, y el tiempo medido por el usuario en respuesta a la búsqueda realizada por los algoritmos.
PREGUNTA Q4	¿Qué algoritmo genera mejor agrupación por fecha en el sistema? La idea es medir de manera general que algoritmo genera mejor agrupación por fecha, analizando lo que piensan cuando usan el sistema.

Tabla 10. Tabla de Preguntas

5.3 CASOS DE SENTENCIAS

1. Un miembro de la Armada Nacional envió mediante cadena de whatsapp fotos íntimas de su expareja sin consentimiento a un grupo creado dentro de la fuerza naval como canal de comunicación de la Escuela de formación de cadetes. La mujer, quien también es oficial de la Armada, presentó denuncia en Fiscalía, antes los superiores jerárquicos de la Armada y ante la Procuraduría para que se impusieran las sanciones de rigor, pero las investigaciones fueron archivadas y no presentan avance alguno. La mujer manifiesta haberle sido vulnerada su intimidad, integridad personal psicológica y sexual.
2. Una mujer que padece de diabetes e insuficiencia renal crónica terminal y por la enfermedad recibe tratamiento de terapia de reemplazo renal con hemodiálisis, cuatro horas al día, cuatro veces por semana, fue calificada por una junta médica de calificación de pérdida de capacidad laboral mediante dictamen que la calificó con pérdida de capacidad

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

3. Una señora que cuenta con 70 años de edad y que en vida de su cónyuge, ella y la hija de ambos, quien padece retardo mental y epilepsia, dependieron económicamente del fallecido, quien cotizó a la seguridad social durante toda su actividad laboral. En compañía de su hija viven en una habitación, que pagan con el subsidio que recibe del Gobierno nacional; afirma que ella padece un cáncer en fase II. El Seguro Social le negó los reconocimientos de la pensión de sobreviviente y de la indemnización solicitada con ocasión de la muerte de su esposo, aduciendo que a éste le había sido negada la pensión de vejez y que los interesados no exigieron en tiempo el pago de la prestación sustitutiva. Solicita que le sea reconocida la pensión de sobrevivientes.
4. El Dpto. del Valle como último empleador del señor, le reconoció una pensión de jubilación a partir de 1995, en cuantía de \$165.007, guardando silencio respecto de las peticiones de indexación de la primera mesada pensional y el pago de los intereses moratorios. El peticionario presentó demanda ordinaria laboral contra el Dpto. del Valle; en la sentencia se accedió a las pretensiones de la demanda, condenando a reajustar la pensión de jubilación. Apelada la decisión por la parte vencida, la sala Laboral del tribunal superior decidió revocar la sentencia apelada y en su lugar inhibirse para fallar de fondo considerando que le corresponde a la jurisdicción contencioso-administrativa. El señor solicitó entonces mediante derecho de petición al Dpto. la indexación de su primera mesada pensional con el pago de los intereses correspondientes, sin que se haya dado respuesta a su solicitud.
5. El señor EEE en virtud de demanda instaurada en 2006 y acreditando ante el Juzgado que se había desempeñado como docente en el sector oficial por 20 años, que su vinculación al mismo se produjo antes del 31 de diciembre de 1980, y que tenía una edad superior a los 50 años, le fue concedida mediante sentencia la pensión gracia. A pesar de la sentencia en mención CAJANAL profirió acto administrativo negando la pensión de gracia de la que es beneficiario el señor EEE , entonces fue instaurado incidente de desacato en contra de la entidad, pero pese a este, CAJANAL continúa renuente a dar cumplimiento a la orden impartida en la sentencia de 2006 alegando que existe una incompatibilidad en la pretensión, pues, es prohibido percibir dos pensiones a cargo de la Nación, hecho que se presenta porque se encuentra incluido en nómina como beneficiario de una pensión ordinaria de vejez.

5.4 RESULTADOS

A continuación, se presentaran los resultados obtenidos después de la evaluación por parte de los usuarios especializados en el área de derecho de distintas universidades de la ciudad de Popayán.

RESULTADOS DE BÚSQUEDA POR PROBLEMA					
	CASO 1	CASO 2	CASO 3	CASO 4	CASO 5
EVALUADOR 1	NO	SI	SI	NO	NO
EVALUADOR 2	NO	SI	SI	NO	NO
EVALUADOR 3	SI	SI	SI	SI	SI
EVALUADOR 4	SI	SI	NO	SI	SI
EVALUADOR 5	NO	SI	SI	SI	SI
EVALUADOR 6	SI	SI	SI	SI	SI
EVALUADOR 7	SI	SI	SI	SI	SI
EVALUADOR 8	SI	SI	SI	SI	SI
EVALUADOR 9	SI	SI	SI	SI	SI

Tabla 11. Solución de cada evaluador por caso

En la tabla 11. Se presentan cada uno de los casos evaluados por los nueve abogados, indicando si se encontró o no un resultado de búsqueda positivo.

En el proceso de evaluación de expertos se realizaron las búsquedas con una base de datos de aproximadamente 8000 sentencias divididas en diferentes años, esto trajo consigo un problema en las primeras dos entrevistas y sobre todo el primer caso ya que no habían sentencias que dieran credibilidad ante los evaluadores, al notar este primer problema se dio paso a aumentar la cantidad de sentencias en la base de datos mejorando con esto la precisión de las búsquedas, de tal manera que fueron satisfactorias para los evaluadores como se puede observar en la tabla 6. Este crecimiento en la cantidad de sentencias encontradas se ve reflejado en la tabla de evaluación extraída de la encuesta para la pregunta Q1, donde las 2 primeras encuestas presentan deficiencia en la búsqueda de estos casos pero que mejora a partir del 3 evaluador donde el dataset constaba de más de 28000 sentencias, se dio una mejora en la percepción de la aplicación.

5.4.1 Pregunta Q1.

PREGUNTA Q1		
USUARIO	DBSCAN	KMEANS
EVALUADOR 1	2	2
EVALUADOR 2	2	3
EVALUADOR 3	3	5
EVALUADOR 4	4	3
EVALUADOR 5	3	4
EVALUADOR 6	4	4
EVALUADOR 7	5	3
EVALUADOR 8	3	4
EVALUADOR 9	4	5
PROMEDIO	3,3	3,7

Tabla 12. Resultados pregunta Q1

Teniendo en cuenta este análisis que se realizó sobre la tabla 12. se puede ver como el Algoritmo KMEANS presento de forma más **eficiente** los resultados posicionando las mejores sentencias según los criterios de búsqueda del usuario en los primeros puestos de la lista y resúmenes, a diferencia de DBSCAN que mostraba buenos resultados, pero en menor medida quedando clara una deficiencia al momento de agrupar las principales sentencias, teniendo en cuenta que la escala de puntuación es 1: Deficiente, 2: Insuficiente, 3: Aceptable, 4: Bueno, 5: Excelente, se nota como éstas agrupaciones fueron mejoradas a partir de las fallos de las primeras 2 evaluaciones, cuando se da promedio de estas notas ya queda en evidencia cuál de los dos posee una diferencia respecto a la pregunta que se hizo con anterioridad la cual se trata de ver cuál es el mejor método de agrupamiento.

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

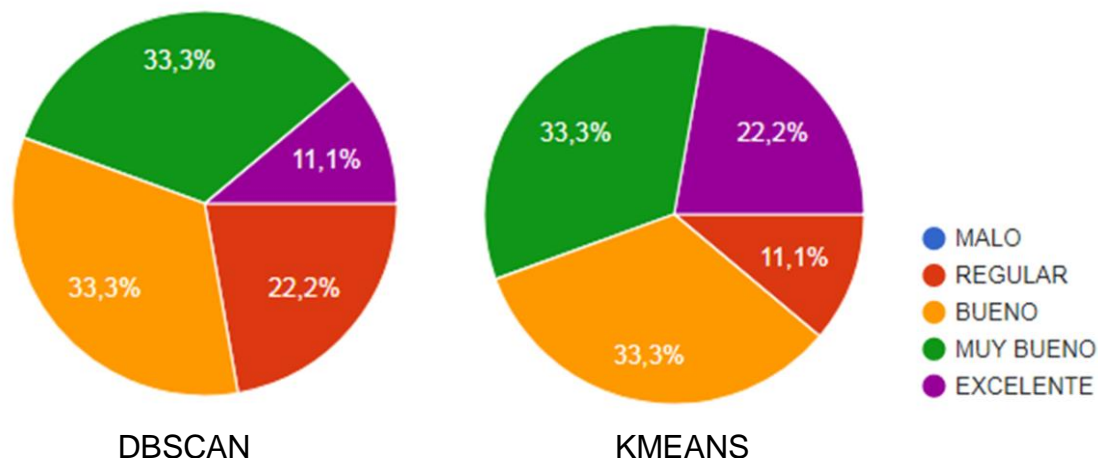


Gráfico 1. Resultados pregunta Q1

La búsqueda de los diferentes casos arroja un análisis donde se determina diferentes categorías para el algoritmo DBSCAN el 22% de los encuestados aseguran que el algoritmo K-MEANS al contrario de DBSCAN se observa que solo el 11% de los encuestados les genera una excelente agrupación, además se determina que el 33% determina que los 2 algoritmos generan un muy buen agrupamiento, por otra parte el algoritmo se destaca que el 22% menciona que DBSCAN genera un agrupamiento regular a contrario del 11% del algoritmo K-MEANS

5.4.2 Pregunta Q2.

PREGUNTA Q2		
USUARIO	DBSCAN	KMEANS
EVALUADOR 1	2	3
EVALUADOR 2	2	3
EVALUADOR 3	4	4
EVALUADOR 4	2	2
EVALUADOR 5	4	4
EVALUADOR 6	4	4
EVALUADOR 7	4	5
EVALUADOR 8	3	5
EVALUADOR 9	3	5
PROMEDIO	3,1	3,9

Tabla 13. Resultados pregunta Q2

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

En la Tabla 13 se muestran los resultados concernientes a la pregunta Q2 cuyos resultados se basan en una escala 1: Deficiente, 2: Insuficiente, 3: Aceptable, 4: Bueno, 5: Excelente; en este caso la pregunta analiza el nivel de agrupamiento que presenta cada uno de los métodos de agrupaciones dados, KMEANS obtiene mejores resultados en el momento de agrupar, esto se reflejaba en la mayoría de las puntuaciones teniendo en cuenta el Promedio que presenta cada uno de estos y en el resultado del contenido donde los usuarios podían observar sus respuestas en las primeras filas de la parte de Resultados y Resumen de Sentencias, estos resultados se distancian en gran parte en los usuarios 8 y 9 sus puntajes fueron excelentes mientras que DBSCAN obtiene un puntaje de bueno en cada uno de ellos respectivamente.

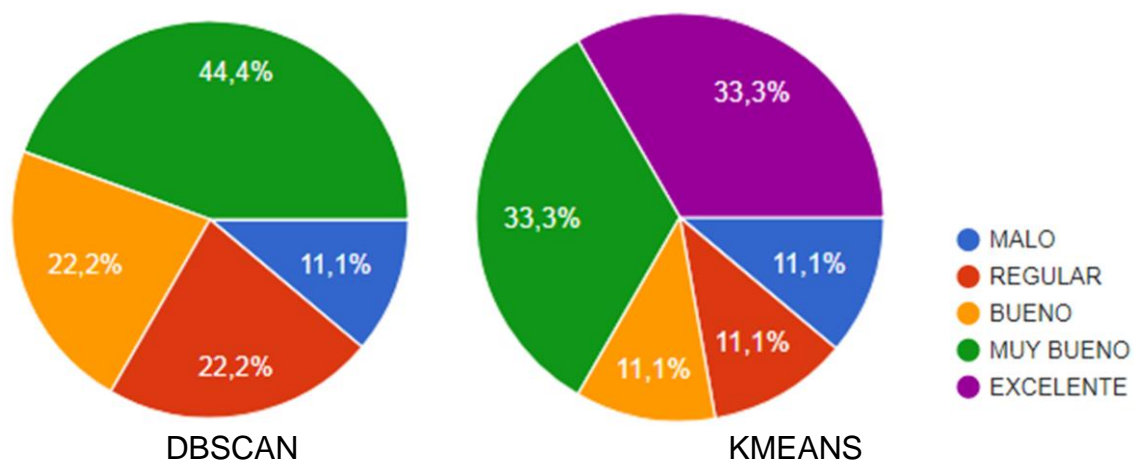


Gráfico 2. Resultados pregunta Q2

En los resultados a nivel de porcentajes se pueden observar las tendencias de forma muy definida con un 33.3% excelente contra un porcentaje de 0% de DBSCAN para este caso, pero aunque KMEANS obtiene la ventaja no significa que DBSCAN esté mal, ambos métodos dieron buenos resultados con evaluaciones de Bueno con el 22% para el DBSCAN y el 11% para el KMEANS, MUY BUENO con el 44% para DBSCAN y 33% para KMEANS.

5.4.3 Pregunta Q3.

PREGUNTA Q3		
USUARIO	DBSCAN	KMEANS
EVALUADOR 1	3	3
EVALUADOR 2	2	2
EVALUADOR 3	3	5
EVALUADOR 4	4	4
EVALUADOR 5	3	4
EVALUADOR 6	4	3
EVALUADOR 7	5	4
EVALUADOR 8	4	3
EVALUADOR 9	4	3
PROMEDIO	3,6	3,4

Tabla 14. Resultados pregunta Q3

La Tabla 9. muestra los resultados sobre la pregunta Q3 donde se analiza la agilidad con la que cada uno de los métodos de agrupamiento entrega los resultados, en la escala donde 1: Deficiente, 2: Insuficiente, 3: Aceptable, 4: Bueno, 5 Excelente, observamos que la mayoría de su comportamiento entra en el rango de Normal - Bueno, todo esto teniendo en cuenta que existen diferentes variables que puedan influenciar este comportamiento, una que se debe destacar es el equipo donde se está ejecutando la herramienta en base a sus características internas por ejemplo una computadora con un procesador A10 con 8 gigas de RAM DDR3 y disco duro HDD

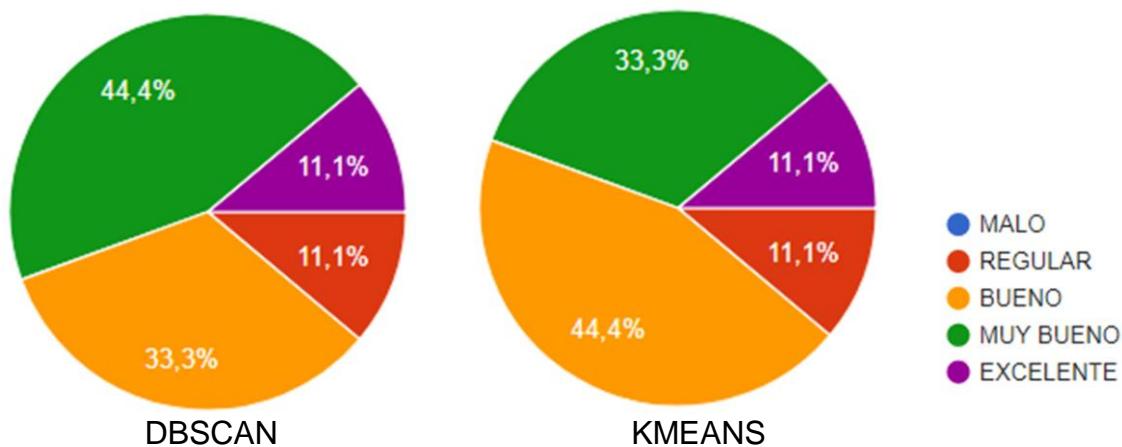


Gráfico 3. Resultados pregunta Q3

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

A nivel de agilidad en tiempo de búsqueda DBSCAN y KMEANS dan un resultado muy similar a los usuarios quedando muy cercano el uno del otro pero donde DBSCAN adquiere una pequeña ventaja en solo dos calificaciones quedando con el 44% DBSCAN y 33% KMEANS en un rango de muy bueno y 33% DBSCAN - 44.4% KMEANS y aunque KMEANS sea superior en el rango de BUENO, DBSCAN tiene la superioridad en KMEANS con un 10.4% de ventaja.

5.4.4 Pregunta Q4

PREGUNTA Q4		
USUARIO	DBSCAN	KMEANS
EVALUADOR 1	4	5
EVALUADOR 2	3	3
EVALUADOR 3	4	5
EVALUADOR 4	5	5
EVALUADOR 5	4	4
EVALUADOR 6	4	4
EVALUADOR 7	4	4
EVALUADOR 8	4	5
EVALUADOR 9	4	5
PROMEDIO	4,0	4,4

Tabla 15. Resultados pregunta Q4

La Tabla 10 muestra muy buenos resultados en la escala donde 1 Deficiente, 2: Insuficiente, 3: Aceptable, 4: Bueno, 5: Excelente, se puede observar con detalle como los usuarios calificaron cada uno de estos métodos en base a la pregunta Q4 la cual evalúa cada algoritmo en base al agrupamiento por fecha dejando en promedio 4 para DBSCAN y 4.4 para KMEANS ubicándolos en una escala de MUY BUENO.

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

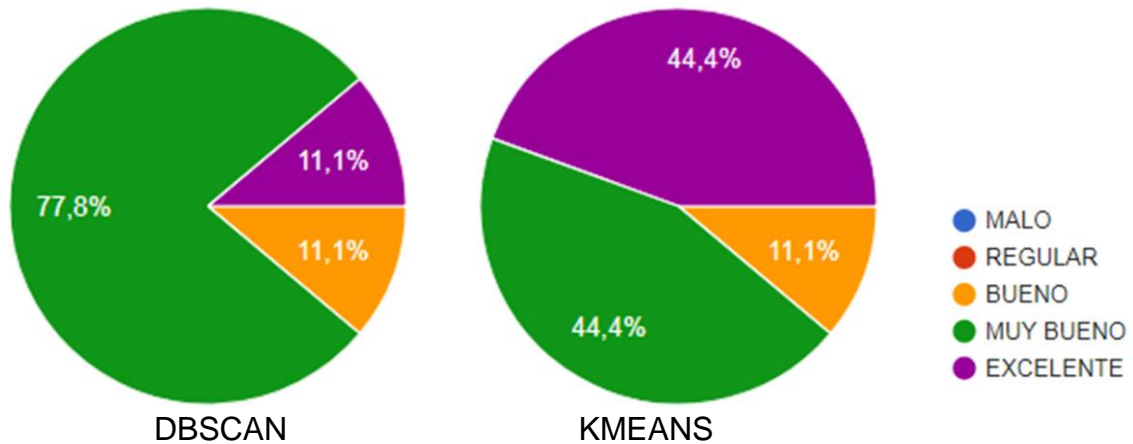


Gráfico 4. Resultados pregunta Q4

Los resultados de ambos algoritmos son similares agrupando de forma correcta según el año solicitado por el usuario como muestra la Ilustración 13. Pero dejando a KMEANS con un 44.4% y solo un 11.1% en el rango de EXCELENTE confirmando una leve superioridad de KMEANS sobre DBSCAN, en ambos casos se obtuvo solo un 11.1% de BUENO en el rango lo cual indica que aún se encuentran algunos errores que deben ser corregidos posteriormente.

Con las pruebas realizadas y cada uno de los casos resueltos se puede observar como cada uno de los métodos de Clustering funcionan correctamente y que su brecha de diferencia es muy baja dejando un espacio para la mejoría a futuro en cada uno de estos métodos, con todo esto también se puede mencionar que el dataset con el que se alimentó es un dataset muy pequeño por lo que si se va alimentando la eficiencia sobre los resultados se hará cada vez mayor.

6. CONCLUSIONES Y RECOMENDACIONES

6.1 CONCLUSIONES

El manejo de herramientas como **stopwords** y **vectorizer** de la librería SKLEARN fueron de gran ayuda para poder categorizar y dar un valor más sólido a cada una de las palabras encontradas e ingresadas, primero porque puede alimentar el diccionario de **stopwords** con palabras clave las cuales fueron de gran ayuda en la separación de las que sirven y las que no sirven para luego dar pesos numéricos a cada una de estas palabras facilitando su uso en cada uno de los métodos antes estudiados.

A medida que se desarrollaron las pruebas del buscador de sentencias jurisprudenciales se logró identificar como los usuarios ingresaban diferentes cadenas de texto para realizar la búsqueda de un mismo caso, la herramienta respondió de manera eficiente ya que aunque se ingresaban diferentes palabras, el sistema arrojó respuestas similares en las búsquedas para llegar a la solución, esto se traduce en el manejo efectivo de la semántica que se da alrededor del derecho y en este aspecto se puede concluir que la implementación arrojó buenos resultados como se muestra en la Ilustración 9 donde el 67% de los usuarios encontraron respuesta del caso uno, el 100% el caso dos, 89% el caso 3 y el 78% en el caso 4 y 5 respectivamente, dando una efectividad de más del 75% en cada uno de los casos pasando un nivel más que aceptable para su uso.

Aunque los algoritmos de agrupamiento KMEANS y DBSCAN fueron eficientes hay ciertos factores que influyen en cada uno de ellos como lo son:

- ✓ Las características de la computadora sobre la cual se está ejecutando la herramienta de búsqueda ya que requiere de un buen procesador y una buena memoria RAM.
- ✓ El número de sentencias que hay en el dataset que se utiliza para realizar las búsquedas, claro efecto de cada uno de estos se ve reflejado en las primeras encuestas, para las dos primeras evaluaciones el dataset tenía alrededor de las 8000 sentencias dando resultados no tan precisos y en algunos casos no se encontraron sentencias que ayudaran al usuario a resolver el problema; este cambio de resultados se elevó de forma positiva

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

cuando se subió la cantidad a más de 32000 sentencias con lo cual las búsquedas se volvieron más efectivas.

- ✓ Debido a que el algoritmo DBSCAN trabaja basado en densidad, a mayor cantidad de sentencias presenta mejor precisión al arrojar los resultados de las búsquedas

Este algoritmo fue creado para comprender datos numéricos debido a su proceso basado en densidad cambiando de forma drástica cuando se trabaja con datos convertidos de texto a valores de vector, donde cada una de las palabras crea un patrón numérico según su uso y su contexto.

- ✓ En el caso de KMEANS se eligieron 8 clúster como una medida optima de grupos de sentencias ya que al elegir menor cantidad de clusters la cantidad de sentencias agrupadas eran de gran cantidad pero generaban resultados poco precisos al momento de buscar.

Al utilizar más de 8 clusters para el agrupamiento se encontraban muy pocos por grupo disminuyendo su precisión; por lo tanto se decidió utilizar un valor de $K=8$ la cual es una medida justa agrupando los mejores resultados.

Teniendo en cuenta las anteriores características que demarcan cada uno de los resultados se pueden observar los resultados de las preguntas Q1, Q2, Q3 y Q4 donde se evidencia que así sea por poco, hay una ventaja del KMEANS ante el DBSCAN como se evidencia en el las respuestas de las preguntas Q1 Tabla 7, donde las calificaciones de DBSCAN aunque estuvieron cerca no rebasaron las encontradas en las calificaciones obtenidas por el KMEANS en este caso la diferencia se hace visible en que DBSCAN obtiene el 11.1% de excelencia frente al 22.2% de KMEANS, mostrando una mayor satisfacción en la búsqueda alrededor de KMEANS por parte de los usuarios como se puede evidenciar en la Ilustración 10 dejando claro que cada uno de los aspectos que fueron anteriormente descritos fueron tratados ayudaron a mejorar cada vez más el buscador.

6.2 SUGERENCIAS PARA FUTURAS INVESTIGACIONES

- ✓ Continuar realizando investigaciones acerca de la semántica alrededor del ámbito de derecho y su manejo a nivel de segmentación creando así herramientas más sólidas que aumenten la eficiencia en la respectiva búsqueda y agrupamiento.
- ✓ Utilizar los algoritmos KMEANS y DBSCAN para realizar agrupamiento y búsquedas más precisas pero añadiendo en la similitud de las palabras sus sinónimos para encontrar resultados que también le sean de ayuda al usuario que utiliza la herramienta.
- ✓ Aumentar el nivel de búsqueda a otras instancias de justicia en Colombia como lo son la Corte Suprema de Justicia, Consejo de Estado, Consejo Superior de la Judicatura creando una herramienta robusta para el área del derecho.
- ✓ Aumentar la investigación sobre el método KMEANS teniendo en cuenta la eficiencia que mostró en resultados respecto al método DBSCAN.

7. BIBLIOGRAFÍA

- [1] M. V. Parra, “El precedente judicial en el derecho comparado,” vol. 4, no. Criterio Jurídico, 2011.
- [2] J. T. Valdez, *Derecho Informático*, Cuarta Edi. Mexico: McGRAW-HILL, 2008.
- [3] Colciencias, “Agenda estratégica de innovación para la justicia / Convocatoria para el Fortalecimiento de los nodos de innovación en TIC - Temática: Justicia en Institución de Estado,” 2014.
- [4] K. Jain, A. N. Murty, M, and J. Flynn, P, “Data clustering: a review,” *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.
- [5] I. F. Aguillo, “Herramientas avanzadas para la búsqueda de información médica en el web,” *Atención Primaria*, vol. 29, no. 4, pp. 246–252, 2013.
- [6] Universitat Pompeu Fabra., “Motores de Búsqueda de información científica y académica.” [Online]. Available: <https://www.upf.edu/hipertextnet/numero-5/motores-busqueda.html>. [Accessed: 25-Mar-2019].
- [7] E. P. Mor, “Diseño Centrado En El Usuario,” *Rev. Q*, p. 14, 2008.
- [8] Y. H. Montero, S. O. Santamar, and I. Apei, “Informe APEI sobre usabilidad,” 2009.
- [9] N. Zhang, Y. Pu, S. Yang, J. Zhou, and J. Gao, “An Ontological Chinese Legal Consultation System,” *IEEE Access*, vol. 5, pp. 18250–18261, 2017.
- [10] Y. Ma, P. Zhang, and J. Ma, “An Ontology Driven Knowledge Block Summarization Approach for Chinese Judgment Document Classification,” *IEEE Access*, vol. 6, pp. 71327–71338, 2018.
- [11] G. Boella, L. Di Caro, L. Humphreys, L. Robaldo, P. Rossi, and L. van der

IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

Torre, “Eunomos, a legal document and knowledge management system for the Web to provide relevant, reliable and up-to-date information on the law,” *Artif. Intell. Law*, vol. 24, no. 3, pp. 245–283, Sep. 2016.

- [12] G. Papastefanatos, “Towards Automatic Structuring and Semantic Indexing of Legal Documents,” in *in Proc. PCI, Patras, Greece*, 2016.
- [13] Y. Wang *et al.*, “Topic Model Based Text Similarity Measure for Chinese Judgment Document,” in *Data Science*, 2017, pp. 42–54.
- [14] S. Chou and T.-P. Hsing, “Text Mining Technique for Chinese Written Judgment of Criminal Case,” in *Intelligence and Security Informatics*, 2010, pp. 113–125.
- [15] S. B. Silveira and A. Branco, “Combining a double clustering approach with sentence simplification to produce highly informative multi-document summaries,” *Proc. 2012 IEEE 13th Int. Conf. Inf. Reuse Integr. IRI 2012*, no. 1, pp. 482–489, 2012.
- [16] L. F. da C. Nassif and E. R. Hruschka, “Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection,” vol. 8, no. 1, pp. 46–54, 2011.
- [17] X. Cai and W. Li, “Ranking through clustering: An integrated approach to multi-document summarization,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 7, pp. 1424–1433, 2013.
- [18] S. Jun, S. S. Park, and D. S. Jang, “Document clustering method using dimension reduction and support vector clustering to overcome sparseness,” *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3204–3212, 2014.
- [19] M. Habibi and A. Popescu-Belis, “Keyword Extraction and Clustering for Document Recommendation in Conversations,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 23, no. 4, pp. 746–759, 2015.
- [20] A. R. Patil and A. A. Manjrekar, “Text Documents Using Feature Extraction

and,” *Int. Conf. Comput. Anal. Secur. Trends*, pp. 371–376, 2016.

- [21] J. E. Judith and J. Jayakumari, “Distributed Document Clustering Analysis Based on a Hybrid Method,” no. February, pp. 131–142, 2017.
- [22] W. Surfhvvlqj *et al.*, “Document Clustering Using Ant Colony Algorithm,” *IEEE Commun.*, vol. 80, pp. 459–463, 2017.
- [23] H. K. Kim, H. Kim, and S. Cho, “Neurocomputing Bag-of-concepts : Comprehending document representation through clustering words in distributed representation,” *Neurocomputing*, vol. 266, pp. 336–352, 2017.
- [24] R. K. Roul, “An effective approach for semantic-based clustering and topic-based ranking of web documents,” *Int. J. Data Sci. Anal.*, 2018.
- [25] C. Herrera and M. Donoso, “Teleduc & Design Thinking: innovación y calidad que trascienden,” 2007.
- [26] E. Loper and S. Bird, “NLTK: The Natural Language Toolkit,” 2002.

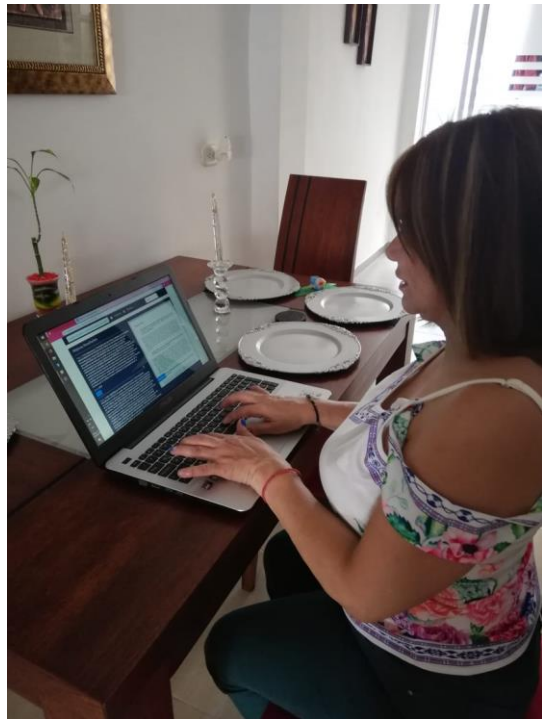
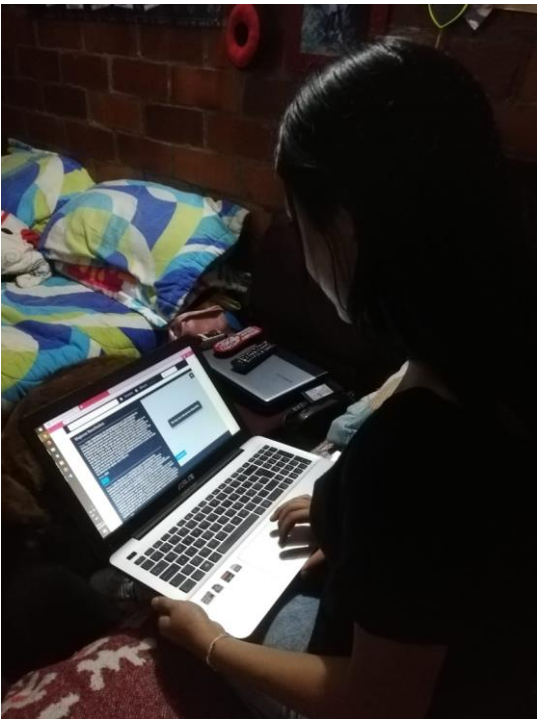
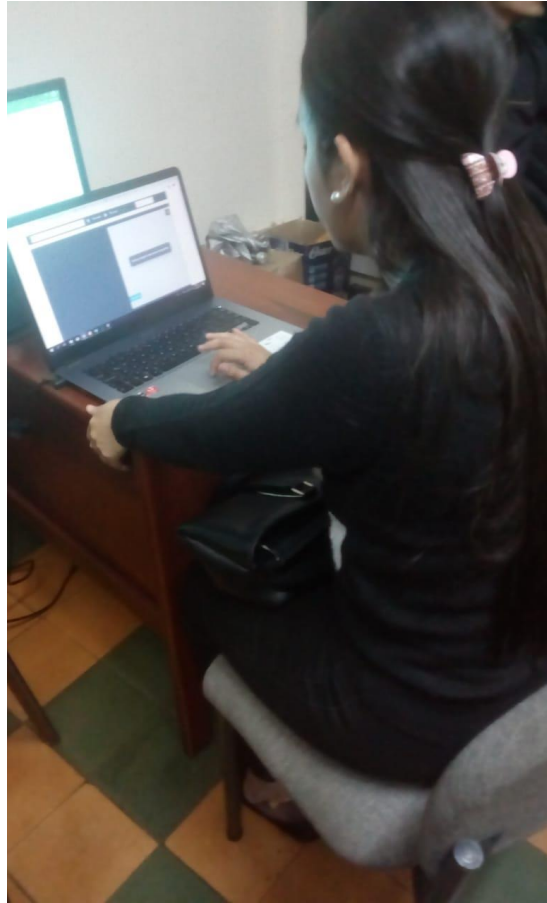
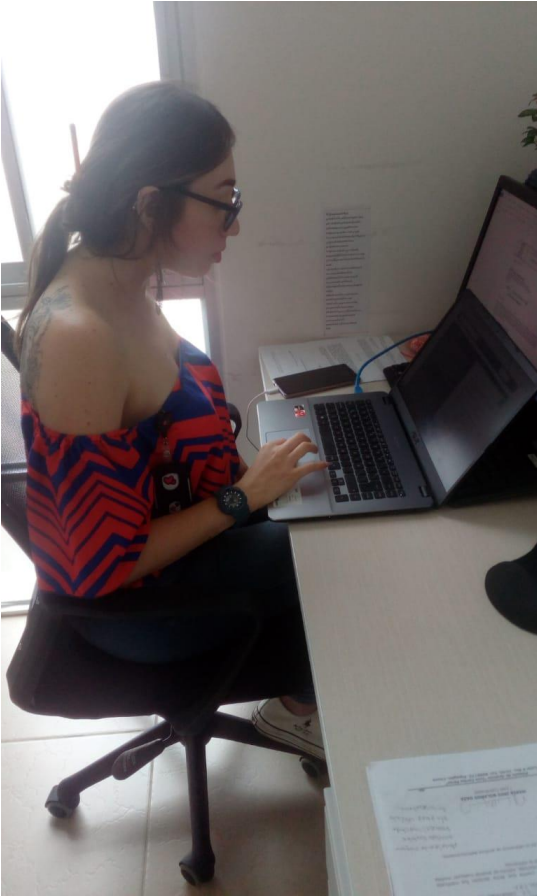
IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

ANEXOS

Evidencias de evaluadores utilizando la herramienta de búsqueda en la cual se implementaron los algoritmos de agrupamiento KMEANS y DBSCAN.



IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES



IMPLEMENTACIÓN Y EVALUACIÓN DE TÉCNICAS DE CLUSTERING PARA UN SISTEMA DE RECUPERACIÓN DE DOCUMENTOS JURISPRUDENCIALES

