UNIVERSIDAD TÉCNICA DE COTOPAXI

FACULTAD DE CIENCIAS DE LA INGENIERÍA Y APLICADAS

CARRERA DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS COMPUTACIONALES

PROPUESTA TECNOLÓGICA

"PROCEDIMIENTO ALGORÍTMICO BASADO EN TÉCNICAS DEL PROCESAMIENTO DEL LENGUAJE NATURAL PARA EL ANÁLISIS DEL CORPUS DE ARTÍCULOS CIENTÍFICOS DE LA PLATAFORMA ECUCIENCIA."

Proyecto de Titulación presentado previo a la obtención del Título de Ingenieros en Informática y Sistemas Computacionales.

Autores:

Chariguaman Morocho Gilson Ariel

Quilumbaquin Tutillo Nataly Lizeth

Tutor:

PhD. Gustavo Rodríguez Bárcenas

Latacunga-Ecuador

Septiembre - 2020

DECLARACIÓN DE AUTORÍA

Nosotros, Chariguaman Morocho Gilson Ariel y Quilumbaquin Tutillo Nataly Lizeth, declaramos ser autores de la presente propuesta tecnológica: "**PROCEDIMIENTO ALGORÍTMICO BASADO EN TÉCNICAS DEL PROCESAMIENTO DEL LENGUAJE NATURAL PARA EL ANÁLISIS DEL CORPUS DE ARTÍCULOS CIENTÍFICOS DE LA PLATAFORMA ECUCIENCIA."**, siendo el PhD. Gustavo Rodríguez Bárcenas tutor del presente trabajo; y eximo expresamente a la Universidad Técnica de Cotopaxi y a sus representantes legales de posibles reclamos o acciones legales.

Además, certifico que las ideas, conceptos, procedimientos y resultados vertidos en el presente trabajo investigativo, son de mi exclusiva responsabilidad.

……	..…….
Chariguaman Morocho Gilson Ariel	Quilumbaquin Tutillo Nataly Lizeth
C.I: 055036477-2	C.I:172467639-8

AVAL DEL TUTOR DEL PROYECTO DE TITULACIÓN

En calidad de Tutor del Trabajo de Investigación sobre el título:

"PROCEDIMIENTO ALGORÍTMICO BASADO EN TÉCNICAS DEL PROCESAMIENTO DEL LENGUAJE NATURAL PARA EL ANÁLISIS DEL CORPUS DE ARTÍCULOS CIENTÍFICOS DE LA PLATAFORMA ECUCIENCIA.", de los estudiantes: CHARIGUAMAN MOROCHO GILSON ARIEL con N° de cedula: 055036477-2, QUILUMBAQUIN TUTILLO NATALY LIZETH con N° de cedula 172467639-8, de la carrera de Ingeniería en Informática y Sistemas Computacionales, considero que dicho Informe Investigativo cumple con los requerimientos metodológicos y aportes científico-técnicos suficientes para ser sometidos a la evaluación del Tribunal de Validación de Proyecto que el Consejo Directivo de la FACULTAD DE CIENCIAS DE LA INGENIERÍA Y APLICADAS de la Universidad Técnica de Cotopaxi designe, para su correspondiente estudio y calificación.

Latacunga, 4 de septiembre del 2020

El Tutor

Firma

PhD. Gustavo Rodríguez Bárcenas

C.I: 175700135-7

APROBACIÓN DEL TRIBUNAL DE TITULACIÓN

En calidad de Tribunal de Lectores, aprueban el presente Informe de Investigación de acuerdo a las disposiciones reglamentarias emitidas por la Universidad Técnica de Cotopaxi, y por la FACULTAD de CIENCIAS DE LA INGENIERÍA Y APLICADAS; por cuanto, los postulantes: Chariguaman Morocho Gilson Ariel y Quilumbaquin Tutillo Nataly Lizeth con el título de Proyecto de titulación: **"PROCEDIMIENTO ALGORÍTMICO BASADO EN TÉCNICAS DEL PROCESAMIENTO DEL LENGUAJE NATURAL PARA EL ANÁLISIS DEL CORPUS DE ARTÍCULOS CIENTÍFICOS DE LA PLATAFORMA ECUCIENCIA"**, han considerado las recomendaciones emitidas oportunamente y reúne los méritos suficientes para ser sometido al acto de Sustentación de Proyecto.

Por lo antes expuesto, se autoriza realizar los empastados correspondientes, según la normativa institucional.

Latacunga, 17 de septiembre del 2020

Para constancia firman:

Lector 1 (Presidente)	**Lector 2**
Ing. Cantuña Flores Karla Susana	**Ing. Villa Quishpe Manuel William**
CC: 050230511-3	**CC:180338695-0**

Lector 3

Ing. Quinatoa Arequipa Edwin Edison

CC:180299840-9

AVAL DE IMPLEMENTACIÓN

PhD. Gustavo Rodríguez Bárcenas en forma legal, CERTIFICO que los señores: CHARIGUAMAN MOROCHO GILSON ARIEL con número de cédula 055036477-2, QUILUMBAQUIN TUTILLO NATALY LIZETH con número de cédula 1724667639-8, estudiantes de la Carrera de Ingeniería en Informática y Sistema Computacionales de la Facultad Ciencias de la Ingeniería y Aplicadas desarrollaron e implementaron la propuesta tecnológica cuyo título **"PROCEDIMIENTO ALGORÍTMICO BASADO EN TÉCNICAS DEL PROCESAMIENTO DEL LENGUAJE NATURAL PARA EL ANÁLISIS DEL CORPUS DE ARTÍCULOS CIENTÍFICOS DE LA PLATAFORMA ECUCIENCIA."**, de acuerdo a los requerimientos establecidos.

Es todo lo cuanto puedo certificar en honor a la verdad y autorizo a los peticionarios hacer uso del presente certificado de la manera ética que estimen conveniente.

Latacunga, 4 de septiembre del 2020

Atentamente:

PhD. Gustavo Rodríguez Bárcenas

C.I: 175700135-7

COORDINADOR DEL PROYECTO REDEC

AGRADECIMIENTO

Primeramente, agradezco a Dios por haberme guiado a lo largo de mi vida y mi carrera para la obtención de mi título profesional de tercer nivel, por ser fortaleza en los momentos de agotamiento y por brindarme una vida llena de experiencias y felicidad.

Le doy gracias a mis padres por el apoyo incondicional brindado a lo largo de mi vida, por ser los mejores, por haber estado conmigo apoyándome en los momentos difíciles, además de contar con excelentes consejos en mi caminar diario y por regalarme la oportunidad de estudiar una carrera universitaria.

Gracias Dr. Gustavo Rodríguez por creer en Lizeth y en mi persona, por habernos brindado la oportunidad de desarrollar nuestra tesis, por todo el apoyo y facilidades que nos brindó para poder cumplir nuestra meta. A los docentes que me han visto crecer como persona, y gracias a sus conocimientos hoy puedo sentirme dichoso y contento.

A la gloriosa Alma Mater Universidad Técnica de Cotopaxi por haberme abierto las puertas y permitirme concluir con una etapa de mi vida.

GILSON.

AGRADECIMIENTO

Como prioridad en mi vida, agradezco a mis padres Hernán y Rosa por darme la vida, su enorme esfuerzo que me ha permitido cumplir una meta más en mi camino, gracias por todos los valores que me han inculcado para lograr vencer las adversidades que se presenten.

A mi familia que han estado presentes con sus ánimos. A mis abuelitos Lorenzo y Transito que con su bendición me han llenado de alegría día tras día y por todos los consejos brindados. Asimismo, agradezco infinitamente a mis Hermanas que con sus palabras me hacían sentir orgullosa de lo que soy y de lo que les puedo enseñar.

A mis docentes y tutor que me han ayudado con sus ganas de transmitirme sus conocimientos a lo largo de la preparación de mi profesión.

De igual manera mis agradecimientos a la Universidad Técnica de Cotopaxi, por permitirme estudiar en la prestigiosa Alma Mater quienes con la enseñanza de sus valiosos conocimientos hicieron que pueda crecer día a día como profesional.

No puedo dejar de agradecerte a ti Gilson, mi compañero de Universidad, de tesis y ahora de corazón. Por haber tenido la paciencia necesaria y motivarme a seguir adelante en los momentos de desesperación y sobre todo por hacer de tu familia, una familia para mí.

LIZETH.

DEDICATORIA

Este trabajo de tesis principalmente decido a mis padres Hugo y Carmen, por el esmero y las metas alcanzadas, el cariño que invierten sus padres en sus hijos se ve reflejada de alguna manera. Gracias a mis padres soy quien soy, mi inspiración. Si bien se ha necesitado de energía y mucha dedicación, sino no hubiese sido posible si finalización sin la colaboración desinteresada de todas y cada una de las personas que me acompañaron en el camino de este trabajo y muchas de las cuales han sido un soporte muy fuerte en todo momento.

A dios por estar conmigo en cada paso que doy, por animar mi corazón e iluminar mi mente y por haber puesto en mi recorrido a aquellas personas que han sido mi sustento y compañía durante todo el periodo de estudio para logar mi título universitario.

Mil veces gracias.

GILSON.

DEDICATORIA

Durante el camino por mi vida universitaria me pude dar cuenta que tenía destrezas y habilidades, aunque lo importantes es que se puede obtener mejores resultados trabajando en compañía, en el desarrollo de esta tesis se presentaron muchos momentos en los cuales pareciera que los deberes no fueran a terminar, pero además entendí en ese instante de dificultad, que la ayuda siempre llega cuando la necesitas.

Por eso quiero dedicar mi tesis a quien han sido un apoyo fundamental, mis padres por las incontables veces que me brindaron su apoyo en cualquier decisión que haya tomado sean buenas o malas. Gracias por darme la libertad de desenvolverme como persona.

¡Que nadie se quede afuera, se los dedico a todos!

A todas las personas que me han apoyado para que el trabajo se realice con éxito en especial a los que nos abrieron las puertas y compartieron sus conocimientos.

Los llevo en el corazón.

LIZETH.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ECUACIONES

UNIVERSIDAD TÉCNICA DE COTOPAXI

FACULTAD CIENCIAS DE LA INGENIERÍA Y APLICADAS

TÍTULO: "PROCEDIMIENTO ALGORÍTMICO BASADO EN TÉCNICAS DEL PROCESAMIENTO DEL LENGUAJE NATURAL PARA EL ANÁLISIS DEL CORPUS DE ARTÍCULOS CIENTÍFICOS DE LA PLATAFORMA ECUCIENCIA."

Autores:

Chariguaman Morocho Gilson Ariel

Quilumbaquin Tutillo Nataly Lizeth

RESUMEN

Hoy en día analizar una cantidad excesiva de documentos en formato electrónico que se encuentran por la web es una tarea complicada y desgastante para cualquier persona, en la plataforma científica ECUCIENCIA al analizar un artículo científico se basa solamente en el título, resumen y palabras claves, existen documentos en formato pdf con mucha más información en el cuerpo del documento, en donde se puede visualizar datos con mayor exactitud ya que estamos viviendo en una era en donde la tecnología y el internet nos ha permitido generar y recopilar grandes volúmenes de información, para el estudio del proyecto se tuvo como objetivo el establecimiento de un procedimiento algorítmico mediante técnicas de procesamiento de lenguaje natural que permitió el análisis del corpus de artículos científicos de los docentes investigadores de la Universidad Técnica de Cotopaxi almacenados en la plataforma ECUCIENCIA; se tuvo dos fases para cumplir el desarrollo del proyecto, se utilizó la metodología KDD(Knowledge Discovery in Databases) para la primera etapa que conduce a la extracción de conocimiento el cual es el proceso metodológico para encontrar un modelo valido, útil y entendible que describa patrones de acuerdo a la información extraída, por otro lado para la segunda etapa se utilizó la metodología scrum el cual permitió una comunicación directa entre el cliente y el equipo de desarrollo teniendo así una mayor calidad del producto final y así el proyecto fue creciendo de iteración en iteración sin problemas y se logró unir la lógica adquirida de la primera etapa con el desarrollo de un módulo, donde se aplicaron librerías de Python que permitió realizar el análisis del corpus de los artículos científicos en formato pdf obteniendo de los mismos la riqueza léxica, frecuencia de palabras, palabras de parada, similitud y distancias de textos de los mismos que se representan mediante gráficos para los usuarios visualicen el contenido del análisis de datos sin dificultad.

PALABRAS CLAVES: EcuCiencia, KDD, Scrum, Python, Similitud, Distancia, Riqueza Léxica.

TECHNICAL UNIVERSITY OF COTOPAXI

FACULTY OF SCIENCE APPLIED ENGINEERING

TITLE: "ALGORITHMIC PROCEDURE BASED ON NATURAL LANGUAGE PROCESSING TECHNIQUES FOR THE ANALYSIS OF THE CORPUS OF SCIENTIFIC ARTICLES OF THE ECUCIENCIA PLATFORM".

Authors:

Chariguaman Morocho Gilson Ariel

Quilumbaquin Tutillo Nataly Lizeth

ABSTRAC

Today to analyze an excessive amount of documents in electronic format that are found on the web is a complicated and tiring task for any person, in the scientific platform ECUCIENCIA when analyzing a scientific article is based only on the title, summary and keywords, there are documents in pdf format with much more information in the body of the document, where it is possible to visualize data with greater accuracy since we are living in an era where technology and the Internet have allowed us to generate and collect large volumes of information. For the study of the project, the objective was to establish an algorithmic procedure through natural language processing techniques that allowed the analysis of the corpus of scientific articles of the research professors of the Technical University of Cotopaxi stored in the ECUCIENCIA platform; There were two phases to fulfill the development of the project, the methodology KDD (Knowledge Discovery in Databases) was used for the first phase that leads to the extraction of knowledge which is the methodological process to find a valid, useful and understandable model that describes patterns according to the extracted information, On the other hand, for the second stage, the scrum methodology was used, which allowed a direct communication between the client and the development team, thus having a higher quality of the final product. In this way, the project grew from iteration to iteration without problems and the logic acquired from the first stage was joined to the development of a module, where Python libraries were applied that allowed the analysis of the corpus of the scientific articles in pdf format obtaining from them the lexical richness, word frequency, stop words, similarity and distances of the texts that are represented by means of graphics for the users to visualize the content of the data analysis without difficulty.

KEY WORDS: EcuCiencia, KDD, Scrum, Python, Similarity, Distance, Lexical Wealth.

AVAL DE TRADUCCIÓN

Universidad Técnica de Cotopaxi

AVAL DE TRADUCCIÓN

En calidad de Docente del Idioma Inglés del Centro de Idiomas de la Universidad Técnica de Cotopaxi; en forma legal **CERTIFICO** que: La traducción del resumen de la propuesta tecnológica al Idioma Inglés presentado por los señores egresados de la Carrera de **INGENIERÍA EN INFORMÁTICA Y SISTEMAS COMPUTACIONALES** de la **FACULTAD CIENCIAS DE LA INGENIERÍA Y APLICADAS:** Chariguaman Morocho Gilson Ariel y Quilumbaquin Tutillo Nataly Lizeth , cuyo título versa **"PROCEDIMIENTO ALGORÍTMICO BASADO EN TÉCNICAS DEL PROCESAMIENTO DEL LENGUAJE NATURAL PARA EL ANÁLISIS DEL CORPUS DE ARTÍCULOS CIENTÍFICOS DE LA PLATAFORMA ECUCIENCIA"**, lo realizaron bajo mi supervisión y cumple con una correcta estructura gramatical del Idioma.

Es todo cuanto puedo certificar en honor a la verdad y autorizo a los peticionarios hacer uso del presente certificado de la manera ética que estimaren conveniente.

Latacunga, 14 de septiembre del 2020

Atentamente,

Msc. Vladimir Sandoval V.
DOCENTE CENTRO DE IDIOMAS
C.C. 0502104219

CENTRO DE IDIOMAS

1. INFORMACIÓN BÁSICA

1.1.PROPUESTA POR:

- Chariguaman Morocho Gilson Ariel
- Quilumbaquin Tutillo Nataly Lizeth

1.2.TEMA APROBADO:

Procedimiento algorítmico basado en técnicas del procesamiento del lenguaje natural para el análisis del corpus de artículos científicos de la plataforma EcuCiencia.

1.3.CARRERA:

Ingeniería en Informática Sistemas Computacionales

1.4.TUTOR DE PROYECTO DE TITULACIÓN:

PhD. Gustavo Rodríguez Bárcenas

1.5.EQUIPO DE TRABAJO:

Proyecto de investigación “Red de Estudios Cienciométricos REDEC”

1.6.LUGAR DE EJECUCIÓN:

Región Sierra, Provincia de Cotopaxi, Ciudad de Latacunga, Parroquia de San Felipe.

1.7.TIEMPO DE DURACIÓN DE PROYECTO:

6 meses

1.8.FECHA DE ENTREGA:

septiembre 2020

1.9.LÍNEA(S) Y SUBLÍNEAS DE INVESTIGACIÓN:

- **LÍNEA DE INVESTIGACIÓN:**
  Tecnologías de la Información y Comunicación (TIC’s)
- **SUBLÍNEA DE INVESTIGACIÓN DE LA CARRERA:**
  Robótica e Inteligencia Artificial.

2. ESTRUCTURA DE LA PROPUESTA

2.1.TITULO DEL PROYECTO

Procedimiento algorítmico basado en técnicas del procesamiento del lenguaje natural para el análisis del corpus de artículos científicos de la plataforma EcuCiencia

2.2.TIPO DE PROPUESTA TECNOLÓGICA:

Propuesta Tecnológica

2.3.ALCANCE

En la propuesta tecnológica a realizarse tiene como objetivo, "Establecer un procedimiento algorítmico mediante técnicas de procesamiento del lenguaje natural para el análisis del corpus de artículos científicos de los docentes investigadores de la Universidad Técnica de Cotopaxi almacenados en la plataforma EcuCiencia." Por lo que se espera que los docentes investigadores puedan analizar el corpus automáticamente de sus artículos científicos y puedan descargarlos al mismo tiempo de conocer las diferentes graficas que representan dicho corpus.

2.4.ÁREA DE CONOCIMIENTO

En conformidad a la Clasificación Internacional Normalizada de la Educación CINE – Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura UNESCO

Área: Ciencias

Sub- área: Informática.

2.5.SINOPSIS DE LA PROPUESTA TECNOLÓGICA

Hoy en día analizar una cantidad excesiva de documentos en formato electrónico que se encuentran por la web es una tarea complicada y desgastante para cualquier persona, en la plataforma científica ECUCIENCIA al analizar un artículo científico se basa solamente en el título, resumen y palabras claves, existen documentos en formato pdf con mucha más información en el cuerpo del documento, en donde se puede visualizar datos con mayor exactitud ya que estamos viviendo en una era en donde la tecnología y el internet nos ha permitido generar y recopilar grandes volúmenes de información, para el estudio del proyecto se tuvo como objetivo el establecimiento de un procedimiento algorítmico mediante técnicas de procesamiento de lenguaje natural que permitió el análisis del corpus de artículos científicos de los docentes investigadores de la Universidad Técnica de Cotopaxi almacenados en la

plataforma ECUCIENCIA; se tuvo dos fases para cumplir el desarrollo del proyecto, se utilizó la metodología KDD(Knowledge Discovery in Databases) para la primera etapa que conduce a la extracción de conocimiento el cual es el proceso metodológico para encontrar un modelo valido, útil y entendible que describa patrones de acuerdo a la información extraída, por otro lado para la segunda etapa se utilizó la metodología scrum al cual permitió una comunicación directa entre el cliente y el equipo de desarrollo teniendo así una mayor calidad del producto final y así el proyecto fue creciendo de iteración en iteración sin problemas y se logró unir la lógica adquirida de la primera etapa con el desarrollo de un módulo, donde se aplicaron librerías de Python que permitió realizar el análisis del corpus de los artículos científicos en formato pdf obteniendo de los mismos la riqueza léxica, frecuencia de palabras, palabras de parada, similitud y distancias de textos de los mismos que se representan mediante gráficos para los usuarios visualicen el contenido del análisis de datos sin dificultad.

3. DESCRIPCIÓN DE PROBLEMA

La clasificación de textos, en entornos en los que el volumen de datos a clasificar es tan elevado que resulta muy costosa la realización de esta tarea por parte de humanos, los documentos en lenguaje natural disponibles en formato electrónico hacen imposible su análisis, los sistemas de extracción de información permiten estructurar esa información para un dominio especifico, lo que convierte el problema de analizar una colección de documentos a consultar una base de datos especifica[1].

Analizar la cantidad excesiva de documentos en formato electrónico que se encuentran por la web es una tarea complicada y desgastante para cualquier persona, al no contar con un sistema de análisis y clasificación de documentos los grupos de investigadores optan por clasificar los textos de forma intuitiva, presentando un margen considerable de error en la relación de los datos ya que únicamente clasifican basándose en partes específicas y no en el documento completo.

El proyecto EcuCiencia de la Universidad Técnica De Cotopaxi al momento de analizar los artículos científicos que existen en el Sitio Web se basan solamente en el título, resumen y palabras claves, existen documentos pdf con mucha más información que se puede extraer como es el cuerpo del trabajo que contiene información que podría ayudar a ser más explícito sobre lo que se trata el artículo. Dado que la información no está completamente clasificada permite buscar, pero no permite clasificar de acuerdo a las áreas de conocimiento y no se sabe que artículos científicos están relacionado con otros para tener un correcto análisis de un corpus.

4. JUSTIFICACIÓN

El análisis de corpus de documentos surge a través de la necesidad de clasificar textos o separar documentos de un tema o área de conocimiento de conjunto de documentos que contienen diferentes artículos científicos conociendo su similaridad. Cuando se logra clasificar los textos o artículos científicos por temas, la búsqueda y el análisis del corpus de toda la información que poseen los documentos se realiza de una manera más rápida y sencilla. Realizar la clasificación de documentos de forma manual, provoca que la tarea sea complicada, entonces el corpus que se pretende crear será analizado a base de librerías NLTK con Python que lee los documentos y los somete al procesamiento de documentos de lenguaje natural, también conocidas como reducción de la dimensionalidad, la cual está compuesta por tres tareas, las cuales son tokenizar el texto, eliminación de stopwords y el enraizamiento de las palabras (Palabras de parada)[2], para lograr obtener documentos limpios en donde nos permitirá realizar un análisis del corpus completo. Una vez que estos procesos son realizados, se procede a realizar la representación vectorial de los documentos en base a las gráficas las cuales estarás aptas para realizar las pruebas correspondientes.

5. BENEFICIARIOS

5.1.DIRECTOS

Se considera como beneficiarios directos del presente proyecto a los docentes investigadores y estudiantes de la Universidad Técnica de Cotopaxi, porque son los que anexan los artículos científicos los cuales serán analizados.

5.2.INDIRECTOS

Se considera a los beneficiarios indirectos a todas aquellas personas que logren acceder de una manera u otra para que puedan nutrirse de la información de la plataforma científica EcuCiencia.

6. OBJETIVOS

6.1.Objetivo general

- Establecer un método de análisis de información en corpus de artículos científicos, mediante algoritmos de clasificación y librerías NLTK en la Plataforma Científica ECUCIENCIA que permita el reporte de métricas de los documentos en formato pdf que se encuentran en la base de datos del sistema.

6.2.Objetivos específicos

1. Establecer el estado del arte relacionado con métodos de análisis de información en corpus de documentos, a partir de fuentes bibliográficas certificadas científicamente que sirva de base teórica para la investigación.
2. Determinar los requerimientos algorítmicos necesarios para el análisis de la información contenidas en el corpus de los artículos científicos recogidos en la base de datos del sistema, que permita la obtención de métricas de los documentos en formato pdf.
3. Implementar los algoritmos para el análisis de la información a través de un módulo en la Plataforma Científica EcuCiencia que permita la validación de las métricas de los documentos en formato pdf recogidos en la base de datos del sistema.

7. ACTIVIDADES Y SISTEMA DE TAREAS EN RELACIÓN A LOS OBJETIVOS PLANTEADOS:

Tabla 1:Actividades y tareas de los objetivos

OBJETIVO	ACTIVIDAD	RESULTADOS ESPERADOS	DESCRIPCIÓN DE LAS ACTIVIDADES
OBJETIVO 1	-Buscar fuentes de información en base de datos -Identificar las terminologías asociadas al objeto de estudio y al campo de acción. -Investigación y recopilación de las fuentes bibliográficas confiables para lograr obtener información adecuada para armas la redacción.	-Fuentes bibliográficas primarias como libros, tesis así mismo como artículos. -Marco teórico y Marco metodológico.	-Recolección de información en los diferentes medios tanto físicos como digitales.
OBJETIVO 2	-Buscar metodologías para el procesamiento de lenguaje natural. -Selección de una metodología para el	-Modelo de algoritmo óptimo para el análisis de corpus de los documentos	-Utilización de la metodología KDD para el modelado del algoritmo de análisis de corpus y

	procesamiento de lenguaje natural. -Modelar el algoritmo de análisis de corpus	científicos de investigadores.	clasificación de documentos.
OBJETIVO 3	-Buscar metodologías de desarrollo ágil. -Selección de una metodología de desarrollo ágil. -Implementar el algoritmo de análisis de corpus.	-Implementación optima del algoritmo creado en la plataforma.	-Utilización de la metodología Scrum para la implementación del algoritmo de análisis de corpus y clasificación de documentos.

Fuente: Los investigadores

8. MARCO TEÓRICO

8.1.ANTECEDENTES

En la Universidad Técnica de Cotopaxi tener proyectos enfocados a la plataforma EcuCiencia puede ser una parte muy importante para el desarrollo de nuevos proyectos tecnológicos. Motivado por estas razones, se han realizado numerosos estudios realizados en la construcción de modelos predictivos en la educación para diversos fines. La presente tesis[3], se relaciona con la investigación en curso, ya que propone un algoritmo de aprendizaje de máquina para analizar un corpus, a través de enunciados claros, objetivos de aprendizaje precisos y una estructura de trabajo que aborda, paso a paso, así mismo mostrado claramente cuál es el principal problema del proyecto y se basa en la inmensa cantidad de información genera un aumento considerable de documentos científicos en formato digital, dando como resultado que no se pueda aplicar una clasificación automática. Teniendo como objetivo principal el desarrollar una plataforma clasificadora automática de textos, con la finalidad de almacenar documentos en formato digital y organizarlos según el área de conocimiento a la que pertenece, mediante la utilización de herramientas y técnicas Open Source[3]. Para ello se utiliza algoritmos de clasificación supervisados que es unos de los métodos más óptimos para el aprendizaje de máquina con lo que se propone realizar una plataforma científica que permita recolectar una cantidad determinada de información más relevante de todo el corpus almacenado y posteriormente implementar un algoritmo clasificador automático de textos que permite estructurar datos relevantes a un dominio específico (clase o categorías), siendo en este

caso los documentos científicos generados por los docentes investigadores de la Universidad Técnica de Cotopaxi para ayudar a solventar favorablemente esta problemática, utilizando una metodología ágil para realizar el trabajo más sencillo con posibilidades de corregir errores. Así mismo se obtiene como resultado clasificar documentos digitales de acuerdo a las áreas de conocimiento a la que pertenece para poder acceder de manera rápida y eficiente en menor tiempo.

El contexto del presente proyecto referenciado [4] muestra que, debido al gran aumento de artículos, libros, proyectos, ponencias entre otros documentos que se requieren almacenar, se pretende implementar una Plataforma científica denominada EcuCiencia, que tiene como objetivo principal la recopilación y visualización de la producción científica y tecnológica a partir de indicadores Cienciométricos. Para poder cumplir con lo que demanda el proyecto, el mismo que es divido por fases y para cumplirlas se desarrollaron métodos aplicando algoritmos de clasificación y minería de datos que realizan el análisis de un conjunto de datos extenso, para obtener como resultado matrices de similaridad y distancia de acuerdo al número de publicaciones de cada usuario pudiendo así comprobar la hipótesis planteada de que si se implementa el uso de algoritmos de clasificación, dentro del sistema EcuCiencia, se logrará representar y visualizar los valores de producción científica de los investigadores de la Universidad Técnica de Cotopaxi que están registrados dentro del mismo. Así mismo para hacer más fácil el desarrollo de los métodos algorítmicos se utilizó la metodología KDD (Knowledge Discovery in Databases) para el proyecto con minería de datos y para el desarrollo de la plataforma se utiliza el modelo iterativo incremental ya que hace el trabajo más fácil porque trabaja mediante ciclos de vida iterativos con etapas lo que da paso al avance de la propuesta tecnológica. Teniendo como resultado que mediante la implementación de los algoritmos de clasificación en la plataforma EcuCiencia se logra representar la simplicidad y distancia de los investigadores de acuerdo a su producción científica con gráficos que se muestra la información de cada uno de ellos.

Tomando como referencia diferentes fuentes se ha obtenido una tesis de la Universidad Técnica de Cotopaxi [5], el mismo que muestra el factor de impacto de las revistas es el índice bibliométrico más utilizado para evaluar y comparar la producción de los países, y ese es el principal problema de la Universidad Técnica de Cotopaxi es una Institución de Educación Superior que busca desarrollar ciencia y tecnología para fomentar el desarrollo de la comunidad universitaria. De modo que, debido al alto volumen de documentos científicos que genera se

vuelve conflictiva verificar el verdadero valor de la investigación que es realizada por los investigadores, así como además existe la necesidad de desarrollar un algoritmo de evaluación que permita visualizar el verdadero valor de la producción científica de forma adecuada. Proponiendo como objetivo principal desarrollar un algoritmo para la evaluación de documentos científicos e investigadores de la Universidad Técnica de Cotopaxi, a través del procesamiento de la información bibliométrica de los documentos de investigación de la plataforma Científica EcuCiencia (REDEC) y la gestión de los indicadores bibliométricos de evaluación. Se pretende desarrollar la presente propuesta tecnológica debido a que proporciona beneficios considerables para toda la comunidad Universitaria, tales como: gestión y planificación de los recursos destinados a la investigación, ya que permite conocer el verdadero rendimiento de la actividad científica, así como su impacto en la sociedad. Teniendo como resultado el análisis de las diferentes aproximaciones de evaluar la calidad de investigación que realiza cada investigador. Por otro lado, la producción científica que cuenta la plataforma científica "EcuCiencia", será organizada de manera adecuada para que mediante la aplicación de un algoritmo de evaluación refleje el nivel de la calidad de producción científica.

En este sentido, la revisión de la literatura permite analizar y reflexionar si la teoría y la investigación anterior sugiere una respuesta a la necesidad e hipótesis propuestas; o bien, si provee una orientación a seguir dentro del planteamiento del estudio en este caso se ha considerado [6], articulo obtenido de SciELO, muestra que la clasificación de textos es importante para el proyecto teniendo como objetivo establecer un proceso tecnológico de soporte, a la farmacovigilancia con base en los Sistemas de Reconocimiento de Patrones (SRP) para identificar la gravedad de la IF, antes y después de la comercialización de un medicamento y así, aproximarse a predecir nuevas interacciones farmacológicas ya que los medicamentos representan un riesgo para la salud de la población, por tal motivo, es imprescindible evaluar el riesgo/beneficio para cada uno de los mismos y para ello se desarrolló un corpus con 540 interacciones farmacológicas que pueden presentarse en los medicamentos más frecuentes del Hospital Universitario de Puebla. Proponen generar un modelo de clasificación en la plataforma Weka para el aprendizaje automático y la minería de datos utilizando un algoritmo Naïve Bayes que predice la posibilidad de una interacción farmacológica, clasificada en leve, moderada o grave[6]. Se realizaron las pruebas con el algoritmo Naïve Bayes utilizando el método de validación cruzada con 10 pliegues. Para obtener mejores resultados se comparó con el algoritmo Random Forest y así poder discutir sobre los resultados de cada uno de ellos. Los materiales que se utilizaron para la construcción del corpus se basaron en la utilización de

diccionarios médicos como iDoctus México, PLM México y Vademécum que tiene formato electrónico y contienen la información de las interacciones farmacológicas de México. El modelo se construyó con el aprendizaje supervisado cumpliendo las siguientes etapas: colección de datos, selección de características, selección del modelo matemático, entrenamiento, pruebas y complejidad computacional, cada uno de estas etapas nos ayudaran para que el modelo sea más exacto. Los resultados obtenidos beneficiarían a la farmacovigilancia para aproximarse a predecir nuevas interacciones farmacológicas antes de la comercialización de un medicamento.

El Modelo de Espacio Vectorial, siendo una aproximación válida a la clasificación de textos, presenta ciertos inconvenientes que se han intentado subsanar, estos intentos han ido encaminados a enriquecer con conocimiento externo la bolsa de palabras, añadiéndole nuevos elementos[1]. En los últimos años, por el tamaño y notoriedad que ha alcanzado, ese conocimiento extra se ha buscado en la Wikipedia. [7], la tesis doctoral obtenida del repositorio institucional de la universidad de Vigo; muestra que con la gran información que se encuentra almacenada en la red con diferentes fuentes y diversos idiomas es casi imposible conseguir información por lo que se requiere que la información esté organizada, clasificada o agrupada de una cierta manera que facilita a los usuarios el acceso a aquella información o documentos que son de su interés de una manera eficaz, eficiente, simple y rápida. Teniendo como objetivo final de la propuesta es la validación de la aplicabilidad y beneficios aportados por el uso de una representación de los documentos basada en conceptos que hace uso de conocimiento enciclopédico en particular de la Wikipedia a diferentes tareas de gestión de información digital multiidioma como la clasificación y clustering de los documentos y la recuperación de información[7]. Esta investigación se ha centrado en la clasificación de documentos modelada como un problema de aprendizaje supervisado, con un algoritmo de clasificación se entrena con un cierto número de ejemplos como: documentos cuya categoría es conocida y posteriormente, el algoritmo entrenado se aplica sobre otro conjunto de documentos cuya categoría es desconocida[7]. Se aplica el modelo de espacio vectorial (Support Vector Machines) para verificar la diversa cantidad de representaciones existentes, se centra principalmente en la clasificación automática de documentos modelada como un problema de aprendizaje máquina supervisado. Con una hipótesis para comprobar sobre la utilización de una representación de los documentos basada en conceptos de la Wikipedia (WikiBoC), obtenidos a través del anotador semántico de propósito general Wikipedia Miner, mejora el rendimiento de las propuestas actuales para la clasificación monolingüe y multilingüe de documentos de texto[7].

Para realizar la investigación se ha optado por seguir la metodología de investigación DSRM (Design Science Research Methodology), debido a las diferencias en las taxonomías de los repositorios integrados, la evaluación del rendimiento de la propuesta presentada fue llevada a cabo utilizando dos estrategias complementarias.

Se ha consultado de diversas fuentes de información en este caso se considera [8]**,** la fuente que se ha considerado para obtener la tesis es el repositorio Pontificia Universidad Católica De Valparaíso, muestra que este proyecto se concentrarán en la investigación y desarrollo de un clasificador de texto basado en agentes inteligentes, definiendo como objetivo realizar un análisis e implementación de una arquitectura utilizando agentes inteligentes para un clasificador de textos, buscando la mejora en el rendimiento del clasificador mediante su reentrenamiento, a través de la evaluación de su conocimiento y un flujo de mensajes. El procedimiento de la determinación mediante las pruebas de t-student, asentadas principalmente en el análisis estadístico del corpus del contenido actual en el agente clasificador contra un conjunto de textos a clasificar, con el fin de observar si existen anomalías suficientes como para promover un reentrenamiento con relación a un conjunto de textos[8]. En las pruebas, se pudo demostrar que la implementación del agente clasificador generó beneficios en el rendimiento. Para la evaluación del rendimiento de los algoritmos de clasificación de textos se han propuesto distintos métodos, los cuales coinciden en que se debe analizar la precisión, recall (recuerdo) y exactitud del clasificador[8]. Para poder realizar estos análisis, se deben determinar los posibles resultados de la clasificación de un documento; y éstos pueden ser representados en una matriz de confusión.

La revisión literaria tiene como objetivo ofrecer un acercamiento a los temas que centran la atención de los investigadores del área y detectar la existencia de algunas líneas de investigación comunes considerando a [9]**,** artículo obtenido de SciELO; el principal problema es la simplificación automática de textos en español con el fin de hacerlos más accesible a las personas con discapacidades cognitivas. El análisis de corpus de artículos originales y artículos simplificados manualmente se ha realizado para identificar y calificar relevantes operaciones que tienen que ser implementadas en el sistema de simplificación de textos. Con 3 partes para analizar la primera es un estudio corpus que analiza cuantitativamente y categoriza las operaciones de simplificación aplicadas por editores humanos con el objetivo de medir el impacto de estas operaciones en un automático sistema de simplificación de texto; la segunda es una comparación de textos originales y simplificados usando tres diferentes grupos de

características (etiquetas POS, sintácticas características y medidas de complejidad) con el objetivo de verificar si los textos originales y simplificados se puede separar automáticamente de acuerdo con estos características; y la tercera parte es la clasificación de dos oraciones experimentos para explorar si el objetivo oraciones para algunas de las operaciones encontradas en la primera parte podría seleccionarse automáticamente usando las funciones y hallazgos de la segunda parte. Se base en el análisis de un corpus de 37 pares de artículos de noticias originales en español (publicado en línea y obtener de la agencia de noticias Servimedia). Y para verificar los resultados todas las clasificaciones se realizaron en Weka con la configuración de validación cruzada 10 veces. Seleccionando primero el algoritmo de selección de atributos CfsSubsetEval implementado en Weka se utilizó para seleccionar un subconjunto de mejores características, después seleccionar toda la clasificación.

El desarrollo de las tecnologías, en la última década, ha dado un impulso notable al internet realizando un recorrido histórico sobre el desarrollo de las nuevas tecnologías para desarrollo de software con el propósito de crear corpus haciendo referencia en [10], articulo de la Universidad Nacional de La Plata, obtenido de Google académico el problema que abarca este proyecto es de extracción de información y minería de datos en textos no estructurados y se busca utilizar técnicas de aprendizaje automático (machine learning), para analizar texto en lenguaje natural y en base a ello determinar la existencia de texto que tengan el "mismo sentido", o bien oraciones/párrafos que estén semánticamente relacionadas entre sí. Generar un corpus etiquetado, para la obtención del corpus se realizó un módulo que permitió tomar un corpus que fue facilitado por National Institute of Standards and Technology para luego poder generarlo, tabularlo, ordenarlo y etiquetarlo. La utilidad de un nuevo corpus etiquetado es vital, ya que servirá como material de entrenamiento a algoritmos de aprendizajes supervisados implementados en el proyecto[10], y también servirá como material para su aplicación otras subáreas de la Inteligencia Artificial. Para poder construir este software se investigaron diversos fenómenos lingüísticos y se los clasificaron en base al tipo de fenómeno presente en un fragmento de texto[11]. En base a ello se identificaron y clasificaron en Fenómenos Léxicos, Morfológicos, Semánticos, y Sintácticos. Esta aplicación permite tomar datos de orígenes de datos estructurados y registrarlos en un único origen de datos estructurado, normalizado para facilitar la búsqueda y análisis de textos.

8.2.Estado del arte

Tenido en cuenta que se investigara el estado de arte del procesamiento de lenguaje natural hay que tomar en cuenta que tiene numerosas definiciones que se clasifican en distintas temáticas.

Una lista extensa de éstas se define en el número 56 de la Revista de Procesamiento de Lenguaje Natural (PLN)[12]:

Modelos lingüísticos, matemáticos y psicolingüísticos del lenguaje, refiriéndonos al PLN enfocado como un sistema de reglas. **Lingüística de corpus**, metodología en la que se estudia el lenguaje a través de ejemplos de textos reales producidos en el mundo real[13].

Desarrollo de recursos y herramientas lingüísticas, en aquellos trabajos que expanden las herramientas PLN disponibles para su futuro uso[13].

Gramáticas y formalismos para el análisis morfológico y sintáctico, refiriéndonos al PLN enfocado como un sistema de reglas lógicas aplicado al análisis morfosintáctico[13].

Semántica, pragmática y discurso. El estudio tradicional del lenguaje natural para cada idioma por parte de lingüistas nos permite abstraer las reglas para su uso y comprensión[13].

Lexicografía y terminología computacional, que busca una sistemática colección y explicación de todas las palabras (o más estrictamente, unidades léxicas) de un lenguaje[13].

Resolución de la ambigüedad léxica, siendo un ejemplo la determinación del sustantivo concreto dada una referencia indirecta[13].

Aprendizaje automático en PLN, es decir la aplicación de algoritmos de aprendizaje no supervisado para sistemas de esta índole[13].

Traducción automática, refiriéndonos a sistemas capaces de realizar la traducción de textos naturales a otro lenguaje natural con poca o nula supervisión humana[13].

Sistemas de búsqueda de respuestas, alternativas a los históricos sistemas expertos, en los que se pretende responder a preguntas del usuario sobre un ámbito reducido[13].

Resumen automático, por el cual obtenemos el resumen de un texto intentando con- servar la información relevante[13].

Sistemas de diálogo, sistemas que se comunican con el usuario utilizando lenguaje natural, de manera similar a mantener una conversación[13].

Análisis de sentimientos y opiniones, actualmente un área del PLN con gran popularidad y que puede considerarse como una de sus aplicaciones industriales de mayor interés[13].

Minería de texto, el procesamiento estadístico de grandes cantidades de texto con la intención de extraer información de alta calidad[13].

Implicación textual y paráfrasis, con la que podemos visualizar información textual de diversas maneras, así como conocer las enunciaciones que se derivan de ellas[13].

Para obtener un enfoque para nuestro trabajo se tomó en cuenta las diferentes áreas temáticas, además se tomando en cuenta variables como factibilidad, tiempo disponible y utilidad, entre otras.

8.2.1. Corpus

Es una colección de textos legibles por máquina que se han producido en un entorno comunicativo natural. Han sido muestreados para ser representativos y equilibrados con respecto a factores particulares; por ejemplo, por género artículos periodísticos, ficción literaria, discurso hablado, blogs y diarios, y documentos legales. Se dice que un corpus es "representativo de una variedad lingüística" si el contenido del corpus puede generalizarse a esa variedad [14].

- **Corpus lingüístico**

Un corpus lingüístico se define como “un conjunto de textos de un mismo origen” y que tiene por función recopilar un conjunto de documentos tales como ensayos, obras de teatro, transcripciones entre otros con el fin de reunir en una misma base de datos o programa el uso de un término de la lengua en un momento dado [15].

Tipos de corpus

A continuación, se describirán los distintos tipos de corpus utilizando como fuente la clasificación proporcionada en “***Diseño de corpus textuales y orales.*** *Joan Torruella y Joaquim Llisterri”*[16]:

Según el porcentaje y la distribución de los diferentes tipos de texto. - Los corpus pueden clasificarse según la distribución y el porcentaje escogido de los diferentes tipos de texto que lo componen. Según estos parámetros tenemos[16]:

1. ***Corpus grande*. -** Corpus que no se plantea el límite del volumen de textos que ha de recoger o que, si se lo plantea, lo cuantifica en un número de palabras muy elevado sin tener en cuenta cuestiones de equilibrio, de representatividad, etc.[17]. Esta característica es, en muchos casos, ambigua, ya que se habla de corpus grandes, pero sin precisar las

dimensiones en número de unidades léxicas que un corpus ha de tener para ser considerado como tal.

2. ***Corpus equilibrado*. -** Corpus que contiene diferentes variedades de textos distribuidos cuantitativamente en proporciones parecidas para cada variedad[17].
3. ***Corpus piramidal*. -** Corpus en que sus componentes, o sea sus textos, están distribuidos en diversos estratos o niveles: un primer estrato que recoge pocas variedades temáticas, pero con muchos textos en cada variedad; un segundo estrato que recoge mayor variedad de textos, pero menos cantidad en cada una de ellas; un tercer estrato compuesto por muchas variedades, pero con pocos textos en cada variedad; y así hasta un número de estratos opcional[17].
4. ***Corpus monitor*. -** Este tipo de corpus es consecuencia de la gran cantidad de palabras que últimamente están incluyendo los corpus. Las grandes dimensiones de los corpus hacen que sean difíciles de controlar y de explotar. Para evitarlo, los corpus monitor quieren tener un volumen textual constante, pero en continua actualización. El conjunto de textos que lo componen se va renovando cada cierto tiempo de manera que siempre se van incluyendo nuevos textos al mismo tiempo que se van excluyendo otros, consiguiendo de este modo un corpus vivo y dinámico como lo es la propia lengua[17].
5. ***Corpus paralelo*. -** Es una colección de textos traducidos a una o varias lenguas. El más sencillo es el que consta del original y su traducción a otra lengua. La dirección de la traducción no es necesario que sea constante, un corpus paralelo puede contener tanto textos traducidos de la lengua A la lengua B como textos traducidos de la lengua B a la lengua A. Este tipo de corpus es de gran utilidad sobre todo en el campo de la traducción, y principalmente de la traducción automática, ya que los programas suelen trabajar con datos probabilísticos que sólo pueden obtenerse a partir de los corpus[17].

Según la especificidad de los textos

Otra clasificación que se puede hacer de los corpus es en función de la especificidad de los textos que lo componen. Atendiendo a este parámetro podemos definir cuatro tipos[17]:

Según la cantidad de texto que se recoge de cada documento

Esta clasificación recoge todos los textos de los documentos que lo constituyen sea como corpus textual, corpus de referencia, corpus léxico[17].

Según la codificación y la anotación. -También se pueden clasificar los corpus atendiendo a las etiquetas descriptivas y analíticas que se han usado en la codificación de los textos. Según estos criterios los corpus serán[17]:

1. ***Corpus simple (o no codificado ni anotado)*. -** Corpus que ha sido guardado en formato neutro (ASCII, también llamado *plain text*), y sin codificación para ninguno de sus aspectos[17].
2. ***Corpus codificado o anotado-*** Corpus formado por textos a los cuales se les ha añadido, ya sea manual o automáticamente, etiquetas declarativas de algunos elementos estructurales de los documentos (indicación de tirulo, de principio de capitulo, de cambio de lengua, etc.) es importante que las etiquetas usadas para codificar y anotar los textos sean fáciles de reconocer y si es necesario eliminarlas fácilmente[17].

8.3.MINERÍA DE DATOS

La evolución digital ha hecho posible que la información digitalizada sea fácil de capturar, procesar, distribuir y transmitir, además de descubrir conocimiento de un volumen inmenso de datos es un reto así que se intentó buscarle sentido a la minería de datos, la tecnología de internet actual necesita el desarrollo de tecnologías de minería de datos más avanzadas para interpretar información del conocimiento de los datos que se distribuyen por el mundo[18].

Beneficios

- Habilidades para extraer información útil, la toma de decisiones y la exploración
- La compresión del gobernante en la fuente de datos

El análisis de datos ha sido tradicionalmente un proceso manual, pero con la ayuda de técnicas estadísticas se proporcionaba resúmenes y generaban informes. Sin embargo, tal enfoque cambio como consecuencia del crecimiento del volumen de datos [19]. Ahí es donde se encuentra la necesidad de buscar metodologías de análisis inteligente de datos, las cuales pueden descubrir conocimiento útil de datos KDD (Knowledge Discovery in Databases) que se refiere a todo el proceso de extracción de conocimiento a partir de base de datos y marca un cambio de paradigma haciendo importante el conocimiento útil que se pueda extraer.

De esta manera la minería de datos hace hincapié en[18]:

- La escalabilidad del número de atributos y de instancias
- Algoritmos y arquitecturas (proporcionando la estadística y el aprendizaje automático los fundamentos de los métodos y las formulaciones)
- La automatización para manejar grandes volúmenes de datos heterogéneos.

A continuación, se muestran las técnicas comunes de minería de datos[19]:

Clasificación. -Clasifica un dato dentro de una de las clases categóricas predefinidas

Regresión. -El propósito de este modelo es hacer corresponder un dato con un valor real de una variable[19].

Clustering. -Se refiere a la agrupación de registros, observaciones o casis en clases de objetos similares[19]. Un clúster es una colección de registros que son similares entre sí y distintos a los registros de otro clúster.

8.4.MACHINE LEARNING

Para construir un modelo sencillo pero que tenga un uso valioso teniendo un alta precisión predictiva se procesa un conjunto de datos, posee áreas de aplicación abundantes como: los bancos analizan sus datos pasados para construir modelos para usar aplicaciones de crédito, la detección de fraude[20]. El aprendizaje automático es parte de la inteligencia artificial, así como un problema de base de datos y nos ayuda a resolver muchos problemas haciendo que cada proceso sea automático.

1. **Aprendizaje supervisado**

Este tipo de aprendizaje es uno de los más óptimos debido a que se encargan de hacer predicciones por sí solos, si se utilizan suficientes datos de entradas este obtendrá mayor capacidad predictiva y lograra establecer una relación dando valores de respuesta significativos[3]. Cabe recalcar que para aplicar el aprendizaje supervisado los datos a utilizar son características o atributos que ayudan a definir a que clase o categoría pertenecen para esa manera guiar el proceso de aprendizaje haciendo que más eficiente su aplicación[3].

2. **Aprendizaje no supervisado**

En el aprendizaje no supervisado no existe ningún supervisor, solo se tiene datos de entrada. El objetivo es encontrar las regularidades en estos datos para ellos existe diferentes métodos para la estimación[20]. La diferencia de este tipo de aprendizaje es que consiste en agrupar un conjunto de datos, pero de igual manera su funcionamiento es efectiva.

8.5.Lenguaje natural

Se conoce como **lenguaje natural** al medio que utilizamos los seres humanos para comunicarnos y expresarnos. Es aquel que ha ido evolucionando a través del tiempo como puede ser el español, inglés, alemán o cualquier otro idioma o dialecto [21].

El **procesamiento del lenguaje natural** (o PLN) es un campo de la inteligencia artificial que hace uso de diferentes algoritmos y análisis estadísticos para aprender, entender y producir contenido en lenguaje humano. Su propósito es ayudar a la interacción entre humanos y ordenadores [22].

8.5.1. Beneficios del procesamiento de lenguaje natural

Existen millones de datos generados diariamente y con el procesamiento de lenguaje natural podemos generar datos instantáneamente en tiempo real, al igual que los motores de búsqueda, dan los resultados adecuados a las personas adecuadas en el momento adecuado[12].

8.5.2. Librerías del procesamiento de lenguaje natural

Hay muchas librerías de PLN de código abierto y estas son algunas de ellas[23]:

- Natural Language toolkit (NLTK).
- Apache OpenNLP.
- Stanford NLP suite.
- Gate NLP library.

8.5.3. NLTK

El **NLTK** es la librería líder para el procesamiento de lenguaje natural que proporciona interfaces fáciles de usar a más de cincuenta corpus y recursos léxicos, junto con un conjunto de bibliotecas de procesamiento de texto para la clasificación, tokenización, el etiquetado, el análisis y el razonamiento semántico [24], además es muy fácil de aprender y utilizar.

8.5.4. Algoritmo

Para implementar la solución de un problema mediante el uso de una computadora es necesario establecer una serie de pasos que permitan resolver el problema, a este conjunto de pasos se le denomina algoritmo, el cual debe tener como característica final la posibilidad de transcribirlo fácilmente a un lenguaje de programación[25].

8.5.5. Algoritmos de Clasificación

Estos algoritmos tratan de clasificar en diferentes categorías una serie de ejemplos o instancias que representan cierta información de un problema[4], en el ámbito del aprendizaje automático, el objetivo de estos sistemas es aprender a decidir cuál es la clase a la que pertenecen los ejemplos nuevos sin etiquetar. Existen dos tipos clasificación:

- Supervisada: En este tipo de clasificación, se tiene un conjunto de datos de los cuales ya sabemos su clasificación, llamados instancias de entrenamiento o conjunto de entrenamiento[26].
- No supervisada: los datos no tienen etiquetas (o no queremos utilizarlas) y estos se clasifican a partir de su estructura interna (propiedades, características)[26].

Clasificación automática de textos Es conocida como categorización de texto o ubicación del tema. Por consiguiente, se puede decir que la clasificación de textos da sus inicios debido al elevado número de documentos en formato digital y por ende resulta tedioso clasificarlo porque involucra tiempo, costo y otros factores que provocan esta problemática[27].

8.5.6. Python

Es un lenguaje orientado a objetos que cuya versatilidad nos permite utilizarlo aplicando diferentes paradigmas de programación[28], lo interesante de Python es que su sencillez nos permite aprender a programar y aprender las bases de un paradigma de mayor complejidad como es la programación orientada a objetos[28].

8.5.7. Sklearn

Es un módulo de Python que integra algoritmos clásicos de aprendizaje automático en el mundo muy unido de los paquetes científicos de Python (numpy, scipy, matplotlib). Su objetivo es proporcionar soluciones simples y eficientes a los problemas de aprendizaje que sean accesibles para todos y reutilizables en diversos contextos: el aprendizaje automático como una herramienta versátil para la ciencia y la ingeniería.[29].

- **Pandas**

Paquete de Python que proporciona estructuras de datos rápidas, flexibles y expresivas diseñadas para trabajar con datos que son fáciles e intuitivos, además tiene como objetivo analizar los datos prácticos y reales de Python, convertirse en una herramienta flexible de análisis, manipulación de datos de código abierto, teniendo características principales de esta librería[30].

- ✓ Alineación automática y explicita de datos: los objetivos pueden alinearse a un conjunto de etiquetas.
- ✓ Realiza operaciones de combinación de aplicación dividida en conjunto de datos de forma potente y flexible.

- ✓ Facilita la conversión de datos irregulares, indexados de manera diferente en otras estructuras de datos de Python[30].

- **Numpy**

Es un módulo muy fundamental para el cálculo científico con Python. Con él se dispone herramientas computacionales para mejorar estructuras con una gran cantidad de datos, diseñados para obtener un buen nivel de rendimiento en su manejo, además posee gran cantidad de métodos que permiten manipular los elementos de array de forma no secuencial[31].

- **TfidfVectorizer**

Convierte una colección de documentos en bruto en una matriz de características TF-IDF.[32]

- **Scipy**.

Es un ecosistema de software de código abierto basado en Python para matemáticas, ciencias e ingeniería[33].

Regresión lineal. – es un método estadístico el cual consiste en crear un modelo de regresión que se pueda analizar con dos variables la relación lineal existente. A la variable dependiente o respuesta se le identifica como Yy y a la variable predictora o independiente como Xx[34].

Error cuadrático. - es el criterio de evaluación más usado para problemas de regresión. Cuando ocupamos aprendizaje automático es en donde tiene relevancia y se usa. Para cada dato histórico podremos indicar el resultado correcto[35].

8.5.8. Herramientas de PLN

Se considero un número de herramientas para el Procesamiento del Lenguaje Natural, las cuales fueron evaluadas según las ventajas y desventajas que proporcionarían al desarrollo del proyecto teniendo en cuenta los objetivos que se deseaban alcanzar[13].

- **Herramienta de PLN elegida**

Una razón importante para la elección de NLTK como herramienta es el gran soporte que tiene, debido a las dimensiones de su comunidad de usuarios. Es una de las herramientas de procesamiento de lenguaje y natural de mayor aceptación en el ámbito científico. Otra razón importante que proporciona el libro Natural Languaje Processing with Python[36]. Con esta información podemos hacer uso de las funcionalidades que contiene la biblioteca de PLN.

- **Preprocesamiento de datos del corpus**

Conocer el preprocesamiento de datos para generar el corpus es muy importante al momento de analizar textos para poder mejorar el rendimiento de cualquier proceso que se realice por lo tanto se necesita limpiar elementos que sean innecesarios para un análisis completo. A continuación, mostramos algunas funciones que se utilizaran.

Stop Words se definen como términos que se consideran irrelevantes para la generación de un corpus limpio de un documento porque no presentar un contenido que sirva como datos repetidos en el texto.

Normalización de frecuencias en el momento en que se realice el proceso de frecuencia de palabras para textos largos, pueden repetirse un número de veces las palabras más relevantes por lo que se ahorra espacio.

Riqueza léxica es el estudio del léxico más conocido como calidad de la escritura, como es sabido, cuenta con diferentes vías de aproximación; las cualitativas y las cuantitativas que permiten conocer la amplitud de vocabulario[37].

Brevemente se mostrarán todas las herramientas que se consideró para lograr realizar el procesamiento de lenguaje natural.

8.6.MÉTRICAS DE SIMILITUD Y DISTANCIA ENTRE DOCUMENTOS

La similitud de los documentos se representa de acuerdo a una cantidad de términos en un documento que se compara con otro, entonces es importante mencionar la semejanza que contiene cada documento ya que existe una enorme cantidad de datos de los cuales se tomaran los que mejor resalten[3]. Se plantea vectores los que se utilizarán para la representación de los documentos. A continuación, se muestran las métricas que se utilizaron para realizar la distancia ente documentos.

8.7.Métricas de distancias

Teniendo en cuenta que son innumerables las maneras que se pueden calcular una distancia se ocupa las distancias que tienen mayor frecuencia las cuales son:

Las funciones de distancia entre dos vectores numéricos u y v. Calcular distancias en una gran colección de vectores es ineficiente para estas funciones.

8.7.1. Distancia Euclidiana

Es la raíz cuadrada de la suma de las diferencias al cuadrado entre los valores de dos casos para cada variable. [38]

8.7.2. Correlación

Se aplica a variables continuas, y usa correlaciones (Pearson, Spearman o Kendall). También se emplea en métodos para jerarquizar variables.

Ecuación 1: Distancia de Correlación

$$1 - \frac{(u - \bar{u}).(v - \bar{v})}{||(u - \bar{u})||_2||(v - \bar{v})||_2}$$

Fuente: [38]

8.7.3. Distancia Chebyshev

Calcula la distancia de Chebyshev entre los puntos[39]. La distancia de Chebyshev entre dos n-vectores u y v es la distancia máxima de la norma-1 entre sus respectivos elementos. Más precisamente, la distancia viene dada por:

Ecuación 2:Distancia Chebychev

$$d(u, v) = \frac{max}{i}|u_i - v_i|$$

Fuente: [38]

8.7.4. Distancia Minskowski

Esta distancia puede considerarse una generalización de las distancias euclideas y Manhattan[40]. Viene definida por la siguiente expresión:

Ecuación 3:Distancia Minskowski

$$dist = (\sum_{k=1}^{n} |p_{k-q_k}|^r)^{\frac{1}{r}}$$

Fuente: [38]

r = 1. Distancia Manhattan Ejemplo típico: Distancia de Hamming: Numero de bits diferentes entre dos arreglos de bits

r = 2. Distancia Euclidiana r → ∞. Distancia “supremo” (norma Lmax o L∞). La máxima diferencia entre los atributos

8.7.5. Distancia Coseno

Una medida de similitud entre dos vectores, mediante la medición del coseno del ángulo entre ellos. El resultante de la función coseno es igual a 1 cuando el ángulo es 0 y es inferior a 1 cuando el ángulo es de cualquier valor[41]. Calcule la distancia del coseno entre matrices 1-D. Se emplea frecuentemente en la búsqueda y recuperación de información representando las palabras (o documento) en un espacio vectorial.1 En minería de textos se aplica la similitud coseno con el objeto de establecer una métrica de semejanza entre textos y da lugar a la siguiente expresión:

Ecuación 4: Distancia Coseno

$$\mathbf{1} - \frac{\boldsymbol{u}.\boldsymbol{v}}{||\boldsymbol{u}||_2 \left|\left|\boldsymbol{v}\right|\right|_2}.$$

Fuente: [38]

La distancia funciona entre dos vectores booleanos (que representan conjuntos) u y v. Como en el caso de los vectores numéricos.

8.7.6. Coeficiente de Jaccard

Coeficiente de Jaccard tiende a ser una razón de similaridad la cual para calcularla se utilizan variables binarias. Calcula la distancia Jaccard entre los puntos. Dados dos vectores, u y v, la distancia Jaccard es la proporción de aquellos elementos u [i] y v [i] que no están de acuerdo.[38]

8.7.7. Índice Dice

Es un estadístico utilizado para comparar la similitud de dos muestras, en este caso se calcula la diferencia de dados entre dos matrices booleanas 1-D con la siguiente expresión:

Ecuación 5:Indice de Dice

$$\frac{C_{TF+}C_{FT}}{2C_{TT+}C_{FT+}C_{TF}}$$

Fuente: [38]

8.7.8. NLTK, Natural Language Toolkit

El NLTK (Natural Language Toolkit) es una biblioteca de Procesamiento de Lenguaje Natural que utiliza el lenguaje de programación Python [36]. Siendo así que es de libre uso para que

estudiantes y personal de distintas ramas académicas ejecuten estudios sin que se preocupen por el estado económico, además de tener la facilidad de reutilizar el código para logar extender requerimientos de cualquier sistema a desarrollarse ya que es de código abierto. El hecho de estar implementada como una biblioteca Python reduce la curva de aprendizaje, y la acerca al mundo académico, cuya mayor parte de integrantes se encuentra familiarizado con este lenguaje de programación[13].

8.8.APLICACIONES WEB

Hoy en día el crecimiento de la tecnología contiene importantes aspectos que hacen que se construyan actividades para que se vaya desarrollando en un orden correcto. En si las aplicaciones web permiten la generación de manera automática algún tipo de contenido o alguna transacción, del mismo modo las herramientas se encuadran dentro de las arquitectura cliente servido que quiere decir que un ordenador solicita un servicio al cliente[3].

8.8.1. Características de una aplicación web

Una aplicación web tiene diferentes características las cuales serán mencionadas seguidamente[3]:

- La portabilidad de la aplicación es dinámica, pudiendo de esta manera ejecutarse en cualquier plataforma, hablamos de dispositivos móviles, computadoras que alojen cualquier sistema operativo e inclusive consolas de videojuegos[3].
- No se necesita instalar la aplicación en el lado del cliente, este accede simplemente a través del navegador web.
- El cliente y el servidor pueden representarse como una sola entidad y también entidades separadas, realizando actividades o tareas independientes.
- Al implementar una aplicación web, no se requieren de sofisticados equipos, lo que aplica una reducción de costos a nivel de infraestructura.
- Para implementar una aplicación web debemos conocer que los recursos de los equipos electrónicos no se ocupan más bien el server es el que aloja todos los procesos.

8.8.2. Lenguaje de programación seleccionado

El código de las herramientas de Procesamiento de Lenguaje Natural desarrolladas en JavaScript y Python es compilado a un código máquina de bajo nivel, lo que le confiere una mayor rapidez de cómputo[3]. Tiene una capacidad alta de procesamiento para realizar comparaciones con las herramientas que Python posee para reducir los procesos y

aprovecharlos al máximo y se decidió utilizar Python para complementar la elección de herramienta de PLN que se hizo, en este caso la biblioteca NLTK para Python, como se mencionó anteriormente [12], [42].

8.8.3. Framework Django

Django es un marco web Python de alto nivel que fomenta un desarrollo rápido con un diseño limpio que fue creado por desarrolladores experimentados, es un framework gratuito y de código abierto que nos puede solucionar problemas rápidamente sin necesidad de reinventar todo el código obteniendo eficientes resultados [43].

Características

- **Ridículamente rápido,** Django fue diseñado para ayudar a los desarrolladores a llevar las aplicaciones desde el concepto hasta su finalización lo más rápido posible.[44]
- **Tranquilizadoramente seguro**, Django se toma la seguridad en serio y ayuda a los desarrolladores a evitar muchos errores de seguridad comunes[44].
- **Extremadamente escalable,** algunos de los sitios más ocupados de la web aprovechan la capacidad de Django para escalar de forma rápida y flexible[44].

8.8.4. Modelo Vista Controlador (MVC)

MVC es un estilo de arquitectura de software que separa los datos de una aplicación así como: la interfaz de usuario, la lógica de control en tres componentes distintos[45], es un modelo muy maduro que demuestra su valides en todo tipo de aplicaciones, de lenguajes y plataformas de desarrollo.

1. **Modelo**

Contiene una representación de los datos que manejan el sistema, su lógica de negocio, siendo responsable de[45]:

- Acceder a la capa de almacenamiento de datos. Teniendo un sistema de almacenamiento debería ser independiente.
- Define las reglas de negocio (la funcionalidad del sistema).
- Lista los registros de los controladores y las vistas del sistema.
- El modelo es en donde se notificará si existen cambios por cualquier agente externo.

2. **Vista**

Es la interfaz de usuario que compone la información que se envía al cliente y los mecanismos interacción con este[45].

- Recibe datos del modelo y la muestra al usuario.
- Administran los registros de todos los controladores que están asociados.
- Pueden dar el servicio de actualización ara que sea invocado por el controlador o modelo.

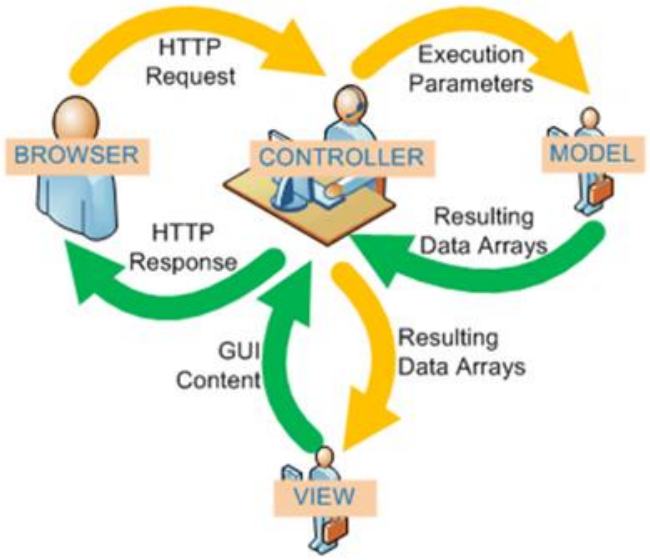
3. **Controlador**

Actúa como intermediario entre el modelo y la vista, gestionando el flujo de la información ente ellos y adaptar los datos a las necesidades de cada uno[45].

- Recibe los eventos de entrada
- Contiene reglas de gestión de eventos.

El flujo que sigue generalmente modelo vista controlador va de la siguiente manera:

Figura 1: Funcionamiento del Modelo Vista Controlador



Fuente: [45]

8.8.5. PyCharm

Es un IDE dedicado de Python y Django que proporciona una amplia gama de herramientas esenciales para los desarrolladores de Python, estrechamente integrados para crear un entorno conveniente para el desarrollo productivo de Python y el desarrollo web.[46]

8.8.6. Base de datos

Una base de datos es un almacén de información, lugar donde se pueden guardar una extensa cantidad de datos de una manera organizada[3]. PostgreSQL es el que ocuparemos en el proyecto.

8.8.7. PostgreSQL

Es una herramienta que sin lugar a duda proporciona resultados positivos en cuanto a los sistemas de gestores de base de datos es libre y de gran alcance[3].

PostgreSQL se ha ganado una sólida reputación por su arquitectura probada, confiabilidad, integridad de datos, conjunto de características sólidas y la dedicación de la comunidad de código abierto detrás del software para ofrecer soluciones de alto rendimiento, además que se ejecuta todos los sistemas operáticos principales[47]. A continuación, conoceremos características que posee PostgreSQL.

Tabla 2: Características de PostgreSQL

Característica	Descripción
Tipos de datos	▪ Primitivas: entero, numérico, cadena, booleano ▪ Estructurado: fecha / hora, matriz, rango, UUID ▪ Documento: JSON / JSONB, XML, valor-clave (Hstore) ▪ Geometría: Punto, Línea, Círculo, Polígono
Integridad de los datos	▪ ÚNICO, NO NULO ▪ Llaves primarias ▪ Llaves extranjeras ▪ Restricciones de exclusión ▪ Cerraduras explícitas, cerraduras consultivas
Concurrencia, rendimiento	▪ Indexación: B-tree, Multicolumn, Expresiones, Parcial ▪ Indexación avanzada: Índices de cobertura, filtros Bloom ▪ Partición de tablas ▪ Recopilación de Just-in-time (JIT)

Confiabilidad, Recuperación de Desastres	▪ Registro de escritura anticipada (WAL) ▪ Replicación: asíncrona, síncrona, lógica. ▪ Espacios de tabla
Seguridad	▪ Autenticación ▪ Sistema robusto de control de acceso ▪ Seguridad de columnas y filas
Extensibilidad	▪ Funciones y procedimientos almacenados. ▪ Lenguajes de procedimiento: PL / PGSQL, Perl, Python, etc. ▪ Contenedores de datos externos: conéctese a otras bases de datos o flujos con una interfaz SQL estándar.
Internacionalización, búsqueda de texto	▪ Soporte para conjuntos de caracteres internacionales. ▪ Búsqueda de texto completo

Fuente:[47],[4].

9. PREGUNTAS CIENTÍFICAS O HIPÓTESIS

FORMULACIÓN DEL PROBLEMA

¿Cómo aportar en la Plataforma Científica EcuCiencia un método capaz de analizar la información en corpus de artículos científicos, donde existen deficiencias en el reporte de métricas y desconocimiento de las relaciones y patrones de comportamiento entre los documentos que se encuentran en la base de datos del sistema?

HIPÓTESIS

Si se establece un procedimiento algorítmico óptimo y eficiente aplicando técnicas de procesamiento del lenguaje natural, se podrá analizar el corpus de artículos científicos de los docentes investigadores de la Universidad técnica de Cotopaxi almacenados en la plataforma EcuCiencia

10. METODOLOGÍA

La esencia del marco metodológico es llevar un lenguaje claro y sencillo con diferentes métodos, técnicas y estrategias con distintos procedimientos e instrumentos que se utilizan por

los investigadores para que se cumplan los objetivos planteados y a su ves justificar utilizando las referencias bibliográficas.

Haciendo referencia a Balestrini dice que metodología es el conjunto de procedimientos lógicos, tecno operacionales implícitos en todo proceso de investigación, con el objeto de ponerlos de manifiesto y sistematizarlos; a propósito de permitir descubrir y analizar los supuestos del estudio y de reconstruir los datos, a partir de los conceptos teóricos convencionalmente operacionalizados.[48]

10.1. Tipos de Investigación

- **Investigación Bibliográfica**

Una investigación bibliográfica o documental es aquella que utiliza textos o grabados como fuentes primarias para obtener sus datos y no se trata solamente de una recopilación de datos contenidos en libros, sino que se centra en la reflexión innovadora y crítica sobre determinados textos y los conceptos planteados en ellos. [48]

- **Investigación de Campo**

A diferencia de la investigación bibliográfica cuya fuente es la biblioteca o fuentes científicas, la investigación de campo exige salir a recabar los datos, sus fuentes pueden ser la naturaleza o la sociedad, pero en ambos casos el investigador debe ir a buscar su fuente para obtener la información[49]. La investigación de campo nos mostrara la situación actual de la base de datos del sistema EcuCiencia implementado en la Universidad Técnica de Cotopaxi en base a los artículos científicos y de esta forma verificar las funcionalidades del algoritmo.

10.2. Diseño de la investigación

Metodología descriptiva y experimental

La metodología descriptiva se utilizara en el presente trabajo porque esta de acorde a las necesidades de la investigación, en donde se describirán las actividades que se realizan al momento de probar el algoritmo para la creación del corpus de los documentos, realizando una observación directa en la plataforma EcuCiencia; esto permitirá conocer el nivel de problema que tiene el fenómeno de estudio, en este casi a que motivos se deben a que los documentos no están clasificados correctamente por sus áreas de conocimiento y similitud que pertenecen.

Y de manera experimental se enfoca en la implementación del módulo procesamiento de datos en la plataforma EcuCiencia, pudiendo realizar pruebas con los artículos científicos de la misma para poder probar la hipótesis.

10.3. Método de investigación

Método empírico

El método se enuncia porque se basa en la experimentación y la lógica empírica, tendrá la comprobación con hechos y su aporte al proceso de investigación es resultado fundamentalmente de la experiencia.

10.4. Enfoque de la investigación

La investigación cualitativa y cuantitativa fueron tomadas en cuenta para nuestro proyecto, ya que para las fases iniciales de los proyectos es ideal la investigación cualitativa mientras que la investigación cuantitativa es muy recomendable para la última parte del proyecto. Se procedió a realizar un reconocimiento de algunos trabajos que fueron significativos en la presente investigación en donde por medio de un algoritmo vamos a poder analizar el corpus de los documentos científicos alojados en la plataforma EcuCiencia siento esta la primera fase que se cumplirá. El desarrollo de software será la segunda fase sigue siendo una de las actividades que más grado de complejidad a la hora de plasmar una idea. Es por eso que existen muchos tipos de metodologías para desarrollar una aplicación, estas van desde las de desarrollo ágil que se usan en la construcción de un sistema.

10.5. Técnicas e instrumentos de investigación

Entrevista

Es un dialogo intencional, una conversación personal que el entrevistador establece con el sujeto investigado, con el propósito de obtener información[50].

Frecuentemente se utiliza la entrevista para realizar un reportaje en los medios de comunicación, programas de opiniones ya que nos permite estar enfocados a una técnica sencilla de recopilación de datos.

Formulario de la entrevista. - Este instrumento permite estableces preguntas partiendo de la hipótesis planteada para obtener respuestas y con ello estableces requerimientos funcionales que nos servirán para el desarrollo de la propuesta tecnológica.

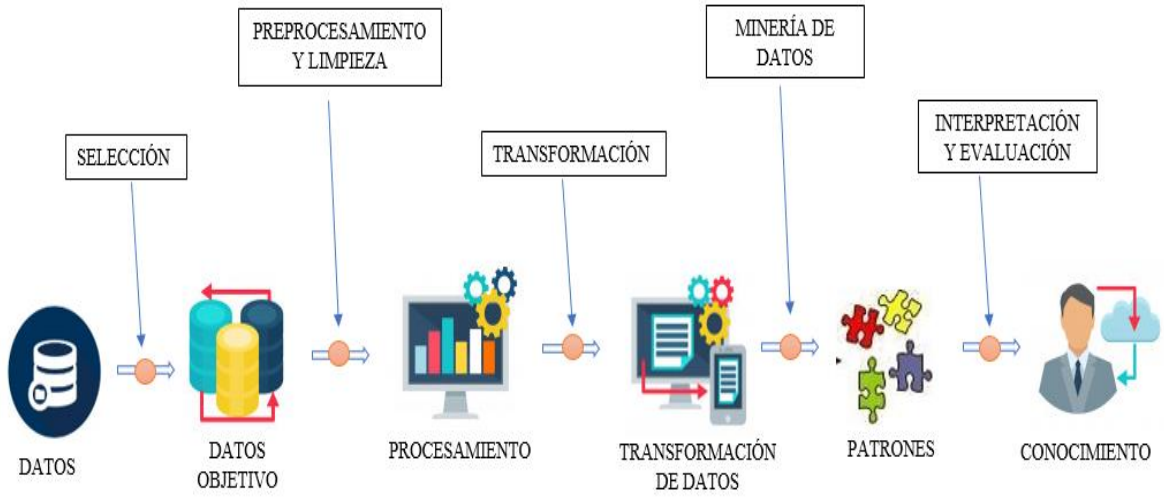
10.6. POBLACIÓN

En los estudios anteriormente realizados se da a conocer la población que comprende el proyecto propuesto para la plataforma EcuCiencia constando del total de docentes que pertenecen a la Universidad Técnica de Cotopaxi [3],[5]. Justificando que nuestro proyecto está orientado en la primera etapa a realizar la modelación de los algoritmos para la creación del corpus de documentos y con ello vamos a experimentar como trabaja y de qué manera se va a procesar el lenguaje natural. Para ellos nos basamos en la base de datos con la cual está trabajando el Sistema EcuCiencia, identificando el total de los artículos científicos digitales, se observó previamente un total de 670 artículos los mismos que serán utilizados para realizar las respectivas pruebas de los algoritmos previamente investigados.

10.7. METODOLOGÍA DE DESARROLLO DE ALGORITMO – METODOLOGÍA PARA EL DESARROLLO DE SOFTWARE

10.7.1. Metodología KDD (Knowledge Discovery in Databases) - proceso de extracción de conocimiento.

Figura 2: Estructura de la Metodología KDD



Fuente: Los investigadores

Para el desarrollo de la presente propuesta tecnológica se tomó como base al proceso de la minería de datos, el mismo que consta de 5 fases, en donde se pretende aplicar el proceso que emite KDD para la realización del algoritmo que nos permitirá el análisis de corpus para poder clasificar de acuerdo a su similitud los artículos científicos de los investigadores de la Universidad técnica de Cotopaxi. La tecnología mencionada se encuentra basado en un proceso

bien definidos para el descubrimiento de conocimiento en grandes volúmenes de datos, en si con la ayuda de KDD lo que se pretende es manejar un proceso iterativo, el cual permite que se mantenga también un proceso de extracción de información de calidad[51]. las fases que se cumplirán son:

Selección de datos. - seleccionar del conjunto de datos originales, un subconjunto apropiado para poder resolver un problema deseado, eliminando variables irrelevantes [52] entonces en esta fase nos permite seleccionar las fuentes de datos para el proceso respectivo que se va a aplicar, como también el tipo de información que se va a utilizar como: corpus del cuerpo del articulo científico y hay que procede a eliminar atributos innecesarios, eliminando tuplas, redefiniendo atributos. En donde lo que se pretende es analizar y digitar un documento.

Limpieza y preprocesamiento de datos. - en esta fase se debería tomar decisiones con respecto a valores faltantes, atípicos, erróneos, etc. Además, se necesita normalizar los valores de las variables o llevar a cabo tareas que se asimilan, este procedimiento es de gran importancia, aunque está un poco descuidada ya que grandes cantidades de datos son recolectados por métodos automáticos. Aunque si se ingresan datos erróneos a los algoritmos de la minería de datos lleva a interponerse en el proceso de aprendizaje o se podría conseguir resultados alejados del comportamiento del algoritmo[52], en esta fase lo que se pretende es obtener la combinación de múltiples datos que pueden ser extensas y heterogéneas además de determinar la confiabilidad, es decir pueden ser datos relacionado o información extraída de cualquier otra fuente de la cual debemos eliminar variables o atributos que no sean útiles para este tipo de tareas como el texto a utilizarse[52].

Transformación de datos. - encontrar características útiles para representar a los datos dependiendo de los objetivos para llevar la continuidad el trabajo con un numero de variables reducido. Eliminar columnas que varían juntas[52]**, e**n esta fase se mejora la calidad de los datos con transformaciones que involucran ya sea reducción de dimensionalidad (disminuir la cantidad de variables del conjunto de datos) para aplicar las diferentes técnicas de modelado, con esta fase lo que se obtiene es que se pierdan las relaciones de integridad y la normalización para poder lograr la obtención de un corpus limpio [51].

Minería de datos. – elegir las herramientas de minería de datos que estén acordes a resolver un problema, teniendo en cuenta siempre el objetivo (predecir, explicar, clasificar, agrupar, etc.). Se procede con el descubrimiento de patrones y relaciones en los datos, para representarlos al usuarios mediante gráficos[52]**,** en esta fase se refiere a elegir el paradigma

apropiado de minería de datos, se aplica el modelo, la técnica y el algoritmo seleccionado. Posteriormente se procede a seleccionar la técnica o algoritmo palabra del patrón y obtener conocimiento y el meta aprendizaje se enfoca en explicar la razón por la que un algoritmo funciona mejor en determinadas problemáticas y para cada técnica existen diferentes posibilidades de como seleccionarlas[53].

Interpretación y evaluación. - en esta fase consiste en reconocer los patrones más importantes. Para finiquitar con la presente técnica mencionada, no está demás decir que KDD es un proceso iterativo e interactivo que permite realizar cambios para conseguir los resultados esperados [51]. Se consolida el conocimiento ganado, probando el modelo creado contra los resultados obtenidos de la aplicación con los que existen en el mundo real, existen características importantes que se detallaran a continuación[52]:

- ✓ La posibilidad de realizar un ciclo entre cualesquiera dos pasos, ya que a veces el conocimiento descubierto puede ser directamente aplicable, y otras veces puede guiar al refinamiento de los objetivos de la minería de datos[52].
- ✓ No hay un proceso determinístico asumido desde un paso a otro. Además, todos los pasos interpretativos y evaluativos, pueden involucrar que no se avance y se quede estancado con los de los pasos anteriores, cualquier número de veces.
- ✓ La incorporación de un agente inteligente, encargado de monitorear las actividades del usuario y brindarle a esta asistencia a lo largo de todo el proceso de KDD.

Para una evaluación confiable del modelo de algoritmo aplicado se procedió a analizarlo con Train/Test Split and Cross Validation in Python (Entrenamiento/prueba dividida y validación cruzada).

En estadísticas y aprendizaje automático, generalmente si divide los datos en subconjuntos para entrenar, valorar y probar, para ajustar el modelo[54].

1. **División de tren/ prueba**

   Los datos que se dividieron se usan como un conjunto de entrenamiento que contiene una salida conocida y el modelo aprende para generalizarlos más adelante, utilizando las diferentes librerías de Python que nos brindan para tener un resultado más amplio y especifico.

2. Validación cruzada

Es un a técnica que se utiliza para la evaluación de los resultados de un análisis estadístico y garantice que son independientes del conjunto de prueba y entrenamiento. Su objetivo principal es la predicción y se quiere estimar la precisión de un modelo que se llevara a probar.

3. Validación Cruzada K-Folds

En este proceso se divide el conjunto de datos en particiones o pliegues iguales a K, para después calcular la precisión de prueba de nuestro modelo.

4. Validación Cruzada (LOOCV)

Este método es muy costoso computacionalmente, por lo que se recomienda utilizar en un conjunto pequeño de datos. En este tipo de validación cruzada, el número de pliegues ósea subconjuntos es igual al número de observaciones que tenemos en el conjunto de datos. Luego promediamos todos los pliegues y construimos nuestro modelo con el promedio.

Se utilizará también para verificar el algoritmo el error cuadrático para utilizar como división entre los valores correctos y los valores de predicción, calculamos el error al cuadrado en lugar del error simple, para que el error siempre sea positivo de esta forma sabemos que el error perfecto es 0, si no elevásemos el error al cuadrado, unas veces el error sería positivo y otras negativo y existe otra posibilidad en utilizar el valor absoluto, en lugar de elevarlo al cuadrado pero no obtendremos una función no-derivable[55].

10.8. METODOLOGÍA DE DESARROLLO ÁGIL

Las metodologías de desarrollo ágil hoy en la actualidad engloban varios conceptos importantes y poseen ciertas propiedades que las hacer totalmente aplicables al dominio del software. Conociendo a estos métodos ágiles como la solución potencial para el desarrollo de software en páginas web[56].

10.8.1. Metodología scrum

La metodología que se aplicara para el desarrollo del módulo en la plataforma EcuCiencia es SCRUM, debido a que presenta grandes beneficios para recolectar información ya que este punto ciertamente es uno de los más esenciales para luego proceder a aplicar el algoritmo para la creación del corpus. Por lo tanto, partiendo de lo mencionado con anterioridad la metodología de desarrollo ágil como es scrum es un proceso en que se aplican de manera regular en donde

se divide el proyecto en pequeños bloques o Sprint, con el objetivo de ir revisando y mejorando la fase anterior y tener un mejor resultado [57].

10.8.2. Roles de la metodología scrum

Product owner es aquella persona que se convierte en la voz del cliente, es decir establece una relación entre el cliente y el equipo de trabajo para trasladar la visión del proyecto al equipo, formaliza las presentaciones e historias a incorporar en el Product backlog (funcionalidades de un sistema).

Scrum master es quien se encarga de coordinar un equipo de trabajo ya que debe estar presente brindando todo el apoyo posible para que se cumplen los procesos de acuerdo a lo establecido siendo el que tiene experiencia para el manejo de la metodología a cumplirse en el proyecto.

Scrum team **es** el equipo de profesionales con los suficientes conocimientos técnicos necesarios y que desarrollan el proyecto. No obstante, dentro del presente equipo se encuentran analistas, diseñadores, programadores y tester[57].

Conociendo el procedimiento o flujo de trabajo de la metodología scrum tenemos:

El Product Backlog es aquel que especifica los requerimientos que estarán comprendidos como las historias de usuario que serán escritas por un usuario normal que luego se comprenderá por un experto, las mismas que estas historias de usuario serán tomadas por la Dirección de Investigación.

El Sprint Plannig en esta fase se mantendrán durante el Product Owner con el Scrum Master para detectar las historias de usuario a realizarlas y seguidamente priorizar, para que en una segunda reunión decidir y organizar como lo van a conseguir cada funcionalidad ya establecida.

Sprint en esta fase se trabaja cada iteración, la cual el team trabaja conjuntamente para lograr que las historias de usuario del Product Backlog sean funcionales y acordes a lo comprometido.

Sprint Backlog son ya los requerimientos directos y no directos que tendrá la plataforma científica, se llevará un registro de todas las tareas que se necesiten cumplir en cada sprint.

- Para las reuniones que se efectuara durante el proyecto tenemos las siguientes:

Reunión del Sprint diario consiste en mantener reuniones diarias entre todos los miembros del equipo de desarrollo para determinar que se hizo hoy, que se va a hacer mañana.

Reunión de retrospectiva consiste en mantener reuniones para determinar si se realizó bien y mal las cosas delegadas por los líderes del equipo de trabajo, se procede a realizar un análisis de retroalimentaciones[57].

11. ANÁLISIS Y DISCUSIÓN DE RESULTADOS

Técnica de investigación

11.1.ENTREVISTA

Obtener más información sobre los requerimientos funcionales que se llevaran para el desarrollo del nuevo módulo en la plataforma EcuCiencia.

1. **¿La plataforma EcuCiencia por cuanto tiempo está funcionando?**

   La plataforma nace a partir de una convocatoria emitida por dirección de investigación de la universidad técnica de Cotopaxi a proyectos de investigación generativa en el año 2017, entonces nace a través de la idea de que en la universidad no existía ningún sistema, plataforma o recurso software que tuviera organizado un grupo de publicaciones, artículos que los docentes hayan realizado y que normalmente se realizaba de forma manual a través de hojas de Excel, etc.

   Se me ocurre esta idea después de que tuve una formación doctoral sobre este tema entonces cumplimento la tesis doctoral hace referencia a este proyecto, poniendo en práctica lo aprendido para poder organizar de cierta manera un sistema que fuera capaz de inferir y de aprender de las cosas que los docentes subirían al sistema, es por eso que se aplica algoritmia. Ese proyecto empezó a financiar con un costo de 20 000 dólares en Enero del 2018 , las primeras cosas como el diseño de base de datos se empiezan a ver en el mes Mayo ya que se tuvo el primer acercamiento hacia los docentes para que pudieran ingresar todo la información de los artículos publicados en el sistema, ya que la primera etapa era aglutinar todos los artículos para que los docentes puedan subirlos mostrándoles el formulario que deberían llenar para subir cada artículo y así para ya después poder aplicar los algoritmos y empezar ya a tener ciertos niveles de inteligencia artificial para mostrar las redes de trabajo entre investigadores siendo esto propósito fundamental del sistema. Ya en la actualidad se han desarrollado varios modulo que han sido de gran importancia para el complemento del sistema.

2. ¿Qué lenguaje de programación fue utilizado para el desarrollado del sistema?

El principal lenguaje de programación utilizado aquí en este sistema es Python porque a través de las investigaciones realizadas con el compañero Alex y mi persona nos percatamos que Python era uno de los lenguajes de programación orientados hacia el análisis de datos porque lo que hace el sistema es analítica de datos enfocada a la parte textual, términos, etc. Vimos la potencialidad que brindaba Python así que rápidamente elegimos la mejor opción.

3. ¿Cuál es el Gestor de Base de Datos con el que está trabajando el sistema?

Igualmente, al analizar las diferentes opciones de base de datos elegimos PostgreSQL porque era un poco más robusto y a medida que iba a creciendo, este gestor de base de datos nos mostraba más eficiencia en todo el proceso de tal manera de que si se seguía creciendo no fuera una traba para continuar con el desarrollo del sistema.

4. ¿Cuál es el objetivo del nuevo módulo de procesamiento de datos de la plataforma EcuCiencia?

El objetico principal de este módulo que se está trabajando es analizar el corpus de los artículos científicos que están en la base de datos de la plataforma EcuCiencia para conocer cuáles son los niveles de similaridad y la posibilidad de obtener patrones que están establecidos dentro de la documentación de los artículos científicos para poder obtener desde el punto de vista léxico semántico resultados para conocer desde la línea de investigación cuales son las área, subáreas especificas en las cuales se han estado trabajando así mismo como las terminologías que se están utilizando ya que es lo que representa al artículo.

5. ¿Qué aporte brindaría a la comunidad universitaria UTC la implementación de este nuevo módulo?

El aporte que brindaría el nuevo módulo a la comunidad universitaria es conocer cales son las principales áreas de conocimiento que se están trabajando desde las líneas de investigación que la universidad técnica de Cotopaxi tiene, así mismo conocer la similitud de artículos por líneas y sublínea de investigación, autores con la facilidad de poder obtener un corpus de los artículos.

6. ¿Cuáles son las funcionalidades del módulo procesamiento de datos?

Las funcionalidades están orientadas hacia:

Determinación de las palabras que tienen mayor frecuencia de aparición en los artículos.

Conocer los niveles de similitud.

Conocer los niveles de distancia.

Conocer las terminologías específicas de cada artículo.

Obtener corpus.

7. ¿Qué resultados espera obtener del sistema ya aplicando el nuevo módulo?

Evidentemente se podría sacar muchos resultados interesantes desde el punto de vista científico en la publicación de artículos con esto se podría determinar uno, dos, tres…, artículos orientados para conocer el corpus de un artículo y el cómo se comportaría un área en un tiempo determinado. Con esto se podría fomentar a la investigación y obtener presupuesto para un área determinada.

11.2.METODOLOGÍA DE DESARROLLO DE ALGORITMO

11.2.1. METODOLOGÍA KDD

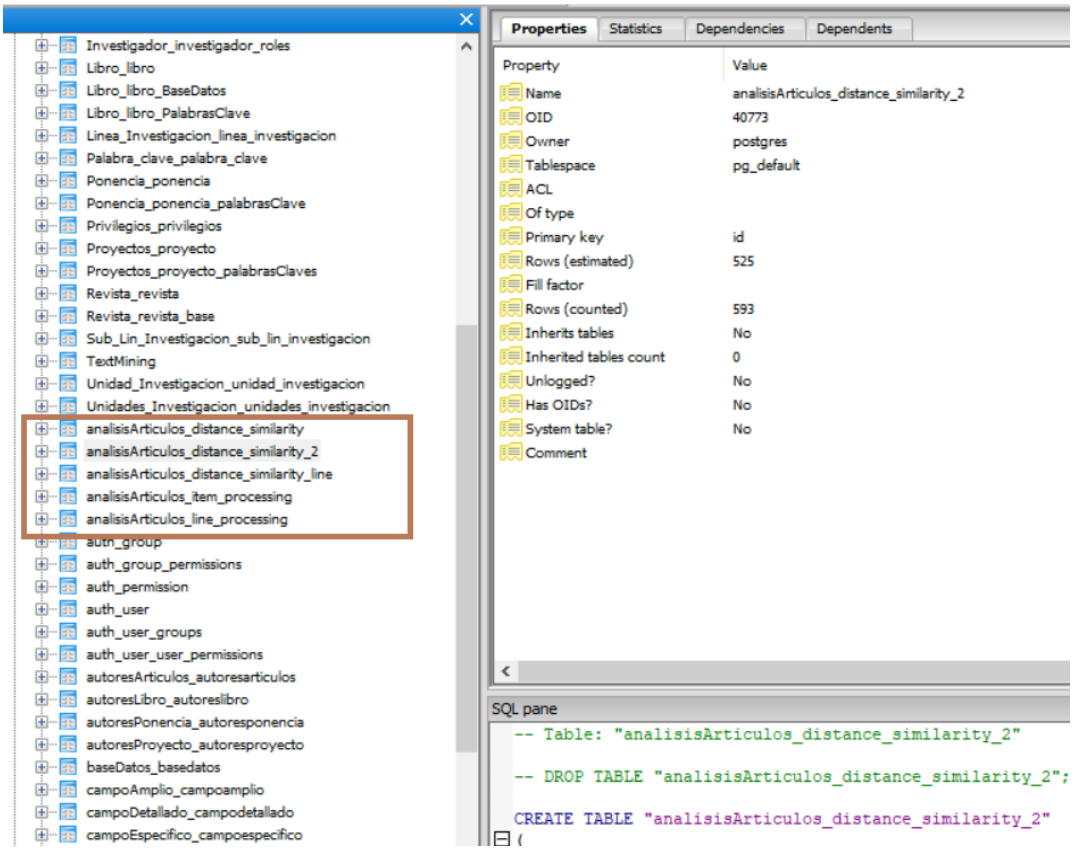
Para lograr un mejor resultado en el desarrollo de proyecto se realizó una entrevista al coordinador del proyecto REDEC el PhD. Gustavo Rodríguez logrando obtener respuestas de gran importancia para tener conocimiento de las herramientas con la que se trabaja en el sistema EcuCiencia.

Entonces para lograr tener un resultado del segundo objetivo específico procedemos a analizar la metodología que se utilizara para el modelado del algoritmo de análisis de corpus, el mismo que mostraremos en diferentes etapas el cumplimiento de cada una, ya que esta fue la primera etapa del proyecto.

1. SELECCIÓN

Logrando cumplir cada etapa de la metodología se logra tomar como evidencia las tablas de la base de datos del sistema EcuCiencia las cuales actualmente están relacionadas entre sí y de las cuales seleccionaremos las que finalmente trabajaremos para cumplir nuestro proyecto.

Figura 3: Proceso de Selección-Metodología KDD

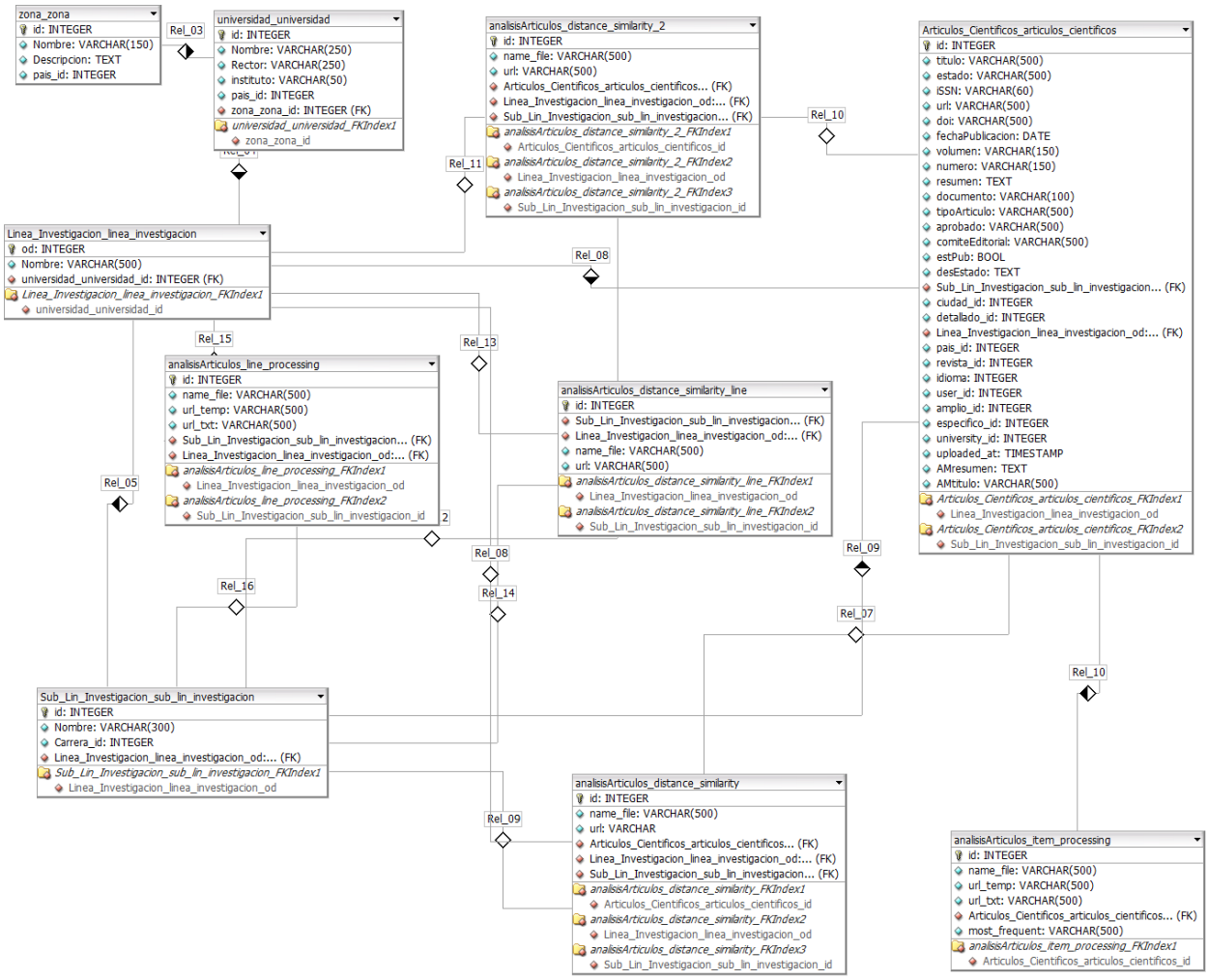


Fuente: Los investigadores

2. PREPROCESAMIENTO Y LIMPIEZA

Una vez conocidas y analizadas todas las tablas de la base de datos de la plataforma EcuCiencia procederemos a seleccionar las distintas tablas que son necesarias para el procesamiento y aplicación de los algoritmos para no tener ningún inconveniente con el desarrollo del proyecto, las mismas que son las siguientes:

Figura 4:Tablas de BBDD-Metodología KDD



Fuente: Los investigadores

3. TRANSFORMACIÓN Y REDUCCIÓN

Para cumplir con esta etapa de transformación y reducción de datos, para el desarrollo del algoritmo se debe tener un tratamiento preliminar de los datos con una estructura apropiada que nos llevará a obtener el corpus limpio para su procesamiento realizamos los procesos que sean necesarios para cumplir sin problema esta etapa para eso tenemos lo siguiente:

Convertir de .pdf a .txt. – se consulta todos los archivos existentes y se convierte de uno en uno todos los artículos científicos que existan y mantengan la extensión .pdf haciendo uso de la librería **fitz. open.**

Figura 5:Codificación Convertir de. Pdf a .txt - Metodología KDD

```
44  # region EEJECUCION PROGRAMADA
45  # crear un archivo de similaridad para cada articulo
46  def pdfToTxtArticles():
47      print('convertir PDF a TXT')
48      title = ''
49      appendFile = ''
50      words_freq = ''
51      # filtramos todos los articulos cientificos
52      articles = articulos_cientificos.objects.all()
53      for i in articles:
54          id_article = i.id
55          title = str(i.titulo).replace("/", " ").replace('"', ' ').replace('´', '')
56          docu = str(i.documento)
57
58          # obtener la extencion del documento
59          ext = os.path.splitext(docu)[1]
60
61          # cambiar el tamaño del titulo
62          if len(title) > 125:
63              title = title[0:125]
64
```

Fuente: Los investigadores

Eliminación de stopwords. - Se procede a realizar una consulta de los documentos con extensión .pdf que se encuentran en la plataforma EcuCiencia mismos que serán transformados a archivos de texto plano que estarán alojados en una carpeta temporal para tokenizar el contenido del texto y así obtener la eliminación de las palabras de parada con ayuda de la función stopwords que nos proporciona la librería NLTK.

Figura 6:Codificación Eliminación de stopwords-Metodología KDD

```
78          if os.path.isfile(URL_TXT + 'analisis/text/' + title + '.txt'):
79              print('el archivo ya existe PDF TO TXT')
80              remove(URL_TXT + 'analisis/text/' + title + '.txt')
81
82          # region CONVERTIR DE .PDF A .TXT
83          # analizar el documento completo y crear un .txt de manera temporal
84          # URL_TEMP = en esta ruta se guardan los archivos .txt creados4
85          salida = open(r'' + URL_TEMP + 'analisis/temp/' + title + '.txt', "wb")
86          for pagina in documento:
87              text = pagina.getText().encode('utf-8', 'ignore')
88              t = text.lower()
89              salida.write(t)
90              salida.write(b"\n")
91          salida.close()
92          # fin creacion de documento .txt
93          # endregion
94
```

Fuente: Los investigadores

Convertir a minúsculas. – previamente teniendo el archivo abierto se utiliza la función lower del lenguaje de programación Python para convertir todo el contenido del texto plano y así tener una mejor limpieza de datos.

Figura 7:Codificación Convertir a minúsculas-Metodología KDD

Fuente: Los investigadores

Elimina el signo de puntuación de cada ficha. – después de convertir en minúsculas todo el texto se procede a utilizar la función Split de Python para dividir el documento en palabras por espacio en blanco y luego utilizar la traducción de cadenas para reemplazar toda la puntuación por nada o eliminarla.

Figura 8:Codificación Eliminar signos de puntuación-Metodología KDD

Fuente: Los investigadores

4. MINERÍA DE DATOS

Después de que se haya realizado la conversión de los archivos .pdf a .txt, la limpieza de signos de puntuación, eliminación de stopwords, convertir a minúsculas todo el texto se obtiene el corpus primeramente representada por una línea de investigación, sublínea de investigación el mismo corpus el cual podremos sacar diferentes análisis tales como: riqueza léxica, número de páginas, numero de palabras, frecuencia de palabras para hacer más fácil el manejo de cantidades masivas de texto que nos brindan la oportunidad de comprender y explicar el contenido de los textos analizados.

5. INTERPRETACIÓN Y EVALUACIÓN

Aplicando un evaluador de nuestro modelo usando una parte y probamos su efectividad en otro conjunto de datos, a continuación, se muestra los resultados arrojados con el método entrenamiento/ prueba dividida y validación cruzada con Python.

I. División de tren/ prueba

Los datos que usamos se dividen en datos de entrenamiento y datos de prueba, en donde el conjunto de entrenamiento tiene una salida y los datos aprende. El conjunto de datos para probar la predicción de nuestro modelo. A continuación, les mostraremos la codificación y las gráficas representados con resultados arrojados.

Figura 9:Codificacion Evaluación del algoritmo

validation.py

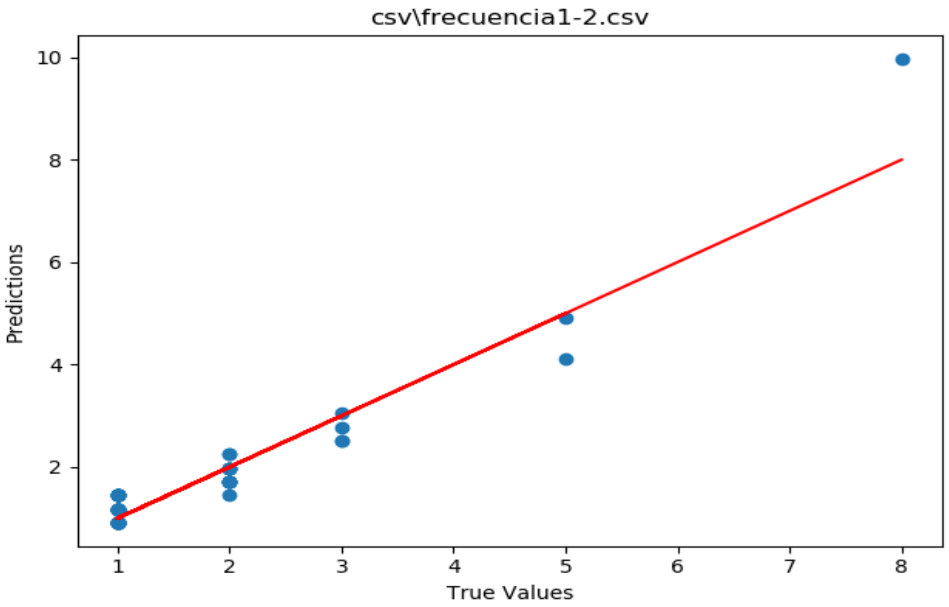
```
8
9   def divTrainTest(file):
10      datos = pd.read_csv(file)
11      col1 = 'Frecuencia_art_1'
12      col2 = 'Frecuencia_art_2'
13      # Load the Diabetes dataset
14      columns = ("Frecuencia_art_1 Frecuencia_art_2").split()
15      df = pd.DataFrame(datos[col1], columns=columns).fillna(value=0)
16      y = datos[col2]
17      X_train, X_test, y_train, y_test = train_test_split(df, y, test_size=0.2)
18      lm = linear_model.LinearRegression()
19      model = lm.fit(X_train, y_train)
20      predictions = lm.predict(X_test)
21      predictions[0:5]
22      plt.scatter(y_test, predictions)
23      plt.title(file)
24      plt.xlabel("True Values")
25      plt.ylabel("Predictions")
26      plt.plot(y_test, y_test, color='red')
27      plt.show()
28      r = model.score(X_test, y_test)
29      return "{0:.3f}".format(r)
```

Fuente: Los investigadores

GRÁFICAS DE FRECUENCIA PARA LA EVALUACIÓN DEL ALGORITMO.

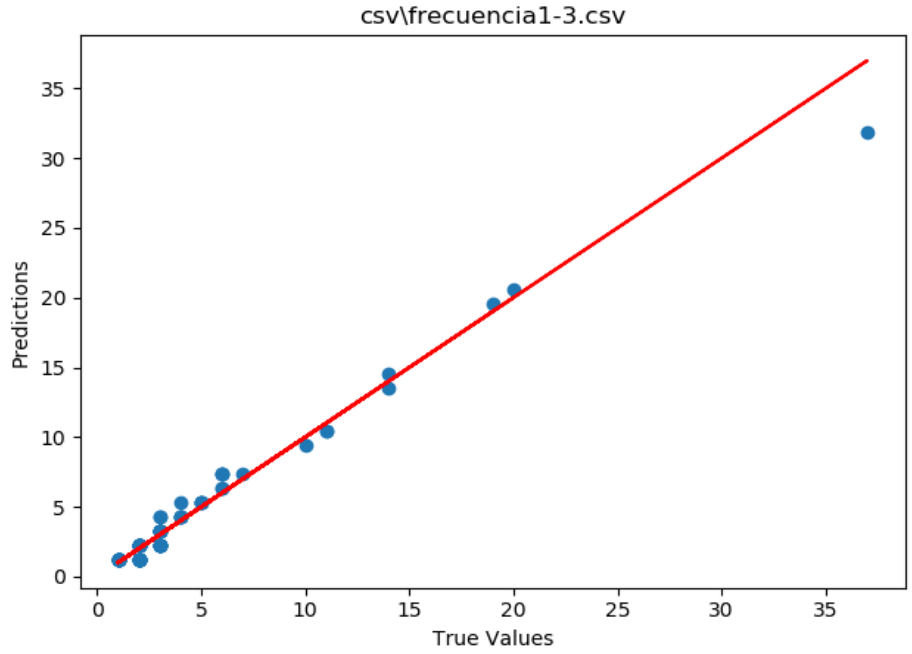
Para evaluar el algoritmo del módulo de procesamiento de datos se pone a prueba con 5 artículos científicos para verificar el entrenamiento del algoritmo y se arroja resultados positivos para avanzar con la implementación del ambiente web.

Figura 10:Frecuencia 1-2 División de tren/prueba



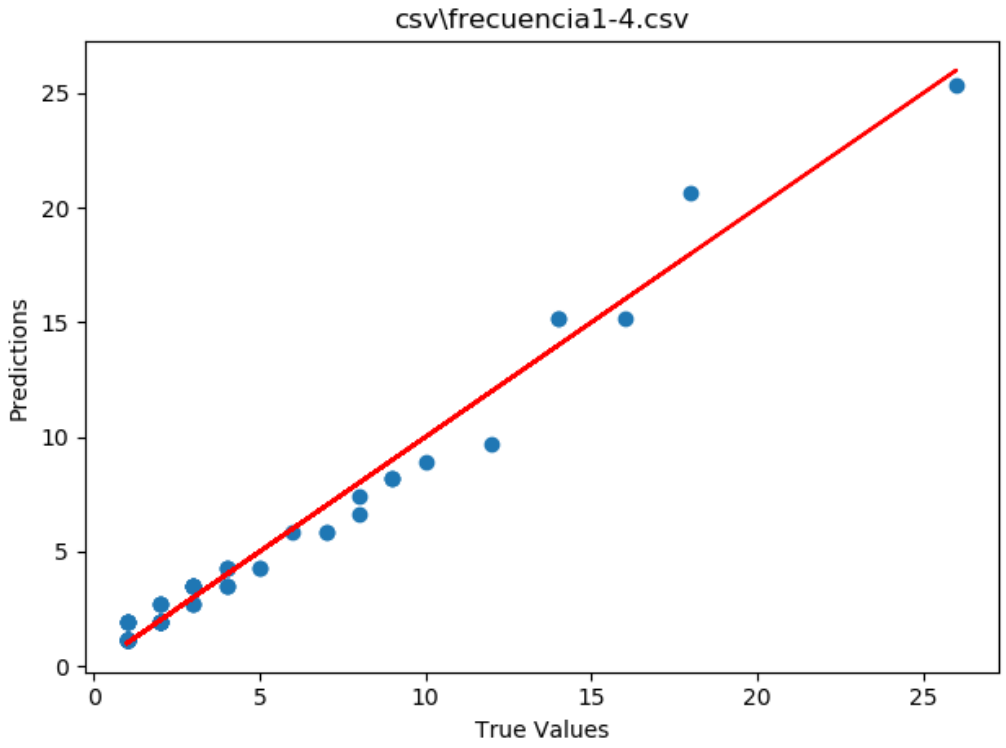
Fuente: Los investigadores

Figura 11: Frecuencia1-3 División de Tren/prueba



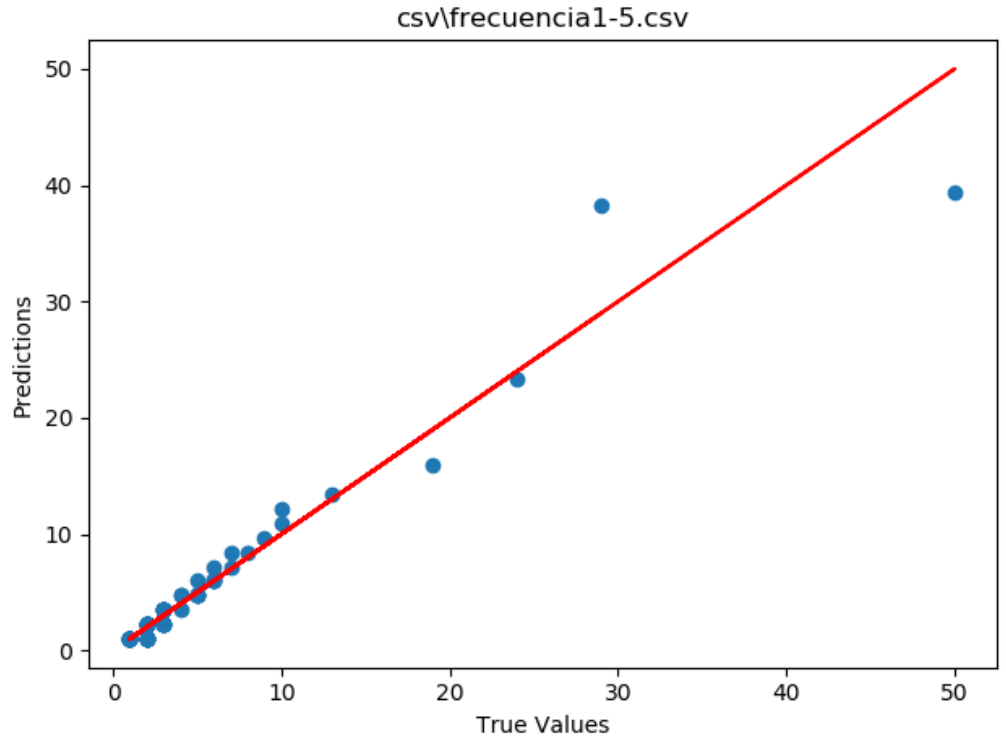
Fuente: Los investigadores

Figura 12: Frecuencia 1-4 División de tren/prueba



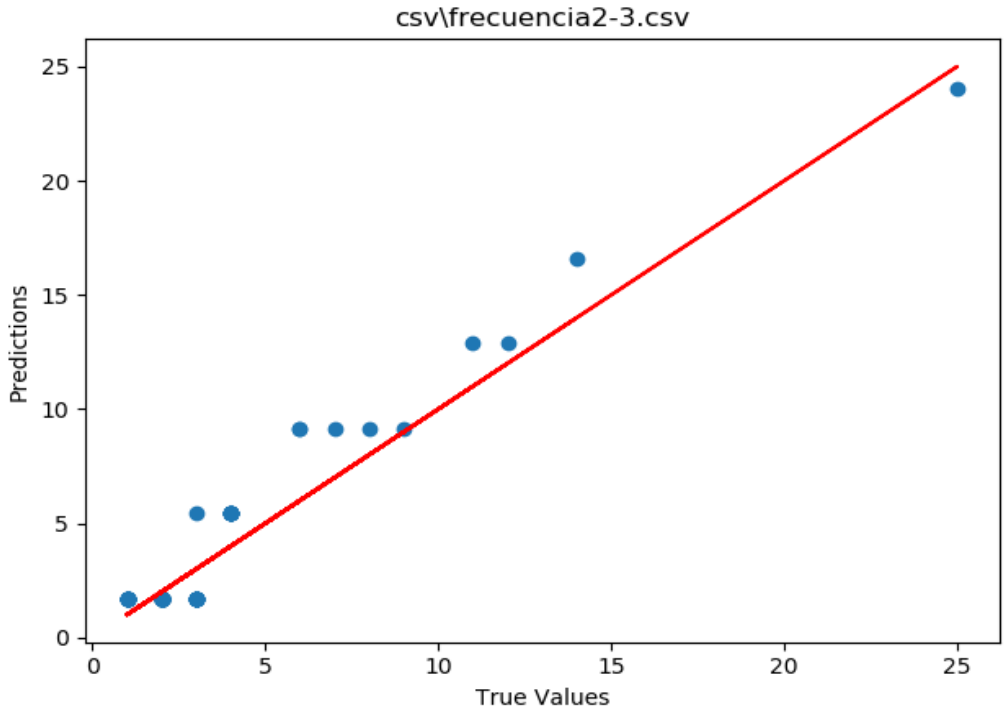
Fuente: Los investigadores

Figura 13:Frecuencia 1-5 División de tren/prueba



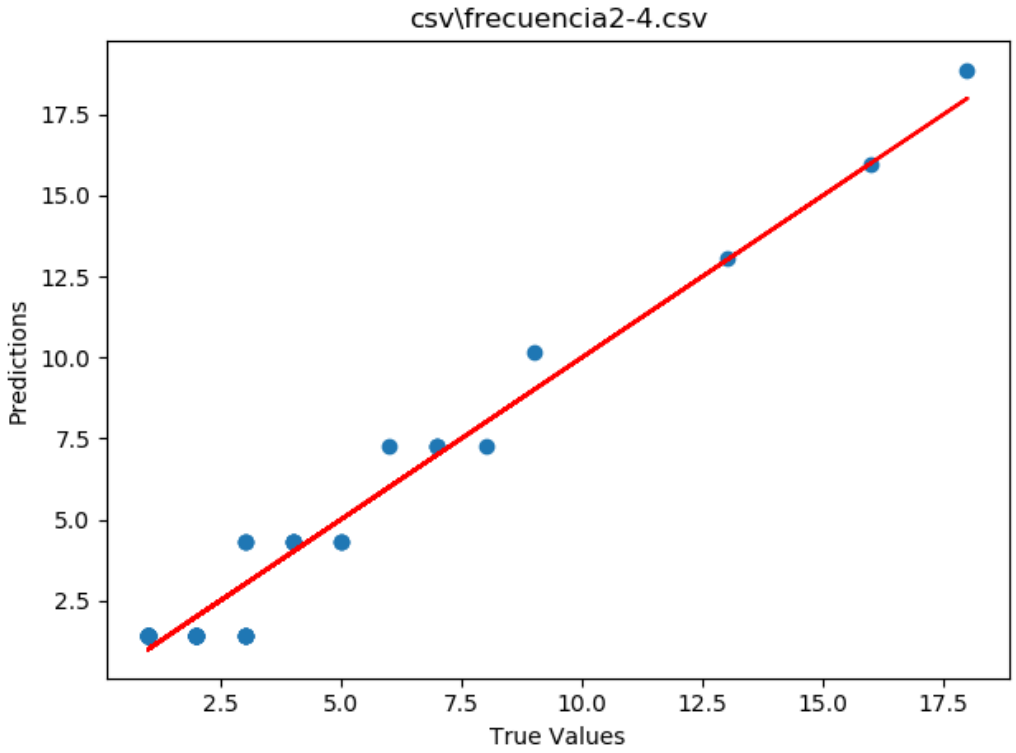
Fuente: Los investigadores

Figura 14: Frecuencia 2-3 División de tren/prueba



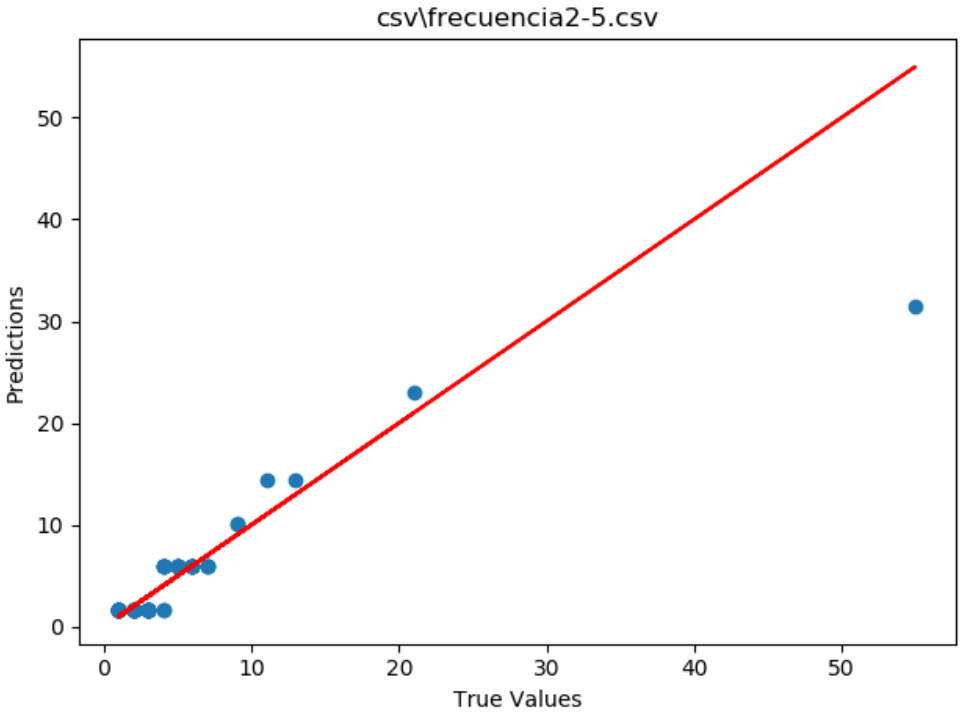
Fuente: Los investigadores

Figura 15:Frecuencia 2-4 División de tren/prueba



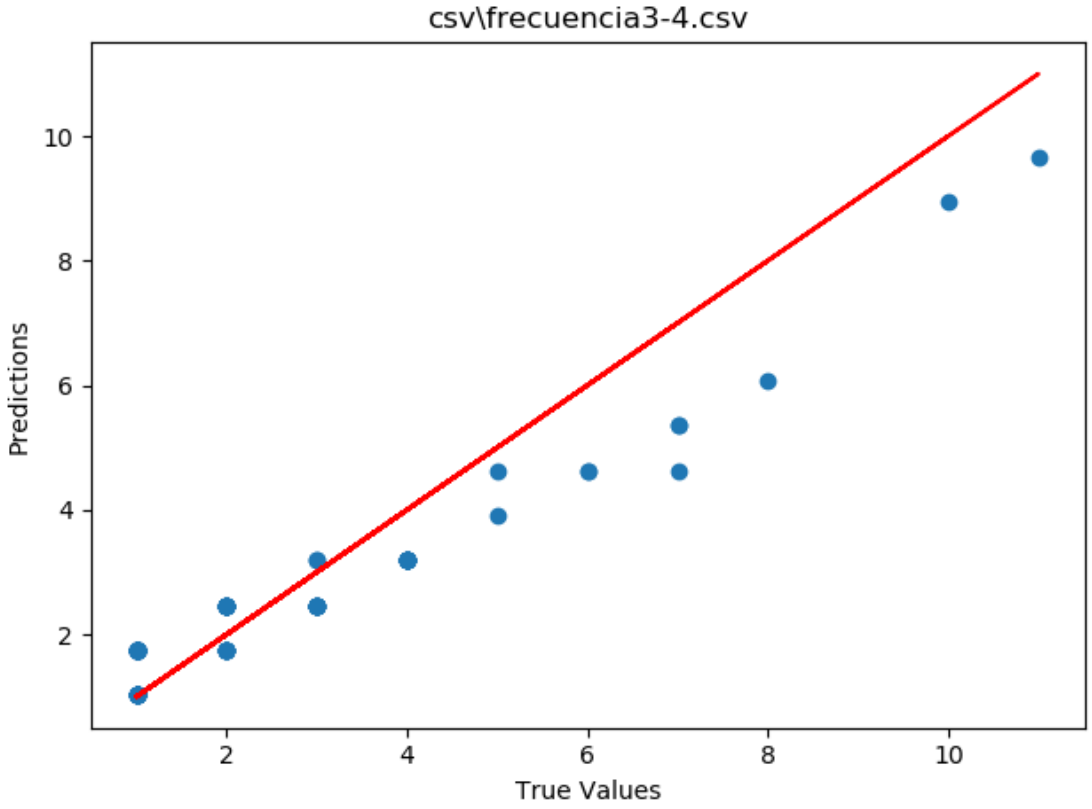
Fuente: Los investigadores

Figura 16:Frecuencia 2-5 División de tren/prueba



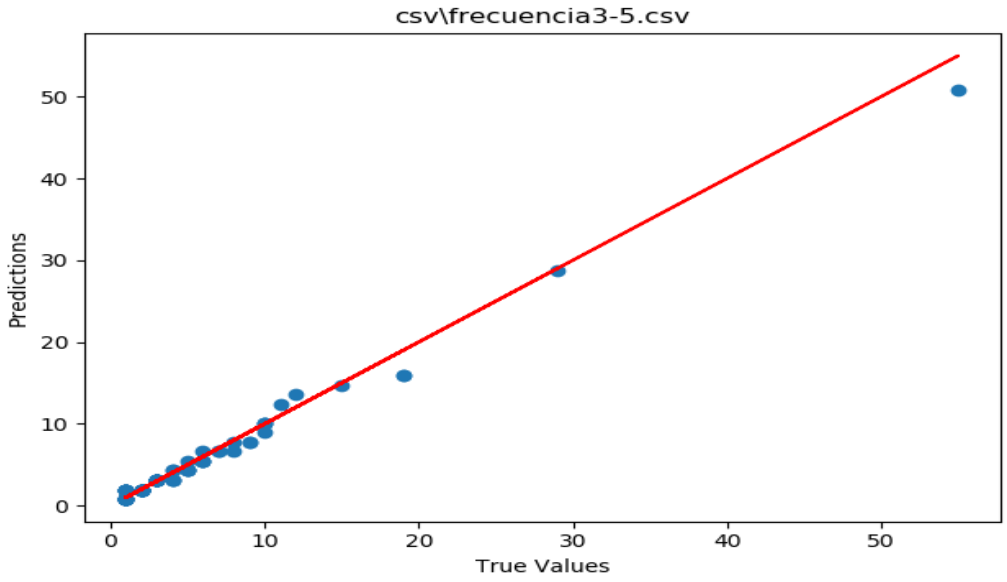
Fuente: Los investigadores

Figura 17: Frecuencia 3-4 División de tren/prueba



Fuente: Los investigadores

Figura 18: Frecuencia 3-5 División de tren/prueba



Fuente: Los investigadores

Los gráficos obtenidos que se generaron a partir de los valores verdaderos (similaridad) y los de predicción usando un modelo de regresión lineal, se evidencia a partir de los valores de predicción: [0.918,0.953,0.962,0.933,0.839,0.941,0.814,0.881,0.823,0.946]. la regresión lineal aproxima la variedad del objetivo minimizando los cuadros de desviaciones, al tener un modelo que tiene coeficientes altos se puede interpretar como unas altas varianzas en los datos.

Figura 19: Codificación valores de predicción/ error cuadrático

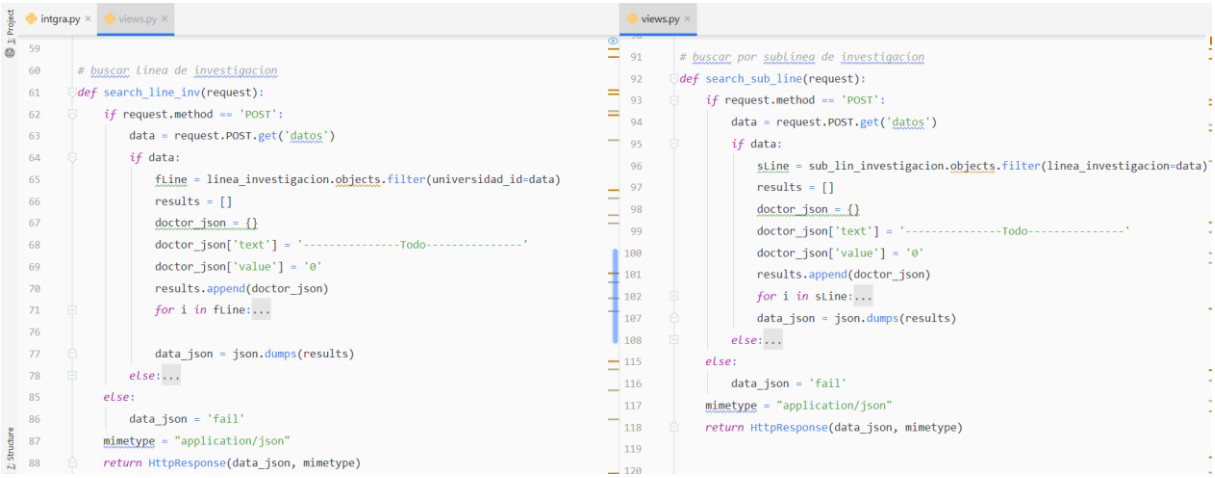
```
validation.py
40
41  def sqError(vt, vp):
42      y_true = vt
43      y_pred = vp
44      mse = mean_squared_error(y_true, y_pred)
45      return mse
46
47
48  vt = [0.133, 0.265, 0.173, 0.233, 0.110, 0.129, 0.130, 0.156, 0.249, 0.161]
49
50  vp = openAllFiles("csv")
51
52  sq = sqError(vt, vp)
53
54  print("Valores Correctos:", vt)
55  print("Valores de Predicción:", vp)
56  print("Error Cuadratico:", sq)
```

Fuente: Los investigadores

Se tiene un error cuadrático igual a 0.56, la interpretación del error cuadrático muestra varianza del estimador y su sesgo, este valor 0.56 muestra una baja variación y demuestran que los datos se estarían ajustando al modelo.

Filtrado de la línea y sub línea de investigación. – para realizar el filtrado de la línea y sub línea de investigación, primeramente, se buscan que estén registradas en la base de datos y filtrar los datos de la tabla según cualquier atributo.

Figura 20:Codificación-Filtrado de la línea y sub línea de investigación.



Fuente: Los investigadores

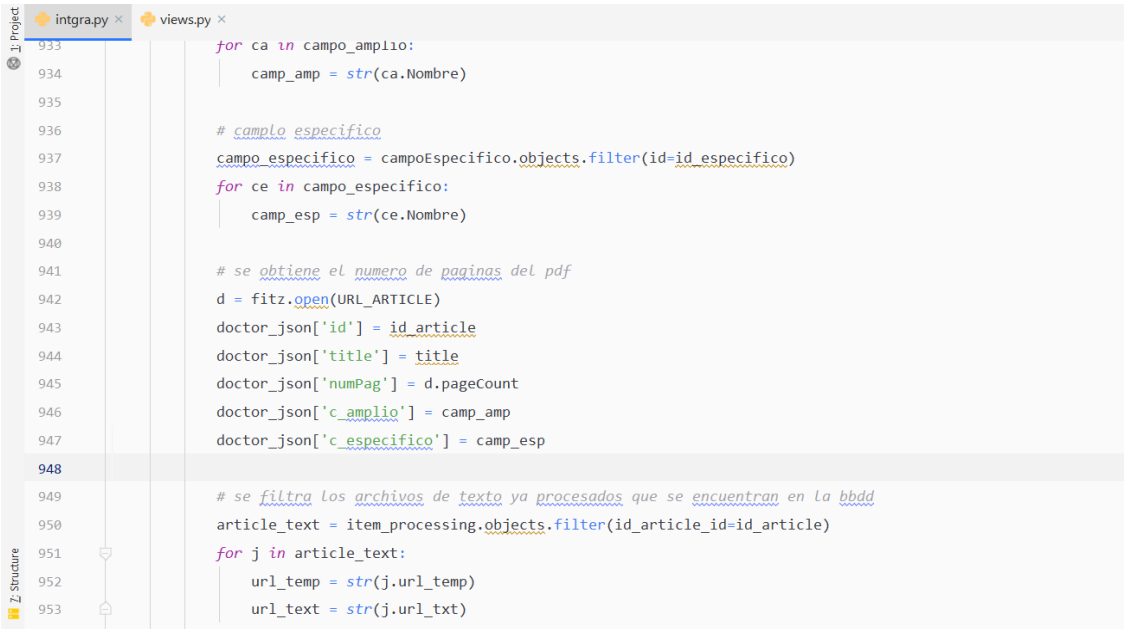
Campo amplio y campo especifico. – después de filtrar toda la información que necesitaremos para comenzar el análisis se busca en la base de datos los atributos referenciados con campo amplio y especifico que nos mostrara que tipo de campo es al que pertenece el articulo científico a analizar.

Figura 21: Codificación-Campo amplio y campo específico.

Fuente: Los investigadores

Número de páginas. – para conocer el número de páginas se utilizó la función pageCount de la librería fitz del lenguaje de programación Python.

Figura 22:Codificación-Número de páginas.



Fuente: Los investigadores

Número de palabras (sin palabras de parada). – se realiza la tokenización del texto plano y se utiliza la función **len** que nos permitirá conocer el contenido del texto sin palabras de parada para no tener problemas al momento de analizar los textos.

Figura 23:Codificación- Número de palabras (Sin palabras de parada)

Fuente: Los investigadores

Riqueza léxica. – la riqueza léxica de un documento analizado se basa en una función de la relación existente entre las palabras totales de un corpus y las palabras diferentes del mismo, dando como resultado un porcentaje. Es decir, cuantas más palabras distintas haya respecto al total, mayor será la riqueza léxica.

Figura 24:Codificación-Riqueza Léxica

Fuente: Los investigadores

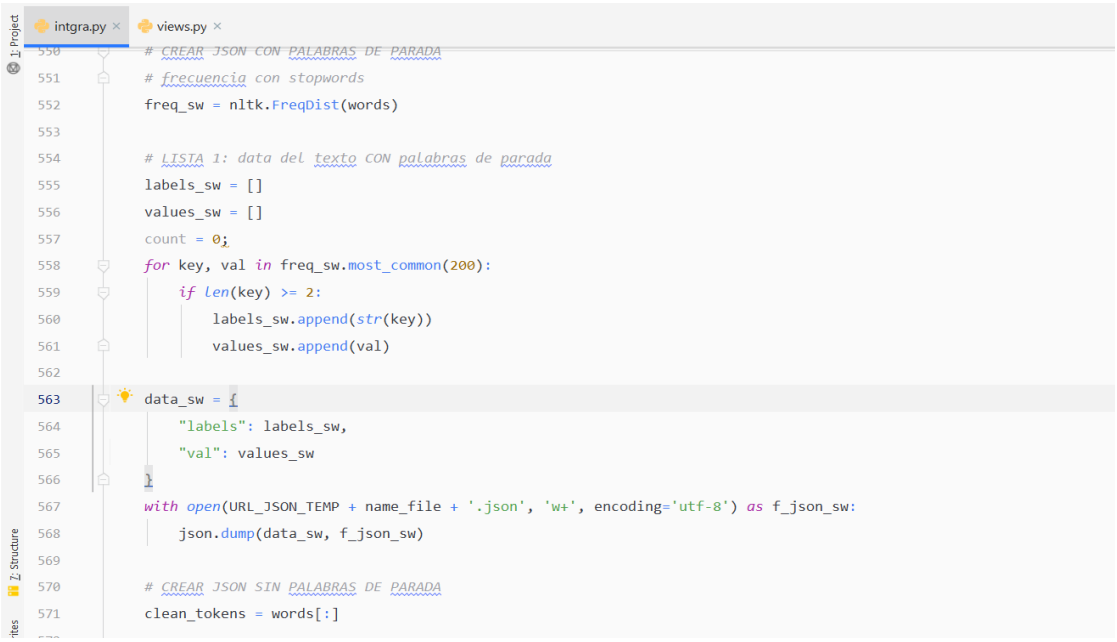
Frecuencia de palabras. – utilizamos la **FreqDist** de la librería **NLTK** para conocer la frecuencia de las palabras que tendría nuestro documento de texto plano sin analizar.

Figura 25:Codificación-Frecuencia de palabras

Fuente: Los investigadores

Graficas con palabras de parada. – para obtener las gráficas de las palabras de parada se realiza una función para especificar todas las palabras que tengan desde 2 caracteres que contiene el documento.

Figura 26:Codificación-Graficas con palabras de parada



Fuente: Los investigadores

Graficas sin palabras de parada. – a diferencia de la gráfica de las palabras con parada este tiene a especificar que se tomarán desde los 4 caracteres de las palabras que serán más precisas para un buen análisis del documento.

Figura 27: Codificación- Graficas sin palabras de parada

Fuente: Los investigadores

11.3.METODOLOGÍA ÁGIL

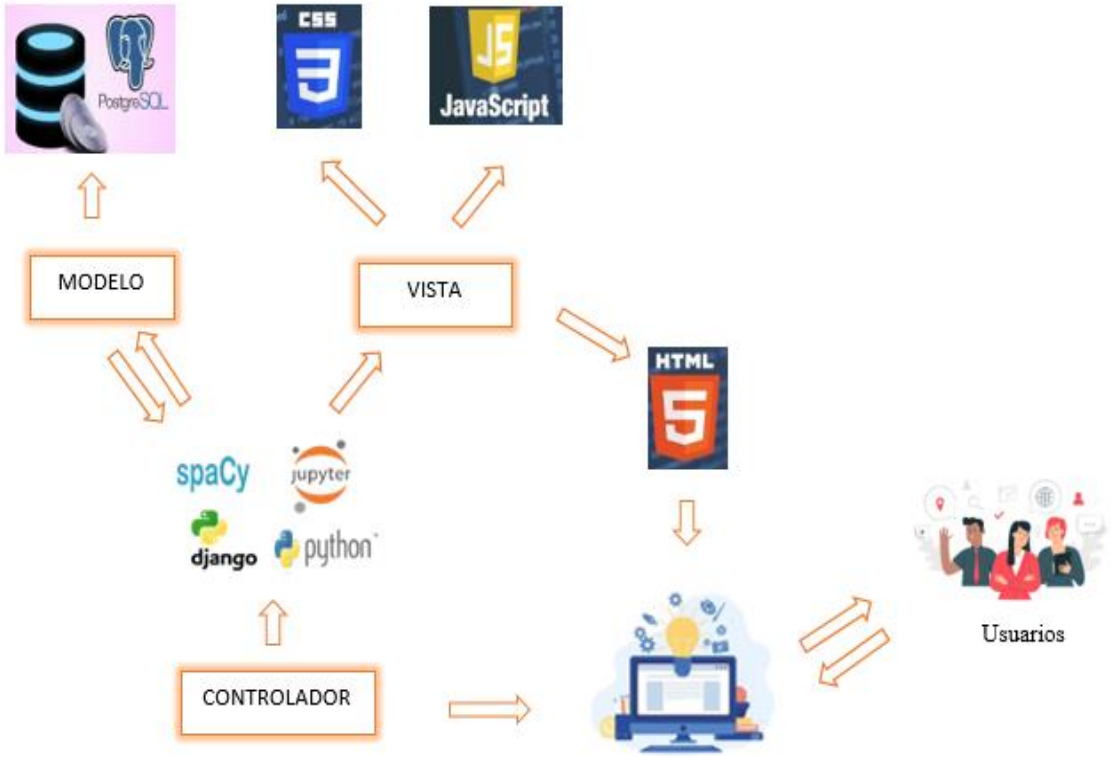
11.3.1. METODOLOGÍA SCRUM

Para complementar el proyecto se utilizó como segunda etapa la metodología scrum para unir la lógica con el desarrollo de software.

Entonces conoceremos las etapas que se lograron desarrollar para el proyecto.

11.3.2. DIAGRAMA DE ARQUITECTURA

Figura 28: Diagrama de arquitectura MVC



Fuente: Los investigadores

11.3.3. ROLES DEL EQUIPO SCRUM

Tabla 3:Roles del equipo SCRUM

Persona	Contacto	Función
PhD. Gustavo Rodríguez	gustavorodriguez@utc.edu.ec	Product Owner
PhD. Segundo Corrales	segundocorrales@utc.edu.ec	Master Scrum
Chariguaman Gilson	gilson.chariguaman4772@utc.edu.ec	Scrum Team
Quilumbaquin Nataly	nataly.quilumbaquin6398@utc.edu.ec	Scrum Team

Fuente: Los investigadores

11.3.4. ARTEFACTOS DEL SCRUM

Para poder realizar el desarrollo del proyecto se necesita conocer las historias de usuario, las cuales estarán definidas seguidamente:

1. Como usuario, quiero revisar el sistema EcuCiencia.
2. Como usuario quiero conocer el filtrado de datos por líneas y sublineas de investigación.
3. Como usuario quiero conocer la búsqueda de datos de un artículo en específico.
4. Como usuario quiero obtener el corpus de mi artículo científico.
5. Como usuario quiero saber el número de palabras tiene un artículo científico.
6. Como usuario saber cuáles son las palabras que se mas repiten en un artículo científico.
7. Como usuario quiero saber cuál es la riqueza léxica de un artículo científico.
8. Como usuario quiero conocer el número de palabras comunes que existe en un artículo científico.
9. Como usuario quiero conocer la distancia y similitud del texto de un artículo científico.
10. Como usuario quiero visualizar las gráficas del contenido de los documentos
11. Como usuario quiero poder descargar el corpus de un artículo científico sin palabras comunes

11.3.5. PRIORIZACIÓN Y ESTIMACIÓN DE TIEMPO

Una vez analizados todas las historias de usuario se procede a elaborar la priorización y estimación de tiempo , para ello utilizaremos la técnica de priorización conocida como MoSCow que se divide en: M (Must) que quiere decir que necesariamente debe ser implementado, S (Should) significa alta prioridad pero no es impredecible, C(Cloud) es un requisito que de alta prioridad pero no es codiciado pero no necesario, W (Won´t) significa que los requisitos están descartados.

Tabla 4: Historias de usuario

Id	Historias de Usuario	M	S	C	W
1	Revisión de código		X		
2	Realizar filtrado de datos	X			
3	Obtener corpus	X			
4	Analizar número de palabras	X			
5	Conocer palabras repetidas	X			
6	Conocer riqueza léxica		X		
7	Conocer las palabras comunes	X			

8	Conocer niveles de distancia	X			
9	Conocer similitud de textos	X			
10	Visualizar gráficos	X			
11	Actualizar información	X			
12	Descargar corpus		X		

Fuente: Los investigadores

Para determinar los requerimientos funcionales que existen en la plataforma, se utilizó la técnica ágil de estimación Planning Poker, la misma que es catalogada como una versión mejorada y moderada en comparación con las técnicas tradicionales, entonces esta estimación se realizara con la serie de Fibonacci.

Tabla 5: Estimación de historia de usuarios

Id	Historia de usuario	CH. G	Q. N	Valor Estimado
1	Realizar filtrado de datos	21	21	21
2	Obtener corpus	21	21	21
3	Conocer distancia y similitud entre textos.	13	8	13
4	Visualizar graficas	21	13	21
5	Actualizar información	8	13	13
6	Descargar corpus	13	8	13

Fuente: Los investigadores

11.3.6. PRODUCT BACKLOG

En esta etapa procederemos a ordenar las prioridades de las historias de usuario que son establecidas ya como funcionalidades del sistema a desarrollarse.

Tabla 6:Prioridades de las historias de usuario

Sprint	Tarea	Estimación de tiempo	Descripción
Prioridad de nivel alto			
1	Realizar filtrado de datos	3 semanas	Para realizar un control de información.
2	Obtener corpus	3 semanas	Para realizar un control de información.

3	Conocer distancia y similitud entre textos.	4 semanas	Para realizar un control de información personal.
4	Visualizar graficas	4 semanas	Se visualizará los gráficos de la información.
5	Actualizar información	3 semana	Para realizar un control de información personal.
6	Descargar corpus	1 semana	Para realizar un control de información académica.

Fuente: Los investigadores

11.3.7. HISTORIAS DE USUARIO DE LOS SPRINT'S

Continuando con el proceso de la metodología scrum, se procederá a analizar cada uno de los sprint´s con la finalidad de realizar una descripción corta y entendible para el usuario.

Tabla 7:Historia de usuario HU-001 Filtrado de datos

Historia de Usuario			
Número:	HU-001	**Usuario:**	Usuario
Nombre de Historia:		Filtrado de datos	
Prioridad:	A	**Responsable:**	Scrum team
Descripción: **Permite al usuario visualizar los artículos científicos mediante líneas, sublíneas de investigación y articulo científico en específico.**			
Observación: **Para realizar el filtrado de datos no se necesita pertenecer al sistema.**			

Fuente: Los investigadores

Tabla 8:Historia de usuario HU-002 Obtener corpus

Historia de Usuario			
Número:	HU-002	**Usuario:**	Usuario
Nombre de Historia:		obtener corpus	
Prioridad:	A	**Responsable:**	Scrum team
Descripción: **Permite al usuario obtener el corpus de un artículo científico.**			
Observación: **Para realizar el análisis del corpus se necesita que el articulo científico tiene que estar transformado a un documento con extensión .txt**			

Fuente: Los investigadores

Tabla 9:Historia de usuario HU-003Distancia y Similitud de textos

Historia de Usuario			
Número:	HU-003	**Usuario:**	Usuario
Nombre de Historia:		Distancia y similitud de textos	
Prioridad:	A	**Responsable:**	Scrum team
Descripción: **Permite al usuario conocer la distancia y similitud de textos analizados del documento.**			
Observación: **Para conocer el número de distancia de textos hay que obtener el corpus del documento.**			

Fuente: Los investigadores

Tabla 10:Historia de usuario HU-004 Visualizar graficas

Historia de Usuario			
Número:	HU-004	**Usuario:**	Usuario
Nombre de Historia:		Visualizar Graficas	
Prioridad:	A	**Responsable:**	Scrum team
Descripción: **Permite al usuario visualizar los gráficos de la información del documento.**			
Observación: **Para poder visualizar los gráficos se debe obtener el procesamiento del corpus.**			

Fuente: Los investigadores

Tabla 11:Historia de usuario HU-005 Actualizar información

Historia de Usuario			
Número:	HU-005	**Usuario:**	Usuario
Nombre de Historia:		Actualizar Información	
Prioridad:	A	**Responsable:**	Scrum team
Descripción: **Permite al usuario descargar el corpus del artículo científico.**			
Observación: **Para poder descargar el corpus del articulo científico se debe obtener el procesamiento del corpus.**			

Fuente: Los investigadores

Tabla 12:Historia de usuario HU-006 Descargar Corpus

Historia de Usuario			
Número:	HU-006	**Usuario:**	Usuario

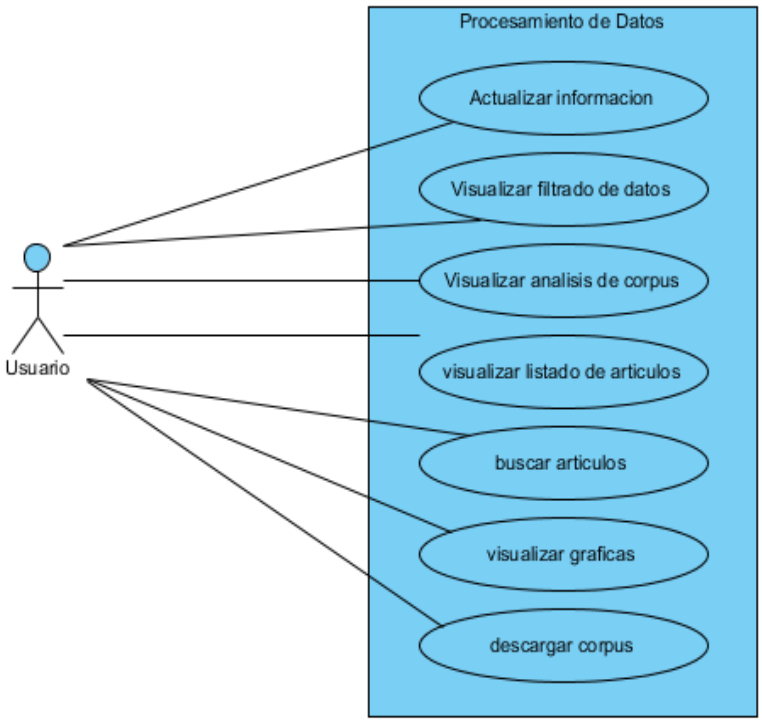
Nombre de Historia:		Descarga del corpus	
Prioridad:	A	**Responsable:**	Scrum team
Descripción: **Permite al usuario descargar el corpus del artículo científico.**			
Observación: **Para poder descargar el corpus del articulo científico se debe obtener el procesamiento del corpus.**			

Fuente: Los investigadores

11.3.8. CASOS DE USO GENERAL

Para generar los casos de uso se tomó en consideración todos los procesos que forman parte de nuestro modulo procesamiento de datos, los cuales serán representados mediante el diagrama de los casos de uso los cuales son: actualizar información, visualizar filtrado de datos, visualizar análisis de corpus, visualizar listado de artículos, buscar artículos, visualizar gráficas, descargar corpus.

Figura 29: Casos de uso general



Fuente: Los investigadores

11.3.9. CASOS DE USO A DETALLE DE LOS SPRINT'S

Con la finalidad de detallar los procesos que se realizan en los sprint´s del sistema se crea una tabla para conocer los responsables y lo que hará cada uno.

Tabla 13: Casos de uso a detalle CU-001 Filtrado de datos

Nº caso	**CU-001**
H. U	HU-001: filtrado de datos
Nombre	filtrado de datos
Autor	Equipo Scrum
Descripción: **Permite al usuario buscar los artículos científicos mediante líneas, sublíneas de investigación y articulo científico en específico.**	
Actores: **Usuario**	
Precondición: **el documento a analizar debe estar en la base de datos del sistema**	
Flujo Normal: Procesamiento de datos por línea de investigación 1. **El usuario ingresa a la plataforma científica.** 2. **La plataforma despliega interfaz de inicio.** 3. **El usuario busca en el menú el módulo procesamiento de datos y da click.** 4. **La plataforma muestra la interfaz del módulo procesamiento de datos.** 5. **El usuario selecciona la opción procesamiento de datos por línea de investigación.** 6. **La plataforma despliega la interfaz inicial del procesamiento de datos.** 7. **El usuario selecciona una opción de zona.** 8. **La plataforma carga las universidades que pertenecen a la zona.** 9. **El usuario selecciona una universidad.** 10. **La plataforma carga las líneas de investigación que tiene la universidad.** 11. **El usuario selecciona una línea de investigación.** 12. **La plataforma carga los todos los artículos que estén relacionados con la línea de investigación seleccionada.** 13. **El usuario selecciona una sublínea de investigación.** 14. **La plataforma carga todos los artículos que estén relacionados con la sublínea de investigación.**	
Flujo Alternativo 1: Procesamiento de datos por articulo científico. 5. **El usuario selecciona la opción procesamiento de datos por articulo científico.** 6. **La plataforma despliega la interfaz inicial del procesamiento de datos.** 7. **El usuario selecciona una opción de zona.** 8. **La plataforma carga las universidades que pertenecen a la zona.** 9. **El usuario selecciona una universidad.** 10. **La plataforma carga las líneas de investigación que tiene la universidad.** 11. **El usuario selecciona una línea de investigación.** 12. **La plataforma carga las sublineas de investigación, artículos que pertenecen a la línea de investigación.**	

13. **El usuario selecciona un artículo especifico** 14. **La plataforma carga el detalle del articulo y muestra los artículos que tengan relación con el documento seleccionado.**

Fuente: Los investigadores

Tabla 14: Casos de uso a detalle CU-02 Obtener corpus

Nº caso	**CU-002**
H. U	HU-002: obtener corpus
Nombre	obtener corpus
Autor	Equipo Scrum
Descripción: **Permite al usuario obtener el corpus de un artículo científico o línea de investigación.**	
Actores: **Usuario**	
Precondición: **El usuario debe realizar el filtrado de datos.**	
Flujo Normal: Análisis del corpus por línea de investigación. **I. La plataforma busca y carga los documentos que estén relacionados con la línea de investigación seleccionada, filtra los documentos de extensión .pdf, transforma los documentos filtrados a archivos. txt y obtiene el corpus de los archivos transformados.** **II. El usuario selecciona la opción análisis de datos.** **III. La plataforma muestra el análisis de datos de la línea de investigación como:** ✓ **Numero de palabras.** ✓ **Número de palabras (sin palabras de parada).** ✓ **Número de palabras (con palabras de parada).** ✓ **Riqueza léxica.**	
Flujo alterno: Análisis del corpus por articulo científico. **1. La plataforma busca y carga los documentos que estén relacionados con la línea de investigación seleccionada y la sublínea de investigación de cada artículo en específico.** **2. El usuario selecciona la opción análisis de datos.** 3. **La plataforma muestra en una tabla el análisis de datos de un artículo científico seleccionado como:** ✓ **Numero de palabras del articulo científico.** ✓ **Número de palabras (sin palabras de parada).** ✓ **Número de palabras (con palabras de parada).**	

✓ **Riqueza léxica del articulo científico.**

Fuente: Los investigadores

Tabla 15: Casos de uso a detalle CU-003 Distancia y Similitud

Nº caso	**CU-003**
H. U	HU-003: Distancia y similitud
Nombre	Distancia y similitud
Autor	Equipo Scrum
Descripción: **Permite al usuario obtener el corpus de un artículo científico.**	
Actores: **Usuario**	
Precondición: **El usuario debe realizar el filtrado de datos.**	
Flujo Normal: **I. La plataforma busca y carga los documentos que estén relacionados con la línea de investigación seleccionada, filtra los documentos de extensión .pdf, transforma los documentos filtrados a archivos. txt y obtiene el corpus de los archivos transformados.** **II. El usuario selecciona el articulo y da click en artículos relacionados.** **III. La plataforma muestra el detalle del artículo en una tabla.**	

Fuente: Los investigadores

Tabla 16: Casos de uso a detalle CU-004 Visualizar gráficas

Nº caso	**CU-004**
H. U	HU-004: Visualizar graficas
Nombre	Visualizar graficas
Autor	Equipo Scrum
Descripción: **Permite al usuario visualizar los gráficos de la información del documento por línea de investigación o articulo científico.**	
Actores: **Usuario**	
Precondición: **El usuario debe realizar el filtrado de datos.**	
Flujo Normal: **1. El usuario selecciona la opción línea de investigación.** **2. La plataforma muestra el filtrado de datos.** **3. El usuario selecciona la opción graficas.** **4. La plataforma muestra las gráficas con palabras de parada y sin palabras de parada, teniendo en cuenta los diferentes tipos de gráficos como:** ✓ **Gráfico de línea.**	

✓ **Gráfico de barras.** ✓ **Gráfico de radar.** ✓ **Gráfico de dona.** ✓ **Gráfico de pastel.** ✓ **Gráfico de área polar.** ✓ **Gráfico Word cloud.** ✓ **Gráfico escalamiento multidimensional.** ✓ **Gráfico similaridad y distancia.**
Flujo Alterno: **5. El usuario selecciona la opción artículo científico.** **6. La plataforma muestra el filtrado de datos.** **7. El usuario selecciona la opción graficas.** **8. La plataforma muestra las gráficas con palabras de parada y sin palabras de parada, teniendo en cuenta los diferentes tipos de gráficos como:** ✓ **Gráfico de línea.** ✓ **Gráfico de barras.** ✓ **Gráfico de radar.** ✓ **Gráfico de dona.** ✓ **Gráfico de pastel.** ✓ **Gráfico de área polar.** ✓ **Gráfico Word cloud.** ✓ **Gráfico escalamiento multidimensional.** ✓ **Gráfico similaridad y distancia.**

Fuente: Los investigadores

Tabla 17: Casos de uso a detalle CU-006 Descargar Corpus

Nº caso	**CU-005**
H. U	HU-006: Descargar corpus
Nombre	Descargar corpus
Autor	Equipo Scrum
Descripción: **Permite al usuario descargar el corpus de un artículo científico o de la línea de investigación.**	
Actores: **Usuario**	
Precondición: **El usuario debe realizar el filtrado de datos.**	

Flujo Normal:

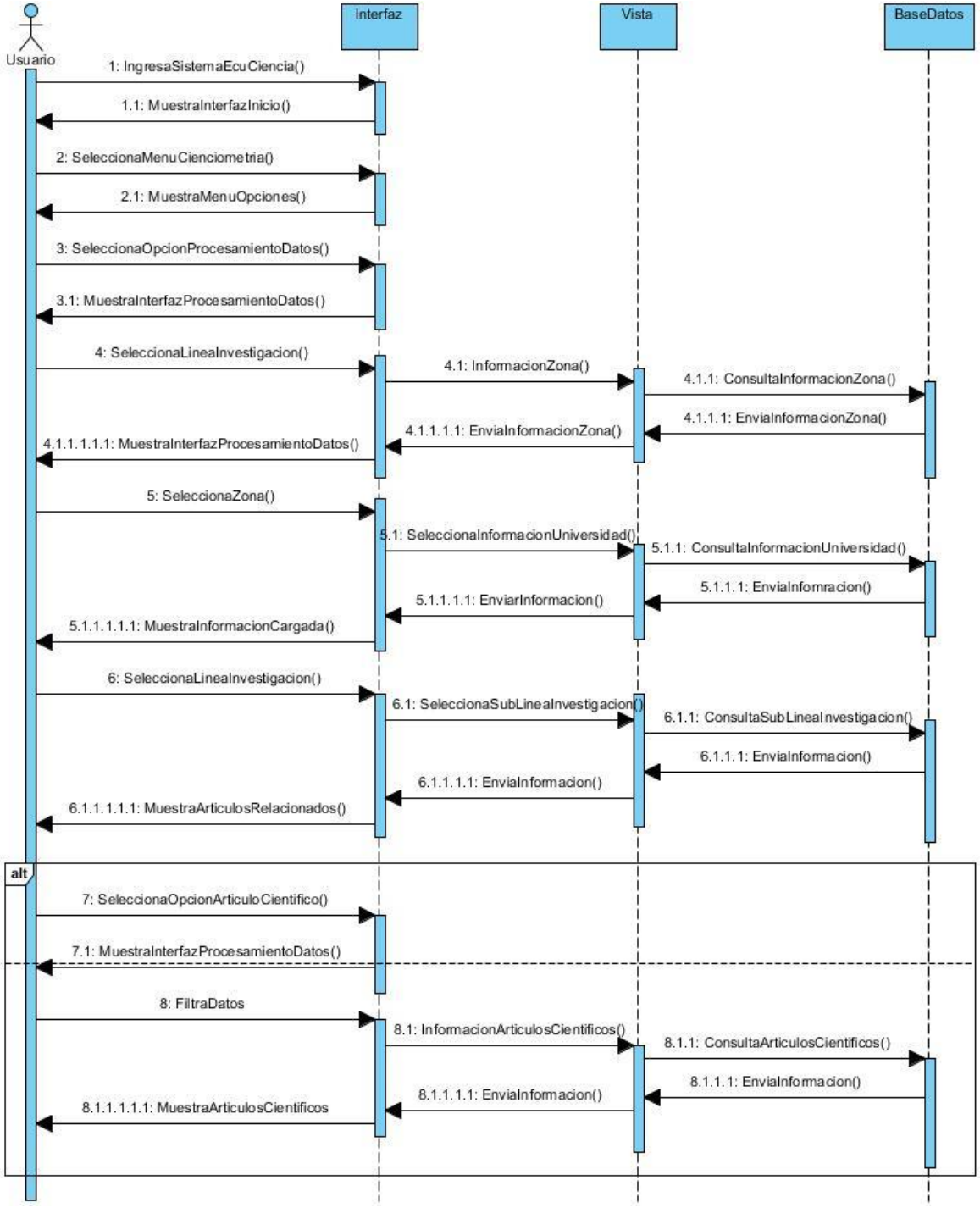
I. **La plataforma busca y carga los documentos que estén relacionados con la línea de investigación o articulo científico seleccionado, filtra los documentos de extensión .pdf, transforma los documentos filtrados a archivos. txt y obtiene el corpus de los archivos transformados.**
II. **El usuario selecciona en la opción ver análisis de datos.**
III. **La plataforma muestra el detalle del artículo en un modal con las opciones descargar articulo o descargar corpus.**
IV. **El usuario da click y se descarga.**

Fuente: Los investigadores

11.3.10. DIAGRAMAS DE SECUENCIA

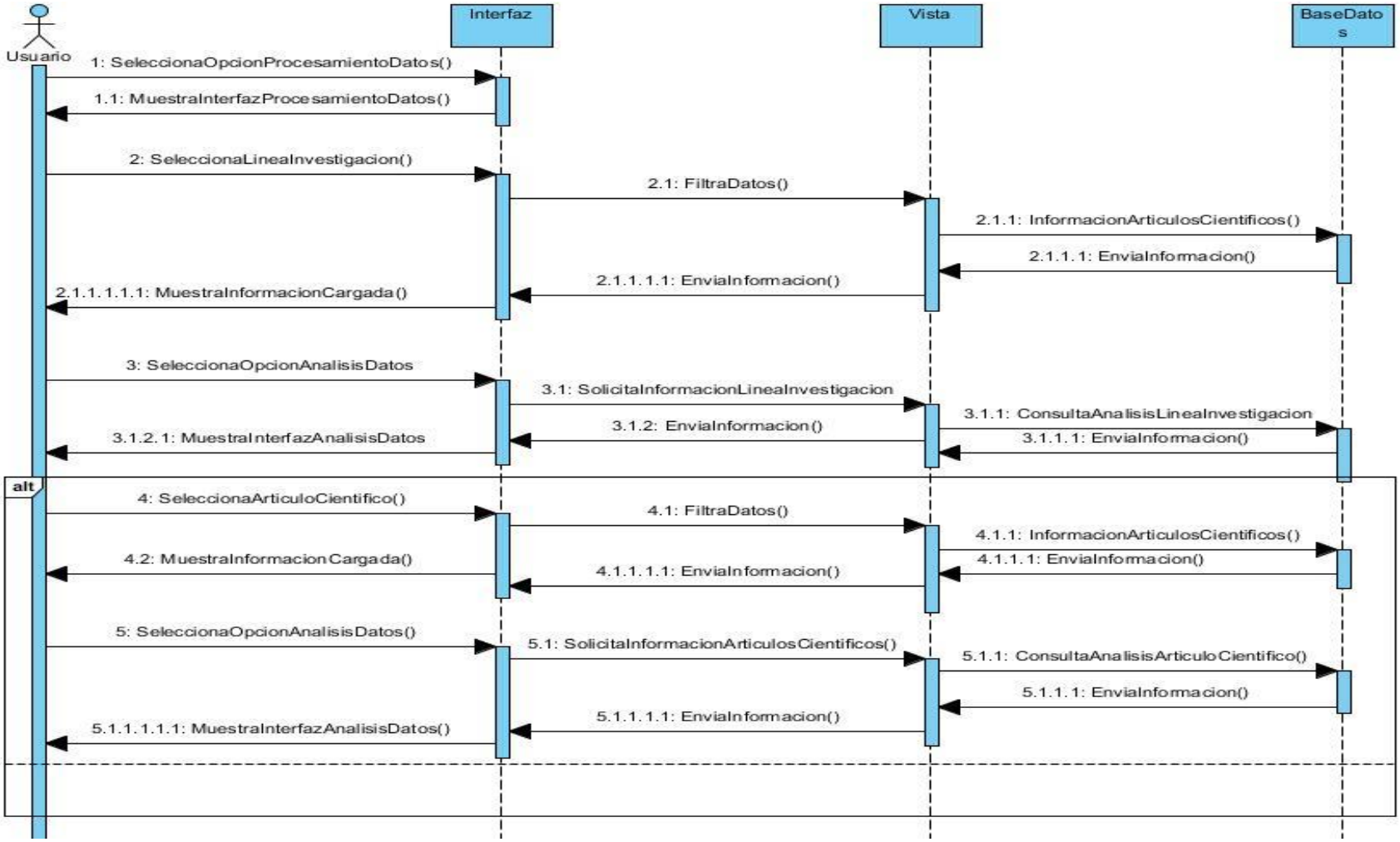
Se procede a realizar los diagramas de secuencia para observar la preparativa cronológica de cada uno de los procesos de cada sprint´s.

Figura 30: Diagrama de secuencia de Filtrado de datos



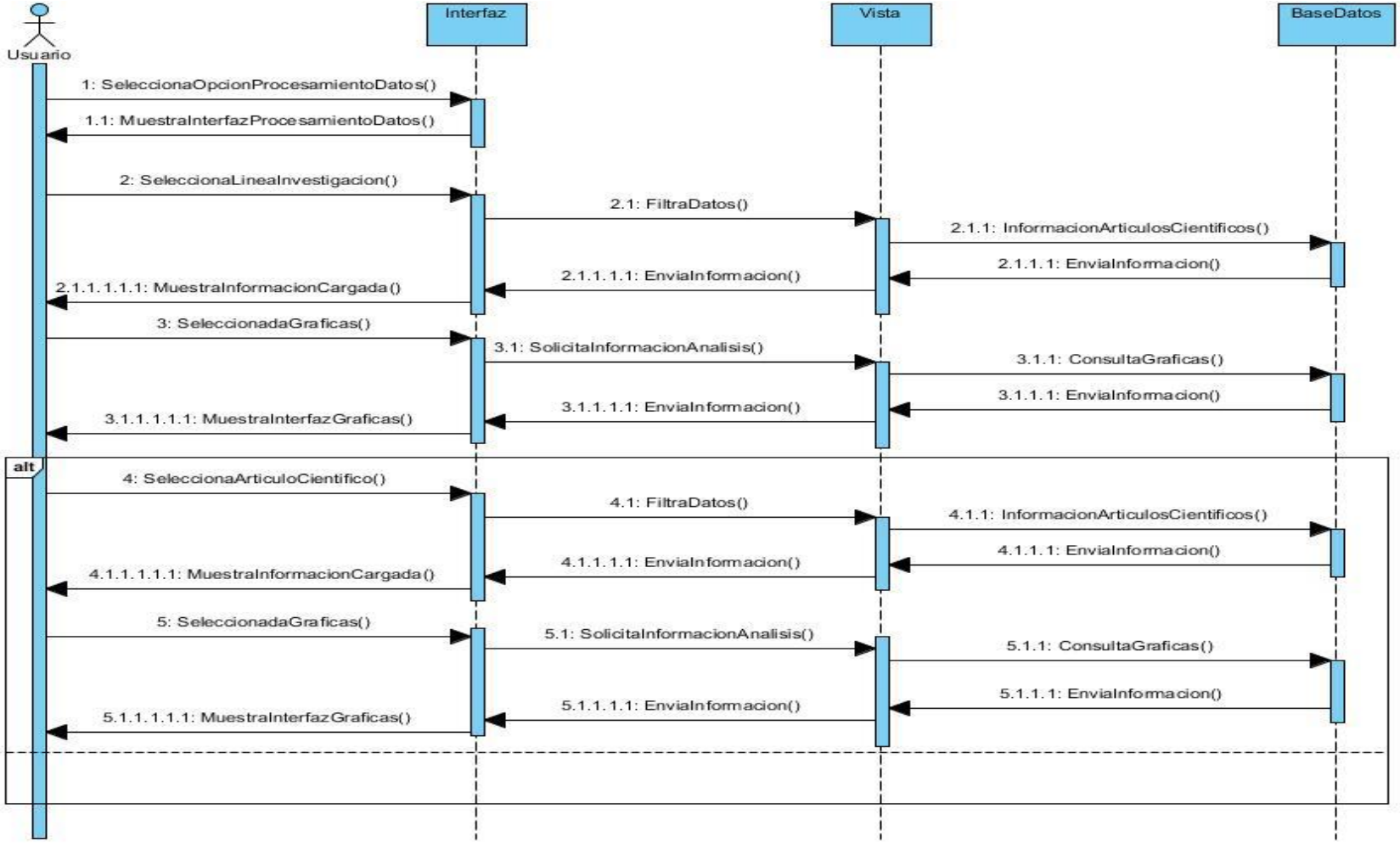
Fuente: Los investigadores

Figura 31: Diagrama de secuencia Obtener corpus



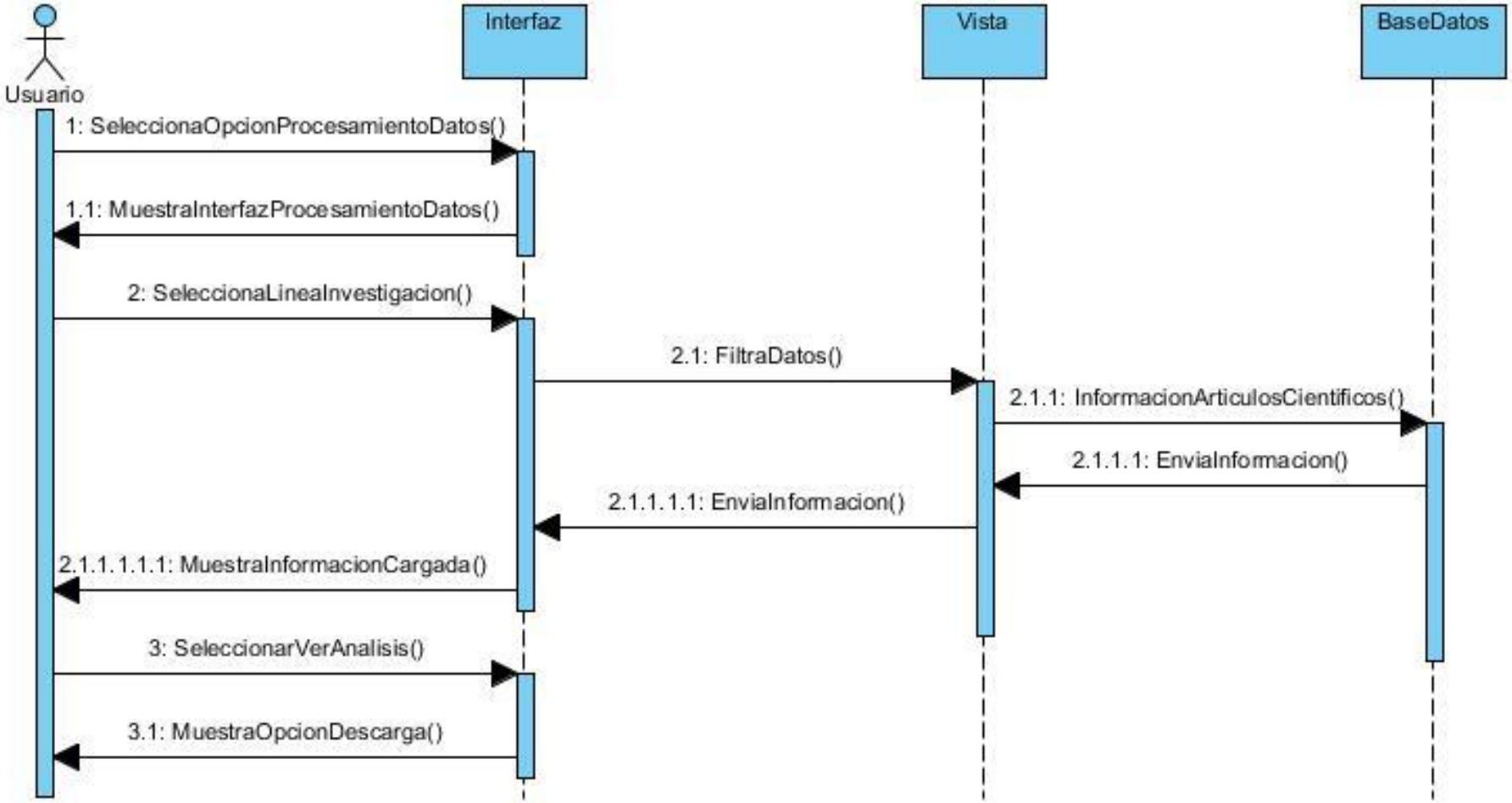
Fuente: Los investigadores

Figura 32: Diagrama de secuencia Visualizar gráficas



Fuente: Los investigadores

Figura 33: Descargar Corpus

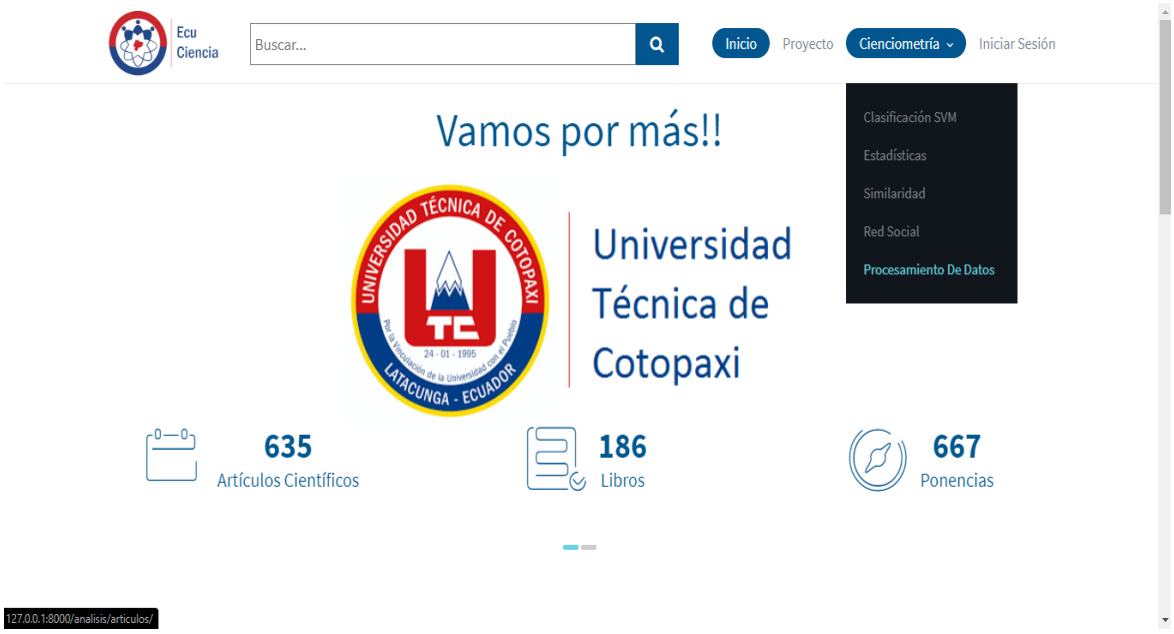


Fuente: Los investigadores

11.3.11. IMPLEMENTACIÓN DE LOS SPRINT'S

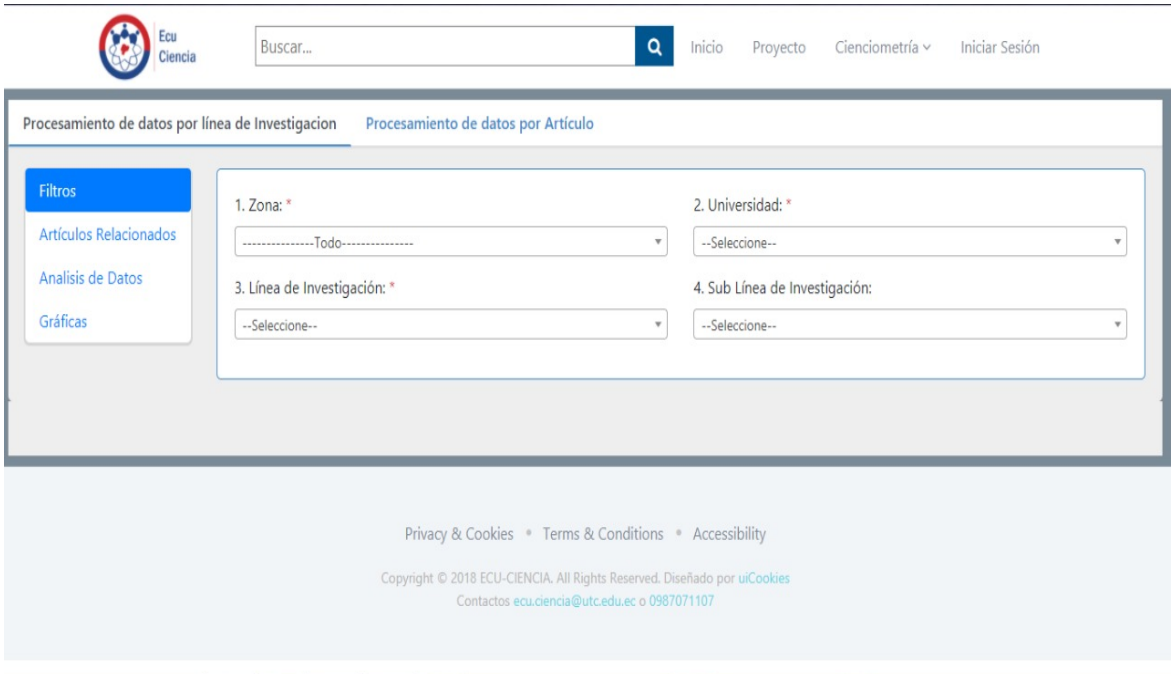
Para darle forma al proyecto que se está desarrollando se muestra la codificación de cada sprint lo que nos llevara a realizar las diferentes pruebas con las funcionalidades correspondientes en el sistema para poder implementar y ejecutar en el sistema.

Figura 34: Pantalla Principal del Sistema EcuCiencia



Fuente: Los investigadores

Figura 35: Modulo de procesamiento de datos-Filtro de datos



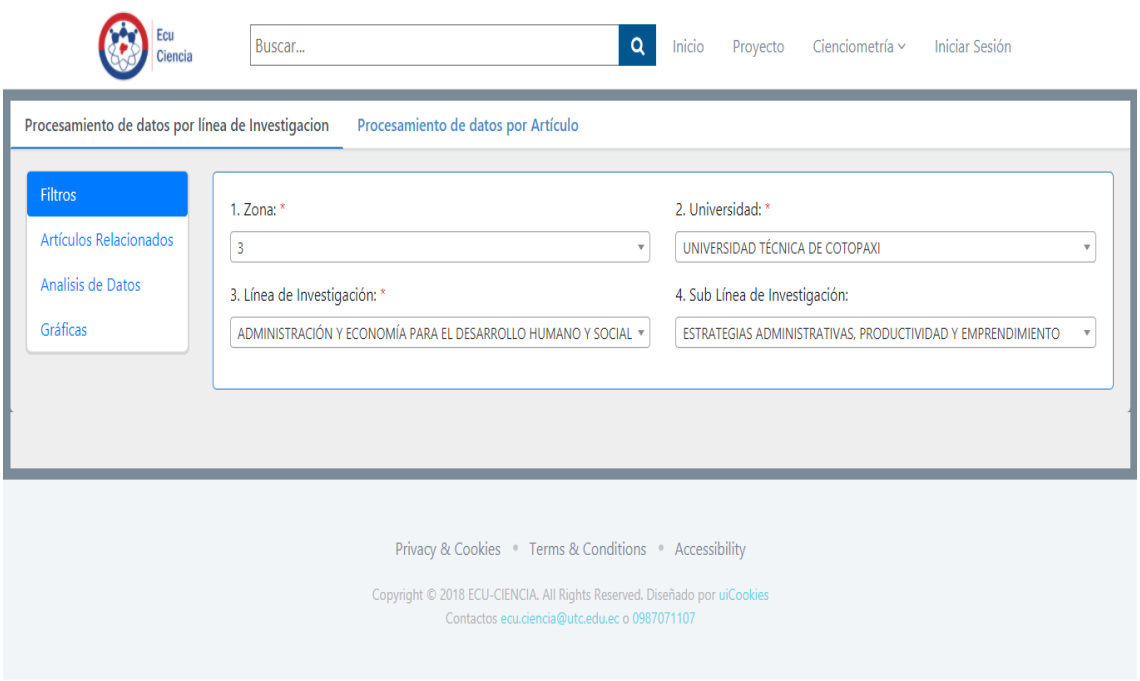
Fuente: Los investigadores

FILTRADO DE DATOS

Figura 36:Desarrollo del código del Sprint 1

Fuente: Los investigadores

Figura 37: Interfaz gráfica del filtrado de datos



Fuente: Los investigadores

En la figura 7 y 8 se muestra la codificación y la interfaz gráfica que se realizó para el filtrado de datos para obtener el corpus.

Figura 38: Interfaz gráfica por línea de investigación

Fuente: Los investigadores

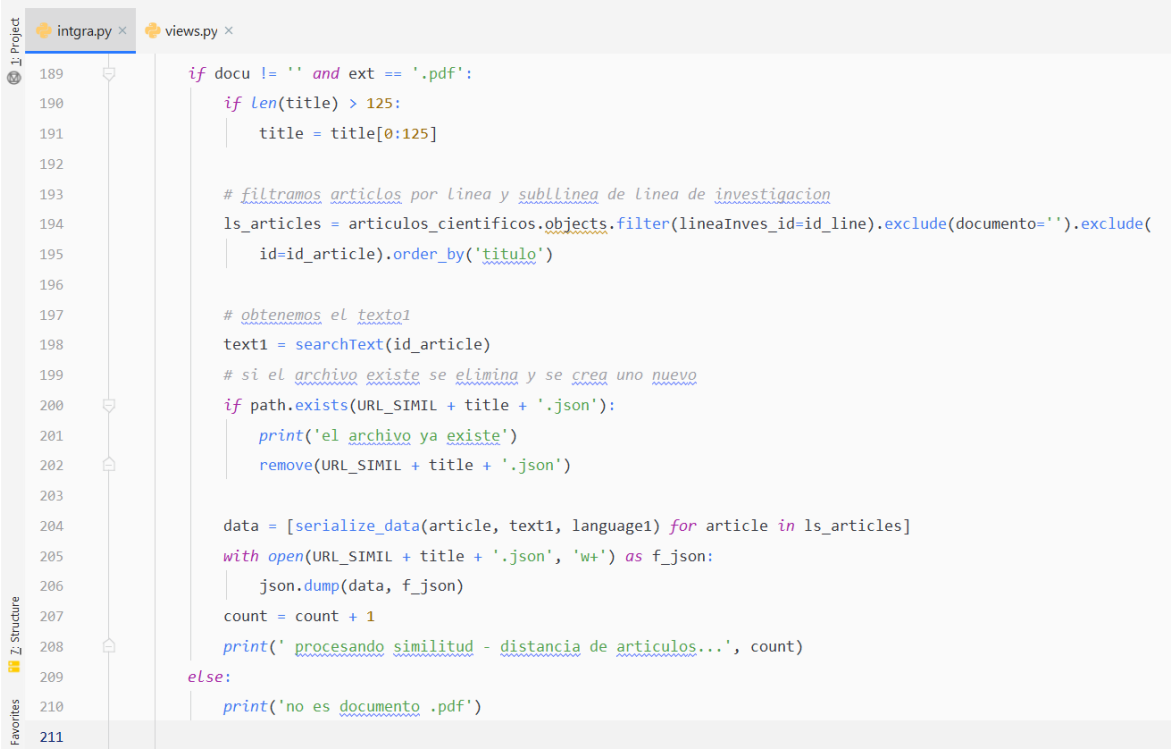
Figura 39: interfaz gráfica por articulo científico.

Fuente: Los investigadores

En la figura 9 y 10 se muestra la interfaz gráfica tanto por línea de investigación como por articulo científico de la obtención de la similitud o relación de los artículos.

LISTADO DE ARTÍCULOS SIMILARES

Figura 40: listado de artículos similares

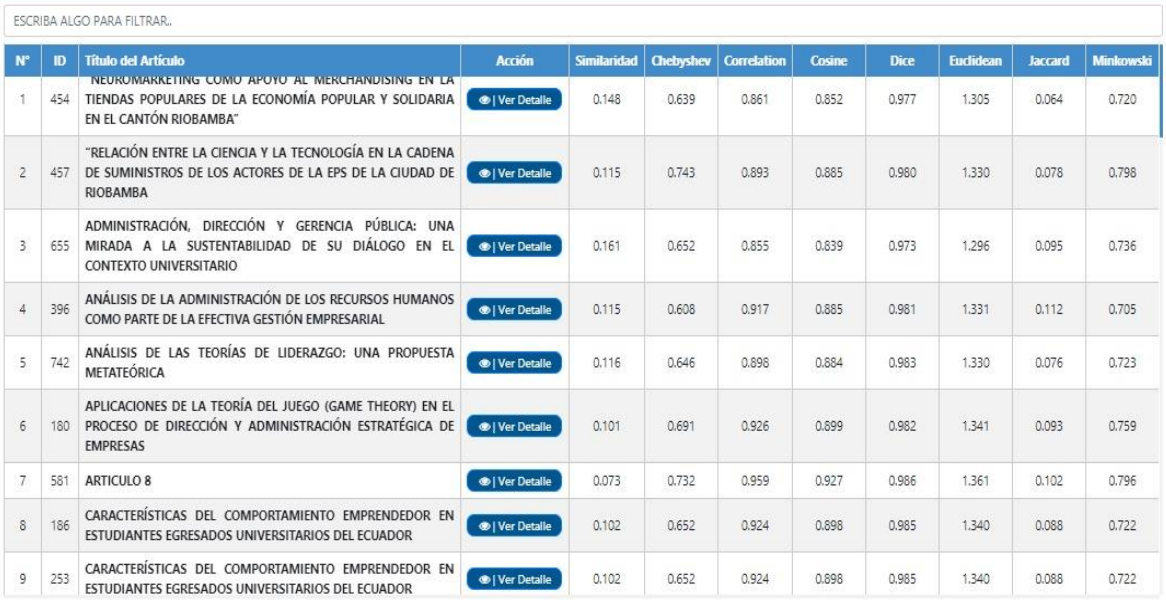


```
intgra.py × views.py ×

189     if docu != '' and ext == '.pdf':
190         if len(title) > 125:
191             title = title[0:125]
192
193         # filtramos articlos por linea y sublinea de linea de investigacion
194         ls_articles = articulos_cientificos.objects.filter(lineaInves_id=id_line).exclude(documento='').exclude(
195             id=id_article).order_by('titulo')
196
197         # obtenemos el texto1
198         text1 = searchText(id_article)
199         # si el archivo existe se elimina y se crea uno nuevo
200         if path.exists(URL_SIMIL + title + '.json'):
201             print('el archivo ya existe')
202             remove(URL_SIMIL + title + '.json')
203
204         data = [serialize_data(article, text1, language1) for article in ls_articles]
205         with open(URL_SIMIL + title + '.json', 'w+') as f_json:
206             json.dump(data, f_json)
207         count = count + 1
208         print(' procesando similitud - distancia de articulos...', count)
209     else:
210         print('no es documento .pdf')
211
```

Fuente: Los investigadores

Figura 41: Interfaz gráfica de listado de artículos similares



ESCRIBA ALGO PARA FILTRAR..

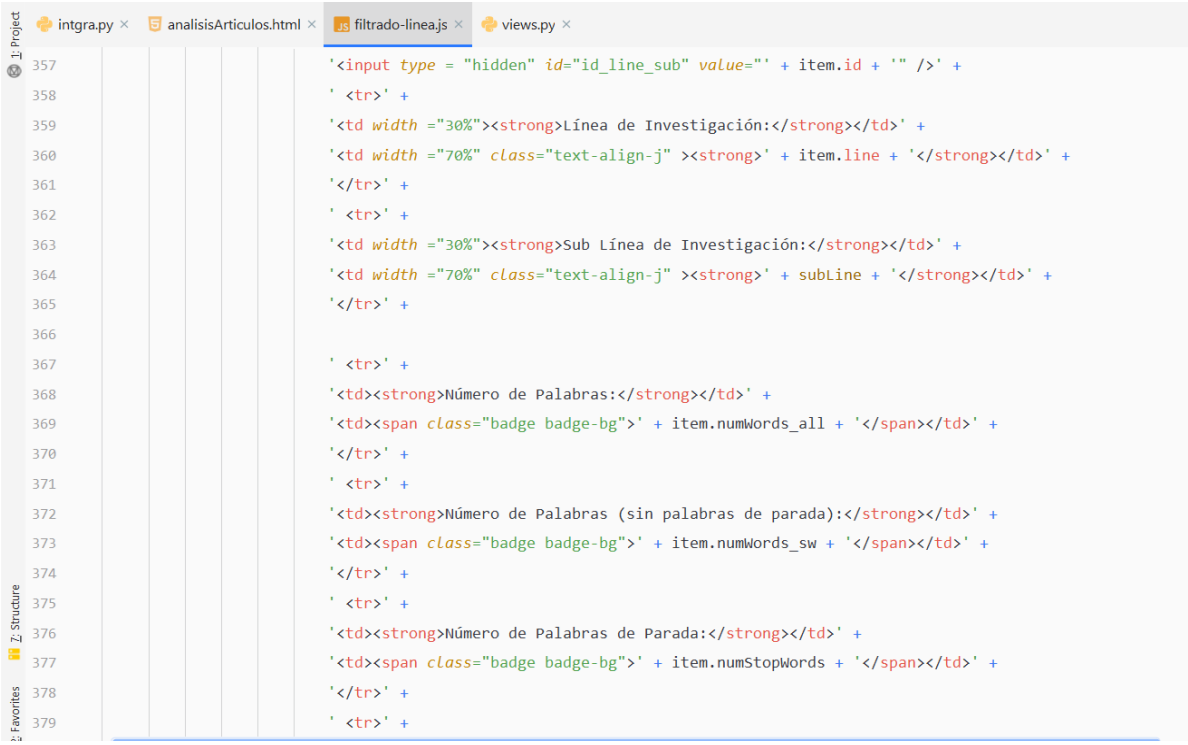
N°	ID	Título del Artículo	Acción	Similaridad	Chebyshev	Correlation	Cosine	Dice	Euclidean	Jaccard	Minkowski
1	454	"NEUROMARKETING COMO APOYO AL MERCHANDISING EN LA TIENDAS POPULARES DE LA ECONOMÍA POPULAR Y SOLIDARIA EN EL CANTÓN RIOBAMBA"	Ver Detalle	0.148	0.639	0.861	0.852	0.977	1.305	0.064	0.720
2	457	"RELACIÓN ENTRE LA CIENCIA Y LA TECNOLOGÍA EN LA CADENA DE SUMINISTROS DE LOS ACTORES DE LA EPS DE LA CIUDAD DE RIOBAMBA	Ver Detalle	0.115	0.743	0.893	0.885	0.980	1.330	0.078	0.798
3	655	ADMINISTRACIÓN, DIRECCIÓN Y GERENCIA PÚBLICA: UNA MIRADA A LA SUSTENTABILIDAD DE SU DIÁLOGO EN EL CONTEXTO UNIVERSITARIO	Ver Detalle	0.161	0.652	0.855	0.839	0.973	1.296	0.095	0.736
4	396	ANÁLISIS DE LA ADMINISTRACIÓN DE LOS RECURSOS HUMANOS COMO PARTE DE LA EFECTIVA GESTIÓN EMPRESARIAL	Ver Detalle	0.115	0.608	0.917	0.885	0.981	1.331	0.112	0.705
5	742	ANÁLISIS DE LAS TEORÍAS DE LIDERAZGO: UNA PROPUESTA METATEÓRICA	Ver Detalle	0.116	0.646	0.898	0.884	0.983	1.330	0.076	0.723
6	180	APLICACIONES DE LA TEORÍA DEL JUEGO (GAME THEORY) EN EL PROCESO DE DIRECCIÓN Y ADMINISTRACIÓN ESTRATÉGICA DE EMPRESAS	Ver Detalle	0.101	0.691	0.926	0.899	0.982	1.341	0.093	0.759
7	581	ARTICULO 8	Ver Detalle	0.073	0.732	0.959	0.927	0.986	1.361	0.102	0.796
8	186	CARACTERÍSTICAS DEL COMPORTAMIENTO EMPRENDEDOR EN ESTUDIANTES EGRESADOS UNIVERSITARIOS DEL ECUADOR	Ver Detalle	0.102	0.652	0.924	0.898	0.985	1.340	0.088	0.722
9	253	CARACTERÍSTICAS DEL COMPORTAMIENTO EMPRENDEDOR EN ESTUDIANTES EGRESADOS UNIVERSITARIOS DEL ECUADOR	Ver Detalle	0.102	0.652	0.924	0.898	0.985	1.340	0.088	0.722

Fuente: Los investigadores

En la figura 11 y 12 muestra la codificación y la interfaz gráfica de cómo se logra obtener el listado de artículos a través de las diferentes librerías y lógica de programación establecidas para realizar el algoritmo y el desarrollo del ambiente web para el usuario.

VISUALIZAR ANÁLISIS DEL CORPUS POR LÍNEA DE INVESTIGACIÓN

Figura 42: Desarrollo del código del Sprint 2



Fuente: Los investigadores

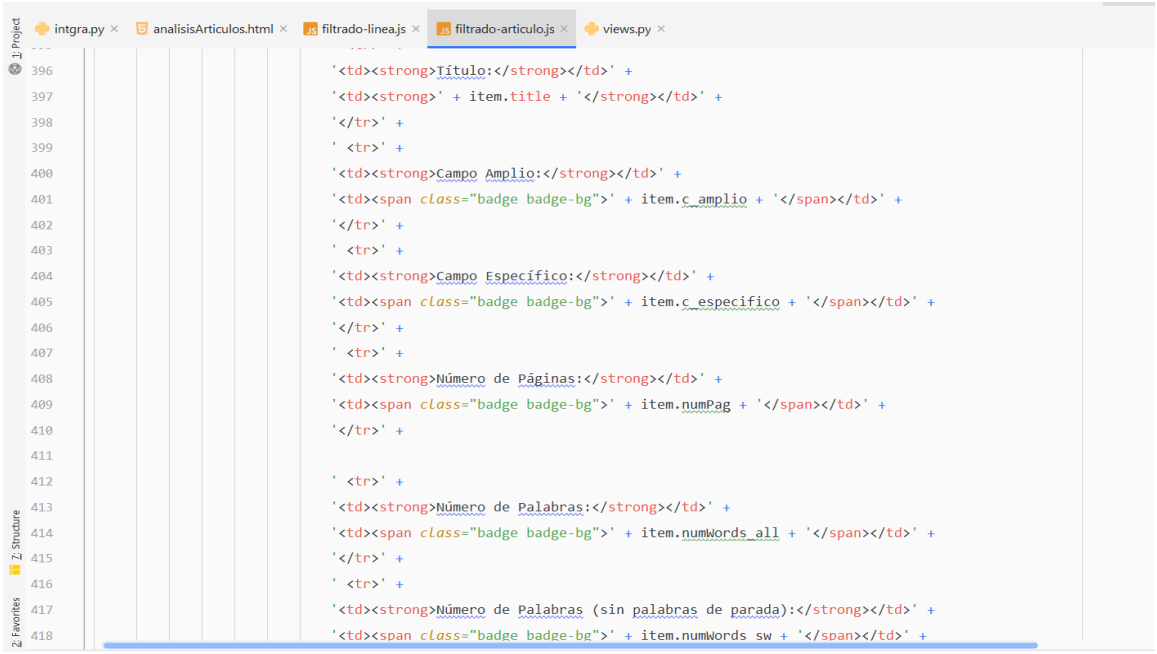
Figura 43: Interfaz gráfica del análisis de datos- línea de investigación

Fuente: Los investigadores

En las figuras 42 y 43 podemos observar el análisis del corpus completo de una línea de investigación elegida aleatoriamente que consta de línea de investigación, sub línea de investigación, numero de palabras sin parada y con parada, riqueza léxica, corpus que estará listo para descargarlo.

VER ANÁLISIS DEL CORPUS POR ARTICULO CIENTÍFICO

Figura 44: Codificación del análisis por articulo científico



Fuente: Los investigadores

Figura 45: Interfaz gráfica del análisis de datos- artículo científico

Fuente: Los investigadores

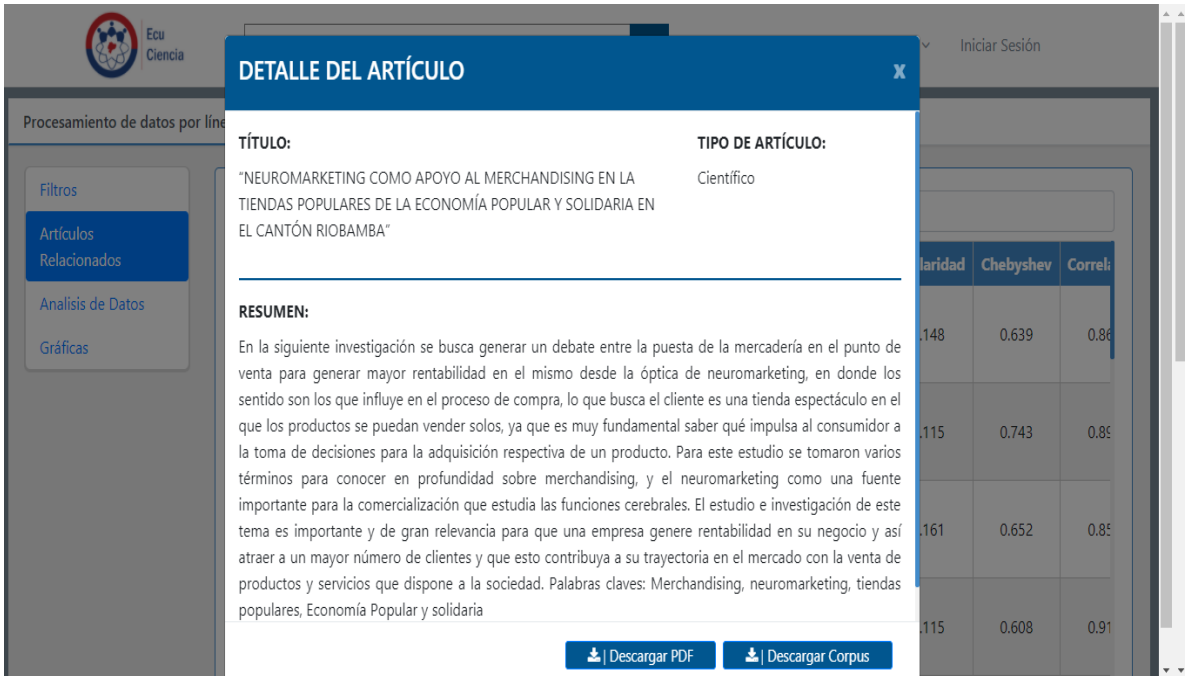
En las siguientes figuras se muestran la información del articulo científico ya analizado obteniendo resultados rápidos y seguros dados por articulo científico.

VER DETALLE DEL ARTICULO CIENTÍFICO Y DESCARGAR CORPUS

Figura 46:Ver detalle del articulo y descarga de corpus

Fuente: Los investigadores

Figura 47: Interfaz Gráfica de Ver detalle del articulo y descarga de corpus



Fuente: Los investigadores

Se observa el detalle de un artículo científico seleccionado que consta de título, tipo de articulo, resumen y a su vez los botones de descargar el articulo y descargar el corpus.

Figura 48: Codificación de Actualizar información (Pantalla de bloqueo)

Fuente: Los investigadores

Figura 49: Interfaz Gráfica de Actualizar información (Pantalla de bloqueo)

Fuente: Los investigadores

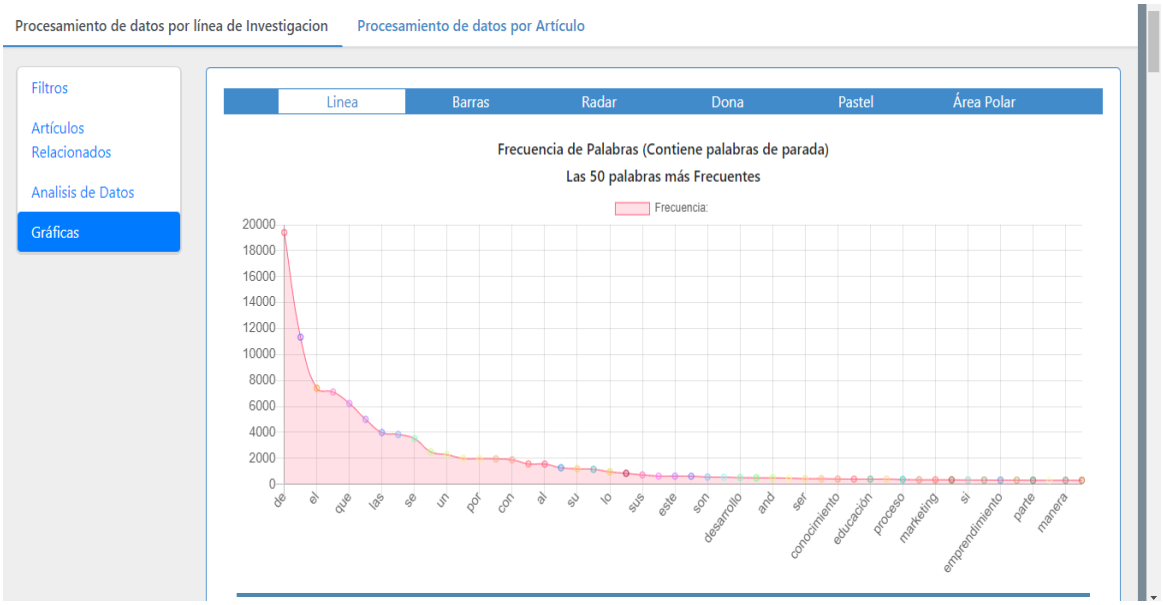
Esta interfaz gráfica se cargará solamente cuando el servidor entre en mantenimiento o actualización de información que se dará diariamente en un horario establecido para que no exista conflictos al momento de verificar o analizar un artículo científico y que el usuario no tenga que esperar el análisis.

VISUALIZAR GRÁFICAS POR LÍNEA DE INVESTIGACIÓN

Figura 50: Codificación de graficas con palabras de parada

Fuente: Los investigadores

Figura 51: Gráfica de línea con palabras de parada



Fuente: Los investigadores

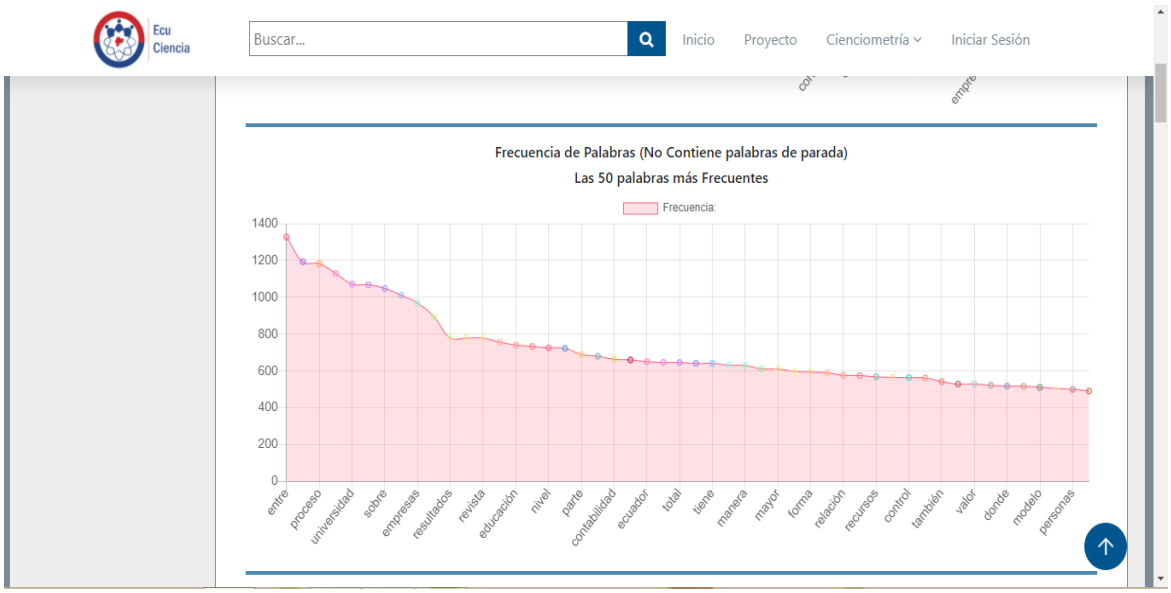
En las siguientes figuras se muestra la codificación y la interfaz gráfica de cómo se mostrará la gráfica de la frecuencia de palabras que contienen palabras de parada utilizando diferentes tipos de gráficos para la interpretación de los resultados.

Figura 52: Codificación de la gráfica sin palabras de parada


```
1145
1146  # funcion para obtener la grafica de las palabras mas frecuentes sin palabras de parada
1147  def line_graphic_2(request):
1148      name_l = ''
1149      if request.method == 'POST':
1150          data = request.POST.get('datos')  # idLinea
1151          data1 = request.POST.get('datos1')  # idSubLinea
1152
1153          if data1 and int(data1) > 0:
1154              corpus_line = line_processing.objects.get(id_line=int(data), id_sub_line_id=int(data1))
1155          else:
1156              corpus_line = line_processing.objects.get(id_line=int(data), id_sub_line_id=None)
1157
1158          name_l = corpus_line.name_file
1159
1160          url = '/media/analisis/json_line_text/' + name_l + '.json'
1161
1162          data = {
1163              "url_json": url
1164          }
1165
1166      return JsonResponse(data)
```

Fuente: Los Investigadores

Figura 53: gráfica de línea sin palabras de parada.



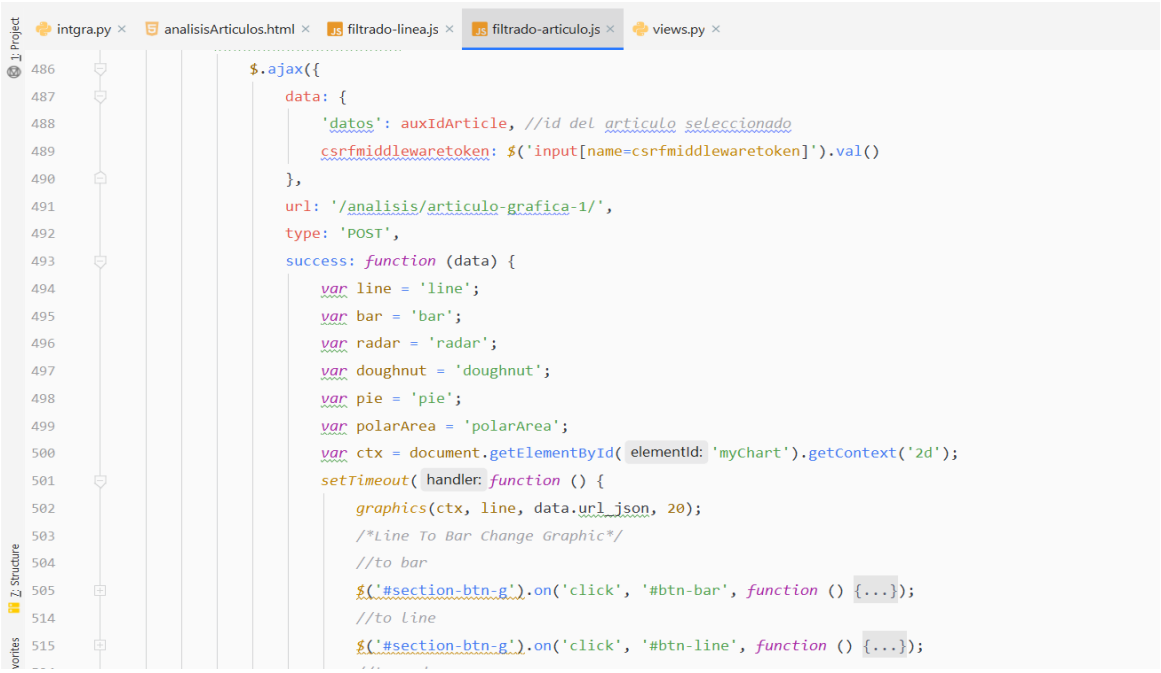
Fuente: Los Investigadores

Para complementar las gráficas de la frecuencia de palabras se utilizó diferentes tipos de graficas para comprender mejor los resultados, tomando en cuenta que para mostrar los datos analizados por línea de investigación se tomó en cuenta las 50 palabras más

relevantes de cada línea de investigación. Teniendo claro esto se muestran las diferentes graficas como: línea, barras, radar, dona, pastel, área polar y un Word cloud.

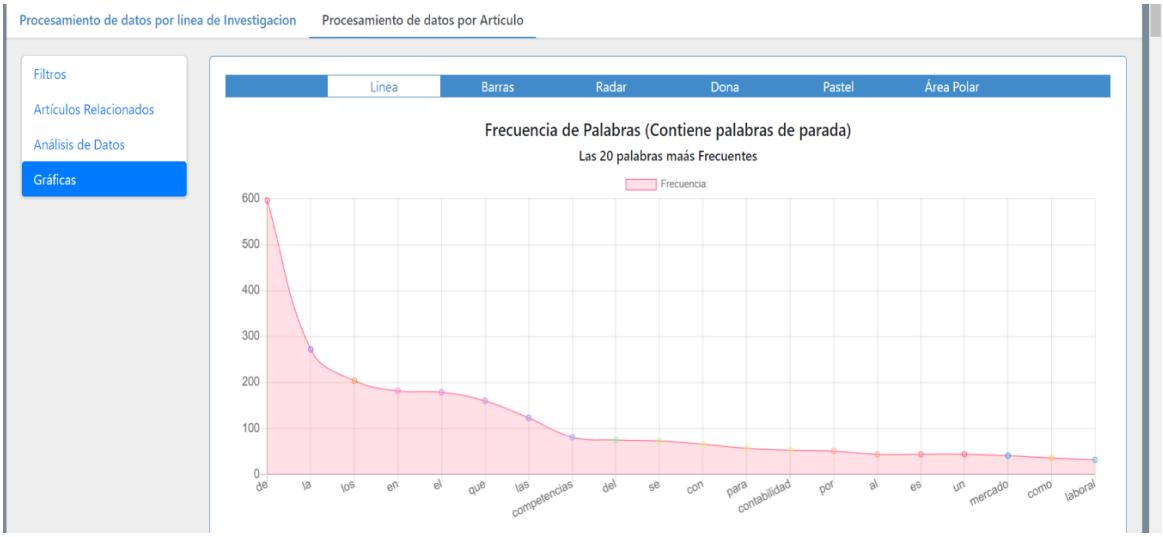
VISUALIZAR GRÁFICAS POR ARTÍCULO CIENTÍFICO

Figura 54: Codificación de la gráfica con palabras de parada



Fuente: Los Investigadores

Figura 55: Gráfica de la frecuencia de palabras con palabras de parada.

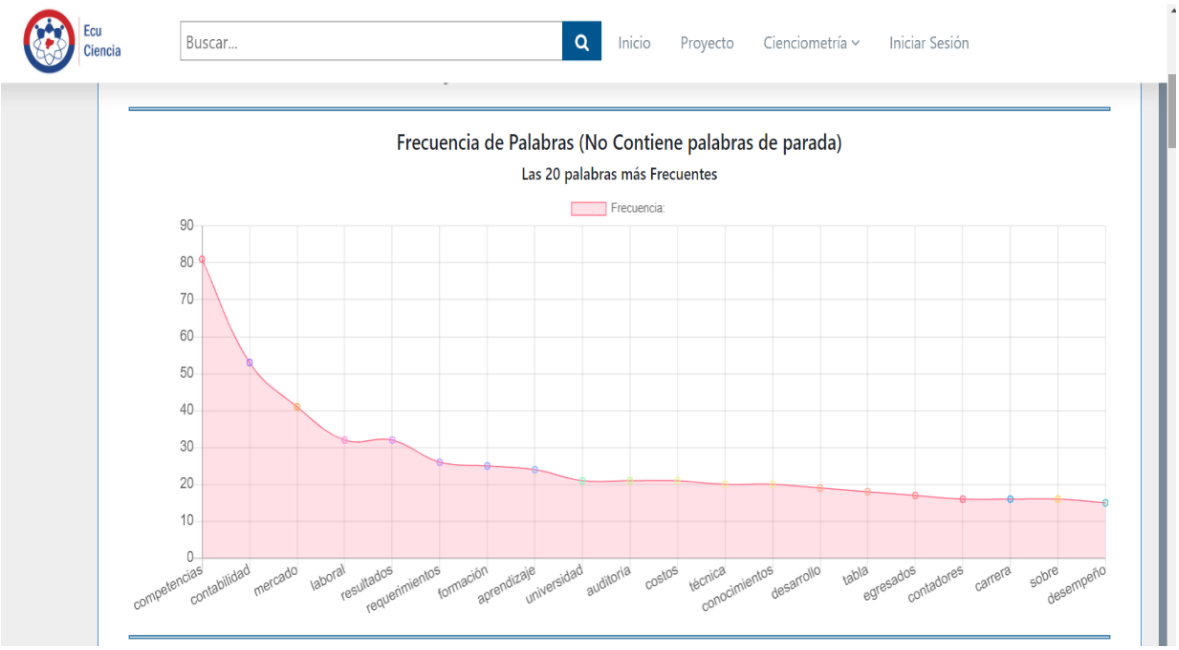


Fuente: Los Investigadores

Figura 56: Codificación de la frecuencia de palabras sin palabras de parada

Fuente: Los investigadores

Figura 57: gráfica de la frecuencia de palabras sin palabras de parada



Fuente: Los investigadores

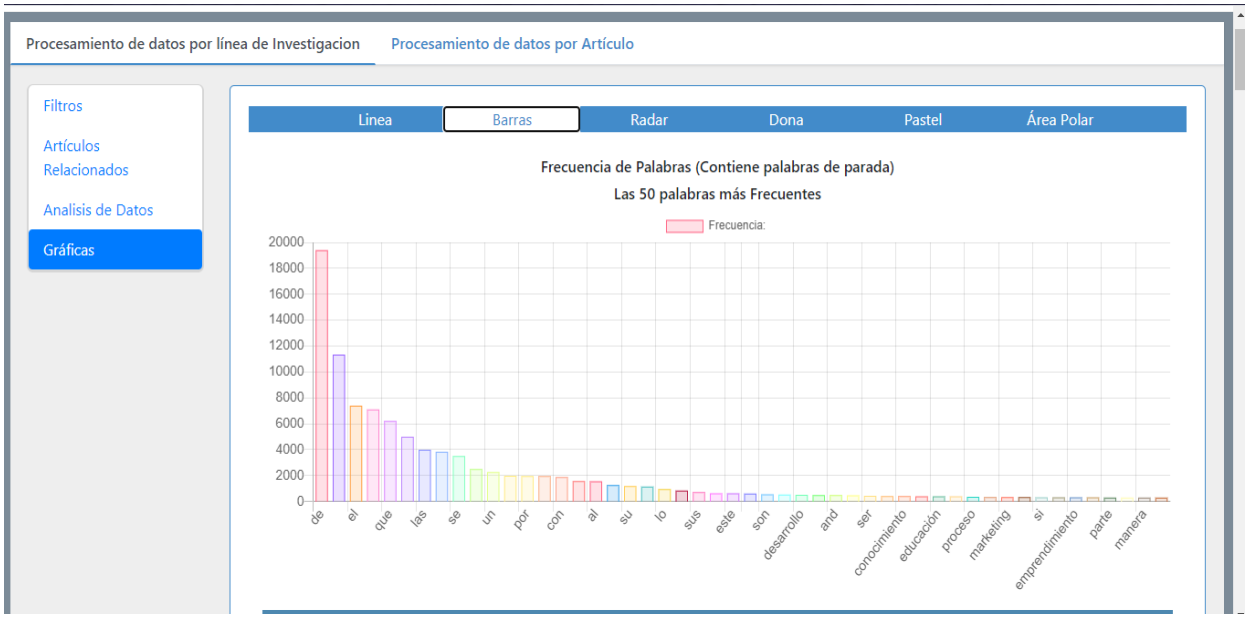
En las siguientes figuras se muestran la codificación y las interfaces graficas de cómo se obtuvieron los resultados para representarlos en una gráfica que se tomaron 20 palabras

más frecuentes de cada artículo científico elegido. Teniendo claro esto se muestran las diferentes graficas como: línea, barras, radar, dona, pastel, área polar y un Word cloud.

VISUALIZAR TIPOS DE GRÁFICAS

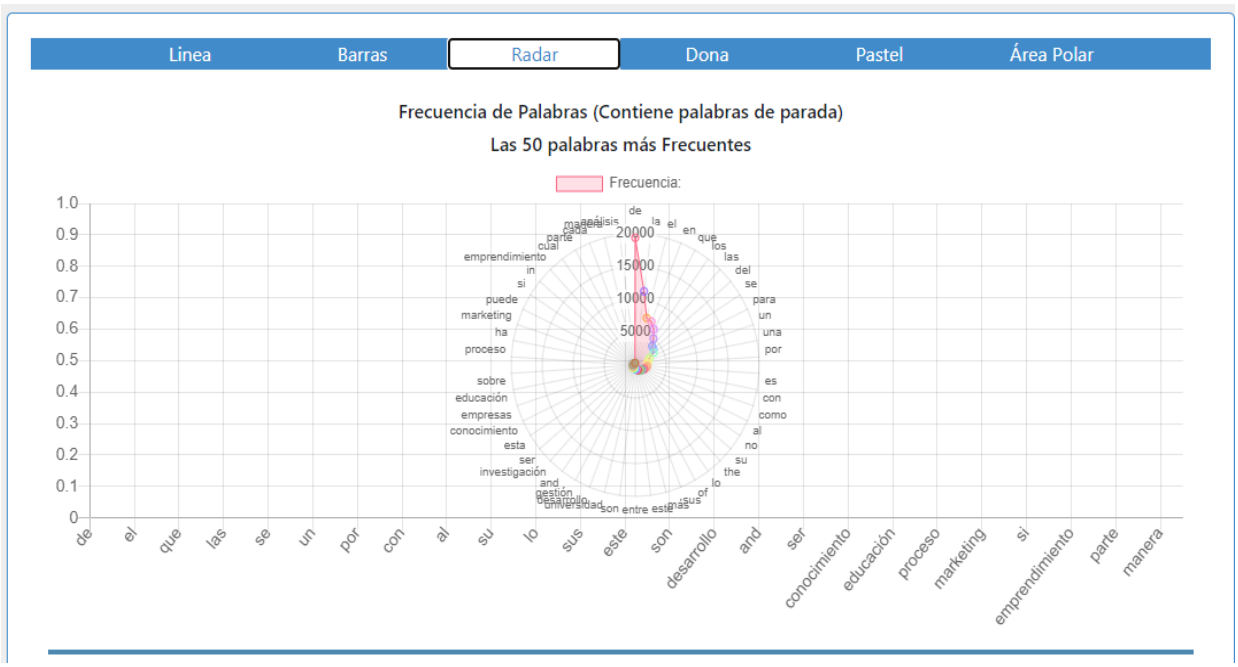
Se visualizarán los diferentes tipos de gráficas que están implementados en nuestro módulo de procesamiento de datos para conocer mejor la interpretación de los resultados.

Figura 58: Gráfica de Barras de la frecuencia de palabras



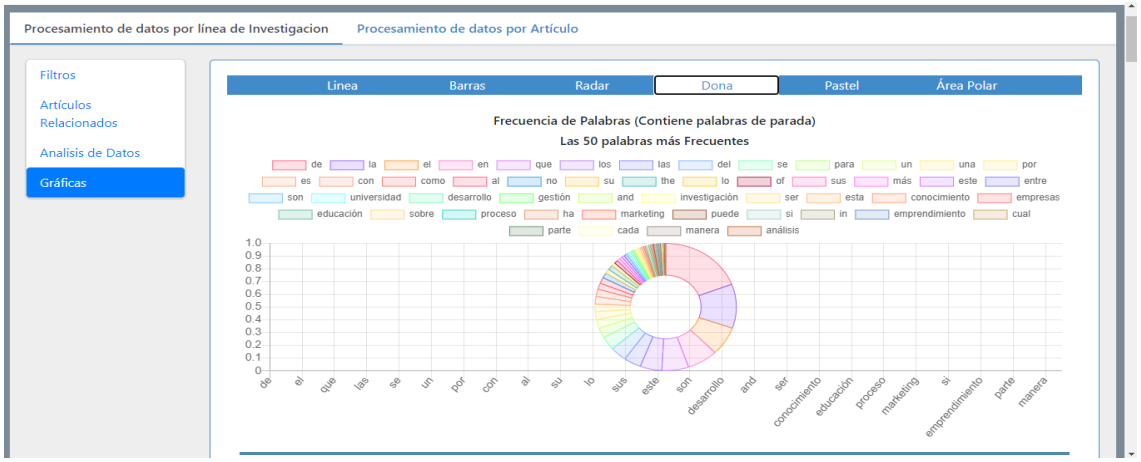
Fuente: Los investigadores

Figura 59: Gráfica de radar de la frecuencia de palabras



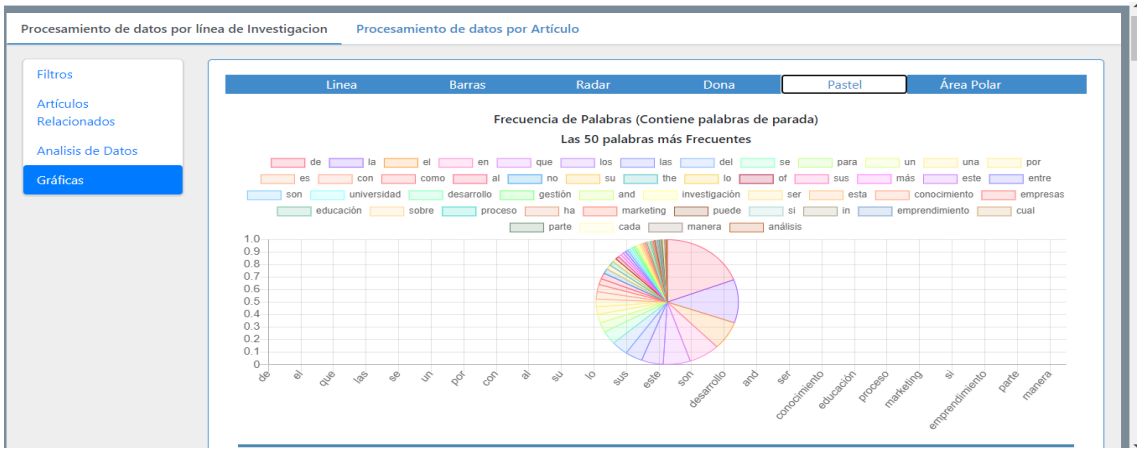
Fuente: Los investigadores

Figura 60:Gráfica de Dona de la frecuencia de palabras



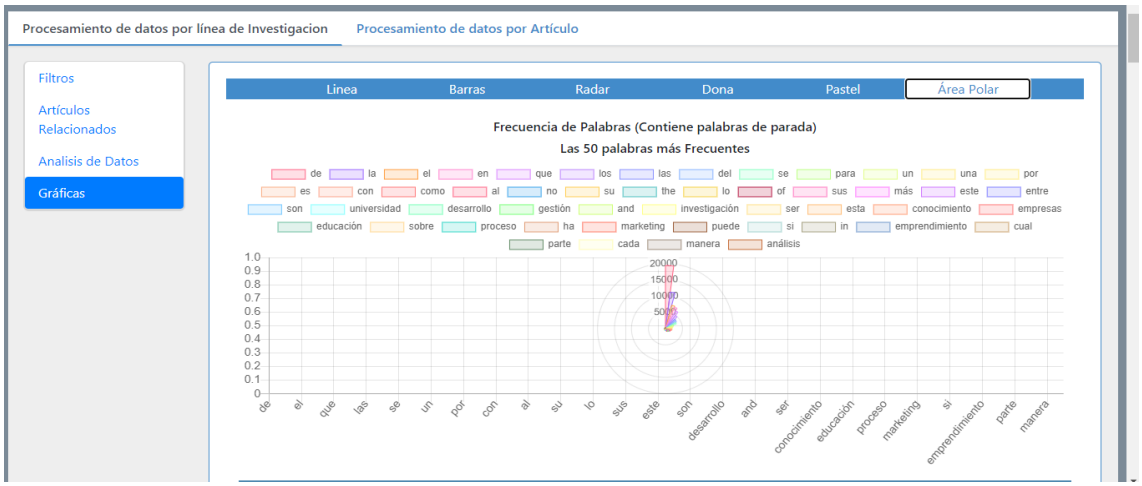
Fuente: Los investigadores

Figura 61:Gráfica de Pastel de la frecuencia de palabras



Fuente: Los investigadores

Figura 62:Gráfica de Área Polar de la frecuencia de palabras



Fuente: Los investigadores

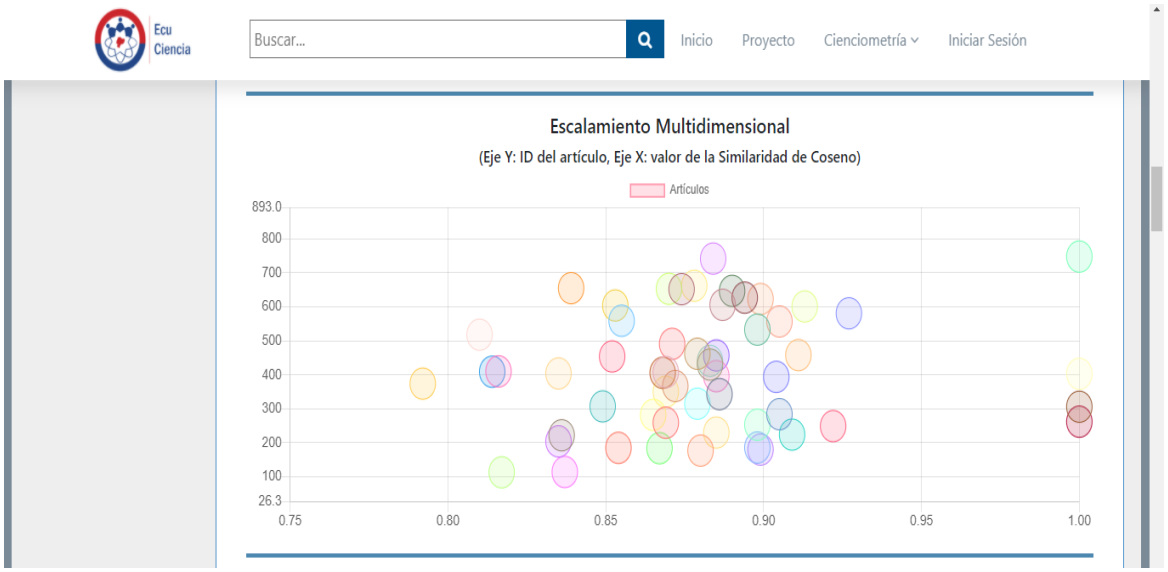
Para esta gráfica del Word Cloud se logra representar las palabras más frecuentes dando a conocer que las de gran tamaño son las más relevantes en cada línea de investigación o articulo científico.

Figura 63:Gráfica de Word Cloud de la frecuencia de palabras

Fuente: Los investigadores

Así mismo tomando en cuenta que se analizó la similitud se logró representar los resultados en graficas obtenido el escalamiento multidimensional que hace referencia a la similitud de los artículos científicos sea por línea de investigación o por artículo en sí.

Figura 64: Grafica de similitud- escalamiento multidimensional

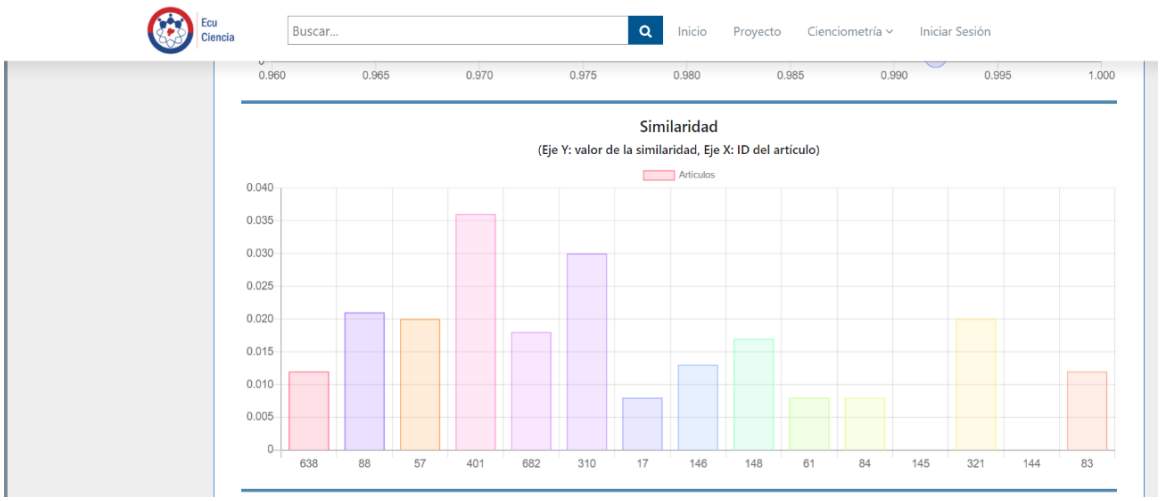


Fuente: Los investigadores

GRÁFICA DE LA SIMILITUD DE TEXTOS

Para lograr analizar los resultados de la similitud de textos para luego representarlas en una gráfica de barras demostrando que artículos son los que se asemejan al artículo analizado en donde en el eje Y se muestran los valores de similitud y en el eje X se muestra el ID de cada artículo científico.

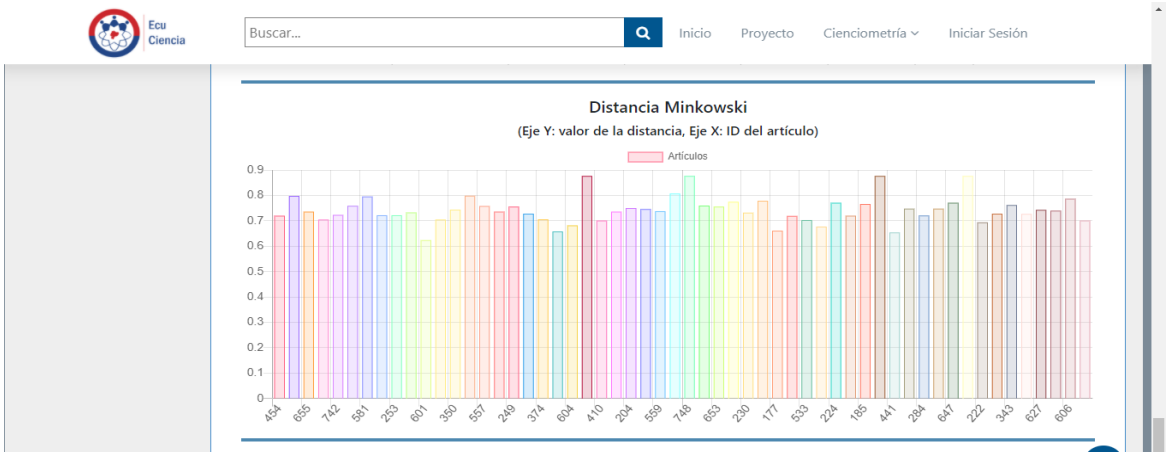
Figura 65:Gráfica de similitud de textos



Fuente: Los investigadores

Para los gráficos de las distancias se tomó diferentes librerías de Python para sacar los resultados se utilizó sklearn para vectorizar los documentos y compararlos además representarlos en graficas de barras las cuales demostraran en que niveles se encuentran cada uno de los artículos en el eje Y se muestran los valores de distancia y en el eje X se muestran los ID de los artículos científicos.

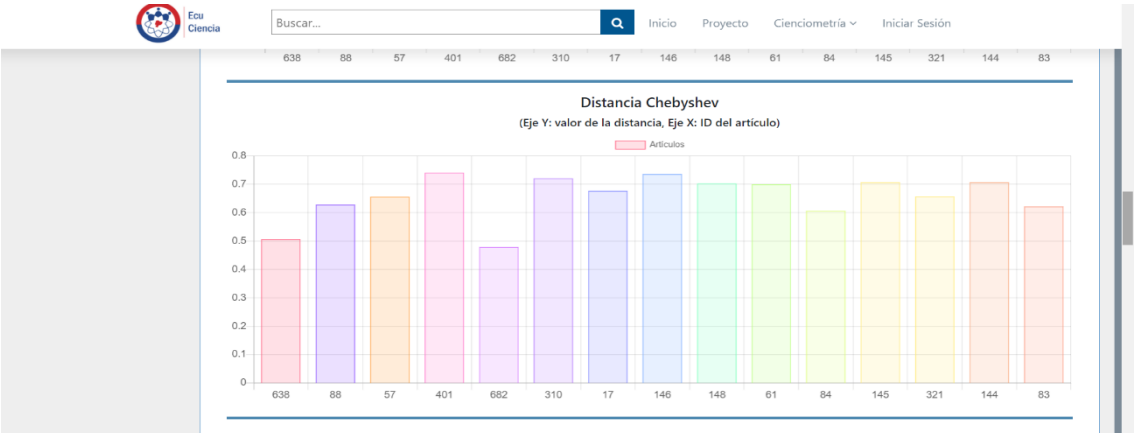
Figura 66:Distancia Minkowski



Fuente: Los investigadores

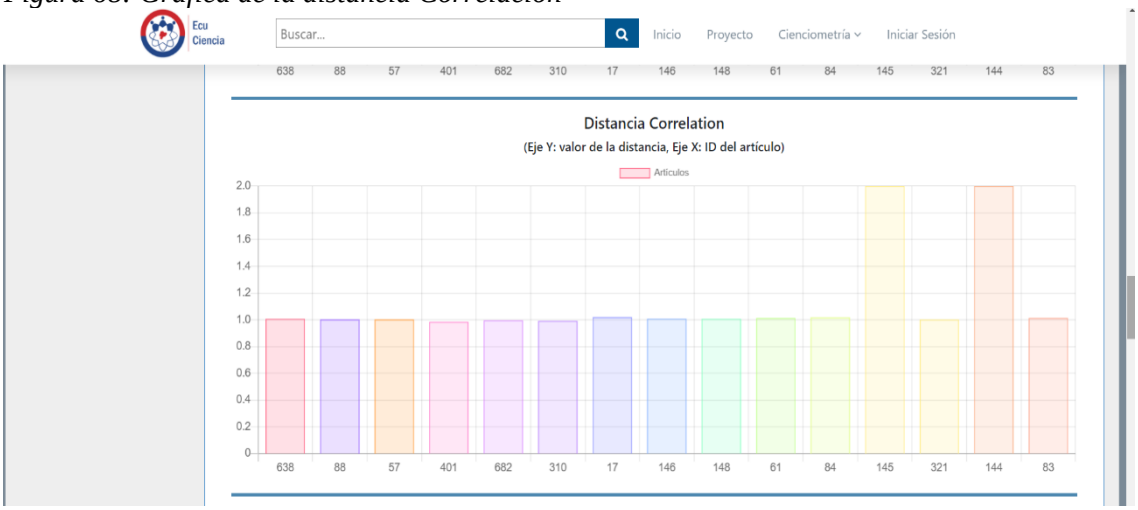
GRÁFICA DE LAS MÉTRICAS DE DISTANCIA

Figura 67: Gráfica de la distancia Chebyshev



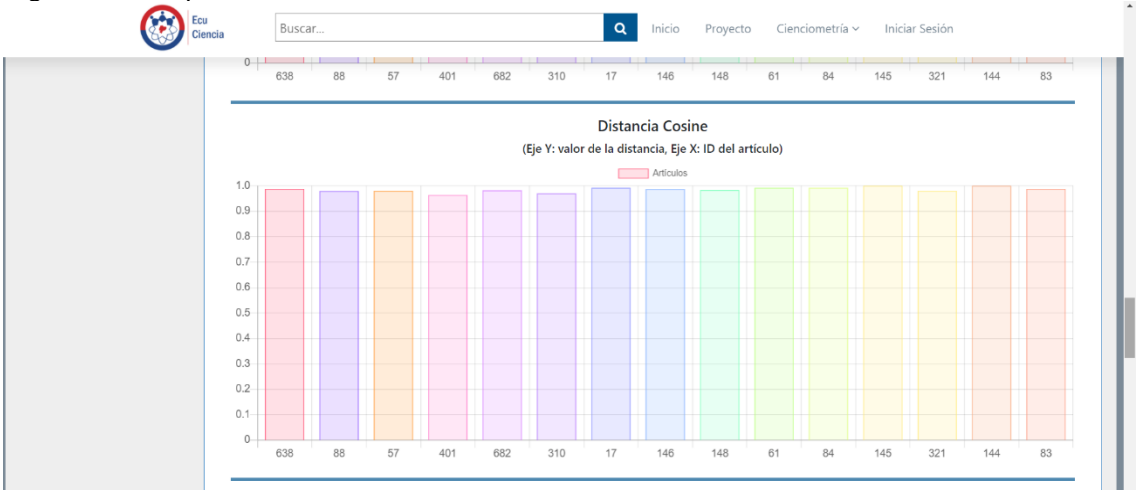
Fuente: Los investigadores

Figura 68: Gráfica de la distancia Correlación



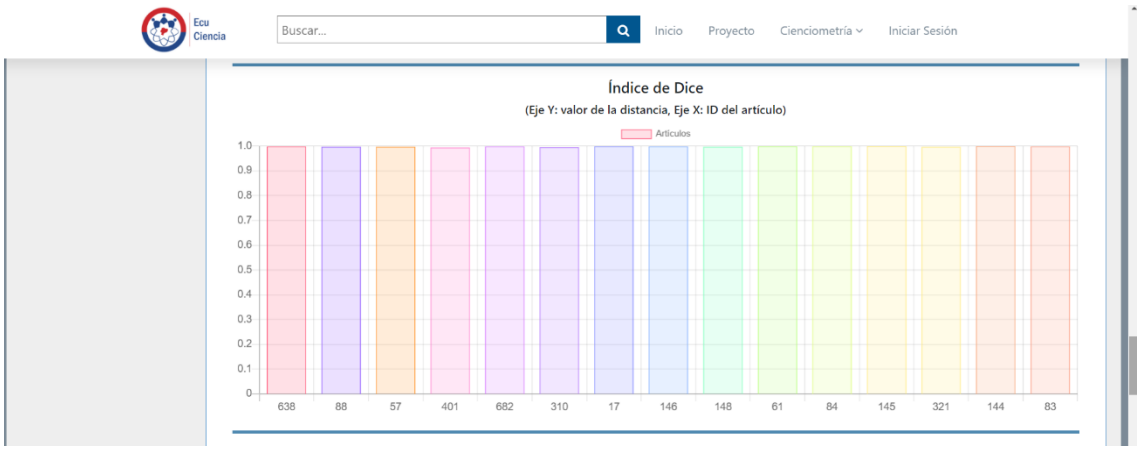
Fuente: Los investigadores

Figura 69: Gráfica de la distancia Coseno



Fuente: Los investigadores

Figura 70: Gráfica de la distancia índice de Dice

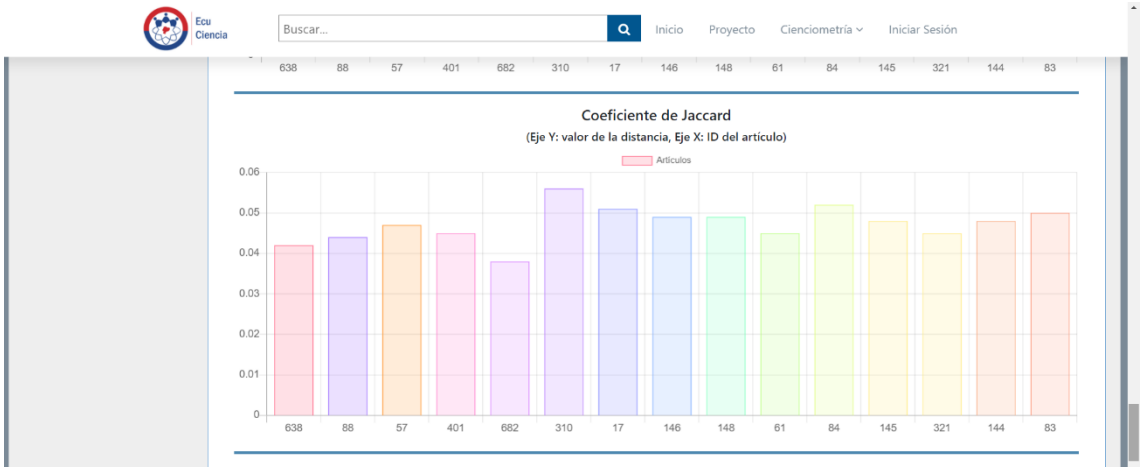


Fuente: Los investigadores

Figura 71: Gráfica de la distancia Euclidiana

Fuente: Los investigadores

Figura 72: Gráfica de la distancia Índice de Jaccard



Fuente: Los investigadores

11.3.12. CASOS DE PRUEBA

Tabla 18: Casos de prueba CP001 filtrado de datos

CP001:	**Filtrado de datos**
H.U:	001
Fecha:	25/07/2020
Responsable	Scrum Team
Descripción	Permite al usuario visualizar los artículos científicos mediante líneas, sublíneas de investigación y articulo científico en específico.
Precondiciones:	El usuario debe estar en la plataforma EcuCiencia.
Resultado esperado 1:	Obtener información completa como zona, universidad, línea y sublínea de investigación para lograr analizar el corpus de cada artículo científico o por línea de investigación.
Evaluación de la prueba:	SUPERADO

Fuente: Los investigadores

Tabla 19: Casos de prueba CP002 Obtener Corpus

CP002:	**Obtener Corpus**
H.U:	002
Fecha:	25/07/2020
Responsable	Scrum Team
Descripción	Permite al usuario visualizar el análisis del corpus por línea de investigación.
Precondiciones:	El usuario debe estar en la plataforma EcuCiencia.
Resultado esperado 1:	Obtener un número de palabras del contenido de los artículos científicos que pertenecen a una línea de investigación.
Resultado esperado 2:	Obtener el número de palabras (sin palabras de parada).
Resultado esperado 3:	Obtener el número de palabras (con palabras de parada).
	Obtener la riqueza léxica del contendido de todos los artículos científicos que pertenecen a una línea de investigación.
Alterno 1:	
Descripción:	Permite al usuario visualizar el análisis del corpus de cada artículo.
Precondiciones:	El usuario debe estar en la plataforma EcuCiencia.

Resultado esperado 1:	Obtener un número de palabras del contenido del artículo.
Resultado esperado 2:	Obtener el número de palabras (sin palabras de parada).
Resultado esperado 3:	Obtener el número de palabras (con palabras de parada).
Resultado esperado 4:	Obtener la riqueza léxica del contendido del articulo científico.
Evaluación de la prueba:	SUPERADO

Fuente: Los investigadores

Tabla 20: Casos de prueba CP003 Distancia y similitud de textos

CP003:	**Distancia y similitud de textos**
H.U:	003
Fecha:	25/07/2020
Responsable	Scrum Team
Descripción	Permite al usuario conocer la distancia y similitud de textos analizados.
Precondiciones:	El usuario debe estar en la plataforma EcuCiencia.
Resultado esperado 1:	Obtener la distancia y similitud de cada artículo para conocer si los artículos son compatibles o no.
Evaluación de la prueba:	SUPERADO

Fuente: Los investigadores

Tabla 21: Casos de prueba CP004 Visualizar gráficas

CP004:	**Visualizar Gráficas**
H.U:	004
Fecha:	25/07/2020
Responsable	Scrum Team
Descripción	Permite al usuario visualizar los gráficos de la información del documento por línea de investigación.
Precondiciones:	El usuario debe estar en la plataforma EcuCiencia.
Resultado esperado 1:	Obtener la gráfica de las 50 palabras más frecuentes del corpus que contiene palabras de parada de la línea de investigación
Resultado esperado 2:	Obtener la gráfica de las 50 palabras más frecuentes del corpus que no contiene palabras de parada de la línea de investigación
Alterno 1:	
Descripción	Permite al usuario visualizar los gráficos de la información del documento por artículo científico.

Resultado esperado 1:	Obtener la gráfica de las 20 palabras más frecuentes del corpus que contiene palabras de parada de los artículos científicos.
Resultado esperado 2:	Obtener la gráfica de las 20 palabras más frecuentes del corpus que contiene palabras de parada de los artículos científicos.
Evaluación de la prueba:	SUPERADO

Fuente: Los investigadores

Tabla 22: Casos de prueba CP005 Actualizar información

CP005:	**Actualizar información**
H.U:	005
Fecha:	25/07/2020
Responsable	Scrum Team
Descripción	Permite al usuario visualizar la información actualizada.
Precondiciones:	El usuario debe estar en la plataforma EcuCiencia.
Resultado esperado 1:	Obtener la información actualizada de los artículos científicos que están alojados en la plataforma EcuCiencia.
Evaluación de la prueba:	SUPERADO

Fuente: Los investigadores

Tabla 23: Casos de prueba CP006 Descargar Corpus

CP006:	**Descargar Corpus**
H.U:	006
Fecha:	25/07/2020
Responsable	Scrum Team
Descripción	Permite al usuario descargar el corpus.
Precondiciones:	El usuario debe estar en la plataforma EcuCiencia.
Resultado esperado 1:	Obtener la descarga del corpus en un archivo de texto plano por línea de investigación o articulo científico.
Resultado esperado 2:	Obtener la descarga del articulo científico.
Evaluación de la prueba:	SUPERADO

Fuente: Los investigadores

12. PRESUPUESTO

- **Gastos Directos**

En la tabla detalle de gastos directos están enlistados todos los gastos que se tomaron en cuenta para el desarrollo del trabajo de tesis como propuesta tecnológica.

Tabla 24: Detalle de Gastos Directos

GASTOS DIRECTOS			
DETALLE	**CANTIDAD/MESES**	**PRECIO**	**TOTAL**
Servicio de Internet	12	25,00	300,00
Impresiones	300	0,10	30,00
Resma de papel	2	3,50	7,00
Esferos	4	0,50	2,00
Anillados	5	1,75	8,75
Pendrive 32gb	1	16.00	16.00
Cuaderno	2	1.25	2.50
Esferos	4	0.50	2.00
TOTAL			**368,25**

Fuente: Los investigadores

- **Gastos Indirectos**

En la tabla detalle de los gastos indirectos están enlistados los gastos que son poco importantes pero necesarios para cumplir con la propuesta tecnológica.

Tabla 25: Detalle de los Gastos Indirectos

GASTOS INDIRECTOS			
DETALLE	**CANTIDAD/ DÍAS**	**PRECIO**	**TOTAL**
Transporte	50	1,00	50,00
Alimentación	120	2,00	240,00
Comunicación	90	0.50	45.00
Imprevistos			50,00
TOTAL			**344,00**

Fuente: Los investigadores

Tabla 26: Gastos totales

GASTOS TOTALES	
GASTOS DIRECTOS	368,25
GASTOS INDIRECTOS	344,00
TOTAL	**712**

Fuente: Los investigadores

13. ANÁLISIS DE IMPACTO

13.1.IMPACTO TECNOLÓGICO

En la actualidad las actividades sobre el desarrollo de proyectos tecnológicos se caracterizan por una creciente relevancia con diferentes cambios tecnológicos modernos en la sociedad es por eso que existen diferentes aportaciones para el estudio de la tecnología, por lo que ahora da lugar al desarrollo y la implementación de la propuesta tecnológica de un nuevo módulo en el plataforma EcuCiencia, "Procedimiento algorítmico basado en técnicas del procesamiento del lenguaje natural para el análisis del corpus de artículos científicos de la plataforma EcuCiencia." Este proyecto conlleva un gran impacto tecnológico porque se busca analizar de una manera más sencilla el corpus de cada documento con la ayuda de algoritmos y lógica de programación que procesara la gran cantidad de información que almacena la plataforma EcuCiencia.

13.2.IMPACTO SOCIAL

El proyecto de tesis realizado ha tenido un gran impacto social ya que se ha obtenido resultados esperados que son positivos porque se ha logrado satisfacer las necesidades de docentes investigadores, gracias a los algoritmos de clasificación de textos que hacen que se pueda generar un análisis automático de corpus de documentos científicos asociados con líneas y sublíneas de investigación permitiéndonos conocer la similitud, distancia, numero de palabras de un artículo científico que hacen que los docentes y estudiantes de la Universidad Técnica de Cotopaxi no se demoren en conocer el análisis de sus documentos.

13.3.IMPACTO ECONÓMICO

Durante el desarrollo e implementación del nuevo módulo en la plataforma EcuCiencia se ha logrado identificar que el proyecto tiene un costo representativo, cabe recalcar que el proyecto de investigación **“Red de Estudios Cienciométricos REDEC”** es muy

amplio por lo cual en el módulo de procesamiento de datos utilizamos las métricas definidas por International Function Point Users Group (IFPUG) para la estimación del proyecto teniendo en cuenta los requerimientos funcionales del nuevo módulo de la plataforma EcuCiencia.

14. ESTIMACIÓN DE LA PROPUESTA TECNOLÓGICA

Para la estimación del esfuerzo, tiempo y costo del proyecto se orientó a utilizar Puntos de Función. En la siguiente Tabla se muestra cada una de las funciones según su tipo y complejidad obtenida de la IFPUG, la cual nos ayudara a definir la complejidad de puntos de función de cada una de las funcionalidades del sistema.

Tabla 27:Funciones según su tipo y complejidad

TIPO/COMPLEJIDAD	BAJA	MEDIA	ALTA
Entrada Externa (EI)	3 PF	4 PF	6 PF
Salida Externa (EO)	4 PF	5 PF	7 PF
Consulta Externa (EQ)	3 PF	4 PF	6 PF
Archivo Lógico Interno (ILF)	7 PF	10 PF	15 PF
Archivo Lógico Externo (EIF)	5 PF	7 PF	10 PF

Fuente: [58]

En la siguiente tabla se muestra las funcionalidades del módulo de trabajo en donde se resaltar el tipo de complejidad a todas las funcionalidades.

Tabla 28:Funcionalidades y su tipo

FUNCIONALIDADES	TIPO	COMPLEJIDAD Media
El sistema permitirá actualizar la información de la línea de investigación	EI	4
El sistema permitirá actualizar la información de la sublínea de investigación	EI	3
El sistema permitirá al usuario visualizar el corpus de los artículos científicos	EO	5
El sistema permitirá al usuario visualizar filtrado de datos	EO	5

El sistema permitirá al usuario buscar artículos científicos	EQ	3
El sistema permitirá al usuario visualizar el listado artículos científicos	EO	5
El sistema permitirá al usuario visualizar las graficas	EO	5
El sistema permitirá al usuario descargar el corpus de los artículos	EO	5
Tablas de la base de datos	ILF	50

Fuente: Los investigadores

En la tabla 26 se dan a conocer el número de funcionalidades que están listas para identificarlas por cada tipo y a su vez tomar el total de los puntos de función sin ajustar teniendo como resultado lo siguiente:

Tabla 29:N° de funcionalidades

FUNCIONALIDADES	N° FUNCIONALIDADES	BAJA	MEDIA	COMPLEJIDAD Alta/Media
Entrada Externa (EI)	2	3	4	7
Salida Externa (EO)	5		5	25
Consulta Externa (EQ)	1	3		3
Archivo Lógico Interno (ILF)	10	5		50
Archivo Lógico Externo (EIF)	0		0	0
	TOTAL			**85**

Fuente: Los investigadores

En la tabla 27 se enlista el factor de ajuste y para poder calificar cada factor se utilizan los valores de 1 a 5.

Tabla 30:Factor de ajuste

FACTOR DE AJUSTE	PUNTAJE
Comunicación de datos	4
Rendimiento	4
Frecuencia de transacciones	2
Entrada de datos on-line	3

Eficiencia del usuario final	4
Actualizaciones on-line	2
Procesamiento complejo	3
Reusabilidad	4
Facilidad de instalación	3
Facilidad de operación	3
Instalación en distintos lugares	3
Facilidad de cambio	2
TOTAL	**37**

Fuente: [58]

Se utiliza la siguiente formula expuesta para calcular el total de los puntos de función ajustados (PFA).

PFA= PFSA*[0.65+(0.01*FA)] -PFA= 85*[0.65+(0.01*37)]

PFA= 85*[0.65+(0.40)]

PFA= 85*1.02

PFA=86.7 Aproximadamente son 87

Realizando el cálculo correspondiente sobre los puntos de función se procede a calcular la estimación del esfuerzo que se debe tener para avanzar con el aplicativo. La siguiente tabla hace referencia al lenguaje por hora y línea de código por puntos de función, entonces tomando en cuenta a que tenemos los lenguajes de cuarta generación con 8 horas de promedio por utilizar el lenguaje Python con 20 líneas de código por puntos de función.

Tabla 31:Lenguaje por horas y línea de código por PF

LENGUAJE	HORAS PF PROMEDIO	LÍNEAS DE CÓDIGO POR PF
Ensamblador	25	300
COBOL	15	100
Lenguajes de 4ta Generación	8	20

Fuente: [58]

Se calculó el valor de hora/hombre que es igual al punto de función ajustado por las horas PF promedio, en este caso como el lenguaje fue Python y se utilizó las estimaciones de

Lenguajes de cuarta generación establecida en la tabla lenguaje por horas y líneas de código por PF[59].

H/H=PFA*Horas PF promedio

H/H=87 *8

H/H=696 Horas Hombre

Para tener el resultado del número de días y meses que se trabajó, se consideró de 5/8 horas productivas de 20/30 días con 2 desarrolladores para el avance de nuestro proyecto.

Horas=(H/H) /Desarrolladores ---Horas=696/2 ---Horas=348 Duración del proyecto horas

Días Trabajo=Horas/5 --- Días Trabajo=348/5 --- Días Trabajo=69.6 aproximado 70

Meses Desarrollo=Días Trabajo/20 --- Meses Desarrollo=70/20

Meses Desarrollo=3.5 meses para desarrollar el software de lunes a viernes 5 horas diarias con 2 desarrolladores. (Estimación de duración del proyecto)

Finalmente, para calcular la estimación total del proyecto se utilizó la siguiente formula[59]:

Costo total del proyecto= (sueldo desarrollador *número de desarrolladores*tiempo en meses) + Otros costos necesarios del Proyecto.

Para conocer el sueldo del desarrollador junior del proyecto se hizo referencia a la tesis "APLICACIÓN MÓVIL CON ADMINISTRACIÓN DE CONTENIDOS WEB, PARA DIFUNDIR INFORMACIÓN DE LOS PRINCIPALES ATRACTIVOS TURÍSTICOS DE LA PROVINCIA DE COTOPAXI."[59] a los 450 dólares mensuales.
Total, del Proyecto= (450*2*3.5) + **712**

Total, del Proyecto= (3 150) + **712**

Total, del Proyecto=3862 dólares (Costo Estimado)

Aplicando las métricas definidas por International Function Point Users Group (IFPUG) para saber la estimación del proyecto se estima un monto de 3862 dólares con 2 desarrolladores con un sueldo de $450.00 dólares, en un tiempo de 3 meses con 5 semanas.

15. CRONOGRAMA

ACTIVIDAD	MAYO	JUNIO				JULIO				AGOSTO				SEPTIEMBRE			
	S	SEMANAS				SEMANAS				SEMANAS				SEMANAS			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Investigación Preliminar	X																
Aprobación del tema		X															
Información del perfil de la propuesta		X															
Información General			X														
Fundamentación general sobre la información de la propuesta tecnológica.			X														
Justificación de la propuesta				X													
Definición del problema				X													
Descripción de la problemática				X													
Planteamiento de Objetivos general y específicos					X												
Planteamiento y tareas de los objetivos de la propuesta					X												
Fundamentación teórica de la propuesta					X												
Planteamiento de la Hipótesis						X											
Desarrollo del algoritmo de análisis de corpus						X	X										
Revisión y corrección de los errores							X										
Desarrollo del sistema web del sistema EcuCiencia							X	X									
Revisión y corrección de errores								X	X								
Plan de pruebas del sistema web									X	X							
Análisis y Discusión de resultados										X	X						
Presupuesto												X					
Conclusiones / recomendaciones												X					
Revisión y corrección de las observaciones del primer encuentro de lectores.													X				
Entrega del documento final														X			
Pre defensa del proyecto															X		
Defensa del proyecto final																X	

16. CONCLUSIONES

- En los últimos años el procesamiento de lenguaje natural es más utilizado por lo que se puede encontrar con la información necesaria en las fuentes bibliográficas certificadas para hacer uso de las mismas y sin lugar a duda comprender y generar nuevos conocimientos a partir de lo estudiado para poder dar una solución a nuestra investigación y alcanzar un conocimiento aceptable para la comprensión de todo el trabajo de titulación.
- Para determinar la eficiencia y la predicción de nuestro algoritmo diseñado se utilizó la validación de división de tren/prueba para lo cual se tomó cinco artículos de manera aleatoria de la base de datos de la plataforma científica EcuCiencia mismos que en las pruebas realizadas dieron resultados óptimos permitiéndonos así continuar con la segunda fase del desarrollo del módulo procesamiento de datos.
- Finalmente, para migración del algoritmo hacia el ambiente web, se utilizó la metodología Scrum ya que al ser una metodología ágil facilitó la implementación del módulo procesamiento de datos en la plataforma científica Ecuciencia.

17. RECOMENDACIONES

- La plataforma científica Ecuciencia al estar en continuo crecimiento debería mantener actualizadas las librerías y herramientas que hacen parte del sistema para evitar conflictos con los módulos existentes y futuros porque puede ocasionar pérdida de tiempo además la búsqueda infalible de información.
- Se recomienda a los estudiantes que formen parte del proyecto REDEC se mantenga la buena organización de cada módulo para así evitar bugs o fallos en el sistema.
- Se recomienda tener documentación específica de cada módulo ya que esto facilitará futuros mantenimientos del sistema.
- Se recomienda para futuros trabajos se ponga en consideración complementar las funcionalidades que tiene el presente módulo de la plataforma EcuCiencia para que de esta manera tenga mayor valor funcional.

18. BIBLIOGRAFÍA

[1] J. M. Quinteiro González, E. Martel Jordán, P. Hernández Morera, J. A. Ligero Fleitas, and A. López Rodriguez, "Clasificación de textos en lenguaje natural usando la Wikipedia," *RISTI Rev. Ibérica Sist. e Tecnol. Informação*, vol. N° 8, no. 1696–9895, pp. 39–52.

[2] K. Aurangzeb, B. Baharum, L. H. Lee, and K. Khairullah, "A Review of Machine Learning Algorithms for Text-Documents Classification," *J. Adv. Inf. Technol.*, vol. 1, no. 0, p. 10.

[3] E. M. Chicaiza Haro and J. J. Allauca Chaquinga, "APLICACIÓN DE ALGORITMO DE EXTRACCIÓN DE TEXTO EN PERFILES DE USUARIO EN CASO DE LOS INVESTIGADORES DE LA UNIVERSIDAD TÉCNICA DE COTOPAXI," 2018.

[4] D. G. Falconí Punguil and J. N. Gualpa Mendoza, "MÉTODO PARA LA DETERMINACIÓN DE SIMILARIDAD Y DISTANCIA ENTRE INVESTIGADORES A PARTIR DE ALGORITMOS DE CLASIFICACIÓN," 2019.

[5] E. V. Gamboy Parrales and O. J. Yunda Cando, "'IMPLEMENTACIÓN DE UN ALGORITMO PARA LA EVALUACIÓN DE LOS DOCUMENTOS CIENTÍFICOS DE LOS INVESTIGADORES DE LA UNIVERSIDAD TÉCNICA DE COTOPAXI,'" 2019.

[6] L. E. Colmenares Guillén, M. Carrillo Ruiz, V. G. Morales Murillo, and J. G. López López, "Validación de un algoritmo de clasificación para la identificación de interacciones farmacológicas," *SciELO*, vol. 20, no. 1405–7743, 2019.

[7] M. A. Mouriño Garcia, "Clasificación multilingüe de documentos utilizando machine learning y la Wikipedia," 2017.

[8] S. E. RODRÍGUEZ ORTIZ, "CLASIFICACIÓN DE TEXTO BASADO EN AGENTES INTELIGENTES," 2015.

[9] Š. Sanja, D. Biljana, and S. Horacio, "Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification," *SciELO*, vol. 17, N°2, no. 1405–5546, 2013.

[10] J. Castillo, M. Cardenas, A. Curti, and O. Casco1, “Creación de corpus para aplicaciones de análisis de texto no estructurado,” *Rev. ARGENTINA Ing.*, vol. 8, 2016.

[11] J. Castillo, M. Cardenas, and A. Curti, “Software para asistencia en la creación de corpus para sistemas de análisis de texto no estructurado,” *https://www.semanticscholar.org/*, 2015.

[12] Sociedad Española para el Procesamiento de Lenguaje Natural, “Procesamiento del Lenguaje Natural,” *Rev. SEPLN*, vol. 56, 2016.

[13] J. M. Hernández Hernández, “Analisis automatico de textos en español utilizando NLTK,” 2016.

[14] J. Pustejovsky and A. Stubbs, *Natural Language Annotation for Machine Learning*. 2013.

[15] F. A. Lopez, “¿Qué es un corpus lingüístico y cuál es su uso?,” *Welun Translations Profesionales de la traducción*, 2018. .

[16] H. González Sanjurjo, “Creación de un Framework para el Tratamiento de Corpus Lingüísticos-Development of a Framework for Corpus Linguistic Analysis.”

[17] J. Torruella and J. Llisterri, “Diseño de corpus Textuales y orales.”

[18] K. S. and H. Kargupta, A. Joshi and Y.Yesha, “Data mining: next generation challenges and future directions,” *MIT/AAAI*.

[19] J. C. Riquelme, R. Roberto, and K. Gilbert, “Minería de Datos: Conceptos y Tendencias,” *Departamento de Lenguajes y Sistemas Informáticos Universidad de Sevilla*, p. 8.

[20] E. Alpaydin, *Introduction to Machine Learning*. 2020.

[21] W. O. Kohan and E. T. José, *CONOCIMIENTO, PENSAMIENTO Y LENGUAJE. UNA INTRODUCCIÓN A LA LÓGICA Y AL PENSAMIENTO CIENTIFICO*. 2006.

[22] J. Hirschberg and C. D. Manning, “Advances in natural language processing,” *Science (80-.).*, vol. 349, no. 6245, pp. 261–266, 2015.

[23] M. Ebrahim, “Tutorial De NLP Con Python NLTK,” *Like geeks*, 2019. .

[24] R. E. Lopez Briega, “Procesamiento del lenguaje natural con Python,” *Matematicas, Analisis de Datos y Python*, 2017. .

[25] F. J. Pinales Delgado and C. E. Velázquez Amador, *Algoritmos resueltos con Diagramas de Flujo y Pseudocódigo*. México: Universidad Autónoma de Aguascalientes, 2018.

[26] X. M. Martin Uriz and M. Galar Idoate, “Aprendizaje de distancias basadas en disimilitudes para el algoritmo de clasificación kNN,” Universidad Pública de Navarra, 2015.

[27] A. Téllez Valero, “Extracción de Información con Algoritmos de Clasificación.”

[28] C. Guagliano, *Programacion en Python I*. 2019.

[29] D. Scholnick, “Machine learning module for Python,” *Kite Your Programming Copilot*, 2020. .

[30] W. M. & P. D. Team, *“pandas: powerful Python data analysis toolkit Release 0.23.4 Wes McKinney & PyData Development Team,*. 2018.

[31] and G. V. S. van der Walt, S. C. Colbert, “‘The NumPy Array: A Structure for Efficient Numerical Computation,’” *Comput. Sci. Eng*, vol. 13, no. 2, pp. 22–30.

[32] S. Developers, “sklearn.feature_extraction.text.TfidfVectorizer.”

[33] S. Developers, “Scipy.” .

[34] J. Amat Rodrigo, “Correlación lineal y Regresión lineal simple,” *cienciadedatos.net*.

[35] J. Martinez Heras, “Error Cuadrático Medio para Regresión,” *IArtificial.net Inteligencia Artificial y Machine Learning en Español*, 2020. .

[36] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. 2009.

[37] H. López Morales, “Los índices de ‘riqueza léxica’ y la enseñanza de lenguas,” *cvc.cervantes*, p. 14, 2015.

[38] T. S. Community, “Distance computations,” *docs.scipy.org*, 2020. .

[39] D. Burrueco, “DISTANCIA DE CHEBYSHEV,” *interactivechaos*, 2020. .

[40] D. Burrueco, “machine learning,” *interactivechaos*, 2020. .

[41] L. C. R, E, “Método de clasificacion Automatica de textos Basado en palabras claves utilizando informacion semntica,” *Aplicación a historias clínicas*, p. 92, 2014.

[42] M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, C. D. Manning, and D. McClosky, “The Stanford CoreNLP Natural Language Processing Toolkit,” in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60.

[43] Threespot, “Django makes it easier to build better Web apps more quickly and with less code.,” *2015*, 2019. .

[44] andrew Mccarthy, “Django facilita la creación de mejores aplicaciones web de forma más rápida y con menos código.,” *Django Software Foundation*, 2020. .

[45] U. de Alicante, “Modelo vista controlador (MVC),” *Servicio de InformáticaASP.NET MVC 3 Framework*, 2020. .

[46] jetbrains, “Conoce a PyCharm - Ayuda | PyCharm,” 2018. .

[47] PostgreSQL, “‘PostgreSQL: About.’” .

[48] C. Sabino, “CAPITULO III. Marco Metodológico de la Investigación.” .

[49] R. D. F. Parot, “Metodologia de la investigacion,” *Madrid Akal Psicol.*

[50] P. . Abril, Victor Hugo, “TÉCNICAS E INSTRUMENTOS DE LA INVESTIGACIÓN,” 2016.

[51] L. Mariñelarena Dondena, M. L. Errecalde, and A. Castro Solano, “Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología,” *SciELO*, vol. 9, N°2, 2017.

[52] O. Nigro, D. Xodo, G. Corti, and D. Terren, “KDD (Knowledge Discovery in Databases): un proceso centrado en el usuario.”

[53] J. Landa, “Tratamiento de los datos,” *Febrero 19*, 2016. .

[54] Bronshtein and Adi, “Train/Test Split and Cross Validation in Python,” *Towards*

Data Science, 2017.

[55] J. Martinez Heras, “Error Cuadrático Medio para Regresión,” *IArtificial.net*, 2020.

[56] F. Cristopher, “METODOLOGÍA ÁGIL,” 2013. .

[57] R. G. Figueroa, C. J. Solís, and A. A. Cabrera, “METODOLOGÍAS TRADICIONALES VS. METODOLOGÍAS ÁGILES,” *Google Acad.*, p. 9, 2013.

[58] M. Coral, J. Guerra, and C. Luza, “LA MÉTRICA DE PUNTO DE FUNCIÓN Y SU APLICACIÓN EN LA ESTIMACIÓN DEL TAMAÑO DEL SOFTWARE,” *Universidad Inca Garcilaso de la Vega*, pp. 1–10.

[59] A. J. Cando Toapanta and J. A. Oñate Cajamarca, “APLICACIÓN MÓVIL CON ADMINISTRACIÓN DE CONTENIDOS WEB, PARA DIFUNDIR INFORMACIÓN DE LOS PRINCIPALES ATRACTIVOS TURÍSTICOS DE LA PROVINCIA DE COTOPAXI,” 2018.

ANEXOS

I. ANEXO GUÍA DE LA ENTREVISTA

UNIVERSIDAD TÉCNICA DE COTOPAXI

FACULTAD DE CIENCIAS DE LA INGENIERÍA Y APLICADAS
CARRERA DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS COMPUTACIONALES

Objetivo. - Obtener más información sobre los requerimientos funcionales que se llevaran para el desarrollo del nuevo módulo en la plataforma EcuCiencia.

Guía de entrevista

1. **¿La plataforma EcuCiencia por cuanto tiempo está funcionando?**
2. **¿Qué lenguaje de programación fue utilizado para el desarrollado del sistema?**
3. **¿Cuál es el Gestor de Base de Datos con el que está trabajando el sistema?**
4. **¿Cuál es el objetivo del nuevo módulo de procesamiento de datos de la plataforma EcuCiencia?**
5. **¿Qué aporte brindaría a la comunidad universitaria UTC la implementación de este nuevo módulo?**
6. **¿Cuáles son las funcionalidades del módulo procesamiento de datos?**
7. **¿Qué resultados espera obtener del sistema ya aplicando el nuevo módulo?**

II. ANEXO PLANTILLA CASOS DE USO

Tabla 32: Plantilla casos de uso

Nº caso	**CU-004**
H. U	
Nombre	
Autor	
Descripción:	
Actores:	
Precondición:	
Flujo Normal:	
Flujo Alterno:	

Fuente: Los Investigadores

III. ANEXO PLANTILLA CASOS DE PRUEBA

Tabla 33: Plantilla casos de prueba

CP002:	**Obtener Corpus**
H.U:	
Fecha:	
Responsable	
Descripción	
Precondiciones:	
Resultado esperado 1:	
Resultado esperado 2:	
Resultado esperado 3:	
Alterno 1:	
Descripción:	
Precondiciones:	
Resultado esperado 1:	
Resultado esperado 2:	
Resultado esperado 3:	
Resultado esperado 4:	
Evaluación de la prueba:	

Fuente: Los Investigadores

IV. ANEXO HOJA DE VIDA

GILSON ARIEL CHARIGUAMAN MOROCHO

Edad: 23 años

Telf.: 0987253981

Email: Gilson258@live.com

ESTUDIOS REALIZADOS

PRIMARIA	Unidad Educativa Nuestra Señora de Pompeya
SECUNDARIA	Colegio Nacional Saquilisí • Bachillerato General Unificado
SUPERIOR	Universidad Técnica de Cotopaxi • Ingeniera en informática y sistemas computacionales (Estudiante)
IDIOMAS	Español: Natal Inglés: intermedio

NATALY LIZETH QUILUMBAQUIN TUTILLO

Edad: 24 años

Telf.: 0939051485

Email: lizethquilumbaquin@gmail.com

ESTUDIOS REALIZADOS

PRIMARIA	Unidad Educativa "José Acosta Vallejo"
SECUNDARIA	Unidad Educativa Particular Mariana de Jesús. Colegio Nacional Técnico Cayambe • Bachiller Técnico en Informática y Administración de sistemas.
SUPERIOR	Universidad Técnica de Cotopaxi • Ingeniería en informática y sistemas computacionales (Estudiante)
IDIOMAS	Español: Natal Ingles: intermedio