UNIVERSIDAD TÉCNICA DE COTOPAXI

DIRECCIÓN DE POSGRADO

MAESTRÍA EN SISTEMAS DE INFORMACIÓN

MODALIDAD: PROPUESTA METODOLÓGICA Y TECNOLÓGICA AVANZADA

Título:

Métodos para el análisis de la información en corpus de artículos científicos con algoritmos de clasificación y librerías NLTK en la Plataforma Científica ECUCIENCIA

Trabajo de titulación previo a la obtención del título de Magíster en Sistemas de Información

Autor

Segundo Humberto Corrales Beltrán Mg.

Tutor

Gustavo Rodríguez Bárcenas PhD.

LATACUNGA –ECUADOR

2020

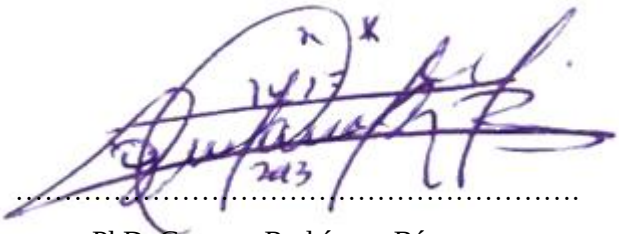
APROBACIÓN DEL TUTOR

En mi calidad de Tutor del Trabajo de Titulación "**Métodos para el análisis de la información en corpus de artículos científicos con algoritmos de clasificación y librerías NLTK en la Plataforma Científica ECUCIENCIA**" presentado por Corrales Beltrán Segundo Humberto, para optar por el Título Magister en Sistemas de Información.

CERTIFICO

Que dicho trabajo de investigación ha sido revisado en todas sus partes y considero que reúne los requisitos y méritos suficientes para ser sometido a la presentación pública y evaluación por parte del Tribunal de Lectores que se designe.

Latacunga, septiembre del 2020



……

PhD. Gustavo Rodríguez Bárcenas

CI.: 1757001357

APROBACIÓN TRIBUNAL

El trabajo de Titulación: "**Métodos para el análisis de la información en corpus de artículos científicos con algoritmos de clasificación y librerías NLTK en la Plataforma Científica ECUCIENCIA**", ha sido revisado, aprobado y autorizada su impresión y empastado, previo a la obtención del Título de Magíster en Sistemas de Información, el presente trabajo reúne los requisitos de fondo y forma para que el estudiante pueda presentarse a la sustentación del trabajo de titulación.

Latacunga, septiembre del 2020

……..

MSc. José Augusto Cadena Moreano

CC: 0501552798

PRESIDENTE DEL TRIBUNAL

…….

MSc. Karla Susana Cantuña Flores

CC: 0502305113

LECTOR 2

…….

MSc. Alex Christian Casa LLano

CC: 0502589864

LECTOR 3

DEDICATORIA

El presente trabajo de investigación lo dedico en primer lugar a Dios, quién me ha dado la oportunidad de culminar con éxito este proceso, guiándome a cada paso hasta conseguir uno de los anhelos más deseados.

A mis queridos padres Leonor y Segundo, quienes con el paso de los años se han convertido en mi mayor apoyo, gracias por inculcar en mí el ejemplo de esfuerzo, responsabilidad y perseverancia, a ti querida hermana y Josecito.

Con todo el cariño dedico este trabajo a mis amados hijos Damián y Randy quienes son el mayor y mejor regalo que Dios y la vida me pudieron dar, son mi más grande fuente de inspiración, la razón principal para seguir luchando cada día y ser un mejor ser humano.

A ella Diana Molina Sáenz quien ha sido un pilar fundamental para culminar esta tesis con éxito y poder alcanzar esta anhelada victoria. Gracias por hacerme saber que su amor es incondicional y por enseñarme a no temer a las adversidades porque Dios está conmigo siempre.

Beto Corrales

AGRADECIMIENTO

Expreso mi sentimiento de profunda gratitud a la Universidad Técnica de Cotopaxi por haberme dado la oportunidad de formarme profesionalmente. Así también a los docentes de la Maestría en Sistemas de la Información por todos los conocimientos impartidos a lo largo de mi formación académica, gracias a todos ustedes por su paciencia, dedicación, apoyo incondicional y amistad.

Mi profundo agradecimiento a mi tutor el PHD Gustavo Rodríguez por su orientación en el desarrollo de esta investigación, gracias por su amabilidad, su tiempo y sus ideas.

Finalmente, mi agradecimiento a todas las personas que fueron parte de esta investigación, amigos, colegas, y familia ustedes han significado un verdadero apoyo y motivación para desarrollarme como ser humano y como profesional del Alma Mater.

Beto Corrales

RESPONSABILIDAD DE AUTORIA

Quien suscribe, declara que asume la autoría de los contenidos y los resultados obtenidos en el presente trabajo de titulación.

Latacunga, septiembre del 2020

……

Segundo Humberto Corrales Beltrán

C.C: 0502409287

RENUNCIA DE DERECHOS

Quien suscribe, cede los derechos de autoría intelectual total y/o parcial del presente trabajo de titulación a la Universidad Técnica de Cotopaxi.

Latacunga, septiembre del 2020

……

Segundo Humberto Corrales Beltrán

C.C: 0502409287

AVAL DEL PRESIDENTE

Quien suscribe, declara que el presente Trabajo de Titulación: "**Métodos para el análisis de la información en corpus de artículos científicos con algoritmos de clasificación y librerías NLTK en la Plataforma Científica ECUCIENCIA**", contiene las correcciones a las observaciones realizadas por los lectores en sesión científica del tribunal.

Latacunga, septiembre del 2020

……

MSc. José Augusto Cadena Moreano

CC: 0501552798

UNIVERSIDAD TECNICA DE COTOPAXI

DIRECCION DE POSGRADO

MAESTRIA EN SISTEMAS DE INFORMACIÓN

TITULO: "**Métodos para el análisis de la información en corpus de artículos científicos con algoritmos de clasificación y librerías NLTK en la Plataforma Científica ECUCIENCIA**"

AUTOR: Segundo Humberto Corrales Beltrán Mg.

TUTOR: Gustavo Rodríguez Bárcenas PhD.

RESUMEN

La plataforma web denominada ECUCIENCIA perteneciente a la Universidad Técnica de Cotopaxi almacena la producción científica de los docentes investigadores, este sistema muestra algunas métricas para los artículos considerando para ello solo el título, resumen y palabras claves, siendo insuficiente si analizamos la riqueza de todo el contenido del documento en formato PDF; se podría extraer información relevante relacionada con las líneas de investigación y otros documentos científicos a partir de la frecuencia de las palabras en cada documento, para solventar esta problemática se estableció un método de análisis de información en corpus de artículos científicos, mediante algoritmos de procesamiento de datos que se encuentran en las librerías NLTK, NUMPY, MATPLOTLIB, PYPDF2, SKLEARN y SCIPY de Python. Se usó la metodología Scrum para el desarrollo del módulo y se validaron los resultados a través de métodos estadísticos. Se obtuvieron datos a partir de un muestreo aleatorio simple y el análisis de la información contenidas en el corpus de los artículos científicos de la muestra seleccionada, pudiéndose obtener información relevante y visualización de datos significativos de las distancias Euclidiana, Correlación, Chebychev, Coseno, Coeficiente de Jaccard y el Índice Dice. La validación de los resultados a través del análisis de la varianza de un factor arrojó el valor de F = 17,621 siendo mayor que el valor crítico para F que fue de 2,412 y la probabilidad menor a 0,05 demostrando que las variables de frecuencias de los artículos se comportan de manera significativa en el proceso de representar métricas de acuerdo al corpus de los artículos.

PALABRAS CLAVE: análisis de información, corpus, artículos científicos, algoritmos de clasificación, NLTK, Ecuciencia.

UNIVERSIDAD TECNICA DE COTOPAXI

DIRECCION DE POSGRADO

MASTER'S DEGREE IN INFORMATION SYSTEMS

TITLE: "Methods for the analysis of information in corpus of scientific articles with classification algorithms and NLTK libraries in the Scientific Platform ECUCIENCIA".

AUTHOR: Segundo Humberto Corrales Beltrán Mg.

TUTOR: Gustavo Rodríguez Bárcenas PhD.

ABSTRACT

The web platform called ECUCIENCIA belonging to the Technical University of Cotopaxi stores the scientific production of the research teachers, this system shows some metrics for the articles considering only the title, summary and keywords, being insufficient if we analyze the richness of all the content of the document in PDF format; relevant information related to research lines and other scientific documents could be extracted from the frequency of the words in each document, to solve this problem, a method of analysis of information was established in corpus of scientific articles, using data processing algorithms found in the NLTK, NUMPY, MATPLOTLIB, PYPDF2, SKLEARN and SCIPY libraries of Python. The Scrum methodology was used for module development and the results were validated through statistical methods. Data was obtained from a simple random sampling and the analysis of the information contained in the corpus of scientific articles of the selected sample, being able to obtain relevant information and visualization of significant data of Euclidean distances, Correlation, Chebychev, Cosine, Jaccard Coefficient and Dice Index were obtained. The validation of the results through the analysis of the variance of a factor yielded the value of F = 17.621 being higher than the critical value for F which was 2.412 and the probability less than 0.05 demonstrating that the frequency variables of the articles behave significantly in the process of representing metrics according to the articles' corpus.

KEY WORDS: analysis of information, corpus, scientific articles, classification algorithms, NLTK, Ecuciencia.

AVAL DE TRADUCCIÓN

Marcia Janeth Chiluisa Chiluisa, con cédula de identidad número: 0502214307. Licenciado/a en: CIENCIAS DE LA EDUCACIÓN ESPECIALIZACIÓN INGLES con número de registro de la SENESCYT: 1020-05-575335; **CERTIFICO** haber revisado y aprobado la traducción al idioma inglés del resumen del trabajo de investigación con el título: "**Métodos para el análisis de la información en corpus de artículos científicos con algoritmos de clasificación y librerías NLTK en la Plataforma Científica ECUCIENCIA"** de: Segundo Humberto Corrales Beltrán, aspirante a magister en Sistemas de Información.

…….....

MSc. Marcia Janeth Chiluisa Chiluisa

CC: 0502214307

INDICE DE CONTENIDOS

CAPÍTULO II

CAPÍTULO III

INDICE DE TABLAS

INDICE DE FIGURAS

INTRODUCCIÓN

La información que representan los contenidos de los perfiles de usuarios, así como la diversidad de artículos científicos que se encuentran en la actualidad en la gran red de redes (Internet) es significativa. El escenario científico devela las potencialidades de las universidades en la producción de nuevos conocimientos a través de la publicación, ponencias en seminarios internacionales, así como la elaboración de libros en muchas áreas de conocimiento según necesidades del contexto, todo forma un compendio relevante, por tanto, obtener nuevos métodos para el análisis de la métrica que envuelven estos escenarios es una necesidad de primer orden.

En la Universidad Autónoma de Madrid, conscientes de la importancia de recopilar la producción científica de sus investigadores, se han desarrollado una serie de plataformas que recogen toda la actividad científica de los investigadores (Portal de Producción Científica) y que almacenan los textos completos de las publicaciones en las que se plasma esta producción, en acceso abierto, a través de su repositorio institucional Biblos-e Archivo.[1]

En las universidades ecuatorianas hasta la década de los años setenta el objetivo fundamental era la docencia, con un componente investigativo casi nulo, un número reducido de bibliografías y escasas publicaciones.[2] A partir del año 2008 ha tenido un efecto positivo en el desarrollo de la actividad científica en las universidades.[3]

En el período 2009-2013, 48 universidades publicaron en la base de datos Scopus, un total de 1.992 artículos, cifra que supera de manera apreciable las del quinquenio 2004-2008, en que se reportan solo 32 instituciones y 866 artículos. Mientras en los años 2014 y 2015 se logró publicar 976 y 1.174 artículos respectivamente.[4] Es importante señalar que en el ranking anual del Ecuador en el año 2015 que realiza la revista estadounidense Nature (una de las más prestigiosas del mundo en el área de ciencias naturales) se destacan tres universidades Ecuatorianas en primer lugar

la Pontificia Universidad Católica, en segundo la Universidad de Investigación de Tecnología Experimental (Yachay) y, por último, la Escuela Politécnica Nacional.[3]

Actualmente el conocimiento constituye un pilar fundamental en el desarrollo de nuevas tecnologías, la labor de repositorios de documentación científica en conjunto con el capital humano, se encuentra orientado al desarrollo de servicios y productos que ofrezcan la posibilidad de gestionar el conocimiento, el mismo que permita realizar una búsqueda y recuperación de información de manera eficiente y rápida, que se construye con la información inteligente puesta a disposición de la comunidad universitaria de la región.[5]

Existen muchos sistemas que de una manera muy eficiente brindan información sobre el dominio científico de distintas regiones del mundo, tal como SCImago Journal & Country Rank, RedSearch, Dataciencias, Redciencias, entre otras que su objetivo fundamental es brindar desde el punto de vista métrico (Cualitativo y cuantitativo) el estado de la ciencia, un inconveniente es que la mayoría de estos sistemas generan una visualización global y a partir de los indicadores utilizados no se puede dar una valoración local.

La visualización científica posibilita reconocer patrones de comportamiento de los datos, ver en una sola imagen o en una secuencia de estas (animación) una gran cantidad de datos y facilita la comprensión de algunos conceptos, sobre todo de tipo abstracto.

Por otro lado se conoce que 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.[6]

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.

En Universidad Técnica de Cotopaxi ubicada en la Av. Simón Rodríguez, barrio El Ejido sector San Felipe, del cantón Latacunga, provincia de Cotopaxi, se está desarrollando una cultura investigativa a través de la creación y recreación de ciencia, tecnología y arte, como la formación científica, generación, difusión y promoción de los saberes y conocimientos, que coadyuven al desarrollo sostenible y sustentable del entorno, con enfoque investigativo progresista y dedicado a promover la sostenibilidad productiva, ambiental y la equidad social de la región y el país.

Todas las investigaciones desarrolladas en la Universidad Técnica de Cotopaxi, son documentadas mediante artículos, libros, ponencias y proyectos; que requieren ser almacenados y visualizados por la comunidad universitaria. Para resolver esta problemática la Dirección de Investigación aprueba la implementación de una plataforma científica denominada Ecuciencia (ecuciencia.utc.edu.ec). La misma que está recopilando la producción científica y tecnológica de todas las disciplinas que se estudian en las distintas facultades existentes en la institución.

Toda la información almacenada en la base de datos de la plataforma Ecuciencia, requiere ser visualizada en herramientas que el usuario pueda entender con facilidad, para ello es necesario realizar tareas como diseñar métodos inteligentes de búsqueda, sobre todo los asociados a la información interna y a la producción científica de la institución.

Partiendo de estas características, surge la necesidad de identificar grupos de investigadores con similares características en la Universidad Técnica de Cotopaxi, los sistemas actuales globales no brindan la posibilidad de entregar con detalles esta información, de manera que se puedan establecer comunidades colectivas de conocimientos, conocer la calidad de los recursos léxicos aportados a través de la producción científica y específicamente de los artículos publicados por los investigadores.

La Plataforma Científica 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.

En tal sentido se define entonces como **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?

Para dar solución a la problemática descrita se plantea como **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.

Los **Objetivos Específicos** que permitirán el tránsito por las distintas etapas de la investigación son:

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.

Tabla 1: Tareas para el cumplimiento de objetivos

Objetivo	**Actividad (tareas)**
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.	• Establecer términos, interrogantes o conceptos fundamentales a investigar. • Buscar información en fuente de consultas confiables. • Utilizar el conocimiento obtenido para aplicarlo de manera oportuna en la investigación.
2. Determinar los requerimientos algorítmicos necesarios para el análisis de la información contenidas en el corpus de los artículos científicos y perfiles de los usuarios investigadores recogidos en la base	• Definir la metodología más adecuada a emplear. • Definir los algoritmos y librerías adecuados en Python y modelar a través de ellos.

de datos del sistema, que permita la obtención de métricas de los documentos en formato PDF.	• Definir las métricas a utilizar.
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.	• Aplicación de la metodología SCRUM. • Implementación de los algoritmos y librería NLTK de Python a través del framework django. • Validación a través de estadística.

Elaborado por: El investigador

- **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 un conjunto de documentos que contienen diferentes artículos científicos. Al lograr clasificar los documentos por temas, la búsqueda y el análisis de información de los mismos se puede realizar de manera sencilla. Realizar la clasificación de documentos de forma manual, provoca que la tarea sea complicada. Entonces el corpus que se pretende crear será analizado con un clasificador que lee los documentos y los somete a técnicas de procesamiento de documentos en lenguaje natural, también conocidas como reducción de la dimensionalidad, la cual está compuesta por técnicas de extracción y selección de características. La extracción de características está compuesta por tres tareas, las cuales son tokenizar el texto, eliminación de las stopwords y el enraizamiento de las palabras (Stemming).[7] Una vez que estos procesos son realizados, se procede a realizar la representación vectorial de los documentos para posteriormente direccionarlos al algoritmo de aprendizaje, y realizar las pruebas correspondientes.

También hay que resaltar que con la investigación planteada se busca obtener el máximo provecho a la información científica procedente de los docentes investigadores de la UTC y de sus publicaciones, por lo cual es fundamental la creación de un métodos para el análisis de la información en corpus de artículos científicos con algoritmos de clasificación y librerías NLTK en la Plataforma Científica ECUCIENCIA de manera que se vean beneficiados en primera instancia los docentes y estudiantes de la UTC, así como todos aquellos usuarios ocasionales que investiguen sobre el campo de la cienciometría.

- **Hipótesis:**

Si se establece un método de análisis de información en corpus de artículos científicos, mediante algoritmos que muestren las frecuencias de palabras, distancias y riquezas léxicas en la Plataforma Científica ECUCIENCIA se podrá obtener reportes de métricas de los documentos en formato PDF que se encuentran en la base de datos del sistema, así como conocimientos cienciométricos de los mismos.

- **Metodología**

En cuanto a la metodología empleada para el desarrollo del trabajo investigativo se puede resaltar que esta propuesta se guía en la modalidad de investigación cuantitativa principalmente. Respecto a la modalidad cuantitativa se puede decir que se tiene variables de análisis como la cantidad de producción científica generada por los docentes investigadores de la UTC referenciada por la cantidad de artículos, libros o ponencias publicadas.

La investigación propuesta busca dar una solución práctica a la problemática expuesta, para lo cual se debe establecer la manera de cómo aplicar el conocimiento científico en una solución tecnológica en concreto, es importante reconocer que la base teórica de una investigación es el sustento consolidado para encaminar el resto

de la investigación, las fuentes usadas en la consulta bibliográfica son de primer orden de bases de datos como Scopus, Scielo, entre otras.

Estar en el lugar de los hechos es indispensable para tener una aproximación real con la problemática detectada y de este modo poder hacer un seguimiento de los procesos en los que se encuentran involucrados los individuos a ser estudiados, en ese sentido es fundamental usar la investigación de campo para el desarrollo del presente trabajo, de igual modo la abstracción del mundo es una de las prácticas más importantes a la hora de analizar las situaciones problemáticas es por esta razón que el método sintético analítico es de mucha utilidad a lo largo del proceso de indagación puesto a que con el trabajo de investigación se busca aportar con nuevas ideas y generar mayores conocimientos hasta conseguir el dominio del contexto para evitar errores a la hora emitir de juicios de valor.

Por otro lado, los métodos de inducción y deducción sirvieron para establecer un razonamiento lógico de los aspectos particulares del objeto de estudio, así como los más generales, ello permitió pasar de un estado de análisis en ambas direcciones, de lo general a lo particular y de lo particular a lo general, pudiendo inferir importantes conclusiones al respecto y como parte de los resultados obtenidos.

La propuesta de usar metodologías de desarrollo ágiles particularmente Scrum para esta investigación nace debido a que los proyectos donde también intervienen el desarrollo de software, frecuentemente se enfrentan a dificultades para entregar prototipos a tiempo. En ese sentido Scrum ayuda a realizar proyectos de calidad en tiempos relativamente cortos, lo cual es posible porque este marco de referencia busca dividir tareas grandes y complejas en subtareas sencillas que pueden ser implementadas en un menor tiempo. Por ende el método de modelación es importante destacar en el desarrollo de cualquier producto informático y tratamiento algorítmico de procesos, en el caso de la presente investigación sirvió principalmente para entender y representar las funcionalidades de los algoritmos a través de un lenguaje de programación orientado hacia analítica de datos y netamente también en el desarrollo del módulo que valida los resultados del análisis

de la información contenida en los artículos científicos ubicados en la Plataforma Científica Ecuciencia.

CAPÍTULO I

1 Fundamentación Teórica

1.1 Antecedentes

En este capítulo se pueden apreciar algunos referentes teóricos importantes que describen las bases de la representación de la información en el contexto científico, plataformas como SCImago, Scielo y otras plataformas existentes en el mundo y que presentan excelente efectividad macro y meso de los distintos dominios científicos a nivel mundial, ello podrá constituir un importante acervo de consulta, así como los referentes hacia los complementos algorítmicos y tecnológicos necesarios en la investigación.

1.1.1 SCImago Journal

SCImago Journal & Country Rank es un portal de indicadores cienciométricos e informétricos que permite a investigadores, editores, especialistas en información y decidores en materia de política científica, en especial de los países subdesarrollados, seguir el comportamiento y el impacto de sus contribuciones a escala internacional. Para esto emplea la amplia colección de literatura disponible en Scopus de Elsevier.[8] Scopus es una base de datos bibliográfica de resúmenes y citas de artículos de revistas científicas y cubre más de 18 mil revistas siendo más del 90% de ellas del tipo arbitradas y pertenecientes a las áreas de ciencias, tecnología, medicina, ciencias sociales, artes y humanidades.[9]

La plataforma ha sido desarrollada por SCImago Research Group, un grupo de investigación de las universidades de Granada, Extremadura, Carlos III de Madrid y Alcalá de Henares de España, y es hoy en día la plataforma más inclusiva disponible para publicaciones. En su plataforma se encuentran ranking de impacto de las revistas y también de las instituciones de donde provienen los autores. SCImago incluye también un mapa que permite visualizar la investigación que se

realiza en los países iberoamericanos y publica todos los años el Ranking de Revistas y Países de SCImago.[9]

1.1.2 SciELO

SciELO (Scientific Electronic Library Online o Biblioteca Científica Electrónica en Línea) es un proyecto de biblioteca electrónica, iniciativa de la Fundación para el Apoyo a la Investigación del Estado de São Paulo, Brasil (Fundação de Amparo à Pesquisa do Estado de São Paulo — FAPESP) y del Centro Latinoamericano y del Caribe de Información en Ciencias de la Salud (BIREME),[10], que permite la publicación electrónica de ediciones completas de las revistas científicas mediante una plataforma de software que posibilita el acceso a través de distintos mecanismos, incluyendo listas de títulos y por materia, índices de autores y materias y un motor de búsqueda.

El proyecto SciELO, que además cuenta con el apoyo de diversas instituciones nacionales e internacionales vinculadas a la edición y divulgación científica,[11] , tiene como objetivo el «desarrollo de una metodología común para la preparación, almacenamiento, diseminación y evaluación de la literatura científica en formato electrónico». Actualmente participan en la red SciELO los siguientes países: Sudáfrica, Argentina, Brasil, Chile, Colombia, Costa Rica, Cuba, España, México, Peru, Portugal, Venezuela; además se encuentran en fase de desarrollo: Bolivia, Paraguay y Uruguay.[12]

1.1.3 Modelo de Espacio Vectorial

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. El trabajo realizado por [13] refleja que en los últimos años, por el tamaño y notoriedad que ha alcanzado este tema, la diversidad y cantidad de información 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.

El trabajo desarrollado por este autor tuvo como finalidad 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. 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. 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 "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". 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.

En la Universidad Técnica de Cotopaxi se han desarrollado varios proyectos de titulación enfocados principalmente a establecer procedimientos orientados hacia las siguientes temáticas dentro de ECUCIENCIA y que sirven como antecedentes importantes para la presente investigación:

Tabla 1.1: Proyectos de Investigación

Proyectos de investigación	**Año**	**Proyecto de Investigación Generativa**
Método para la determinación de similaridad y distancia entre investigadores a partir de Algoritmos de Clasificación	2018	ECUCIENCIA
Aplicación de algoritmo de extracción de texto en perfiles de usuario en caso de los investigadores de la Universidad Técnica de Cotopaxi	2018	ECUCIENCIA
Implementación de un algoritmo para la evaluación de los documentos científicos de los investigadores de la Universidad Técnica de Cotopaxi	2019	ECUCIENCIA
Implementación de un algoritmo de lógica difusa en la identificación de incentivos y recomendaciones para los investigadores de la Universidad Técnica de Cotopaxi	2019	ECUCIENCIA
Módulo de certificación y autoevaluación del investigador en la Plataforma Científica ECUCIENCIA	2020	ECUCIENCIA
Métodos analíticos de redes para identificar relaciones y colaboraciones científicas entre investigadores de la Universidad Técnica de Cotopaxi	2020	ECUCIENCIA
Visualización de información cienciométrica mediante mapeo autoorganizado en datos de producción científica de los docentes - investigadores de la Universidad Técnica de Cotopaxi	2020	ECUCIENCIA
Sistema georreferencial de análisis de ubicación geográfica, área de conocimiento y producción científica de los docentes investigadores de la UTC	2020	ECUCIENCIA

Elaborado por: El investigador

1.2 Fundamentación epistemológica

1.2.1 Cienciometría

La cienciometría estudia los aspectos cuantitativos de la ciencia como disciplina o actividad económica, forma parte de la sociología de la ciencia y encuentra aplicación en el establecimiento de las políticas científicas, donde incluye entre otras las de publicación. Ella emplea, al igual que las otras dos disciplinas estudiadas, técnicas métricas para la evaluación de la ciencia (el término ciencia se refiere, tanto a las ciencias naturales como a las sociales), y examina el desarrollo de las políticas científicas de países y organizaciones. [14]

A mediados de la década de 1970, se comenzó a reconocer la importancia del análisis cuantitativo de las actividades de ciencia y tecnología como un instrumento útil y eficaz en el aparato público ligado a la política y la planificación. La evaluación de la investigación a través de indicadores cuantitativos ha llegado a ser parte constitutiva de la agenda de la política científica en todo el mundo.[15]

La cienciometría es la ciencia que se encarga del estudio de la producción científica y tecnológica, a través de indicadores que permiten medir y analizar el impacto que genera en la sociedad las investigaciones desarrolladas. Para ejecutar el proceso de la obtención de similitud y distancia entre investigadores, es necesario realizar un estudio previo de la cienciometría, para en base a sus indicadores seleccionar las características correctas de los objetos de estudio.

1.2.2 La Cienciometría en América Latina

Como punto de “inflexión” del proceso de estructuración de la cienciometría en América Latina, se podría establecer el año 1995, cuando se creó la Red Iberoamericana de Indicadores de Ciencia y Tecnología (RICYT), auspiciada por el Programa Iberoamericano de Ciencia y Tecnología para el Desarrollo (CYTED) programa perteneciente a la UNESCO y la OEA. Su objetivo central era y sigue

siendo el de apoyar técnicamente a los países integrantes para que mejoren en materia de información en el ámbito de la ciencia, la tecnología y la innovación.[15]

En América Latina, el hecho de no haber podido avanzar de forma adecuada en materia de cienciometría se ha convertido en una de las mayores debilidades de los sistemas de ciencia, tecnología e innovación. Carecer de canales formales de interacción que promovieran objetivos colectivos, que apuntarán a un progreso sostenido de esos países utilizando como plataforma el diseño de políticas públicas basadas en información adecuada para tomar decisiones "confiables" en el avance de las actividades tecnocientíficas, se ha transformado en una de las causas de su atraso.[15]

Para corroborar los antecedentes previamente plasmados, se realizó una investigación de las plataformas de visualización científicas dentro de América Latina, a continuación de mencionan las varias de ellas:

a) DataCiencia

Es una plataforma de visualización de las dimensiones de la producción científica de Chile la que pretende relevar y visualizar la actividad científica de un modo comprensivo y sistémico. En este contexto, no se trata de un ranking de instituciones ni de personas.[16]

Esta herramienta permite visualizar, cuantificar y caracterizar la producción científica chilena segmentada en cuatro grandes categorías: Investigadores, Territorio (Regiones), Instituciones y Revistas Científicas. Todo esto a partir de la base de datos Web of Science (WoS) de Thomson Reuters que contiene la producción científica nacional del período 2008-2016.[16]

b) RedCiencia

Es un canal de comunicación y encuentro entre quienes viven la ciencia. Un espacio para destacar y diseminar el quehacer de investigadores, estudiantes y profesionales de todas las áreas del conocimiento, tanto a nivel nacional como internacional. Un

lugar donde encontrar oportunidades de crecimiento profesional, desde una mirada colaborativa e inclusiva.[17]

c) RedSearch

Es una herramienta que permite visualizar las relaciones de coautoría de documentos científicos chilenos del período 2008-2016 indizada en Web of Science. Mediante el análisis de estas relaciones de coautoría, la RedSearch entrega al usuario varias métricas relacionadas con la red pero también con los autores que la conforman. [18]

d) Redalyc.org

Es un proyecto académico para la difusión en Acceso Abierto de la actividad científica editorial que se produce en Iberoamérica. Es, en principio, una hemeroteca científica en línea de libre acceso y un sistema de información científica, que incorpora el desarrollo de herramientas para el análisis de la producción, la difusión y el consumo de literatura científica.[19]

El nombre Redalyc viene de Red de Revistas Científicas de América Latina, el Caribe, España y Portugal. El proyecto, impulsado por la Universidad Autónoma del Estado de México (en colaboración con cientos de instituciones de educación superior, centros de investigación, asociaciones profesionales y editoriales iberoamericanas), surge en el año 2003 como iniciativa de un grupo de investigadores y editores preocupados por la escasa visibilidad de los resultados de investigación generados en y sobre la región. Se ha propuesto, desde su creación, ser un punto de encuentro para los interesados en reconstruir el conocimiento científico de y sobre Iberoamérica.[19]

1.3 Fundamentación del Estado del Arte

1.3.1 Metodología de Minería de Datos KDD

El Descubrimiento de conocimiento en bases de datos (KDD, del inglés Knowledge Discovery in Databases) es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. [20]

Las etapas que componen el KDD hacen que el desarrollo sea iterativo e interactivo. Es iterativo, ya que dependiendo a la salida que se obtengan en cada etapa se puede regresar a un paso anterior, también porque a menudo son necesarias varias iteraciones para extraer conocimiento de alta calidad. Es importante mencionar que es interactivo porque involucra al usurario en la toma de muchas decisiones.

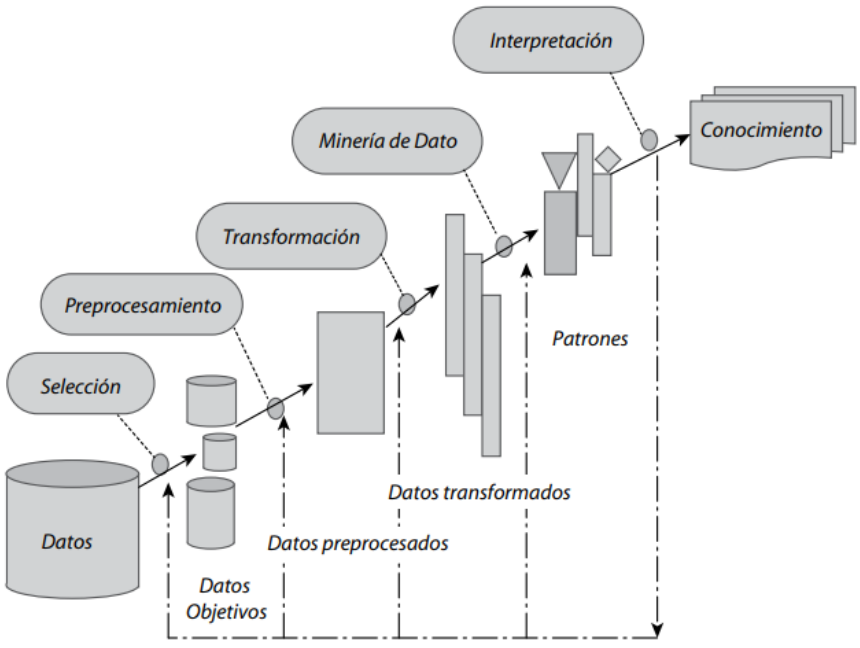


Figura 2.1: Etapas del Proceso KDD

Fuente: [20]

Etapas del Proceso KDD

- Selección
- Preprocesamiento/limpieza.
- Transformación/reducción.
- Minería de datos (data mining).
- Interpretación/evaluación

1.3.2 Técnicas de Minería de Datos

"Las técnicas de minería de datos constituyen un enfoque conceptual y, habitualmente, son implementadas por varios algoritmos".[21] Estas pueden clasificarse, según su utilidad, como se indica a continuación:

- **Las técnicas de predicción:** Permiten obtener pronósticos de comportamientos futuros a partir de los datos recopilados.[22] Estas técnicas resultan útiles, por ejemplo, en aplicaciones para predecir el parte meteorológico o en la toma de decisiones por parte de un cliente en determinadas circunstancias.

- **Las técnicas de Clustering:** El análisis de conglomerados o Clustering, es una técnica que permite analizar y examinar datos que no se encuentran etiquetados, formando conjuntos de grupos a partir de su similitud.[23] El principal objetivo de esta técnica es dividir un conjunto de objetos en dos o más grupos, dependiendo de las características que tengan en común cada uno de ellos.

  La similitud puede medirse a través de funciones de distancia, las cuales juegan un papel crucial, ya que individuos cercanos deberían ir para el mismo grupo. Se agrupan los objetos de acuerdo a todas las variables y por ello, una variable irrelevante puede generar ruido en los resultados obtenidos.[23]

- **Las técnicas de reglas de asociación:** Permiten establecer las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes; pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros.[21]

- **Las técnicas de clasificación:** Definen unas series de clases, en que se pueden agrupar los diferentes casos. Dentro de este grupo se encuentran las técnicas de árboles de decisión y reglas de inducción.[22]

Todas estas técnicas citadas, tiene como objetivo principal el análisis de datos extensos, para obtener como resultado información que ayude a interpretar el comportamiento de los objetos de estudio, que ayude a tomar decisiones. Dichas técnicas se aplican mediante algoritmos probados e implementados en soluciones de minería de datos.

- **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.[24]

- **Algoritmos de Clasificación**

Estos algoritmos tratan de clasificar en diferentes categorías una serie de ejemplos o instancias que representan cierta información de un problema. 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: [25]

1. **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.
2. **No supervisada:** los datos no tienen etiquetas (o no queremos utilizarlas) y estos se clasifican a partir de su estructura interna (propiedades, características).

- **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.[26]

1.3.3 Procesamiento del 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.[27]
El **procesamiento del lenguaje natural** (o LNP) 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.[28]

1.3.4 Corpus

Un **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.[29]

- **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.[30]

- **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"*:

A. 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:

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.

   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.

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.

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.

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 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.

6. **Corpus comparables:** Son corpus que seleccionan textos parecidos en cuanto a sus características en más de una lengua o en más de una variedad. Una de las principales finalidades de este tipo de corpus es poder comparar el comportamiento de diferentes lenguas o de diferentes variedades de una lengua en circunstancias de comunicación parecidas, pero evitando las inevitables distorsiones lingüísticas introducidas en las traducciones recogidas en los corpus paralelos.

7. **Corpus multilingües:** cuando se recopilan textos de diferentes lenguas sin que sean traducciones unos de otros y sin compartir criterios de selección, como lo hacen los textos que componen un corpus comparable, habría que hablarse de corpus multilingües.

8. **Corpus oportunista:** Corpus que recoge textos que encuentra disponibles

sin seguir ningún criterio de selección. Esto normalmente está motivado por la poca disponibilidad de textos en soporte electrónico (aunque cada vez se pueden encontrar en mayor cantidad) y por el elevado número de palabras necesarias para poder realizar muchos trabajos de investigación y la falta de recursos para obtenerlas.

B. 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:

1. **Corpus general:** Corpus que, al pretender reflejar la lengua común en su ámbito más amplio, se interesa por recoger cuantos más tipos de géneros mejor. Este tipo de corpus es útil para describir la lengua común de una colectividad, el lenguaje que utilizan los hablantes en situaciones comunicativas normales.

2. **Corpus especializado:** Se opone al corpus general. El corpus especializado recoge textos que puedan aportar datos para la descripción de un tipo particular de lengua. El corpus especializado es diferente al corpus que contempla una o más variedades de la lengua general (subcorpus); un corpus que recoja conversaciones de la calle no es un corpus especializado, como tampoco lo es uno que recoja el lenguaje de los periódicos; sí que lo sería, por ejemplo, un corpus que solo recogiera textos poéticos.

3. **Corpus genérico:** Corpus condicionado por el género de los textos que contiene, interesándose solo por algunos de ellos; por ejemplo, una recopilación de textos de revistas científicas especializadas o la selección de textos poéticos.

4. **Corpus canónico:** Corpus formado por todos los textos que configuran

lo obra completa de un autor, independientemente de los géneros.

5. **Corpus periódico o cronológico:** Corpus que recoge textos de unos años determinados o de unas épocas concretas.

6. **Corpus diacrónico:** Corpus que incluye textos de diferentes etapas temporales sucesivas en el tiempo con el fin de poder observar evoluciones en la lengua.

C. Según la cantidad de texto que se recoge de cada documento

1. **Corpus textual:** Corpus que recoge íntegramente todos los textos de los documentos que lo constituyen. Se entiende como textos enteros las series de frases y/o párrafos coherentes, homogéneos estilísticamente y completos en sí mismos. Las novelas, por ejemplo, son un prototipo de texto que cumple estos requisitos, pero hay otros tipos de documentos que también se adaptan a esta definición, se considera un texto entero a las recopilaciones de pequeños anuncios de periódico o colecciones de poemas cortos de un mismo autor. A veces incluso todos los artículos de un periódico o de una revista se han considerado como un solo texto, aunque es más razonable considerar como un solo texto los diversos artículos de una misma sección (economía, deportes, editoriales, etc.)

2. **Corpus de referencia:** Corpus formado por fragmentos de los textos de los documentos que lo constituyen. En este caso no interesa tanto el texto en sí sino el nivel de lengua que representan. En este tipo de corpus son muy importantes los aspectos de equilibrio y representatividad en la selección de los fragmentos.

3. **Corpus léxico:** Corpus que recoge fragmentos de textos muy pequeños y de longitud constante de cada documento. En este caso el interés de los diseñadores del corpus está en el léxico.

D. 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:

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.

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 título, de principio de capítulo, de cambio de lengua, etc.) - codificación- o etiquetas analíticas de algunos aspectos lingüísticos (indicación de frase subordinada, de aspectos pragmáticos, etc.). - anotación - De todos modos, es importante que las etiquetas usadas para codificar y anotar los textos sean siempre extratextuales, de manera que se puedan reconocer y, si es necesario, eliminar fácilmente. También es importante que se usen sistemas de codificación estándares para asegurar la transportabilidad y reusabilidad de los textos.

E. Según la documentación que acompaña a los textos

Otra clasificación que se puede hacer de los corpus es en función de si los textos que los componen están documentados o no.

1. **Corpus documentado:** Corpus en el que cada documento que lo compone lleva asociado un archivo DTD (Document Type Definition) o una cabecera “header” de descripción de su filiación y sus constituyentes.

2. **Corpus no documentado:** Corpus en el que sus textos constituyentes no disponen de ningún apartado o archivo relacionado donde se describan

sus elementos o su filiación.

1.3.5 Herramientas y Librerías

- **Anaconda**

Es una distribución libre y Open Source de los lenguajes de programación Python y R muy usada en computación científica (Data ScienceData Science, Machine Learning, Ciencia, Ingeniería, analítica predictiva,Big Data, etc.).[31]

- **Python**

Es un lenguaje orientado a objetos que cuya versatilidad nos permite utilizarlo aplicando diferentes paradigmas de programación. 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.[32]

i. Librerías de Python

En este apartado se comentarán algunas librerías de Python, que sirven de apoyo para la minería de datos y la programación, junto con una breve descripción de cada una de ellas.

A. Librería

En el ámbito de la programación, una librería es un conjunto de archivos que implementan un grupo de funciones, codificadas en un lenguaje de programación concreto, preparadas para ser utilizadas de forma fácilmente accesible al programar en dicho lenguaje.[33]

- **NLTK, Natural Language Toolkit**

El NLTK (Natural Language Toolkit) es una biblioteca de Procesamiento de Lenguaje Natural que utiliza el lenguaje de programación Python.[34] NLTK es software libre, lo que permite a estudiantes y al personal académico realizar estudios con la herramienta sin necesidad de realizar una inversión económica. Esta herramienta es también de código abierto, lo que lo hace ideal para expandir sus funcionalidades en caso de necesitarlo. 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.

Una razón 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 Natural de mayor aceptación en el ámbito científico.

Otra razón importante es el apoyo que proporciona el libro Natural Language Processing with Python.[34] Este libro tiene por autores a los creadores del NLTK, haciéndolo idóneo para comprender y utilizar todas las funcionalidades que aporta esta biblioteca.

El **NLTK** es la librería líder para el procesamiento de lenguaje natural. Proporciona interfaces fáciles de usar a mas de cincuenta corpus y recursos léxicos, junto con un conjunto de bibliotecas de procesamiento de texto para la clasificación, tokenizacion, el etiquetado, el análisis y el razonamiento semántico.[35]

- **Pandas**

Pandas es un paquete de Python que proporciona estructuras de datos rápidas, flexibles y expresivas diseñadas para trabajar con los datos "relacionales" o "etiquetados" son fáciles e intuitivos. Su objetivo es el análisis de datos prácticos y reales en Python. Además, tiene el objetivo más amplio de convertirse en una

herramienta flexible de análisis / manipulación de datos de código abierto disponible en cualquier idioma. Pandas es adecuado para diferentes tipos de datos:[36]

- Datos tabulares con columnas de tipo heterogéneo, como en una tabla de SQL o una hoja de cálculo de Excel.
- Datos de series de tiempo ordenados y desordenados (no necesariamente de frecuencia fija).
- Datos matriciales arbitrarios (tipificados homogéneamente o heterogéneos) con etiquetas de fila y columna.
- Cualquier otra forma de conjuntos de datos observacionales / estadísticos. Los datos realmente no necesitan ser etiquetados en absoluto para ser colocados en una estructura de datos pandas.

Estas son solo algunas de las cosas que los pandas hacen bien:[36]

- Alineación automática y explícita de datos: los objetos pueden alinearse explícitamente a un conjunto de etiquetas, o el usuario puede simplemente ignorar las etiquetas y deje que Series, DataFrame, etc., alinee automáticamente los datos en los cálculos.
- Potente y flexible grupo por funcionalidad para realizar operaciones de combinación de aplicación dividida en conjuntos de datos, tanto para agregar como para transformar datos.
- Facilita la conversión de datos irregulares, indexados de manera diferente en otras estructuras de datos de Python y NumPy a Objetos DataFrame.
- Rebanado inteligente basado en etiquetas, indexación elegante y subconjunto de grandes conjuntos de datos.
- Combinación intuitiva y unión de conjuntos de datos.
- Etiquetado jerárquico de ejes (posible tener múltiples etiquetas por tic)

- **Numpy**

NumPy (acrónimo de Numeric Python) es un módulo fundamental para el cálculo científico con Python. Con él se dispone de herramientas computacionales para manejar estructuras con una gran cantidad de datos, diseñadas para obtener un buen nivel de rendimiento en su manejo. El módulo incorpora un nuevo tipo de dato, el array, similar a una lista, pero que es computacionalmente mucho más eficiente. Además posee una gran cantidad de métodos que permiten manipular los elementos del array de forma no secuencial, lo que se denomina vectorización, y que ofrece un alto grado de rendimiento.[37]

- **Math**

El módulo math agrega las funciones trigonométricas seno, coseno y tangente que se representan, respectivamente mediante sin, cos y tan. Por defecto, esas funciones asumen que los ángulos se miden en radianes. Por su parte, el logaritmo natural (o neperiano) de base e se llama en Python log y la exponencial (para calcular e elevado a un número) se llama exp. Pero, además de funciones, a menudo los módulos de Python contienen otro tipo de objetos. Por ejemplo, el módulo math contiene valores aproximados de las constantes matemáticas π y e, entre otras.[38]

- **TfidfVectorizer**

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

- **Librería SKlearn**

EL proyecto SKlearn es una librería para aprendizaje automático de código abierto escrita en Python, eficientes para la minería de datos y el análisis de datos. Cuenta con varios algoritmos de clasificación, regresión y clustering incluyendo máquinas de vectores soporte, regresión logística, Naives Bayes, k-medias, etc. y está

diseñado para interoperar con las bibliotecas numéricas y científicas de Python NumPy y Scipy. Es un proyecto muy popular en Github, presentando en agosto de 2014.[39]

B. Algoritmos Clustering

El proceso de clustering consiste en la división de los datos en grupos de objetos similares. Para medir la similitud entre objetos se suelen utilizar diferentes formas de distancia: distancia euclidiana, de Manhattan, de Mahalanobis, etc. Clustering es una técnica más de Aprendizaje Automático, en la que el aprendizaje realizado es no supervisado. Desde un punto de vista práctico, el clustering juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la información y minería de texto, aplicaciones sobre bases de datos espaciales y aplicaciones web.[40]

La agrupación de datos sin etiquetar se puede realizar con el módulo sklearn.cluster, Una cosa importante a tener en cuenta es que los algoritmos implementados en este módulo pueden tomar diferentes tipos de matriz como entrada.

A continuación se realizara una breve descripción de varios algoritmos que vienen incorporados al módulo sklearn.cluster

- **K-Means**

El K-Means algoritmo que agrupa los datos al tratar de separar muestras en n grupos de igual varianza, minimizando un criterio conocido como la inercia o la suma de cuadrados dentro del clúster. Este algoritmo requiere que se especifique la cantidad de grupos. Se adapta bien a un gran número de muestras y se ha utilizado en una amplia gama de áreas de aplicación en muchos campos diferentes.[41][42]

De acuerdo a la literatura [43] se pueden identificar cuatro pasos en el algoritmo:

Inicialización: Se definen un conjunto de objetos a particionar, el número de grupos y un centroide por cada grupo. Algunas implementaciones del algoritmo estándar determinan los centroides iniciales de forma aleatoria; mientras que algunos otros procesan los datos y determinan los centroides mediante de cálculos.

Clasificación: Para cada objeto de la base de datos, se calcula su distancia a cada centroide, se determina el centroide más cercano, y el objeto es incorporado al grupo relacionado con ese centroide.

Cálculo de centroides: Para cada grupo generado en el paso anterior se vuelve a calcular su centroide.

Condición de convergencia: Se han usado varias condiciones de convergencia, de las cuales las más utilizadas son las siguientes: converger cuando alcanza un número de iteraciones dado, converger cuando no existe un intercambio de objetos entre los grupos, o converger cuando la diferencia entre los centroides de dos iteraciones consecutivas es más pequeño que un umbral dado. Si la condición de convergencia no se satisface, se repiten los pasos dos, tres y cuatro del algoritmo.

- **MiniBatchKMeans (Mini Lote K-Means)**

El MiniBatchKMeans es una variante del K-means algoritmo que utiliza mini-lotes para reducir el tiempo de cálculo, mientras que todavía intentar optimizar la misma función objetivo. Los mini lotes son subconjuntos de los datos de entrada, muestreados aleatoriamente en cada iteración de entrenamiento. Estos mini lotes reducen drásticamente la cantidad de cálculos necesarios para converger a una solución local.[44]

El algoritmo itera entre dos pasos principales. En el primer paso, las muestras se extraen al azar del conjunto de datos, para formar un mini-lote. Estos se asignan al centroide más cercano. En el segundo paso, se actualizan los centroides. En contraste con k-means, esto se realiza en base a cada muestra. Para cada muestra en

el mini-lote, el centroide asignado se actualiza tomando el promedio de transmisión de la muestra y todas las muestras anteriores asignadas a ese centroide. Esto tiene el efecto de disminuir la tasa de cambio de un centroide a lo largo del tiempo. Estos pasos se realizan hasta que se alcanza la convergencia o un número predeterminado de iteraciones.[41][42]

MiniBatchKMeans converge más rápido que K-means, pero se reduce la calidad de los resultados. En la práctica, esta diferencia en la calidad puede ser bastante pequeña.[41]

- **MeanShift (Cambio de media)**

MeanShift, el agrupamiento apunta a descubrir manchas en una densidad suave de muestras. Es un algoritmo basado en centroide, que funciona mediante la actualización de los candidatos para que los centroides sean la media de los puntos dentro de una región determinada. Estos candidatos luego se filtran en una etapa de procesamiento posterior para eliminar los duplicados cercanos para formar el conjunto final de los centroides.[44]

El algoritmo establece automáticamente el número de grupos, en lugar de depender de un parámetro bandwidth, que determina el tamaño de la región por la que se realiza la búsqueda. Este parámetro se puede configurar manualmente, pero se puede estimar utilizando la función estimate_bandwidth, que se llama si el ancho de banda no está configurado.[41][42]

El algoritmo no es altamente escalable, ya que requiere varias búsquedas de vecinos más cercanos durante la ejecución del algoritmo. Se garantiza que el algoritmo converge, sin embargo, el algoritmo dejará de iterar cuando el cambio en los centroides sea pequeño.[42]

- **Spectral Clustering (Agrupamiento espectral)**

SpectralClustering hace una incorporación de baja dimensión de la matriz de afinidad entre muestras, seguida de un KMeans en el espacio dimensional bajo. Es especialmente eficiente si la matriz de afinidad es escasa y el módulo pyamg está instalado. SpectralClustering requiere que se especifique la cantidad de clusters. Funciona bien para una pequeña cantidad de grupos, pero no se recomienda cuando se usan muchos grupos.[42]

La agrupación espectral se utiliza para los datos que están conectados, pero no necesariamente aislados de una manera que puede tener lugar la optimización convexa. El objetivo básico es dividir los puntos de datos de un gráfico dado en grupos de puntos similares que son diferentes de otros grupos. Esto produce un número específico de grupos de puntos que son matemáticamente similares. Para implementar la agrupación espectral, es necesario determinar el número de clusters (agrupaciones).[45]

- **Agglomerative Clustering (Agrupamiento jerárquico)**

La agrupación jerárquica es una familia general de algoritmos de agrupación en clústeres que crean agrupaciones anidadas al fusionarlas o dividirlas sucesivamente. Esta jerarquía de grupos se representa como un árbol (o dendrograma). La raíz del árbol es el único grupo que reúne todas las muestras, las hojas son los grupos con una sola muestra.[41]

En un clustering jerárquico se utiliza principalmente el método de aglomeración, donde cada elemento comienza como un clúster individual. En cada etapa se va construyendo un árbol jerárquico, donde los dos grupos más cercanos se van uniendo en un mismo nodo hasta finalmente terminar en un nodo único superior. La agrupación jerárquica es representada por un esquema en dos dimensiones llamado dendrograma o árbol jerárquico, el cual muestra las uniones entre grupos

realizadas en cada etapa. Un dendrograma tendrá 2N-1 nodos, donde N es el número de elementos a agrupar.[46]

C. Técnicas de preprocesamiento de textos

Es muy importante el preprocesamiento de los textos al momento de realizar la clasificación automática de textos, causando una mejora en el rendimiento del proceso de la clasificación. Podemos decir que se mejora porque existe la eliminación de elementos que estén redundantes en los documentos sin perder significancia en los procedimientos necesarios. Algunas técnicas de preprocesamiento de textos propuestas por [7] son:

- **Stop Words** se definen como términos que se consideran irrelevantes para la clasificación del documento, ya sea porque no presentan un contenido relevante que ayude al clasificador o por las posibles ocurrencias repetidas en el texto.
- **Stemming** esta técnica consiste en el enraizado de palabras comunes y agruparlas en un mismo grupo, eliminado así posibles redundancias por alcance de significados.
- **Normalización de frecuencias** en el momento en el que se realice el proceso de frecuencia de palabras para textos largos, puede ocurrir que palabras relevantes para el clasificador se repitan un número considerable de veces, por lo que no es raro realizar una normalización de estas frecuencias para ahorrar espacio de recursos computacionales, junto con una mejor representación de estos.
- **Categorización de las características** es una técnica de categorizar ciertas características de textos para realizar una transformación de características similares agrupándolas en una categoría de características madre.

D. Framework django

Ahora bien, Python trabaja conjuntamente con el framework Django debido a que presenta grandes beneficios para aquellos usuarios que hagan uso del mismo, tales como: posee ORM incluido, un sistema de plantillas, trabaja con MVC, servidor incluido para hacer pruebas, ya que con unos cuantos scriptings y se logra obtener eficientes resultados que sin lugar a duda será de gran utilidad para simplificar el proceso de desarrollo de la plataforma científica, es por ello que según la página oficial de Django.[47]

Características:

- Genera información de manera estructurada y que sirve para administrar adecuadamente cualquier proyecto en desarrollo.
- Proporciona un sistema extensible de plantillas.
- Permite la normalización de URLs.
- Ofrece una portabilidad eficiente para diferentes Sistemas Operativos.
- La documentación que emite es completamente fácil de comprender por todos aquellos usuarios que se inclinen por el ámbito de la programación.
- Escalable.[47]

E. Métricas de distancias

Existen diversas distancias y maneras de calcularlas, las que se aplican con mayor frecuencia 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.

- **Distancia Euclidiana:** Es la raíz cuadrada de la suma de las diferencias al cuadrado entre los valores de dos casos para cada variable.

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

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

- **Distancia Chebychev:** Calcula la distancia de Chebyshev entre los puntos. [4] 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:

$$d(u, v) = \max_i |u_i - v_i|$$

- **Distancia Minskow** SKLEARN**:** Esta distancia puede considerarse una generalización de las distancias euclideas y Manhattan.[5] Viene definida por la siguiente expresión:

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

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

- **Distancia Coseno:** 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. 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:

$$1 - \frac{u \cdot v}{||u||_2 ||v||_2}.$$

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

F. Coeficiente de Jaccard

Conocido como razón de similaridad, se aplica en muchos casos a 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.

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:

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

1.4 Conclusiones

- EL presente capítulo tiene un rol fundamental en el desarrollo de la investigación planteada, ya que en el mismo se ha recopilado una gran cantidad de información de libros, artículos científicos, entre otras fuentes científicas de consulta que aportaron invaluablemente con la problemática analizada.
- Se ha considerado también los aspectos principales que conciernen al uso del corpus en el estudio lingüístico, que sirvió para el análisis de los contenidos de los PDF de los artículos científicos de los docentes de la UTC.
- Las metodologías ágiles particularmente Scrum permite aplicar de manera regular un conjunto de buenas prácticas para trabajar colaborativamente, en equipo, y obtener el mejor resultado posible de un proyecto optimizándolo tanto en tiempo como recursos.

CAPÍTULO II

2 Propuesta

2.1 Metodología KDD

Para el procesamiento y análisis de la información se ha seguido lo establecido por la metodología KDD compuesta por varias etapas, estas son:

1. **Abstracción del escenario:**

La plataforma científica ECUCIENCIA hospedada en http://ecuciencia.utc.edu.ec/ presenta una base de datos en PostgreSQL en su Versión 12.3, contiene toda la información relacionada a los investigadores de la Universidad Técnica de Cotopaxi (UTC) incluyendo su producción científica que contempla la publicación de artículos científicos, libros o capítulos de libros y ponencias, este cúmulo de información permite tener una fuente idónea para realizar diversas actividades de clasificación y minería, entre otros.

El propósito de esta plataforma científica en su primera etapa era almacenar toda la actividad científica de la UTC, su segunda etapa es la aplicación de lógicas para recuperar informaciones relacionadas con la cantidad de artículos, ponencias y visualizar estos resultados estadísticos y su tercera etapa es la implementación de algoritmos de inteligencia artificial y minería de datos para descubrir conocimientos, este proyecto se enfoca hacia una parte de esta última etapa.

La base de datos contiene los registros de palabras claves, título y resúmenes como campos considerados para analizar, muchos algoritmos utilizados clasifican en estos tres indicadores, pero actualmente en la plataforma no se cuenta con procedimientos ni algoritmos que se encarguen de analizar el corpus contenidos dentro de los PDF, siendo el objeto fundamental de la presente investigación, de

manera que se pueda asociar a las líneas de investigación y establecer relaciones entre estas y los artículos, así como entre los propios artículos.

2. **Selección de los datos:**

Los datos que se pretenden seleccionar son los que están relacionados con la tabla de artículos y para ello se usarán como herramienta para exportar los datos asociados a ella el software pgAdmin en su versión 4.21, las instrucciones a utilizar de SQL una vez dentro de la base de datos será:

- SELECT * FROM public."Articulos_Cientificos_articulos_cientificos";

Con ellos seleccionar todos los campos de la tabla "Articulos_Cientificos_articulos_cientificos"

3. **Preprocesamiento/limpieza:**

En esta etapa se determina la confiabilidad de la información, es decir, realizar tareas que garanticen la utilidad de los datos. Para esto se hace la limpieza de datos (tratamiento de datos perdidos o remover valores atípicos). Esto implica eliminar variables o atributos con datos faltantes o eliminar información no útil.

La tabla "Articulos_Cientificos_articulos_cientificos" tiene diversos campos, no todos serán usados, solo lo que sean pertinentes al objeto principal de estudio, para ello se seguirán los siguientes pasos:

a) Optimizar el nombre del csv descargado de la base de datos realizada en el paso 2 selección de datos.
b) Analizar y seleccionar los campos pertinentes a documentos que representan los artículos científicos en formato PDF.
c) Identificar caracteres especiales que se encuentran en el csv de análisis para su descarte.

d) Eliminar elementos vacíos como por ejemplo campos que no hayan sido llenado por los investigadores.

4. Transformación/reducción:

En esta etapa 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) o bien transformaciones como por ejemplo convertir los valores que son números a categóricos (discretización).

En esta etapa se usarán las potencialidades de Python como herramienta de análisis de datos, para determinar estadísticas y parámetros relevantes en el proceso.

5. Muestreo:

En la base de datos en el momento de análisis se tiene una población de 636 artículos científicos, esto como se aprecia para el procesamiento de los 636 se hará una vez que se desarrolle el módulo siguiendo las premisas de los algoritmos que se usarán, se tomarán 5 artículos de los 636, para un 8% de representatividad, usando los criterios del muestreo aleatorio simple, debido que unos de los principios de este muestreo es de que cada muestra tiene la misma posibilidad de ser elegida, este tipo de muestreo se llevó a cabo para obtener los resultados 5 artículos, su selección de manera aleatoria y luego generalizar la lógica hacia el resto de los artículos en formato PDF a través del desarrollo de un módulo para la plataforma científica ECUCIENCIA.

6. Minería de datos (data mining):

Fase en la que se refiere a elegir el paradigma apropiado de Minería de Datos, ya sea la clasificación, regresión o agrupación, según los objetivos que se haya planteado para la investigación (predicción o descripción), la primera ocupada para

encontrar un modelo que sea utilizada para casos futuros y desconocidos; mientras que la segunda solo para observar su comportamiento.

En este sentido serán analizados algunos algoritmos de clasificación, procesamiento del lenguaje natural y librerías asociadas a Python, las principales son:

- NLTK.
- NUMPY.
- MATPLOTLIB.
- PYPDF2.
- SCIKIT-LEARN.
- SCIPY.
- SKLEARN.

7. Análisis algorítmicos para los pdf de los artículos científicos:

a) Descripción de las bases algorítmicas del modelo espacio vectorial

Podemos considerar una base de documentos (U), compuesta por documentos u_i, donde contienen un conjunto de términos (T), formado por n términos tj, en la que cada documento ui contiene un número de términos. De esta forma, es posible representar a cada documento como un vector perteneciente a un espacio n-dimensional, siendo n el número de términos ingresados en el documento que forman el conjunto T:

$$u_i = (t_{i1}; t_{i2}; t_{i3}; \dots\dots\dots; t_{in})$$

Donde cada uno de los elementos tij de este vector puede representar la presencia, ausencia o relevancia del término tj en el usuario ui en su perfil.

La representación de cada vector-documento tendrá n componentes, de los cuales los que estén referenciados tendrán un valor diferente de 0, mientras que los que no estén referenciados tendrán un valor nulo o 0.

Un primer enfoque se basa en contar las ocurrencias de cada término en un documento, medida que se denomina frecuencia del término i-ésimo en el documento j-ésimo, y se nota como $tf_{i,j}$.

Una segunda medida de la importancia del término o palabra es la conocida como frecuencia documental inversa de un término en la colección, conocida normalmente por sus siglas en inglés: idf (inverse document frequency), como reflejan y que responde a la siguiente expresión:

$$w_{i,j} = tf_{i,j} \times Log\left(\frac{N}{n_i}\right)$$

Donde N es el número de documentos de la colección, y n_i el número de documentos donde se menciona al término i-ésimo, si asociamos al caso de la presente investigación a N con U como el número de documentos de la base de datos en la tabla de artículos científicos, y n_i como el número de documentos que contienen en su corpus el término i, entonces es posible determinar la importancia, peso o frecuencia de cada término en el documento.

Finalmente se tendría una matriz de vectores-documentos por términos como se muestra a continuación:

	t_1	t_2	t_3		t_n
Doc_1	**1**	**2**	**1**		n
Doc_2	**1**	**1**	**1**	⋯	n
Doc_3	**0**	**2**	**1**		n
⋮	⋮	⋮	⋮	⋱	⋮
Doc_n	n	n	n	⋯	n

La lógica usada con Python sería el algoritmo TfidfVectorizer que realiza todos los pasos anteriores en python:

- from sklearn.feature_extraction.text import TfidfVectorizer

b) Para la similitud entre los documentos de la tabla artículos científicos de la base de datos del sistema Ecuciencia:

Se tiene en consideración el cálculo de similitud entre los vectores que componen el peso-frecuencia, que en esencia son los vectores-documentos. Aquí se establece un modelo matemático basado en el cálculo del coeficiente de similaridad entre vectores. Este modelo de cierta forma responde a las necesidades del presente estudio, ya que para obtener el grado de relevancia de un documento u_i según su corpus con respecto a los demás que componen la tabla de artículos, es posible establecer la similaridad entre los vectores de esta matriz, o sea cada vector lo constituirá un documento en pdf por cada artículo de la tabla objeto de estudio y será posible determinar la similitud de cada documento con respecto a los demás. El sistema toma un valor real que será tanto mayor cuanto más similares sean los documentos que se analizan.

El modelo vectorial hace la suposición básica de que la proximidad relativa entre dos vectores es proporcional a la distancia semántica de los documentos. Existen diferentes funciones para medir la similitud entre vectores, todas ellas están basadas en considerar a ambos como puntos en un espacio ndimensional como se describen a continuación:
Producto escalar:

$$Producto\ escalar\ (A,B) = \sum_{j=1}^{n} A_j \cdot B_j$$

donde A_j y B_j son, respectivamente, los pesos asociados al término t_j en la representación de los documentos A y B.
Función del coseno en python from sklearn.metrics.pairwise import cosine_similarity:

$$F\ \cos(A,B) = \frac{\sum_{j=1}^{n} A_j \cdot B_j}{\sqrt{\sum_{j=1}^{n} A_j^2 \cdot \sum_{j=1}^{n} B_j^2}}$$

Índice de Dice (ID) en python import scipy.spatial.distance as distance:

$$ID\ (A,B) = \frac{2 \cdot \sum_{j=1}^{n} A_j \cdot B_j}{\sum_{j=1}^{n} A_j^2 + \sum_{j=1}^{n} B_j^2}$$

Índice de Jaccard (IJ) en python from jaccard_index.jaccard import jaccard_index:

$$IJ\ (A,B) = \frac{\sum_{j=1}^{n} A_j \cdot B_j}{\sum_{j=1}^{n} A_j^2 + \sum_{j=1}^{n} B_j^2 - \sum_{j=1}^{n} A_j \cdot B_j}$$

Una matriz de similitud puede quedar representada simétricamente, donde cada elemento δ_{ij} de M representa la similaridad entre el estímulo i y el estímulo j como se muestra a continuación:

$$M = \begin{pmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \delta_{23} & \cdots & \delta_{2n} \\ \delta_{31} & \delta_{32} & \delta_{33} & & \delta_{3n} \\ & \vdots & & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \delta_{n3} & \cdots & \delta_{nn} \end{pmatrix}$$

De esta manera queda determinada la matriz de similitud de los documentos, de forma tal que pueden ser identificados los niveles de similaridad entre los documentos en PDF de los artículos científicos de la plataforma Ecucienca partiendo de su corpus. Todo esto también brinda la posibilidad de establecer conglomerados de documentos, así como la posibilidad de ser representados en un escalamiento multidimensional o por sus siglas en inglés Multidimensional Scaling (MDS).

c) Escalamiento multidimensional para graficar grupos de artículos:

El MDS es una técnica de representación espacial que trata de visualizar sobre un mapa un conjunto de estímulos cuya posición relativa se desea analizar. El propósito del MDS es transformar los juicios de similitud o preferencia llevados a cabo por una serie de individuos sobre un conjunto de objetos o estímulos en distancias susceptibles de ser representadas en un espacio multidimensional. El MDS está basado en la comparación de objetos o de estímulos, de forma que si un individuo juzga a los objetos A y B como los más similares entonces las técnicas

de MDS colocarán a los objetos A y B en el gráfico de forma que la distancia entre ellos sea más pequeña que la distancia entre cualquier otro par de objetos.
Para la introducción de técnicas de Escalamiento Multidimensional son precisos dos requisitos esenciales estos son:

i. Partir de un conjunto de números, llamados proximidades o similaridades, que expresan todas o la mayoría de las combinaciones de pares de similaridades dentro de un grupo de objetos.
ii. Contar con un algoritmo para llevar a cabo el análisis.

El punto de partida es una matriz de similaridad entre n objetos, con el elemento δ_{ij} en la fila i y en la columna j, que representa la similaridad del objeto i al objeto j. También se fija el número de dimensiones, p, para hacer el gráfico de los objetos en una solución particular.

Se sigue el siguiente algoritmo:

1. Arreglar los n objetos en una configuración inicial en p dimensiones, esto es, suponer para cada objeto las coordenadas (x_1, x_2, ..., x_p) en el espacio de p dimensiones.

2. Calcular las distancias euclidianas entre los objetos de esa configuración, esto es, calcular las d_{ij} que son las distancias entre el objeto i y el objeto j.

$d(O_i, O_j) = \sqrt{\sum_{k=1}^{n}(x_k(O_i) - x_k(O_j))^2}$ distancia euclidiana.

Donde O_i y O_j son los objetos para los cuales se desea calcular la distancia, n es el número de características de los objetos del espacio y x_k(Oi), x_k(O_j) es el valor del atributo k-ésimo en los objetos O_i y O_j, respectivamente.

De tal manera también debe verificarse los tres axiomas siguientes:

- $d(x,y) \geq 0 \quad \forall\, x,y \in X, y\ d(x,y) = 0$ *si y solo si* $x = y$
- $d(x,y) = d(y,x)\ \forall\, x,y \in X$ *(simetría)*
- $d(x,z) \leq d(x,y) + d(y,z)\ \forall\, x,y,z \in X$ *(desigualdad triangular)*

3. Hacer una regresión de d_{ij} sobre δ_{ij}. Esta regresión puede ser lineal, polinomial o monótona. Utilizando el método de los mínimos cuadrados se obtienen estimaciones de los coeficientes a y b, y de ahí puede obtenerse lo que genéricamente se conoce como una "disparidad".

$$\hat{d}_{ij} = \hat{a} + \hat{b}\delta_{ij}$$

Si se supone una regresión monótona, no se ajusta una relación exacta entre d_{ij}, sino se supone simplemente que, si δ_{ij} crece, entonces d_{ij} crece o se mantiene constante.

4. A través de algún estadístico conveniente, se mide la bondad de ajuste entre las distancias de la configuración y las disparidades. Existen diferentes definiciones de este estadístico, pero la mayoría surge de la definición del llamado índice de esfuerzo (en inglés: STRESS).

Uno de los criterios más utilizados es el siguiente:

$$STRESS1 = \sqrt{\frac{\sum\sum(d_{ij} - \hat{d}_{ij})^2}{\sum\sum d_{ij}^2}}$$

Todas las sumatorias sobre i y j van de 1 a p y las disparidades dependen del tipo de regresión utilizado en el tercer paso del procedimiento.

El STRESS1 es la fórmula introducida por Kruskal quien ofreció la siguiente guía para su interpretación en la tabla 2.1:

Tabla 2.1: Interpretaciones del Stress

Tamaño del STRESS1	Interpretación
0.2	Pobre
0.1	Regular
0.05	Bueno
0.025	Excelente
0.00	Perfecto

Fuente: [48]

Las coordenadas (x_1, x_2, ..., x_t) de cada objeto se cambian ligeramente de tal manera que la medida de ajuste se reduzca.

Algoritmos establecidos para este procesamiento con Python son:

- El algoritmo scipy.spatial.distance
- Del módulo jaccard_index.jaccard el algoritmo jaccard_index
- Del módulo sklearn.metrics.pairwise usar el algoritmo cosine_similarity

En el marco teórico en la sesión de métricas de distancias se abordan las formas de obtener las distancias, coeficiente e índices.

d) Análisis algorítmicos para el tokenizado de los documentos PDF:

El proceso de tokenizado forma parte del procesamiento del lenguaje natural. La tokenización consiste en dividir cadenas de texto más largas en piezas más pequeñas o tokens.

Para el tokenizado se siguen los siguientes pasos:

1. Extraer de la base de datos los PDFs de cada uno de los artículos científicos.
2. Conversión de los PDFs en TXT.
3. Subdivisión del contenido del documento TXT en palabras separadas (tokenizado) para ello del módulo nltk.corpus importar los stopwords en español.
4. Determinar la frecuencia de las palabras (incluyendo palabras de paradas o stopwords) con el algoritmo TfidfVectorizer.
5. Eliminar palabras de parada.
6. Determinar la frecuencia de las palabras (sin palabras de paradas o stopwords) con el algoritmo TfidfVectorizer.

El diagrama de flujo en la figura 2.1 describe la lógica que se sigue:

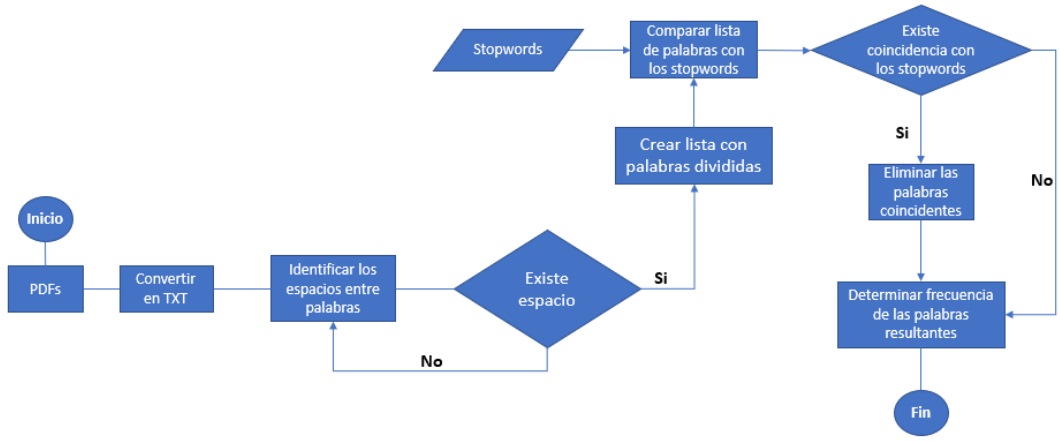


Figura 2.1: Diagrama de flujo que describe el proceso de tokenizado

Elaborado por: El investigador

8. Interpretación/evaluación:

En esta etapa se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones, también se puede incluir la visualización de los patrones extraído. Por otra parte, se consolida el conocimiento descubierto para incorporarlo en otro sistema para posteriores acciones o, simplemente, para documentarlo y reportarlo a las partes interesadas; también para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto.[49]

En esta etapa se usó los principios de validación cruzada para lo cual se toma la las frecuencias de los 5 artículos de muestra aplicando para ello la función sklearn.model_selection.train_test_split ya permite dividir un dataset en dos bloques, típicamente bloques destinados al entrenamiento y validación del modelo (se llamará a estos bloques "bloque de entrenamiento " y "bloque de pruebas" para mantener la coherencia con el nombre de la función.

La función train_test_split de la librería sklearn.model_selection, añade por defecto al bloque de entrenamiento el 75% de los registros y al bloque de pruebas el 25% restante, pero con el parámetro test_size se puede ajustar este valor, para el caso de la presente investigación se usa 0.2 significando el 20% para el bloque de prueba y 80% para el bloque de entrenamiento.

Se usaron un total de 450 registros, 360 registros para cada conjunto de Train y 90 registros para el conjunto de test de forma aleatoria.

Para la evaluación del algoritmo de medición se aplicó el error cuadrático medio donde se consideraron como valores reales las similaridades derivadas de las frecuencias como valores reales y como estimador del modelo se usó linear_model o lo que es lo mismo un modelo de regresión lineal.

Todo esto descrito se evalúa y se verifica si cumple con los objetivos de visualización de similitud y distancia entre investigadores, para posteriormente incluirlo como parte de los módulos de la Plataforma Científica Ecuciencia.

2.2 Implementación de los resultados en la etapa de KDD en la Plataforma Científica ECUCIENCIA.

A. Metodología de Desarrollo Ágil: 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 clasificador de textos. Por lo tanto, partiendo de lo mencionado con anterioridad la metodología de desarrollo ágil, como es Scrum es un proceso en el que se aplican de manera regular un conjunto de buenas prácticas para trabajar colaborativamente, en equipo y obtener el mejor resultado posible [50] esta metodología está compuesta por las siguientes etapas:

a) 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** aquel encargado de liderar el equipo de trabajo, siempre estando presente, apoyando y verificando que se cumplan las reglas y procesos de la metodología, ya que es experto en el manejo de la metodología que se está tratando.
- **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. [50]
- **El product Backlog** determinaran los requerimientos que son denominados como las historias de usuario descritos en el lenguaje no técnico por el usuario normal, 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** dentro de 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** estos ya vienen a ser las funcionalidades directas y no directas que va a contener la plataforma científica. Como el proyecto de investigación a trabajar es extenso, se enlistará solo las listas de tareas necesarias para llevar a cabo las historias de cada sprint.

Entre las reuniones que se van dando durante el desarrollo del proyecto se mantienen entre:

- **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, es decir presenta los resultados de cada Sprint que va desarrollando y esta reunión dura 15 minutos como mínimo.

- **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 retroalimentación, donde si llega a presentarse algún tipo de modificaciones en cuanto al desarrollo de cada sprint se debe de mejorar y esta reunión dura entre 2 – 4 horas con todo el equipo de trabajo. [50]

2.3 Validación de los resultados obtenidos mediante métodos estadísticos

En esta etapa se pretende tomar los resultados que el sistema (módulo implementado en la plataforma científica ECUCIENCIA) y describir el comportamiento del conjunto de dos variables, se tomarán como variables de análisis las distancias euclidianas, para ello el uso de diagrama de Pareto, desviación estándar y análisis de varianza, considerando la frecuencia como elemento base de donde se desprenden resultados de como las distancias entre artículos científicos y entre las líneas de investigación.

La varianza se define como el grado de distanciamiento de un conjunto de valores respecto a su valor medio.

El diagrama de Pareto, la desviación estándar y el análisis de varianza, consisten en la aplicación del método estadístico y obtener gráficas de variables para un conjunto de datos.

Para ello se seguirán los siguientes pasos:

Paso 1: Determinar cuál es la situación.
Paso 2: Determinar las variables a estudiar.
Paso 3: Recolectar los datos de las variables.
Paso 4: Procesar los datos recolectados.
Paso 5: Determinar los criterios relevantes encontrados en el procesamiento.
Paso 6: Interpretar los resultados obtenidos y su relación para contrastar hipótesis.

2.4 Conclusiones

- La metodología “KDD”, nos permitió establecer una estructura metodológica adecuada para poder obtener conocimientos relevantes acerca de los datos contenidos en los documentos en PDF del sistema ECUCIENCIA.
- La metodología Scrum nos permitió organizar el trabajo en cuatro etapas como son la planificación, desarrollo, revisión y retroalimentación cada una de ellas posee sus respectivos artefactos de entrada y salida los cuales se constituyen en la documentación del desarrollo del módulo de procesamiento de datos en la Plataforma Científica ECUCIENCIA.
- En base a la validación de los resultados se opta por el método estadístico para validar la investigación, el diagrama de Pareto, la varianza y la desviación estándar permiten como herramientas estadísticas contrastar los resultados que se obtendrán.

CAPÍTULO III

3 Aplicación y/o Validación de la Propuesta

3.1 Resultados de la Metodología KDD

Como se aseveró la plataforma científica ECUCIENCIA hospedada en http://ecuciencia.utc.edu.ec/ presenta una base de datos en PostgreSQL en su Versión 12.3, contiene toda la información relacionada a los investigadores de la Universidad Técnica de Cotopaxi (UTC) incluyendo su producción científica que contempla la publicación de artículos científicos, los metadatos relacionados a la autoría y el formato en PDF del artículo.

La unidad de análisis principal de la presente investigación está dada por los contenidos que se encuentran en estos documentos en PDF que presenta mayor riqueza en cuanto a las terminologías que se encuentran en el cuerpo de este.

Los análisis que el sistema realiza en la actualidad son asociados a los campos de las palabras claves, el título y el resumen de los artículos lo que lleva no a una suficiente representación para establecer niveles de frecuencia, similaridad y distancias entre los artículos en relación a las líneas de investigación y entre los propios artículos.

Una vez conocidas todas las tablas de la base de datos se procedió a seleccionar las tablas necesarias para el procesamiento y aplicación de los algoritmos las mismas que son las de la siguiente figura 3.1:

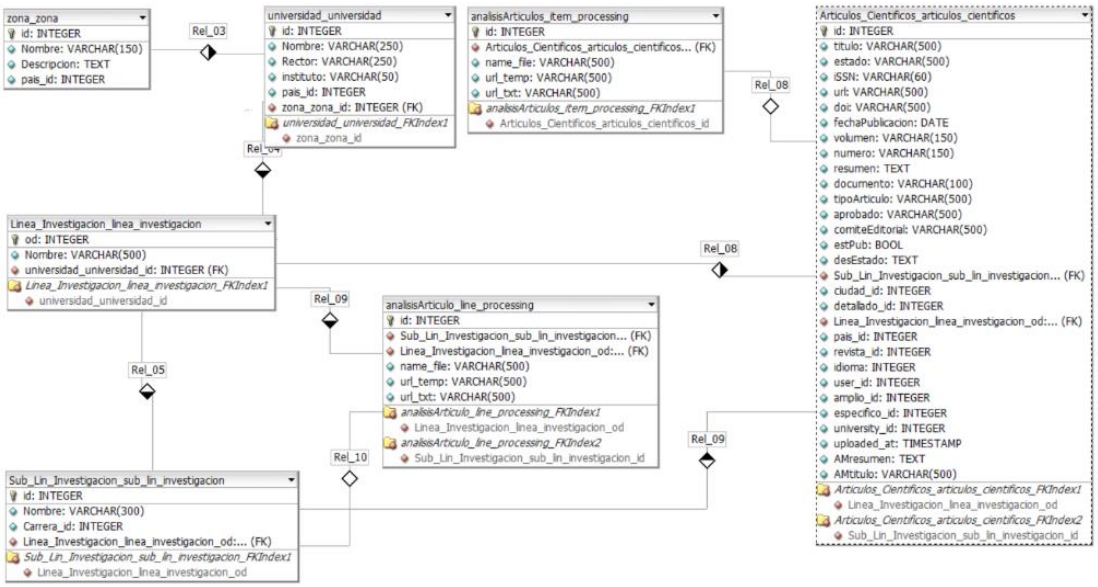


Figura 3.1: Modelo que representan las tablas que tributan a la de los artículos científicos

Elaborado por: El investigador

En la tabla 3.1, se evidencian un resumen de algunas estadísticas presentadas por el sistema de gestión de base de datos al momento de obtener la información.

Tabla 3.1: Algunas estadísticas de la tabla articulos_cientificos en la base de datos, obtenida en el momento de su lectura

CRITERIO		VALOR
Sequential Scans	Escaneos secuenciales	1680
Sequential Tuples Read	Tuplas secuenciales leídas	1068392
Index Scans	Escaneos de índice	396087
Index Tuples Fetched	Tuplas de índice obtenidas	343412
Tuples Inserted	Tuplas insertadas	7
Tuples Updated	Tuplas Actualizado	35
Tuples HOT Updated	Tuplas HOT Actualizado	20
Live Tuples	Tuplas vivas	7
Dead Tuples	Tuplas muertas	11
Heap Blocks Read	Heap Blocks Leer	177

Heap Blocks Hit	El golpe de los bloques de escombrera	634241
Index Blocks Read	Bloques de índice leídos	75
Index Blocks Hit	El golpe de los bloques de índice	835151
Toast Index Blocks Read	Bloques de índice de tostado Leer	84
Table Size	Tamaño de la tabla	1384 kB
Indexes Size	Tamaño de los índices	808 kB

Elaborado por: El investigador

1. **Selección de los datos:**

Para la selección de los datos usando la herramienta pgAdmin se ejecutó la siguiente codificación: SELECT * FROM public."Articulos_Cientificos_articulos_cientificos"; obteniéndose el resultado de la figura 3.2 luego de exportar se obtuvo la tabla 3.2.

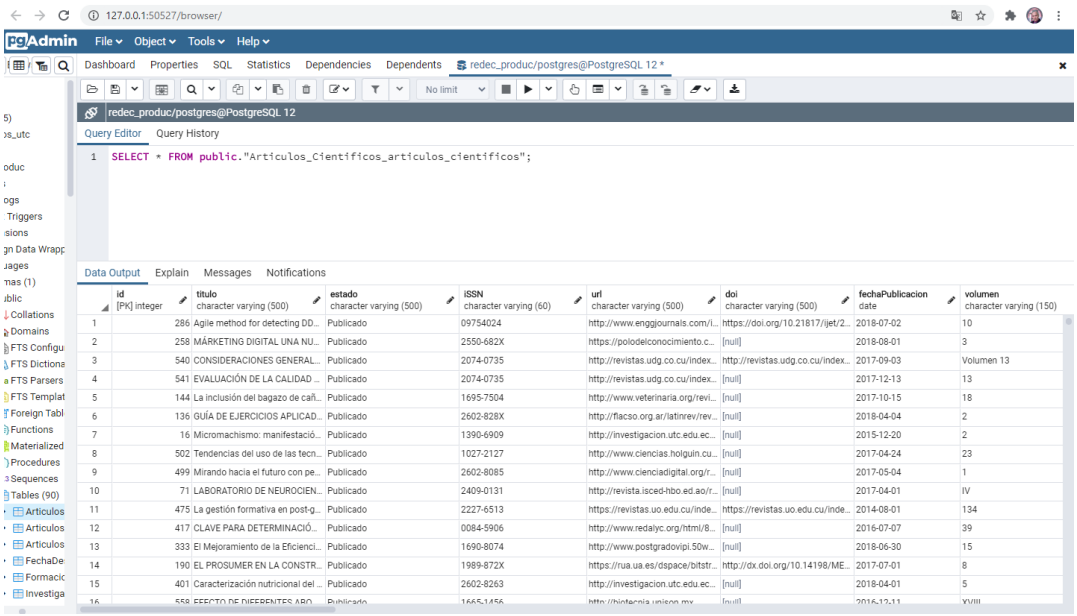


Figura 3.2: Datos obtenidos de la tabla Articulos_Científicos_articulos_cientificos.

Elaborado por: El investigador

Tabla 3.2: Porción de la tabla Articulos_Cientificos_articulos_cientificos y algunos campos importantes como el de documento.

id	titulo	estado	iSSN	url	doi	fechaPublica	volumen	numero	resumen	documento
286	Agile method for detecting DDoS	Publicado	9754024	http://www.	https://doi.c	2/7/2018	10	3	DDoS	articulo/IJET18-10-03-125.pdf
258	MÃRKETING DIGITAL UNA NUEVA	Publicado	2550-682X	https://polo	NULL	1/8/2018	3	8	El marketing	articulo/MARKETING_DIGITAL_ZDdqquR.pdf
540	CONSIDERACIONES GENERALES S	Publicado	2074-0735	http://revist	http://revist	3/9/2017	Volumen 13	3	La	articulo/CONSIDERACIONES_GENERALES_SOBRE_EL_PROCESO.pdf
541	EVALUACIÃ"N DE LA CALIDAD NU	Publicado	2074-0735	http://revist	NULL	13/12/2017	13	4	La presente	articulo/EVALUACION_DE_LA_CALIDAD_NUTRITIVA_DE_UN_ENSILADO.pdf
144	La inclusiÃ³n del bagazo de caÃ±a	Publicado	1695-7504	http://www.	NULL	15/10/2017	18	10	El aumento	articulo/La_inclusiÃ³n_del_bagazo_de_caÃ±a.PDF
136	GUÃA DE EJERCICIOS APLICADO A	Publicado	2602-828X	http://flacso	NULL	4/4/2018	2	2	El presente e	articulo/articulo-presentado.pdf
16	Micromachismo: manifestaciÃ³n	Publicado	1390-6909	http://inves	NULL	20/12/2015	2	3	Este artÃ-cul	articulo/Micromachismo_Magaly_Gina.pdf
502	Tendencias del uso de las tecnol	Publicado	1027-2127	http://www.	NULL	24/4/2017	23	2	La presente	articulo/Tendencias_del_uso_de_las_tecnologÃ-as_30-04-2017.pdf
499	Mirando hacia el futuro con pens	Publicado	2602-8085	http://www.	NULL	4/5/2017	1	1	Los cambios	articulo/PublicaciÃ³n_1.pdf
71	LABORATORIO DE NEUROCIENCIA	Publicado	2409-0131	http://revist	NULL	1/4/2017	IV	1	El presente a	articulo/4.1._articulo_neuromarketing_rJJbk9n.pdf
475	La gestiÃ³n formativa en post-gra	Publicado	2227-6513	https://revis	https://revis	1/8/2014	134	134	Resumen	articulo/La_gestiÃ³n_formativa_en.pdf
417	CLAVE PARA DETERMINACIÃ"N D	Publicado	0084-5906	http://www.	NULL	7/7/2016	39	1	Se provee un	articulo/clave_musgos__ABV.pdf
333	El Mejoramiento de la Eficiencia	Publicado	1690-8074	http://www.	NULL	30/6/2018	15	15	El estudio	articulo/VENEZUELA_PDF.pdf
190	EL PROSUMER EN LA CONSTRUCC	Publicado	1989-872X	https://rua.u	http://dx.do	1/7/2017	8	2	Las plataform	articulo/ReMedCom_08_02_18.pdf
401	CaracterizaciÃ³n nutricional del p	Publicado	2602-8263	http://inves	NULL	1/4/2018	5	1	El objetivo d	articulo/CARACTERIZACIÃ"N_DEL_PALMISTE.pdf
558	EFECTO DE DIFERENTES ABONOS	Publicado	1665-1456	http://biote	NULL	11/12/2016	XVIII	3	Entre los	articulo/333-734-1-SM_dxKJBvM.pdf
569	Artículo	Publicado	2550-682X	https://stack	https://stack	20/11/2018	3	11	CientÃ-fico	articulo/798-2253-1-PB.pdf
610	La calidad de la educaciÃ³n en las	Publicado	2256-1536	https://revis	https://revis	30/12/2016	5	12	Los problem	articulo/documento_publicado_sonia_gonzalo_gp86gKU.pdf
550	LUPIN PEST MANAGEMENT IN TH	Publicado	1743-1034	https://www	https://doi.c	1/12/2017	NULL	NULL	The Andean	articulo/S5.pdf
543	BIOMETRIC SIGNS OF AN AMARA	Publicado	0321-0499	https://scho	NULL	0015-01-01	3	113	Biometric sig	articulo/Articulo_emerson_vdnaba_2015_3_26.pdf
599	NIVEL DE SATISFACCIÃ"N DE EGR	Publicado	2224-2643	https://dialn	NULL	1/12/2018	IX	4	En el	articulo/Dialnet-NivelDeSatisfaccionDeEgresadosDeLaCarreraDeIngenie-6716273.p
74	MARKETING DIGITAL; UNA VISIÃ"	Publicado	2409-0131	http://www.	http://www.	7/2/2018	V	1	En la actualid	articulo/7.1._articulo_marketing_digital_5xVBioy.pdf
648	EVALUACIÃ"N DE FACTORES DE R	Publicado	2477-9253	http://geo1.	NULL	8/8/2019	4	8	Actualment	articulo/EVALUACIÃ"N_DE_FACTORES_DE_RIESGOS_PSICOSOCIALES.pdf
605	Una mirada del proceso de regul	Publicado	2631-2603	https://journ	NULL	1/12/2017	4	1	La	articulo/Panchi_Una_mirada_del_proceso_de_regulaciÃ³n_contable2017.pdf
271	ANÃLISIS DE EQUIDAD DE GENER	Publicado	2224-2643	http://runac	NULL	6/7/2017	8	3	En el presen	articulo/472017_ANALISIS_DE_LA_EQUIDAD_DE_GENERO.pdf
227	ESTIMACIÃ"N DE DATOS FALTANT	Publicado	2588-0764	https://revis	https://doi.c	27/12/2017	2	3	Se evaluaron	articulo/ART_1.pdf
734	Factors that Influence Undergrad	Publicado	2319-8613	http://www.	http://www.	6/12/2018	10	6	University de	articulo/IJET18_GFifQ49.pdf
521	LA LINGUISTICA APLICADA A LA E	Publicado	266-1536	http://revist	NULL	11/11/2016	6	3	El lenguaje	articulo/la_linguistica_aplicada.pdf

Elaborado por: El investigador

En la tabla 3.2 se puede observar la tabla objeto de estudio, donde contiene los artículos científicos y los diferentes campos que ofrecen información del título, las palabras claves, resumen y el documento en PDF que es la cual se usará para realizar análisis del contenido que se encuentra en su cuerpo.

2. Preprocesamiento/limpieza:

En la tabla 3.2, se evidencia información ruidosa con campos donde se evidencian caracteres especiales, en el proceso de limpieza de los datos se procedió a eliminar toda la información que no aportaba a la investigación, para ello se usó como herramienta el blog de notas para buscar y reemplazar muchos de estos caracteres como se muestra en la siguiente figura 3.3, en el anexo 1 se encuentra la tabla de datos luego de realizar limpieza, eliminar valores vacíos y optimizar los campos necesarios (por la magnitud de la tabla solo se pondrán filas iniciales y finales con el número, ID y documento.

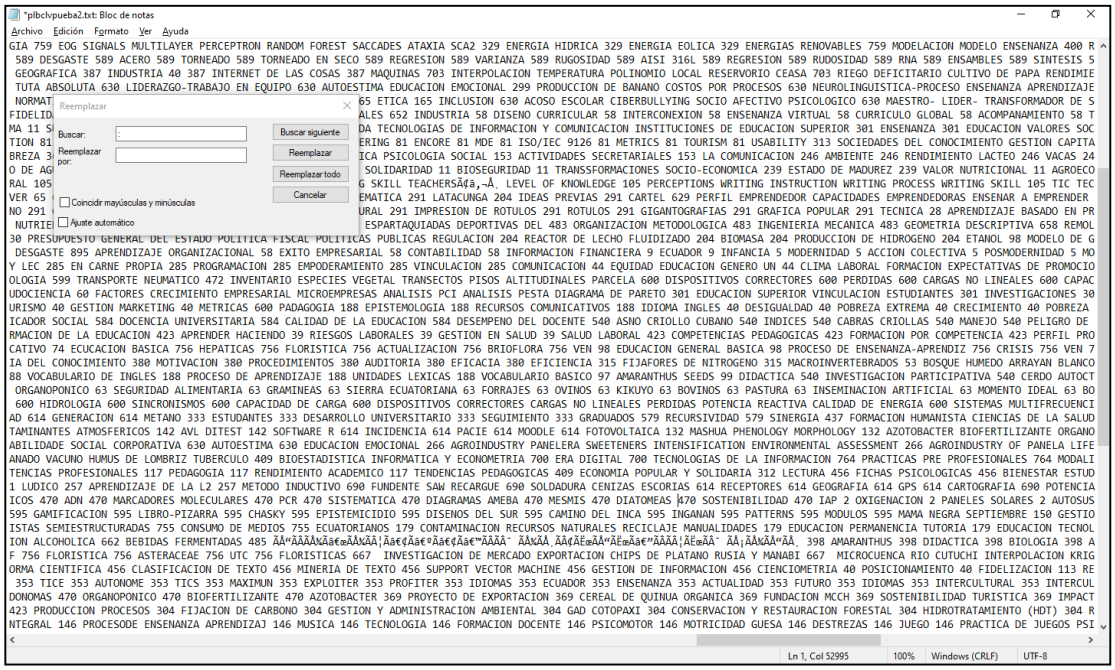


Figura 3.3: Uso del blog de notas para limpiar los datos de análisis

Elaborado por: El investigador

Fueron corregidas palabras que contenían los siguientes caracteres:

- Ã“, Ã, â€™, Ã±, â€œ, Ã□, Ã‰, Ã, Â´S, ¬â€œ, Â, Ã, ¬â€, ¬Å¸ Ã|

3. Transformación/reducción:

Para experimentar con los algoritmos y librerías de Python se toma como alternativa solo realizar la experimentación con una porción de la población de artículos que se encuentran en la base de datos.

Muestreo:

Según el sistema se encuentra una población de 636 artículos científicos, para ello se hará una selección de una muestra aleatoria simple de 5 artículos para un 8% de representación, se considerarán los ID de los artículos para la generación de números aleatorios usando Python y el módulo random de la biblioteca estándar. Este módulo ofrece una serie de funciones que generan números aleatorios de manera diferente, con la siguiente instrucción en Python:

Importamos en el Shell de Python la librería random:

- *import random*

Se obtiene de la base de datos en la tabla de artículos científicos el mayor valor del ID de los artículos registrado con el propósito de generar desde el 1 hasta ese número y a través con la instrucción:

- *SELECT max(int2(ID)) FROM public."Articulos_Cientificos_articulos_cientificos"*

Se pudo conocer que el mayor valor del ID en la tabla es: 806.

Con un ciclo for en un rango de 5 números desde el 1 hasta el 806 se imprimieron los números aleatorios:

- Con el *for i in range(5): print(random.randrange(1, 806))* se generaron los siguientes números aleatorios:
  - ✓ 796.
  - ✓ 4.
  - ✓ 358.
  - ✓ 514.
  - ✓ 501.

Se procedió a extraer de la base de datos los PDF asociados a los ID con los números 796; 4; 358; 514; 501, para conocer que artículos están asociados a ID se usó la siguiente sentencia de SQL:

- *SELECT ID FROM public."Articulos_Cientificos_articulos_cientificos" WHERE ID=796; (este para el caso del ID 796).*

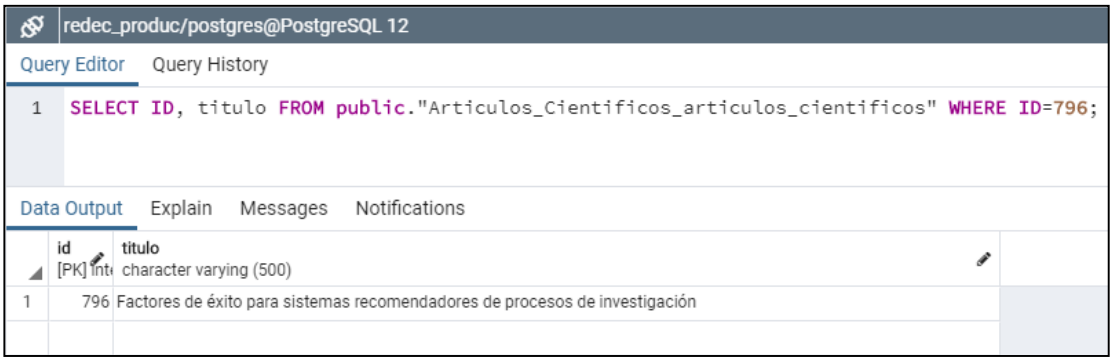


Figura 3.4: Sentencia SQL para determinar el artículo asociado al ID, caso del 796

Elaborado por: El investigador

Quedando como muestras seleccionadas los siguientes artículos:

- 796: Factores de éxito para sistemas recomendadores de procesos de investigación.
- 4: Finca agroecológica sostenible de la Universidad de Granma.
- 358: Guía virtual interactiva en Android a través de códigos QR en el Museo de la Escuela Fiscal Isidro Ayora del Ecuador.
- 514: Optimización con Colonia de Hormigas para la Planificación Óptima de la Fuerza de Trabajo.
- 501: Utopía o realidad de aplicaciones informáticas en la educación. Caso universidad ecuatoriana.

4. Minería de datos (data mining):

Para esta etapa se usaron las siguientes librerías en Python:

- import os (Acceso portable a funciones específicas del sistema operativo)
- import fitz (PyMuPDF también conocido como "fitz": enlaces Python para MuPDF, que es un visor de PDF y XPS ligero, como se trabaja con PDF es importante el uso de los algoritmos de esta librería)
- from os import remove (En la librería OS existe el algoritmo relacionado con remover archivos)
- import re (El uso más común para re es buscar patrones en texto).

- import string (El módulo string contiene funciones útiles que manipulan cadenas).
- from nltk.corpus import stopwords (Librería para procesamiento del lenguaje natural, análisis de corpus y los stopwords o palabras de parada)
- import nltk (Librería que contiene varios algoritmos para el procesamiento del lenguaje natural).
- import pandas as pd (Pandas es una librería de python destinada al análisis de datos, que proporciona unas estructuras de datos flexibles y que permiten trabajar con ellos de forma muy eficiente).
- from sklearn.feature_extraction.text import TfidfVectorizer (La biblioteca scikit-learn ofrece herramientas fáciles de usar para realizar tanto la tokenización como la extracción de características de sus datos de texto; TfidfVectorizer es para calcular las frecuencias de las palabras, y el método más popular es el llamado TF-IDF. Este es un acrónimo que significa Frecuencia de Término – Frecuencia Inversa de Documento que son los componentes de las puntuaciones resultantes asignadas a cada palabra).
- import scipy.spatial.distance as distance (Scipy es el paquete científico más completo, que incluye interfases a librerías científicas y distance permite determinar distancias entre vectores).
- from jaccard_index.jaccard import jaccard_index (El índice de Jaccard o coeficiente de Jaccard mide el grado de similitud entre dos conjuntos, sea cual sea el tipo de elementos).
- from sklearn.metrics.pairwise import cosine_similarity (Cosine Similarity mide el ángulo de los vectores, cuanto menores, serán más similares).
- import pandas as pd (Pandas es una librería de python destinada al análisis de datos, que proporciona unas estructuras de datos flexibles y que permiten trabajar con ellos de forma muy eficiente).
- import matplotlib.pyplot as plt (Matplotlib es una librería para generar gráficas a partir de datos contenidos en listas, vectores).
- from wordcloud import WordCloud (Algoritmo que permite generar nubes de palabras).

```
import os
import fitz
from os import remove
import re
import string
from nltk.corpus import stopwords
import nltk
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
import scipy.spatial.distance as distance
from jaccard_index.jaccard import jaccard_index
from sklearn.metrics.pairwise import cosine_similarity
import pandas as pd
import matplotlib.pyplot as plt
from wordcloud import WordCloud
```

Figura 3.5: Librerías empleadas para el análisis del corpus de los artículos de muestra

Elaborado por: El investigador

Procedimiento seguido con Python:

a) Creando variables para obtener el camino a los artículos de muestras:

```
PATH_ROOT = 'D:/01-TRABAJO-UTC/Investigacion/01-Proyecto RECI/Tesis de maestria/Beto/procesamiento de datos2ok/procesamiento de
doc1 = 'Factores_de_exito_para_sistemas_recomendadores_de_procesos_de_investigacion'
doc2 = 'Finca_agroecologica_sostenible_de_la_Universidad_de_Granma'
doc3 = 'Guia_virtual_interactiva_en_Android_a_través_de_codigos_QR_en_el_Museo_de_la_Escuela_Fiscal_Isidro_Ayora_del_Ecuador'
doc4 = 'Optimizacion_con_Colonia_de_Hormigas_para_la_Planificacion_optima_de_la_Fuerza_de_Trabajo'
doc5 = 'Utopia_o_realidad_de_aplicaciones_informaticas_en_la_educacion_Caso_universidad_ecuatoriana'

extPdf = ".pdf"
extTxt = ".txt"
```

Figura 3.6: Ruta de los artículos y variables con cada artículo en formato PDF

Elaborado por: El investigador

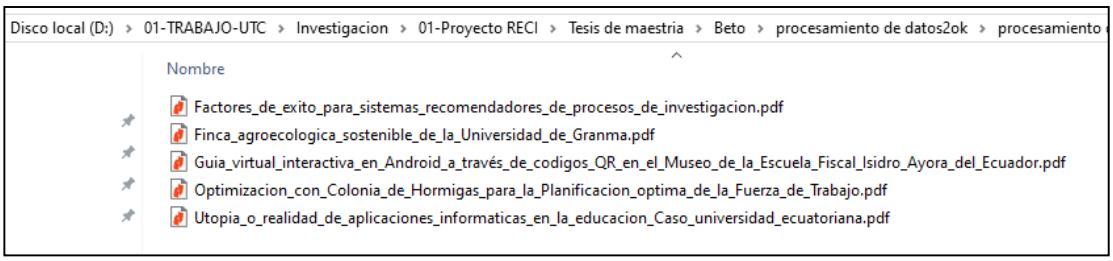


Figura3.7: Directorio con los artículos de muestra en formato PDF.

Elaborado por: El investigador

En las figuras 3.8, 3.9, 3.10, 3.11, 3.12, 3.13, se muestran los códigos que describen los algoritmos usados para determinar frecuencia, riqueza léxica, tokenizado, número de stopwords, estandarizar a minúsculas, similaridad y distancia.

```
def freq(text, num, title):
    aux = text.lower()
    tokens = [t for t in aux.split()]
    words = [word for word in tokens if word.isalpha()]
    labels = []
    value = []
    freq = nltk.FreqDist(words)
    for key, val in freq.most_common(num):
        if len(key) >= 4:
            value.append(str(val))
            labels.append(str(key))

    freq.plot(num, cumulative=False, title=title)
    print('***FRECUENCIA DE PALABRAS***')
    print('palabras: ', labels)
    print('frecuencia: ', value)
```

Figura 3.8: Creación de función para determinar la frecuencia de los documentos.

Elaborado por: El investigador

```
def lexical_wealth(tokens):
    tokens_conjunto = set(tokens)
    palabras_totales = len(tokens)
    palabras_diferentes = len(tokens_conjunto)
    riqueza_lexica = palabras_diferentes / palabras_totales
    return round(riqueza_lexica, 2)
```

Figura 3.9: Función para calcular la riqueza léxica, usando tokenizado

Elaborado por: El investigador

```
def numStopWords(a, b):
    sw = 0
    if a > b:
        sw = a - b
    else:
        sw = b - a

    return sw
```

Figura 3.10: Función para restar siempre del mayor número de stopwords (palabras de parada) el menor

Elaborado por: El investigador

```
def data(lineTemp, lineText):
    # file temp
    t = lineTemp.lower()
    tkns = [t for t in t.split()]
    # file text
    temp = lineText.lower()  # convertir todo el texto a minusculas
    tokens = [t for t in temp.split()]

    print('***Número de palabras***')
    print('Con palabras de parada:', len(tokens))
    print('Sin palabras de parada:', len(tkns))
    print('************************')
    print('Numero de palabras de parada:', numStopWords(len(tokens), len(tkns)))
    print('************************')
    print('Riqueza léxica: ', lexical_wealth(tkns))
```

Figura 3.11: Función para convertir todo el texto de los artículos en minúscula y poder procesarlo, también se muestra número de palabra, longitud y riqueza léxica.

Elaborado por: El investigador

```
def similarity_distance(text1, text2):
    x = [text1]
    y = [text2]

    df1 = pd.DataFrame(data={'x': x, 'y': y})
    df1['xy'] = df1.apply(lambda x: x['x'] + ' ' + x['y'], axis=1)

    clf1 = TfidfVectorizer(ngram_range=(1, 1))
    clf1.fit(df1['xy'])

    tfidf_x = clf1.transform(df1['x']).todense()
    tfidf_y = clf1.transform(df1['y']).todense()

    a = []
    b = []
    for z in range(len(tfidf_x)):
        a = tfidf_x[z]
        b = tfidf_y[z]

    cosine_simil = cosine_similarity(a, b)[0][0]
    chebyshev_dst = distance.chebyshev(a, b)
    correlation_dst = distance.correlation(a, b)
    cosine_dst = distance.cosine(a, b)
    dice_dst = distance.dice(a, b)
    euclidean_dst = distance.euclidean(a, b)
    jaccard_dst = jaccard_index(text1, text2)
    minkowski_dst = distance.minkowski(a, b, p=4)

    print('similarity: ', "{0:.3f}".format(cosine_simil))
    print('chebyshev: ', "{0:.3f}".format(chebyshev_dst))
    print('correlation: ', "{0:.3f}".format(correlation_dst))
    print('cosine: ', "{0:.3f}".format(cosine_dst))
    print('dice: ', "{0:.3f}".format(dice_dst))
    print('euclidean: ', "{0:.3f}".format(euclidean_dst))
    print('jaccard: ', "{0:.3f}".format(jaccard_dst))
    print('minkowski: ', "{0:.3f}".format(minkowski_dst))
```

Figura 3.12: Función para determinar similaridad y distancia

Elaborado por: El investigador

```
def generateWordCloud(data, title):
    stop_words_sp = set(stopwords.words('spanish'))
    wordcloud = WordCloud(background_color='white',
                          stopwords=stop_words_sp,
                          max_words=20,
                          max_font_size=200,
                          scale=3,
                          random_state=3).generate(str(data))

    wordcloud.recolor(random_state=1)
    plt.figure(figsize=(20, 15))
    plt.title(title, fontsize=20, color='blue')
    plt.imshow(wordcloud)
    plt.axis('off')
    plt.show()
```

Figura 3.13: Función para generar nubes de palabras

Elaborado por: El investigador

5. Interpretación/evaluación:

En las figuras 3.14, 3.15, 3.16, 3.17, 3.18, se muestran las frecuencias de las palabras de parada de los artículos 1, 2, 3, 4 y 5, como se puede evidenciar las palabras comunes más usadas en este análisis son “de, la, el, y en”; son palabras de paradas que mayor frecuencia tienen. Desde el punto de vista semántico, la mayor parte de las preposiciones no tienen significado léxico, pues constituyen marcas de función que contribuyen a señalar la relación sintáctica entre un núcleo y su complemento. En definitiva, en las preposiciones que se evidencia como resultado de los artículos, predomina el significado relacional, o sea ponen en contado dos elementos, el elemento del que depende el sintagma o grupo y el término de la preposición, también tienen un carácter relacional, porque permiten unir palabras y establecen relaciones entre ellas.

Esas preposiciones principalmente están colocadas delante de sustantivos, adjetivos, pronombres, adverbios y verbos en infinitivo.

Algunos ejemplos:

- Para el desarrollo de la investigación…
- Algoritmos de inteligencia artificial…

- …perteneciente a la Universidad de Granma…
- El museo de la Escuela Isidro Ayora…
- …la educación y la tecnología…

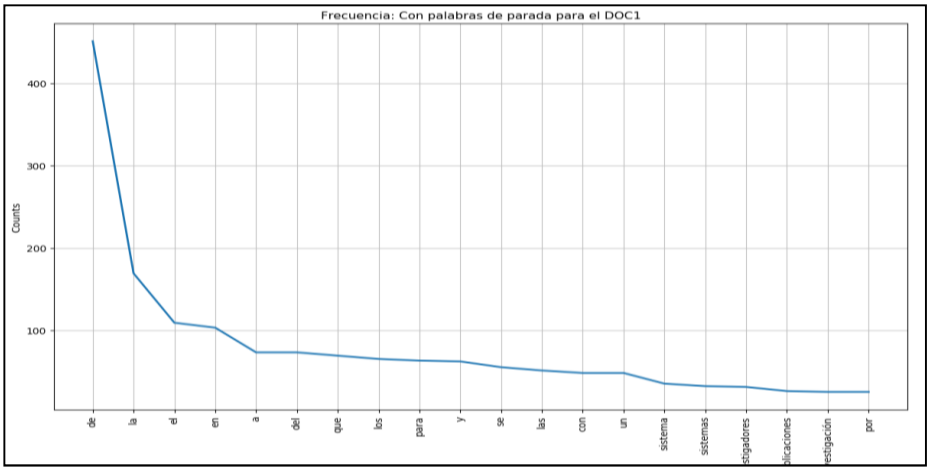


Figura 3.14 Frecuencia de las palabras de parada del artículo número 1
Elaborado por: El investigador

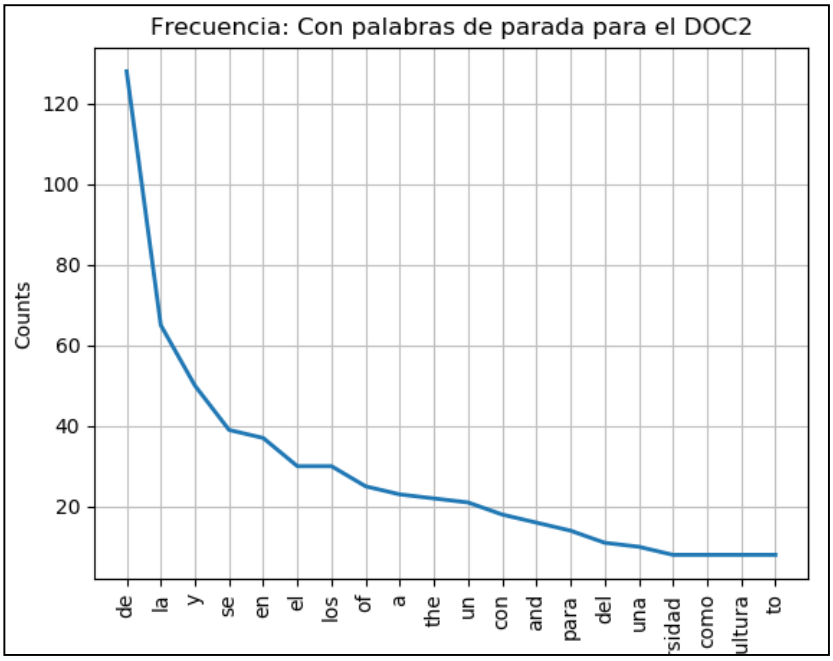


Figura 3.15: Frecuencia de las palabras de parada del artículo 2

Elaborado por: El investigador

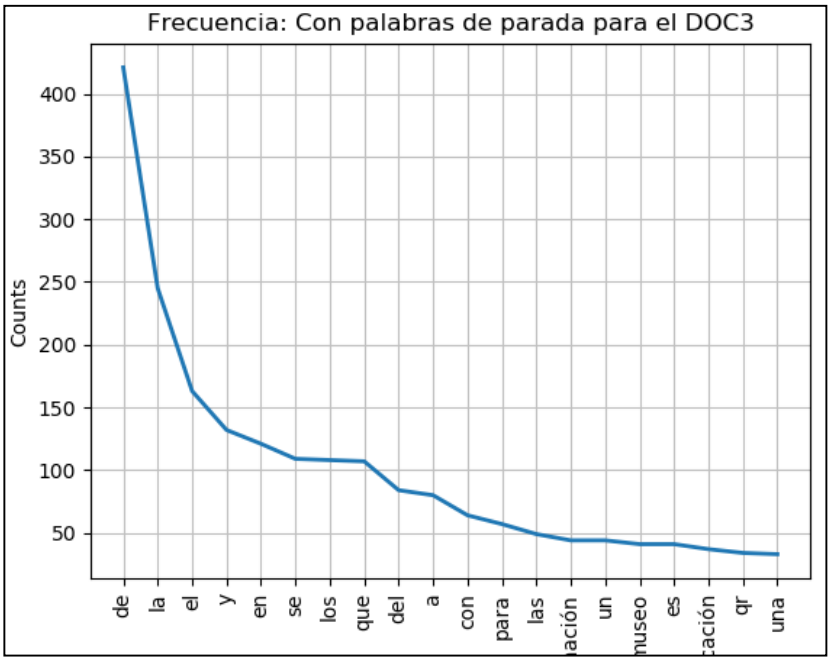


Figura 3.16: Frecuencia de las palabras de parada del artículo 3

Elaborado por: El investigador

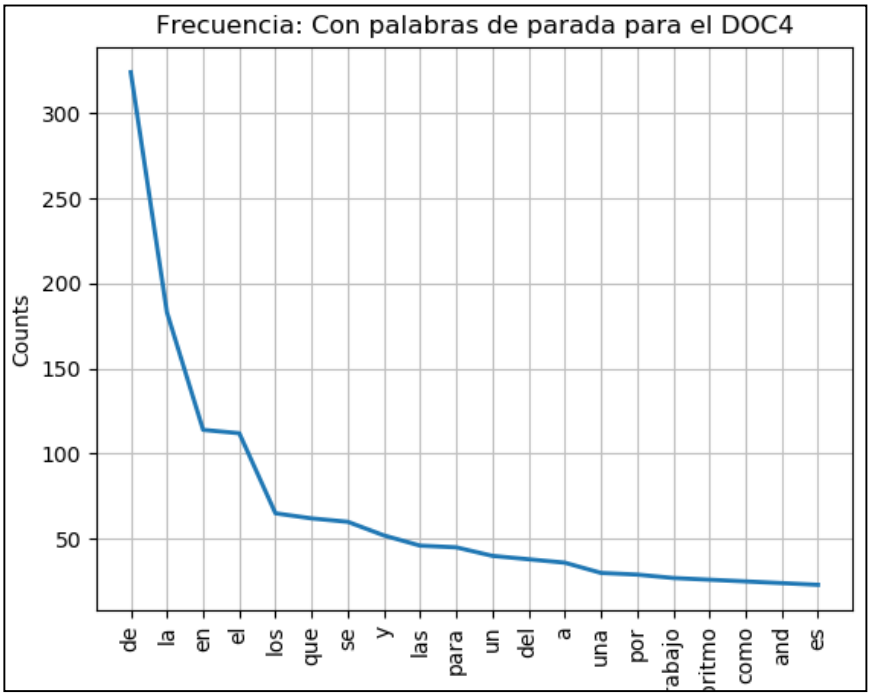


Figura 3.17. Frecuencia de las palabras de parada del artículo 4

Elaborado por: El investigador

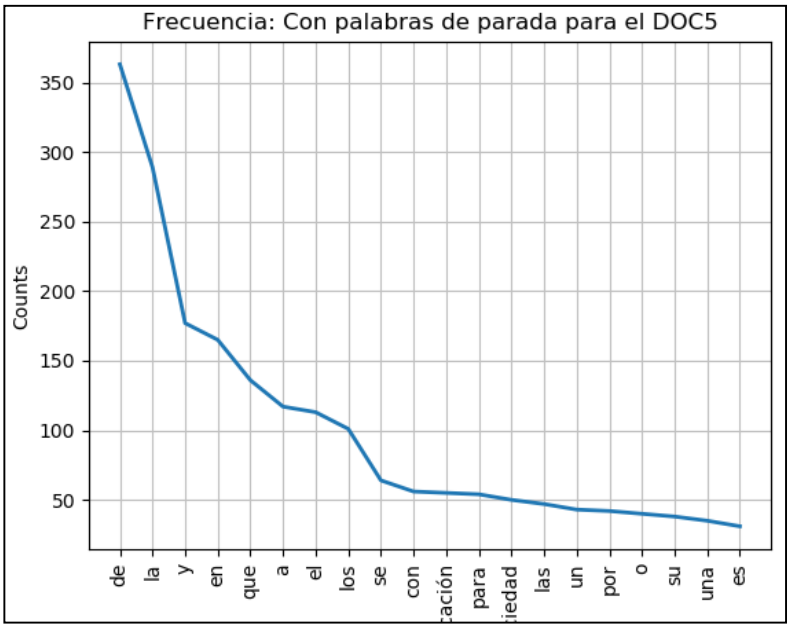


Figura 3.18: Frecuencia de las palabras de parada del artículo 5

Elaborado por: El investigador

En las figuras 3.19, 3.20, 3.21, 3.22, 3.23, se muestran graficadas las frecuencias sin palabras de parada.

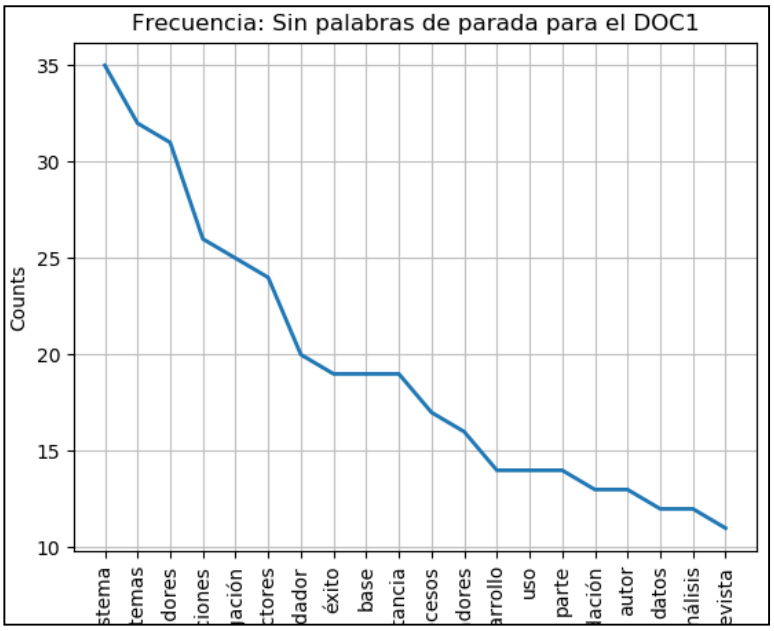


Figura 3.19: Frecuencia de las palabras sin stopwords del artículo 1

Elaborado por: El investigador

Palabras y frecuencia de palabras para el artículo 1 (DOC1):

- palabras: ['sistema', 'sistemas', 'investigadores', 'publicaciones', 'investigación', 'factores', 'recomendador', 'éxito', 'base', 'importancia', 'procesos', 'recomendadores', 'desarrollo', 'parte', 'recomendación', 'autor', 'datos', 'análisis', 'revista']
- frecuencia: ['35', '32', '31', '26', '25', '24', '20', '19', '19', '19', '17', '16', '14', '14', '13', '13', '12', '12', '11']
- Interpretación: En el artículo se evidencia que su contenido según las palabras más repetidas trata sobre sistemas recomendadores, investigaciones, autores, publicaciones en revistas, asociado a la línea de investigación Tecnologías de información y comunicación (TIC).

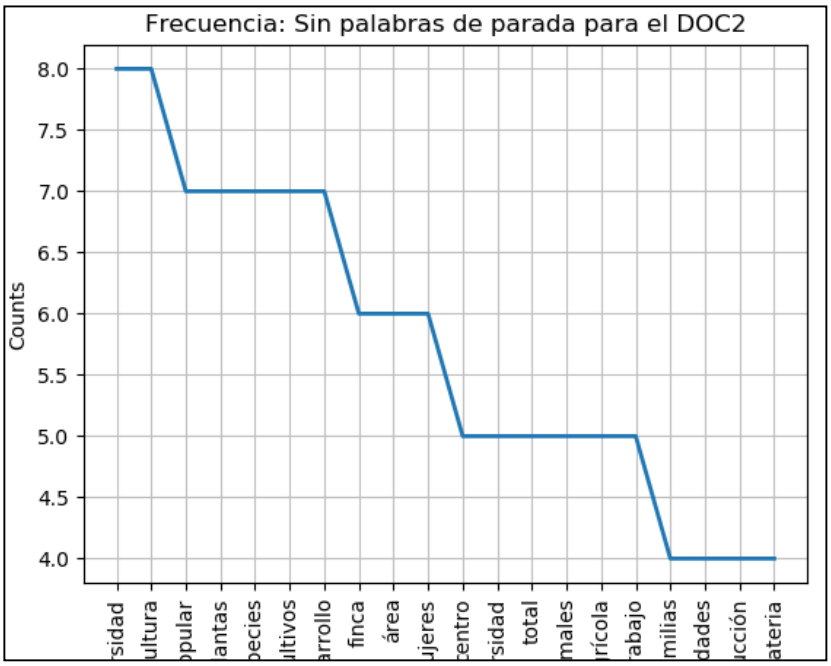


Figura 3.20: Frecuencia de las palabras sin stopwords del artículo 2

Elaborado por: El investigador

Palabras y frecuencia de palabras para el artículo 2 (DOC2):

- palabras: ['biodiversidad', 'agricultura', 'popular', 'plantas', 'especies', 'cultivos', 'desarrollo', 'finca', 'área', 'mujeres', 'centro', 'universidad', 'total', 'animales', 'agrícola', 'trabajo', 'familias', 'variedades', 'introducción', 'materia']

- Frecuencia: ['8', '8', '7', '7', '7', '7', '7', '6', '6', '6', '5', '5', '5', '5', '5', '5', '4', '4', '4', '4']
- Interpretación: Se muestra un gráfico escalonado descendente donde las palabras biodiversidad y agricultura son repetida 8 veces, dando a entender que el campo de estudio del artículo esté vinculado con la línea de investigación Análisis, conservación y aprovechamiento de la biodiversidad local.

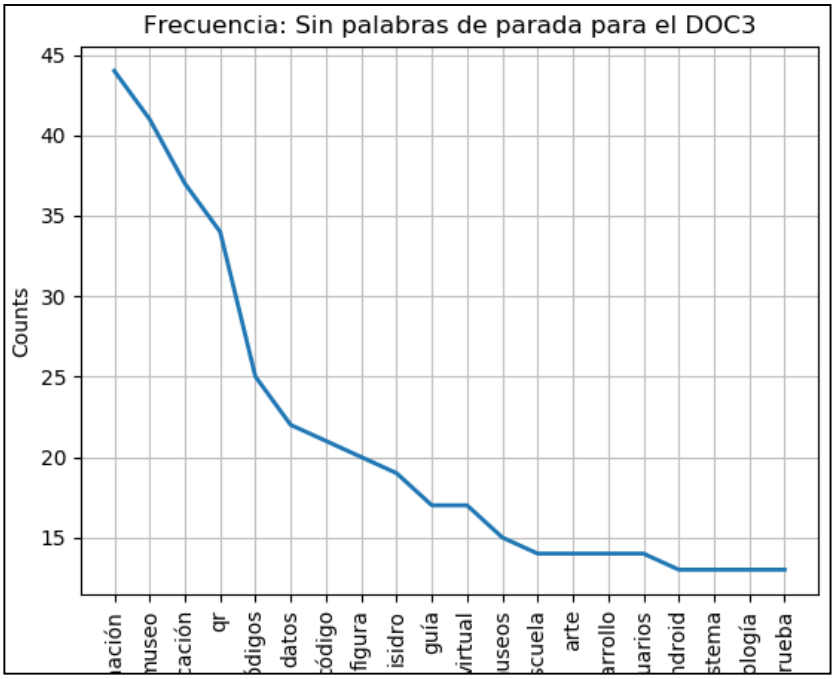


Figura 3.21: Frecuencia de las palabras sin stopwords del artículo 3

Elaborado por: El investigador

Palabras y frecuencia de palabras para el artículo 3 (DOC3):

- palabras: ['información', 'museo', 'aplicación', 'códigos', 'datos', 'código', 'figura', 'isidro', 'guía', 'virtual', 'museos', 'escuela', 'arte', 'desarrollo', 'usuarios', 'android', 'sistema', 'metodología', 'prueba']
- frecuencia: ['44', '41', '37', '25', '22', '21', '20', '19', '17', '17', '15', '14', '14', '14', '14', '13', '13', '13', '13']
- Interpretación: Las palabras información y museo son las que más frecuencia de repetición tienen por lo que evidencia que el artículo tiene que ver con la información de sobre/o en museos, según estas frecuencias de las

palabras ubica al artículo en la línea de investigación TIC y guarda relación con el desarrollo de aplicación móvil, la generación de códigos QR para brindar información de museos.

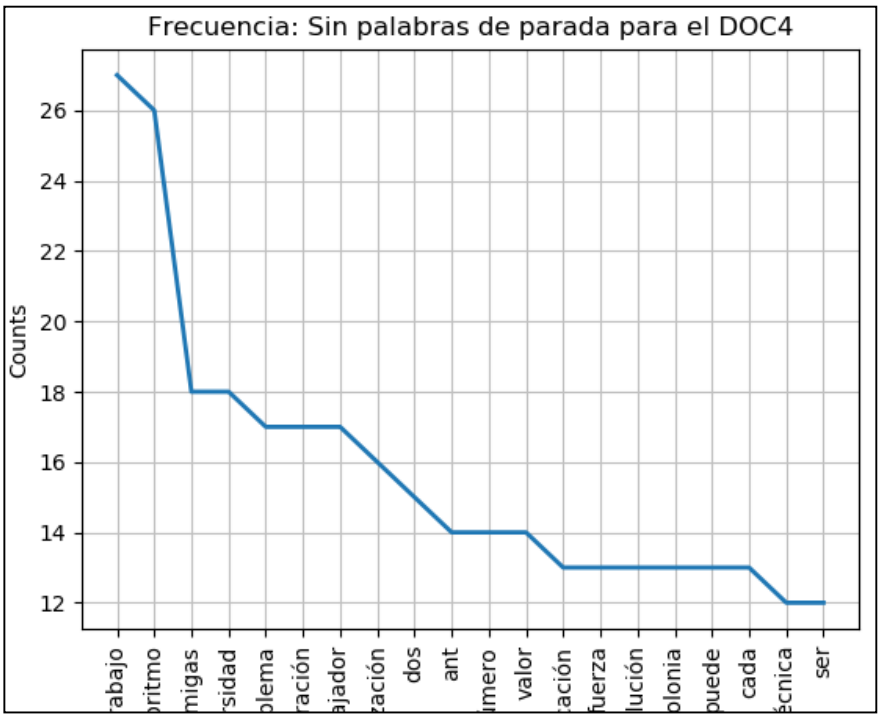


Figura 3.22: Frecuencia de las palabras sin stopwords del artículo 4

Elaborado por: El investigador

Palabras y frecuencia de palabras para el artículo 4 (DOC4):

- palabras: ['trabajo', 'algoritmo', 'hormigas', 'universidad', 'problema', 'exploración', 'trabajador', 'optimización', 'número', 'valor', 'planificación', 'fuerza', 'solución', 'colonia', 'puede', 'cada', 'técnica']
- frecuencia: ['27', '26', '18', '18', '17', '17', '17', '16', '14', '14', '13', '13', '13', '13', '13', '13', '12']
- Interpretación: Las palabras 'trabajo' y 'algoritmo' son las que más frecuencia de repetición tienen por lo que evidencia que el artículo tiene que ver con el trabajo de algoritmos que modelan colonias de hormigas para modelar planificación de fuerzas de trabajo, donde también se ha aplicado la optimización, según estas frecuencias de las palabras ubica al artículo en

las líneas de investigación TIC, Procesos Industriales o Gestión de la Calidad y Seguridad Laboral.

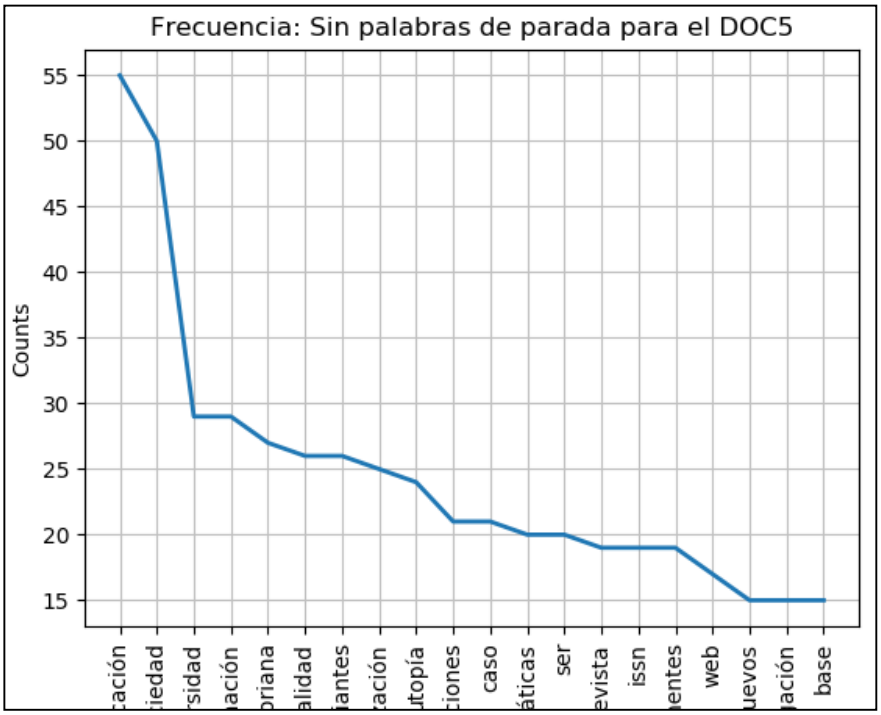


Figura 3.23: Frecuencia de las palabras sin stopwords del artículo 5

Elaborado por: El investigador

Palabras y frecuencia de palabras para el artículo 5 (DOC5):

- palabras: ['educación', 'sociedad', 'universidad', 'información', 'ecuatoriana', 'realidad', 'estudiantes', 'utilización', 'utopía', 'aplicaciones', 'caso', 'informáticas', 'revista', 'issn', 'componentes', 'nuevos', 'investigación', 'base']
- frecuencia: ['55', '50', '29', '29', '27', '26', '26', '25', '24', '21', '21', '20', '19', '19', '19', '15', '15', '15']
- Interpretación: Las palabras educación y sociedad, seguida de universidad, información y ecuatoriana son las que más frecuencia de repetición tienen por lo que evidencia que el artículo guarda relación con el campo de la informática en la educación y la sociedad ecuatoriana, según estas frecuencias de las palabras ubica al artículo en las líneas de investigación Educación, Comunicación y Diseño Gráfico para el Desarrollo Humano y Social.

En las figuras 3.24, 3.25, 3.26, se muestran las nubes de palabras según las frecuencias de las palabras determinadas, las palabras que más resaltan son las que mayor valor presentan.

Figura 3.24: Nubes de palabras tomando la frecuencia del artículo 1 y 2

Elaborado por: El investigador

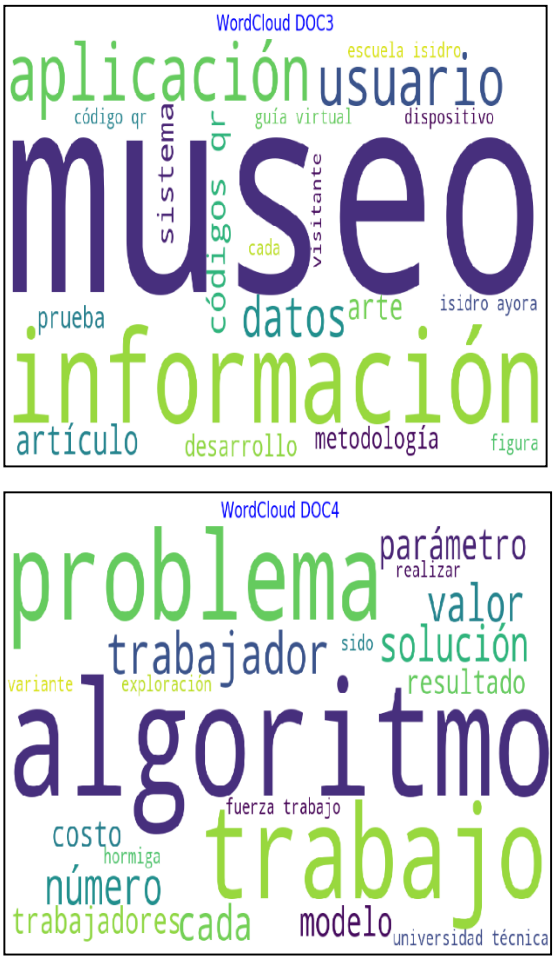


Figura 3.25: Nubes de palabras tomando la frecuencia del artículo 3 y 4

Elaborado por: El investigador

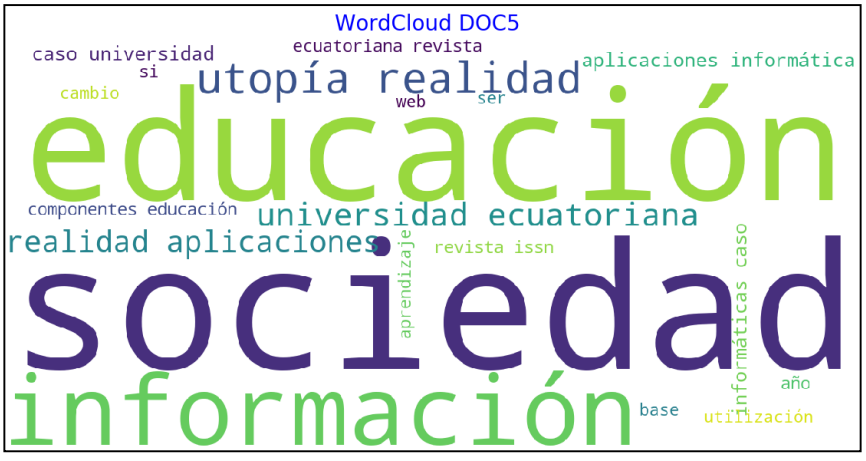


Figura 3.26: Nubes de palabras tomando la frecuencia del artículo 5

Elaborado por: El investigador

En el anexo 2 se puede encontrar el resto del análisis de los datos de los artículos de muestra seleccionados a partir del muestreo aleatorio simple, uno de los resultados se refiere a la riqueza léxica como se evidencia en la figura 3.27, esta riqueza léxica toma como referente los tokens (se refiere a la división de oraciones y palabras del cuerpo del texto en tokens de oraciones o tokens de palabras respectivamente) la longitud de los tokens, en otras palabras la *diversidad léxica* y la *densidad léxica: **la diversidad** léxica* se refiere al número de palabras diferentes utilizadas en un texto, un rango mayor indica una diversidad mayor, mientras que la ***densidad léxica*** en tanto, se entiende como la relación entre el total de palabras léxicas o de contenido semántico (verbos, nombres, adjetivos y algunos adverbios) comparado con las llamadas palabras gramaticales o funcionales (artículos, preposiciones, conjunciones, entre otros).

Como se puede observar en el gráfico, el artículo que presenta mayor riqueza léxica es el artículo 2 "*Finca agroecológica sostenible de la Universidad de Granma*", seguido del artículo 3 "*Guía virtual interactiva en Android a través de códigos QR en el Museo de la Escuela Fiscal Isidro Ayora del Ecuador*", luego el artículo 4 "*Optimización con Colonia de Hormigas para la Planificación Óptima de la Fuerza de Trabajo*", el artículo 1 "*Factores de éxito para sistemas recomendadores de procesos de investigación*" con un valor de 0,42 de riqueza léxica y por último el artículo 5 "*Utopía o realidad de aplicaciones informáticas en la educación. Caso universidad ecuatoriana*".

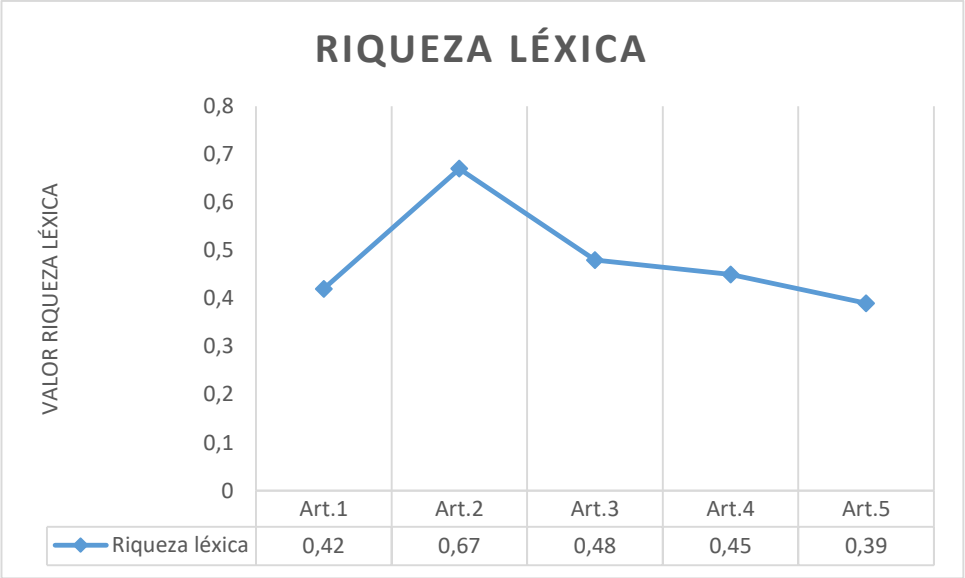


Figura 3.27: Valor de la riqueza léxica de los artículos seleccionados

Elaborado por: El investigador

Como resultados de la validación cruzada se obtuvo lo siguiente:

1. Los gráficos obtenidos en el anexo 2, literal A que se generaron a partir de los valores verdaderos (similaridades) 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 variable objetivo minimizando los cuadros de las desviaciones, al tener un modelo que tiene coeficientes altos se puede interpretar como unas altas varianzas en los datos.

2. Se tiene un error cuadrático igual a 0.53, la interpretación del error cuadrático muestra la varianza del estimador y su sesgo, este valor de 0.53 muestra una baja variación y demuestra que los datos se estarían ajustando al modelo.

3.2 Resultados de la metodología de desarrollo ágil: metodología scrum

A. Diagrama de arquitectura:

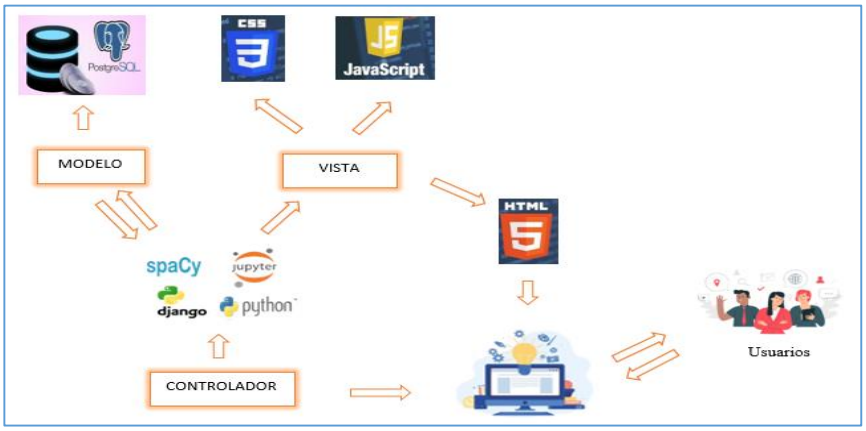


Figura 3.28: Diagrama de arquitectura

Elaborado por: El investigador

B. Roles del equipo scrum

Tabla 3.2: 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

Elaborado por: El investigador

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

D. Product backlog

En esta etapa procederemos a ordenar las prioridades de las historias de usuario

Tabla 3.3: Prioridades de las historias de usuarios

Sprint	Tarea	Estimación de tiempo	Descripción
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	Analizar número de palabras	2 semanas	Para realizar un control de información.

4	Conocer palabras repetitivas	2 semanas	Para realizar un control de información.
5	Conocer la riqueza léxica	2 semanas	Para realizar un control de información.
6	Conocer las palabras comunes	2 semanas	Para realizar un control de información.
7	Conocer distancia y similitud entre textos.	2 semanas	Para realizar un control de información personal.
8	Visualizar graficas	3 semanas	Se visualizará los gráficos de la información.
9	Actualizar información	1 semana	Para realizar un control de información personal.
10	Descargar corpus	1 semana	Para realizar un control de información académica.

Elaborado por: El investigador

E. Historias de usuario de los sprint's

Tabla 3.4: Historia de usuario HU-001

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.**			

Elaborado por: El investigador

Tabla 3.5: Historia de usuario HU-002

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**			

Elaborado por: El investigador

Tabla 3.6: Historia de usuario HU-003

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.**			

Elaborado por: El investigador

Tabla 3.7: Historia de usuario HU-004

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.**			

Elaborado por: El investigador

Tabla 3.8: Historia de usuario HU-005

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.**			

Elaborado por: El investigador

Tabla 3.9: Historia de usuario HU-006

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.**			

Elaborado por: El investigador

F. Casos de uso general

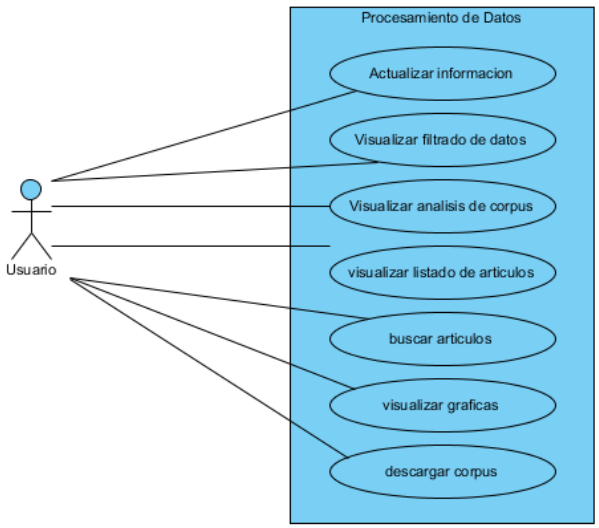


Figura 3.29: Casos de uso general

Elaborado por: El investigador

G. Casos de uso a detalle de los sprint's

Tabla 3.10: Sprint's

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.	

1. El usuario selecciona la opción procesamiento de datos por articulo científico.
2. La plataforma despliega la interfaz inicial del procesamiento de datos.
3. El usuario selecciona una opción de zona.
4. La plataforma carga las universidades que pertenecen a la zona.
5. El usuario selecciona una universidad.
6. La plataforma carga las líneas de investigación que tiene la universidad.
7. El usuario selecciona una línea de investigación.
8. La plataforma carga las sublineas de investigación, artículos que pertenecen a la línea de investigación.
9. El usuario selecciona un artículo especifico
10. La plataforma carga el detalle del articulo y muestra los artículos que tengan relación con el documento seleccionado.

Elaborado por: El investigador

En el **anexo 3** se encuentra el resto de los resultados del proceso de desarrollo del módulo de procesamiento de datos.

3.3 Validación de los resultados

Para la validación de los resultados se considera la hipótesis planteada cuando dice que si se establece un método de análisis de información en corpus de artículos científicos, mediante algoritmos que muestren las frecuencias de palabras, distancias y riquezas léxicas en la Plataforma Científica ECUCIENCIA se podrá obtener reportes de métricas de los documentos en formato PDF que se encuentran en la base de datos del sistema, así como conocimientos cienciométricos de los mismos, para ello se tratarán algunos estadísticos importantes resultante del análisis de los datos considerados para los artículos de muestra seleccionados. Es importante entender que los valores de distancias toman como base para su análisis los valores de la frecuencia.

En la tabla 3.11 se evidencian las métricas tales como el máximo valor de frecuencia de los artículos, siendo el artículo 5 el que presenta un máximo valor y el artículo 2 el mínimo valor de frecuencia con un valor de 3.
Otros valores determinados son la media armónica, la geométrica y la aritmética, al observar con detenimiento se evidencia que se cumple con la relación $H \leq G \leq \bar{X}$ siendo H la media armónica, G la media geométrica y $\bar{X}$ la media aritmética, aunque son diferentes medias la aritmética deberá siempre ser mayor a la geométrica y está a la vez mayor que la armónica.

Por otro lado, se tiene la mediana de la frecuencia de cada artículo, representando que la mediana en el conjunto de datos no ordenado de las frecuencias, muestran que el 50% de los elementos son menores o iguales y el otro 50% mayores o iguales, en este caso para el artículo 1 la mediana de 10 significa que este valor de frecuencia sería el punto medio de las distribuciones y se ordenarían equitativamente los valores de frecuencia superiores a este como valores iguales o mayores, de igual manera las frecuencias iguales o menores a 10 se ordenarían en sentido contrario; para el resto de los artículos se comportaría de igual manera.

Tabla 3.11: Métricas de las frecuencias de palabras de los artículos.

	ART.1	**ART.2**	**ART.3**	**ART.4**	**ART.5**
Máximo valor	35	8	44	27	55
Mínimo valor	7	3	6	7	8
Media armónica	10,76	3,94	10,93	9,93	12,66
Media Geométrica	11,80	4,14	12,06	10,53	14,04
Mediana	10,00	4,00	11,00	9,00	12,00
Media aritmética	13,27	4,38	13,76	11,29	16,13

Elaborado por: El investigador

Análisis de la varianza de un factor

Si se consideran las siguientes hipótesis:

- Ho. La variable frecuencia de los artículos y la media de esta no se comportan de manera significativa en el proceso de representar métricas de acuerdo al corpus de los artículos.
- H1. La variable frecuencia de los artículos y la media de esta se comportan de manera significativa en el proceso de representar métricas de acuerdo al corpus de los artículos.

Para ello se determina la varianza de un factor, obteniéndose el promedio y la varianza del artículo como se muestra en la tabla 3.12.

Tabla 3.12: Resumen del análisis de la varianza de un factor.

RESUMEN				
Grupos	***Cuenta***	***Suma***	***Promedio***	***Varianza***
Artículo 1	45	597	13,26666667	53,01818182
Artículo 2	45	197	4,377777778	2,422222222
Artículo 3	45	619	13,75555556	73,27979798
Artículo 4	45	508	11,28888889	22,43737374
Artículo 5	45	726	16,13333333	104,5272727

Elaborado por: El investigador

En la tabla 3.13 se obtuvo el valor F, según los conceptos de que la varianza (S^2) mide la dispersión de los datos de una muestra ($X_1,X_2,\ldots,X_N$) respecto a la media (X), calculando la media de los cuadrados de las distancias de todos los datos; en la varianza siempre se cumple que esta es mayor o igual que cero ($S2 \geq 0$). Ésta es cero cuando todos los datos son el mismo (ejemplo: {1,1,1,1,1}), otra cualidad positiva es que el valor de F sea mayor al valor crítico para F, en el caso de la tabla 3.13 se tiene a F 17,621 y el valor crítico para F es 2,412, cumpliéndose la superioridad cuantificada de F, esto demuestra que se cumple H1. La variable frecuencia de los artículos y la media de esta se comportan de manera significativa en el proceso de representar métricas de acuerdo al corpus de los artículos. La probabilidad es menor a 0,05, siendo un excelente resultado que valida adecuadamente a H1.

Tabla 3.13: Análisis de la varianza de un factor.

ANÁLISIS DE VARIANZA						
Origen de las variaciones	***Suma de cuadrados***	***Grados de libertad***	***Promedio de los cuadrados***	***F***	***Probabilidad***	***Valor crítico para F***
Entre grupos	3604,3822	4	901,09555	17,621	0,00	2,412
Dentro de los grupos	11250,133	220	51,13696			
Total	14854,515	224				

Elaborado por: El investigador

Desviación estándar

Se obtuvo como desviación estándar el siguiente resultado en la tabla 3.14:

Tabla 3.14: Desviación estándar y media de las frecuencias de los artículos.

Criterio	**σ**	$\bar{X}$
Artículo 1	7,200000	13,26
Artículo 2	1,538959	4,38
Artículo 3	8,464713	13,76
Artículo 4	4,683884	11,29
Artículo 5	10,109621	16,13

Elaborado por: El investigador

La desviación estándar es la medida de dispersión más común, que indica qué tan dispersos están los datos con respecto a la media. Mientras mayor sea la desviación

estándar, mayor será la dispersión de los datos, atendiendo este concepto el artículo con mayor valor de **σ** es el artículo 5 y el de menor valor es el 2.

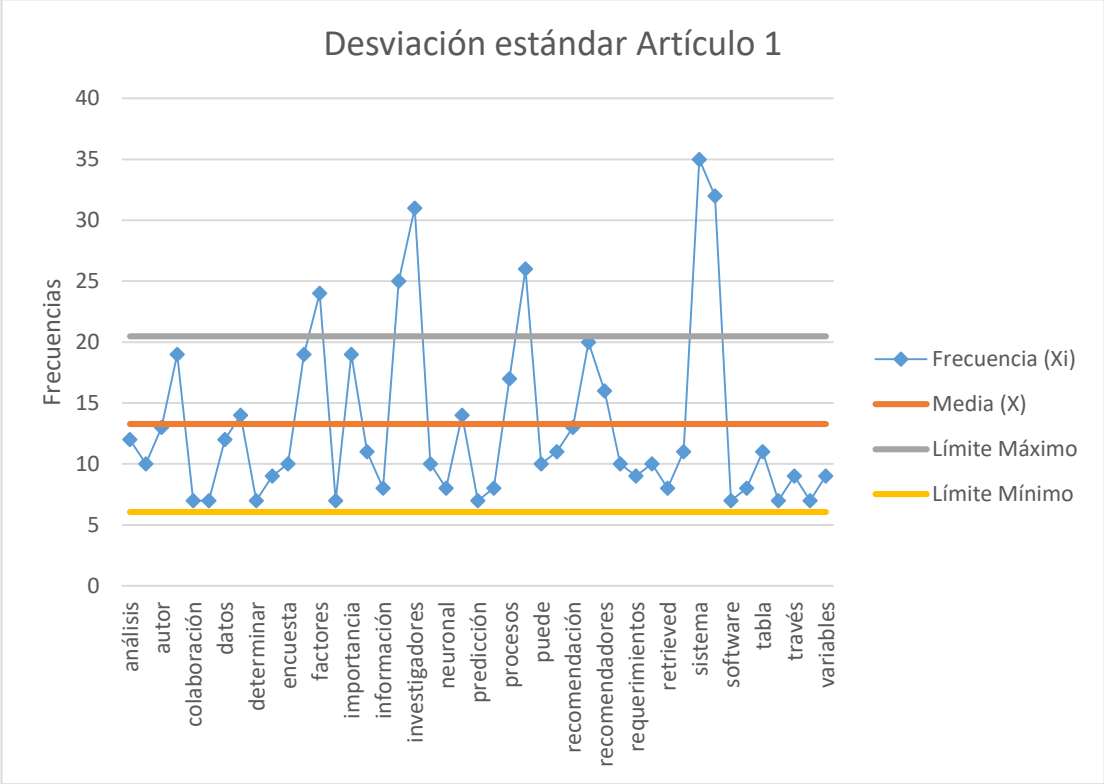


Figura 3.30: Gráfico de σ con límite máximo, mínimo, media y frecuencia

Elaborado por: El investigador

En el gráfico de la figura 3.30 se muestra los resultados de la desviación estándar con el límite máximo que es la suma de la media más el valor de la desviación estándar y el límite es la resta entre estos valores, mostrando que en su mayoría las palabras que están entre estos dos límites se encuentran en el rango de mayor significancia para el artículo, existen algunas palabras como factores, investigación, investigadores, publicaciones, sistema y sistemas que aunque se encuentran fuera de este rango y así alejadas de la media, pero estas constituyen las de mayor frecuencia, comparadas con las que se encuentran dentro de estos límites significa que aunque están alejadas respecto a $\bar{X}$ superan el límite máximo, o sea de manera positiva para el caso de la investigación estas palabras son más fuertes en el documento que el resto, otro aspecto importante es que ninguna se encuentra por

debajo del límite mínimo, en el anexo 4 y 5 se encuentra el resto de los datos y los gráficos de **σ** asociados.

Diagrama de Pareto

El diagrama permite mostrar gráficamente el principio de Pareto (pocos vitales, muchos triviales), es decir, que hay muchos problemas sin importancia frente a unos pocos muy importantes, aplicando este concepto en los artículos de muestra seleccionados se puede interpretar del gráfico 3.31 que el 20% del peso en importancia lo ocupan las palabras de mayor frecuencia en el artículo y el resto que son el 80 % lo ocupan el resto de las demás palabras, en el anexo 5 está el resto de los gráficos.

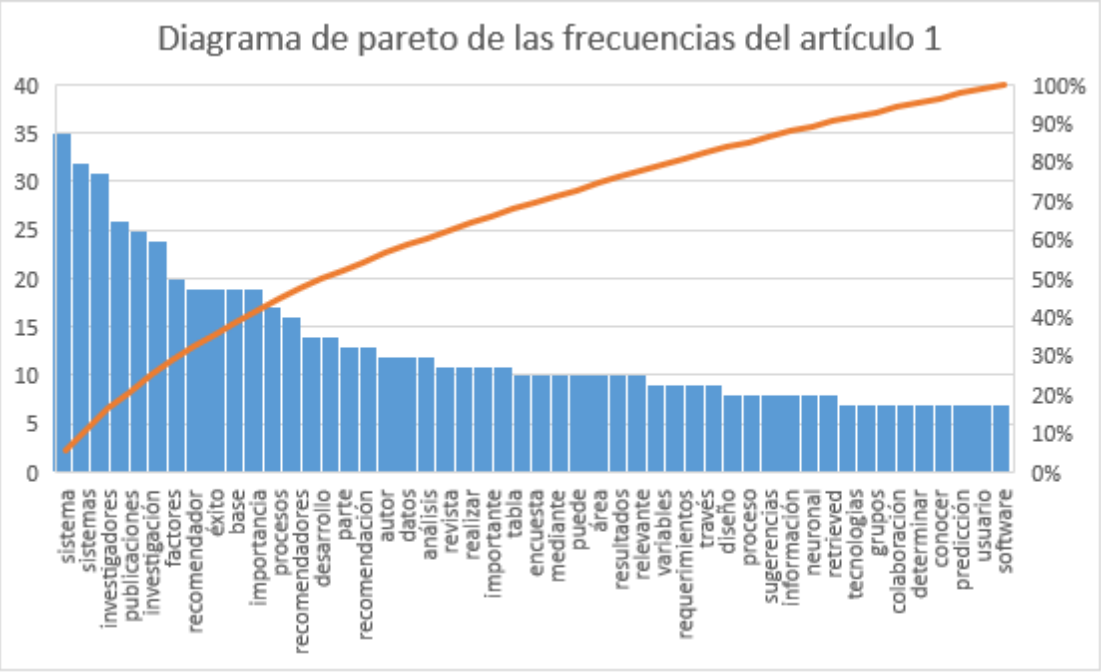


Figura 3.31: Diagrama de Pareto

Elaborado por: El investigador

3.4 Resultados de la valoración económica y, tecnológica.

El presupuesto se realiza considerando el proceso de implementación de los algoritmos y librerías a través del desarrollo de un módulo en la plataforma científica ECUCIENCIA detallados a continuación.

Como parte de los gastos directos se tiene un conjunto de herramientas de software que son necesarias para el desarrollo del módulo de análisis de datos del corpus de los artículos, donde para evitar el incremento excesivo del presupuesto requerido, el investigador considera oportuno emplear herramientas de software libre con licenciamiento gratuito, sin embargo, uno de los valores más elevados dentro de los gastos directos es el costo derivado del proceso de desarrollo.

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.

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 ayudará a definir la complejidad de puntos de función de cada una de las funcionalidades del sistema.

Tabla 3.15: 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

Elaborado por: El investigador

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

Tabla 3.16: Funcionalidades y su tipo.

FUNCIONALIDADES	**TIPO**	**COMPLEJIDAD Alta**
El sistema permitirá actualizar la información de la línea de investigación	EI	6
El sistema permitirá actualizar la información de la sublínea de investigación	EI	6
El sistema permitirá al usuario visualizar el corpus de los artículos científicos	EO	7
El sistema permitirá al usuario visualizar filtrado de datos	EO	7
El sistema permitirá al usuario buscar artículos científicos	EQ	6
El sistema permitirá al usuario visualizar el listado artículos científicos	EO	7
El sistema permitirá al usuario visualizar las graficas	EO	7
El sistema permitirá al usuario descargar el corpus de los artículos	EO	7
Tablas de la base de datos	ILF	105

Elaborado por: El investigador

En la Tabla 3.17 se presenta el número de funcionalidades por cada tipo, en donde se calculó el total de puntos de función sin ajustar (PFSA) dando como resultado:

Tabla 3.17: Número de funcionalidades.

FUNCIONALIDADES	No FUNCIONALIDADES	ALTA	COMPLEJIDAD Alta
Entrada Externa (EI)	2	6	12
Salida Externa (EO)	5	7	35
Consulta Externa (EQ)	1	6	6
Archivo Lógico Interno (ILF)	7	10	70
Archivo Lógico Externo (EIF)	0	0	0
TOTAL			**123**

Elaborado por: El investigador

En la Tabla 3.18 se puede apreciar el factor de ajuste (FA) según IFPUG en donde se establece valores de 1 a 5 por cada factor.

Tabla 3.18: Factor de ajuste

FACTOR DE AJUSTE	PUNTAJE
Comunicación de datos	5
Rendimiento	5
Frecuencia de transacciones	4
Entrada de datos on-line	3
Eficiencia del usuario final	5
Actualizaciones on-line	3
Procesamiento complejo	5
Reusabilidad	4
Facilidad de instalación	5
Facilidad de operación	4
Instalación en distintos lugares	4

Facilidad de cambio	4
TOTAL	**51**

Elaborado por: El investigador

Para el cálculo del total de puntos de función ajustado (PFA) se utiliza la siguiente formula:

- PFA= PFSA*[0.65+(0.01*FA)]
- PFA= 123*[0.65+(0.01*51)]
- PFA= 123*[0.65+(0.51)]
- PFA= 123*1.16
- PFA=142.68 Aproximadamente son 143

A continuación, se procedió a calcular la estimación del esfuerzo requerido el cual consiste en estimar la cantidad de esfuerzo para desarrollar la aplicación. En el Tabla 3.19 se presenta las líneas de código por punto de función y las horas promedio de punto de función según la IFPUG. La cual se toma como referencia los Lenguajes de Cuarta Generación con 8 horas de promedio por punto de función y 20 líneas de código por punto de función.

Tabla 3.19: Lenguaje por horas y línea de código por PF.

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

Elaborado por: El investigador

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.

- H/H=PFA*Horas PF promedio
- H/H=143 *8

- H/H=1144 Horas Hombre

Para calcular el número de días y meses de trabajo se tomó como referencia 5 horas productivas de las 8 y al mes 20 días. Además, se tomó en cuenta 2 desarrolladores.

- Horas=(H/H) /Desarrolladores
- Horas=1160/2
- Horas=572 Duración del proyecto horas
- Días Trabajo=Horas/5
- Días Trabajo=572/5
- Días Trabajo=114.4 aproximado 114
- Meses Desarrollo=Días Trabajo/20
- Meses Desarrollo=114/20

Meses Desarrollo=5.7 meses para desarrollar el software de lunes a viernes 5 horas diarias con 2 desarrolladores. (Estimación de duración del desarrollo del módulo). Finalmente, para calcular la estimación total del proyecto se utilizó la siguiente formula:

- 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.” [51] a los 450 dólares mensuales.

- Total, del Proyecto= (450*2*5.7) + **848,25**
- Total, del Proyecto= (5.131) + **848,25**
- Total, del Proyecto= 5.978 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 5.978 dólares

con 2 desarrolladores con un sueldo de $450.00 dólares, en un tiempo de 5 meses con 7 semanas.

Valoración tecnológica

El sistema de procesamiento de datos representa un aporte tecnológico significativo para la transmisión del conocimiento científico entre los docentes investigadores de la Universidad Técnica de Cotopaxi donde se ha incorporado tecnologías de análisis y procesamiento del lenguaje natural librerías de alto nivel en lenguaje de programación que forma parte de las élites en analítica de datos a nivel mundial como lo es Python, la aplicación de algoritmos de procesamiento de texto, Scikit-learn, biblioteca para aprendizaje automático de software libre para el lenguaje de programación que incluye varios algoritmos de clasificación, regresión y análisis de grupos, todas las herramientas tecnológicas usadas hacen que el proyecto tenga una gran relevancia en el ámbito que ha sido aplicado.

3.5 Conclusiones

- La aplicación de los algoritmos y librerías de análisis de datos, procesamiento de texto y otros en una muestra aleatoria simple de artículos científicos que se encuentran en la base de datos de la Plataforma Científica ECUCIENCIA permitió identificar la frecuencia de las palabras de paradas y de las palabras sin los stopwors permitiendo así obtener valores de un conjunto de métricas asociadas a la distancia euclidiana, chebyshev, Coseno, índice de jaccard, entre otros, pudiéndose representar estas métricas a través de nubes de palabras, de barras, de línea, etc.
- La obtención de los requerimientos de software se lo realizo directamente con los usuarios y beneficiarios contribuyendo así a que el desarrollo del sistema se enfoque en las necesidades de los docentes investigadores del Alma Mater.
- El método estadístico permitió contrastar la hipótesis planteada de que la variable frecuencia de los artículos y la media de esta se comportan de manera

significativa en el proceso de representar métricas de acuerdo al corpus de los artículos científicos analizados.

CONCLUSIONES

- Se estableció un método de análisis de información en el corpus de artículos científicos, de la Plataforma Científica ECUCIENCIA, mediante la aplicación de algoritmos de procesamiento de datos que se encuentran en las librerías NLTK, NUMPY, MATPLOTLIB, PYPDF2, SKLEARN y SCIPY, permitiendo con ello su implementación a través de un módulo web que reporta métricas y visualiza datos de frecuencia, de distancias y nubes de palabras en distintos tipos de gráficos.
- Se pudo constatar con las bibliografías científicas consultadas los principales sustentos teóricos sobre métodos de análisis de información en corpus de documentos, dando como principales resultados los algoritmos de procesamiento de datos que se encuentran en las librerías NLTK, NUMPY, MATPLOTLIB, PYPDF2, SKLEARN y SCIPY, permitiendo su aplicación como parte de los métodos empleados en la investigación.
- Se determinaron los requerimientos algorítmicos necesarios a partir de las librerías estudiadas para el análisis de la información contenidas en el corpus de los artículos científicos seleccionados con el muestreo aleatorio simple dando como resultado información relevante y visualización de datos significativos de las distancias Euclidiana, Correlación, Chebychev, Coseno, Coeficiente de Jaccard y el Índice Dice con los principales algoritmos analizados en Python tales como *TfidfVectorizer, cosine_similarity, scipy.spatial.distance y jaccard_index.*
- La validación de los resultados a través de métodos estadísticos y específicamente con el análisis de la varianza de un factor, arrojó como valor de F = 17,621 siendo mayor que el valor crítico para F que es de 2,412, la probabilidad fue menor a 0,05, demostrando que la variable frecuencia de los artículos y la media de esta se comportan de manera significativa en el proceso de representar métricas de acuerdo al corpus de los artículos, de igual manera la validación cruzada corroboró este resultado.

- Se implementó la lógica del método de análisis de información en el corpus de artículos científicos en formato PDF de la Plataforma Científica ECUCIENCIA, usándose para ello las buenas prácticas que modela SCRUM, obteniéndose un módulo de calidad sobre web, quedando disponible para todos los beneficiarios directos e indirectos de la investigación y sirviendo como una herramienta capaz de ser de referencia en el ámbito de la cienciometría.

RECOMENDACIONES

- Continuar investigando las potencialidades de Python como herramienta de análisis de datos para implementar nuevos algoritmos de procesamiento del lenguaje natural que sirvan para continuar nutriendo a la Plataforma Científica ECUCIENCIA de mayor autonomía e inteligencia en el ámbito de la cienciometría.
- Garantizar el mantenimiento de los distintos módulos de la Plataforma Científica ECUCIENCIA.
- Poner en producción el módulo desarrollado luego de pasar por procesos de testing para verificar el comportamiento del servidor y del resto de los módulos ya implementados.
- Recomendar a la dirección de investigación se use el módulo de procesamiento de datos desarrollado como herramienta para la generación de reportes e informes relacionado a la producción científica de la UTC.

BIBLIOGRAFÍA

[1] UAM_Biblioteca, "Producción científica: Producción Científica de la UAM," 2018. [Online]. Available: https://biblioguias.uam.es/produccion_cientifica. [Accessed: 16-Nov-2019].

[2] E. Ayala Mora, "La investigación científica en las universidades ecuatorianas," *Anales. Rev. la Univ. Cuenca*, vol. 3, no. 57, pp. 61–72, 2015.

[3] C. G. Rivera García, J. M. Espinosa Manfugás, and Y. D. Valdés Bencomo, "La investigación científica en las universidades ecuatorianas. Prioridad del sistema educativo vigente," *Rev. Cuba. Educ. Super.*, vol. 36, no. 2, pp. 113–125, 2017.

[4] Scimago Institutions Rankings, "SIR liber 2015, Rank output 2009-2013," *Scopus*, 2010. [Online]. Available: https://www.scimagoir.com/. [Accessed: 18-Nov-2018].

[5] M. del P. Fernández Díaz, S. Martínez Bernal, C. Rivalta Bermúdez, M. Díaz Rios, and G. Jiménez Santander, "Repositorio de búsquedas y recuperación de la información científica en ciencias de la salud," *EDUMECENTRO*, vol. 5, no. 2, pp. 198–211, 2013.

[6] 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.

[7] 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.

[8] R. Cañedo Andalia and A. J. Dorta Contreras, "SCImago Journal & Country Rank, una plataforma para la evaluación del comportamiento de la ciencia según fuentes documentales y países," *ACIMED*, vol. 21, no. 3, pp. 310–320, 2010.

[9] "SCImago," *Form. Univ.*, vol. 5, no. 5, pp. 1–1, 2012.

[10] SciELO - Scientific Electronic Library Online, "SciELO." [Online]. Available: http://www.scielo.org.ve/. [Accessed: 19-Oct-2018].

[11] T. Dalgleish *et al., Open Access Indicators and Scholarly Communications in Latin America*, 1a ed. Buenos Aires: CLACSO, 2014.

[12] C. Canales Bojo, "La red SciELO (Scientific Electronic Library Online): perspectiva tras 20 años de funcionamiento," *HAD*, vol. 1, no. 4, pp. 211–220, 2017.

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

[14] A. Ochoa Contreras, A. Muñoz García, and H. Morales López, "Perspectivas de la Bibliometría en las Ciencias Médicas," *Arch. en Med. Fam.*, vol. 17, no. 1, pp. 1–3, 2016.

[15] I. De la Vega, "El uso de la cienciometría en la construcción de las políticas tecnocientíficas en américa latina: una relación incierta," *Redes*, vol. 15, no. 29, pp. 217–240, 2009.

[16] DATACIENCIA- Dimensiones de la Producción Científica Nacional, "Programa de Información Científica CONICYT." [Online]. Available: https://dataciencia.conicyt.cl/interfaz/. [Accessed: 22-Nov-2018].

[17] RedCiencia, "Información| Redciencia." [Online]. Available: http://www.redciencia.net/contact. [Accessed: 22-Nov-2018].

[18] REDSEARCH, "REDSEARCH - AYUDA." [Online]. Available: https://redsearch.conicyt.cl/help.php. [Accessed: 22-Nov-2018].

[19] Redalyc.org, "Acerca de Redalyc.org," 2017. [Online]. Available: http://www.redalyc.org/redalyc/media/redalyc_n/Estaticas3/mision.html. [Accessed: 22-Nov-2018].

[20] I. Timarán-Pereira, S. R. Hernández-Arteaga, S. J. Caicedo-Zambrano, A. Hidalgo-Troya, and J. C. Alvarado- Pérez, "El proceso de descubrimiento de conocimiento en bases de datos.," in *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional*, Bogotá: Ediciones Universidad Cooperativa de Colombia, 2016, pp. 63–86.

[21] J. M. Molina López and J. García Herrero, "técnicas de análisis de datos aplicaciones prácticas utilizando microsoft excel y weka," Universidad Carlos III. Madrid, 2006.

[22] Y. Aranda Robles and A. R. Sotolongo, “Integración de los algoritmos de minería de datos 1R, PRISM E ID3 A POSTGRESQL,” *J. Inf. Syst. Technol. Manag.*, vol. 10, no. 2, pp. 389–406, Aug. 2013.

[23] J. Hernández Cáceres, “Clustering technique based on k- means algorithm for the identification of clusters of surgical patients,” *Univ. St. Tomás, Secc. Bucaramanga*, pp. 1–8, 2016.

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

[25] 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.

[26] N. Morato, “Tutorial Anaconda: Qué es, cómo instalarlo y cómo se usa,” *IKKARO*, vol. 0, no. 0, 2019.

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

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

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

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

[31] N. Morato, “Tutorial Anaconda: Qué es, cómo instalarlo y cómo se usa,” *IKKARO*, vol. 0, no. 0, 2019.

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

[33] A. Vara Serrano, “PREDICCIÓN DE VISITAS MEDIANTE GEOLOCALIZACIÓN A TRAVÉS DE DISPOSITIVOS MÓVILES,” Universidad de Barcelona, 2017.

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

[35] R. E. Lopez Briega, “Procesamiento del lenguaje natural con Python,”

Matematicas, Analisis de Datos y Python, 2017.

[36] Wes McKinney & PyData Development Team, “pandas: powerful Python data analysis toolkit Release 0.23.4 Wes McKinney & PyData Development Team,” 2018.

[37] S. van der Walt, S. C. Colbert, and G. Varoquaux, “The NumPy Array: A Structure for Efficient Numerical Computation,” *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, Mar. 2011.

[38] N. Aguilera, *Matemáticas y programación con Python*. 2014.

[39] scikit-learn, “Aprendizaje automático en Python: documentación de scikit-learn 0.20.1,” 2017. [Online]. Available: https://scikit-learn.org/stable/index.html. [Accessed: 12-Dec-2018].

[40] M. Garre, J. J. Cuadrado, M. Sicilia, D. Rodríguez, and R. Rejas, “Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software,” *Rev. Española Innovación, Calid. e Ing. del Softw.*, vol. 3, no. 1, pp. 6–22, 2007.

[41] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[42] L. Buitinck *et al.*, “API design for machine learning software: experiences from the scikit-learn project,” Sep. 2013.

[43] J. Pérez, L. Cruz, G. Reyes, and A. Mexicano, “Mejora al algoritmo de agrupamiento K-means mediante un nuevo criterio de convergencia y su aplicación a bases de datos poblacionales de cáncer,” *Liver- 2do Taller Lat. Iberoam. Investig. Operaciones "la IO Apl. a la solución Probl. Reg.*, no. August 2015, pp. 1–7, 2007.

[44] Z. Yang, R. Algesheimer, and C. J. Tessone, “A Comparative Analysis of Community Detection Algorithms on Artificial Networks,” *Sci. Rep.*, vol. 6, no. 1, p. 30750, Nov. 2016.

[45] W. G. Witt, “Quantifying the Structure of Misfolded Proteins Using Graph Theory,” East Tennessee State University, 2017.

[46] C. E. Román Godoy, “Identificación de fibras cerebrales cortas basada en Clustering jerárquico a partir de base de datos HARDI,” UNIVERSIDAD DE CONCEPCIÓN, 2017.

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

[48] Kruskal. (1964). Multidimensional scaling by opti-mizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29, pp. 1-27.

[49] I. Timarán-Pereira, S. R. Hernández-Arteaga, S. J. Caicedo-Zambrano, A. Hidalgo-Troya, and J. C. Alvarado- Pérez, “El proceso de descubrimiento de conocimiento en bases de datos.,” in *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional*, Bogotá: Ediciones Universidad Cooperativa de Colombia, 2016, pp. 63–86.

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

[51] 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

Anexo 1. Tabla de datos luego de realizar limpieza, eliminar valores vacíos y optimizar los campos necesarios (por la magnitud de la tabla solo se pondrán filas iniciales y finales con el número, ID y documento).

NO.	ID	TITULO	DOCUMENTO
1	286	Agile method for detecting DDoS attacks in the application layer based on userâ€™s dynamism	articulo/IJET18-10-03-125.pdf
2	258	MÃ • RKETING DIGITAL UNA NUEVA ESTRATEGIA PARA LOS EMPRENDEDORES	articulo/MARKETING_DIGITAL_ZDdqquR.pdf
3	540	CONSIDERACIONES GENERALES SOBRE EL PROCESO DE ELABORACIÃ“N DE SILOS	articulo/CONSIDERACIONES_GENERALES_SOBRE_EL_PROCESO.pdf
4	541	EVALUACIÃ“N DE LA CALIDAD NUTRITIVA DEUN ENSILADO PARA LA ALIMENTACIÃ“N DE GANADO LECHERO A PARTIR DE LOS RESIDUOS PROVENIENTES DEL TRILLADO DE QUINUA (CHENOPODIUM QUINOA WILLD) Y SANGORACHE (AMARANTHUS HYBRIDUS L.	articulo/EVALUACION_DE_LA_CALIDAD_NUTRITIVA_DE_UN_ENSILADO.pdf
5	144	La inclusiÃ³n del bagazo de caÃ±a en la raciÃ³n de cuyes (Cavia porcellus) de engorde	articulo/La_inclusiÃ³n_del_bagazo_de_caÃ±a.PDF
6	136	GUÃ • A DE EJERCICIOS APLICADO A LA PERIODIZACIÃ“N TÃ • CTICA; EN EL RENDIMIENTO DEPORTIVO DEL CLUB UTC	articulo/articulo-presentado.pdf
7	16	Micromachismo: manifestaciÃ³n de violencia simbÃ³lica	articulo/Micromachismo_Magaly_Gina.pdf
8	502	Tendencias del uso de las tecnologÃas y conducta del consumidor tecnolÃ³gico	articulo/Tendencias_del_uso_de_las_tecnologÃas_30-04-2017.pdf
9	499	Mirando hacia el futuro con pensamiento complejo en la educaciÃ³n Superior	articulo/PublicaciÃ³n_1.pdf
10	71	LABORATORIO DE NEUROCIENCIAS APLICADO A Ã • REAS ADMINISTRATIVAS: NEUROMARKETING EN EDUCACIÃ“N SUPERIOR	articulo/4.1._articulo_neuromarketing_rJJbk9n.pdf
11	475	La gestiÃ³n formativa en post-grados	articulo/La_gestiÃ³n_formativa_en.pdf
12	417	CLAVE PARA DETERMINACIÃ“N DE MUSGOS EPÃ • FITOS DE LA SERRANÃ • A DEL LITORAL; CORDILLERA DE LA COSTA VENEZOLANA	articulo/clave_musgos__ABV.pdf
13	333	El Mejoramiento de la Eficiencia y la Productividad: Plan de Mejora como respuesta al sistema productivo.	articulo/VENEZUELA_PDF.pdf
14	190	EL PROSUMER EN LA CONSTRUCCIÃ“N DEL DISCURSO RADIOFÃ“NICO: ANÃ • LISIS DE CASO DE LAS RADIOS	articulo/ReMedCom_08_02_18.pdf

		ECUATORIANAS DE COTOPAXI Y TUNGURAHUA	
15	401	CaracterizaciÃ³n nutricional del palmiste (Elaeis guineensis jacq.) procedente de dos extractoras de aceite	articulo/CARACTERIZACIÃ“N_DEL_PALMISTE.pdf
16	558	EFECTO DE DIFERENTES ABONOS ORGÃ • NICOS EN LA PRODUCCIÃ“N DE TOMATE (Solanum lycopersicum; L)	articulo/333-734-1-SM_dxKJBvM.pdf
17	569	Articulo	articulo/798-2253-1-PB.pdf
18	610	La calidad de la educaciÃ³n en las instituciones de educaciÃ³n superior bajo una nueva reforma curricular	articulo/documento_publicado_sonia_gonzalo_gp86gKU.pdf
19	550	LUPIN PEST MANAGEMENT IN THE ECUADORIAN ANDES: CURRENT KNOWLEDGE AND PERSPECTIVES	articulo/S5.pdf
20	543	BIOMETRIC SIGNS OF AN AMARANTHUS IN THE CONDITIONS OF ECUADOR	articulo/Articulo_emerson_vdnaba_2015_3_26.pdf
21	599	NIVEL DE SATISFACCIÃ“N DE EGRESADOS DE LA CARRERA DE INGENIERÃ • A EN MARKETING DE LA UNIVERSIDAD â€œELOY ALFAROâ€ • DE MANABÃ • CAMPUS BAHÃ • A DE CARÃ • QUEZ	articulo/Dialnet-NivelDeSatisfaccionDeEgresadosDeLaCarreraDeIngenie-6716273.pdf
22	74	MARKETING DIGITAL; UNA VISIÃ“N DESDE LA ACADÃ‰MIA	articulo/7.1._articulo_marketing_digital_5xVBioy.pdf
23	648	EVALUACIÃ“N DE FACTORES DE RIESGOS PSICOSOCIALES EN LOS SERVIDORES DE UNA EMPRESA PÃšBLICA; SANTO DOMINGO DE LOS TSÃ • CHILAS; ECUADOR	articulo/EVALUACIÃ“N_DE_FACTORES_DE_RIESGOS_PSICOSOCIALES.pdf
24	605	Una mirada del proceso de regulaciÃ³n contable internacional en el contexto de globalizaciÃ³n	articulo/Panchi_Una_mirada_del_proceso_de_regulaciÃ³n_contable2017.pdf
25	271	ANÃ • LISIS DE EQUIDAD DE GENERO EN LA UNIVERSIDAD TÃ‰CNICA DE COTOPAXI	articulo/472017_ANALISIS_DE_LA_EQUIDAD_DE_GENERO.pdf
26	227	ESTIMACIÃ“N DE DATOS FALTANTES DE PRECIPITACIÃ“N EN LA SUBCUENCA DEL RÃ • O PATATE	articulo/ART_1.pdf
27	734	Factors that Influence Undergraduate University Desertion According to Students Perspective	articulo/IJET18_GFifQ49.pdf
28	521	LA LINGUISTICA APLICADA A LA ENSEÃ‘ANZA DE UNA SEGUNDA LENGUA	articulo/la_linguistica_aplicada.pdf
29	668	Emprendimientos innovadores a partir de competencias cognitivas en estudiantes universitarios	articulo/articulo_4.pdf
30	561	Resultados del proyecto formativo de rediseÃ±o curricular de la carrera de EducaciÃ³n BÃ¡sica de la Universidad TÃ©cnica de Cotopaxi	articulo/879-3170-4-PB.pdf

31	257	EL DOCENTE LÃ • DER TRANSFORMADOR DE SUEÃ‘OS	articulo/ARTICULO_EL_DOCENTE_LIDER.pdf
32	420	CONTRIBUCIÃ“N AL CONOCIMIENTO DE LAS EPÃ • FILAS DE VENEZUELA.	articulo/epifilas_de_venezuela.pdf
33	519	â€œRelaciÃ³n de los indicadores de la calidad y la edad en dos variedades de Megathyrsus maximusâ€ • ;	articulo/Carta_aceptaciÃ³n_manuscrito_02.pdf
…	…	….	….
171	82	ESTRUCTURA Y RELACIONES GENÃ‰TICAS DEL CERDO CRIOLLO DE ECUADOR	articulo/CERDO_CRIOLLO.pdf
172	740	ProducciÃ³n de leche como respuesta a la fertilizaciÃ³n y riego en ganaderÃas de ecosistemas andinos en Ecuador	articulo/Redvetmayo2018.pdf
173	460	Balance forrajero; de energÃa y nitrÃ³geno en pastizales arborizados con Algarrobo (Prosopis juliflora (S.W.) DC.) bajo pastoreo de vacas lecheras.	articulo/Articulo_blance_forrajero.pdf
174	714	Manejo de asociaciones gramÃneas-leguminosas en pastoreo con rumiantes para mejorar su persistencia; la productividad animal y el impacto ambiental en los trÃ³picos y regiones templadas	articulo/GramineasLeguminosasvacas.pdf
175	178	Mathematical modeling of the natural solar drying process in lateritic mineral deposits	articulo/26._ArtÃculo_Autor_Solar_drying_IJM_2017.pdf
176	298	ESTRUCTURA Y COMPOSICIÃ“N DE LA FLORA Y FAUNA EN LA PARROQUIA SANGAY; MORONA SANTIAGO; ECUADOR: IMPLICACIONES AMBIENTALES	articulo/ARTICULO_FLORA_Y_FAUNA_SANGAY.pdf
177	379	Infancia vulnerable y condiciones de acceso a la educaciÃ³n en el sector rural: el caso de la parroquia San JosÃ© de PoalÃ³; de la provincia de Cotopaxi	articulo/Articulo_Vulnerabilidad.pdf
178	230	Factores determinantes en la planeaciÃ³n estratÃ©gica	articulo/2016_utciencia.pdf
179	600	Life Cycle Analysis of the panela agroindustry: Intensification for its development	articulo/Art._Life_Cycle_Analysis_of_the_panela_agroindustry.pdf
180	79	CaracterizaciÃ³n genÃ©tica de la cabra Criolla Cubana mediante marcadores microsatÃ©lites	articulo/193015664002.pdf
181	386	El desarrollo de la Ciberradio en la regiÃ³n central de Ecuador	articulo/576-1613-1-PB.pdf
182	580	Recovery of Heavy Metals from the Spent Catalyst of the Hydrotreating Unit (HDT) for the Use of the Impregnation of Supported Catalysts	articulo/ArtÃculo_-_Recovery_of_Heavy_-_JapÃ³n_CW3YXrM.pdf
183	368	CONCEPCIÃ“N GENÃ‰TICA DE LA DIRECCIÃ“N DEL APRENDIZAJE DE LA HISTORIA	articulo/ConcepciÃ³n_GenÃ©tica_de_la_DirecciÃ³n_del_Aprendizaje_de_la_Historia..pdf
184	85	ESTRUCTURA GENÃ‰TICA Y CARACTERIZACIÃ“N	articulo/45661-279583-2-PB.pdf

		MOLECULAR DEL CERDO CRIOLLO (SUS SCROFA DOMESTICA) DE ECUADOR; UTILIZANDO MARCADORES MICROSATÃ‰LITE	
185	634	FormaciÃ³n acadÃ©mica en la industria grÃ¡fica de Riobamba - Ecuador	articulo/Certificado_de_publicacion.pdf
186	343	Perfil emprendedor de los maestros en la Provincia de Cotopaxi	articulo/Perfil_emprendedor_maestros_artÃculo.pdf
187	741	Milk production and sustainability of the dairy livestock systems with a high calving concentrate pattern at the early spring	articulo/MilKproductionredvet.pdf
188	70	LAS NEUROCIENCIAS. UNA VISIÃ“N DE SU APLICACIÃ“N EN LA EDUCACIÃ“N	articulo/3.1._articulo_neurociencia_3RLB88O.pdf
189	184	Estudio de Caso en las Ciencias Empresariales	articulo/Dialnet-ElEstudioDeCasoEnLasCienciasEmpresariales-5924581.pdf
190	472	CONSIDERACIONES GENERALES ACERCA DE LA FORMACIÃ“N HUMANISTA EN LAS CIENCIAS DE LA SALUD: UN PLANTEAMIENTO TEÃ“RICO.	articulo/742-2502-1-PB_Ax4A2v1.pdf
191	429	RESPONSABILIDAD SOCIAL DE LA UNIVERSIDAD DE COTOPAXI ANTE LAS EXIGENCIAS DEL DESARROLLO EN EL ECUADOR	articulo/Articulo_2015_AFkDW3S.pdf
192	338	Sistema de sedimentaciÃ³n para la recuperaciÃ³n de aguas residuales del proceso de lavado de Ã¡ridos en la UEB del Jobo	articulo/1001-1978-1-SM.pdf
193	793	GESTIÃ“N PARTICIPATIVA DEL ASEGURAMIENTO DE LA CALIDAD EN LA UNIVERSIDAD TÃ‰CNICA DE COTOPAXI	articulo/07CA201902.pdf
194	242	AUTOESTIMA; EDUCACIÃ“N EMOCIONAL Y SU INCIDENCIA EN EL PROCESO DE ENSEÃ‘ANZA APRENDIZAJE DE LOS ESTUDIANTES EN LAS INSTITUCIONES EDUCATIVAS	articulo/ARTICULO_AUTOESTIMA_EDUCACION_EMOCIONAL.pdf
195	252	LA NUEVA ADMINISTRACIÃ“N DEL SIGLO XXI	articulo/LA_NUEVA_ADMINISTRACIÃ“N_DEL_SIGLO_XXI_QHsZ8GW.pdf
196	238	METODOLOGÃ • A DE ENSEÃ‘ANZA DEL SISTEMA DE COSTOS POR PROCESO	articulo/metodologia_costos.pdf
197	156	ComparaciÃ³n de dos mÃ©todos para la predicciÃ³n de la rugosidad superficial en el torneado del acero inoxidable AISI 316L	articulo/art10.pdf
198	249	El empoderamiento de la mujer en la AsociaciÃ³n de Artesanos de la Victoria; PujilÃ; provincia de Cotopaxi	articulo/Empoderamiento_Lavictoria_Utciencia.pdf
199	430	LA UNIVERSIDAD DE COTOPAXI COMO INSTITUCIÃ“N CAPAZ DE	articulo/cotopaxi.pdf

		DAR RESPUESTA AL ENCARGO SOCIAL UNIVERSITARIO	
200	432	MEJORA CONTINUA DE LOS PROCESOS DE GESTIÃ“N DEL CONOCIMIENTO EN INSTITUCIONES DE EDUCACIÃ“N SUPERIOR ECUATORIANAS.	articulo/ARTICULO_2017.pdf
201	644	Andamiaje MetodolÃ³gico en los Estudios Organizacionales: aplicaciÃ³n en Liderazgo Organizacional	articulo/articulo_publicado_en_SAPIENZA_ULA_Venezuela_i8fLXP1.pdf

Anexo 2: Análisis completo de los artículos de muestra

A. DOC-1:

Factores_de_exito_para_sistemas_recomendadores_de_procesos_de_investigacion.pdf

- Número de páginas: 11

a) FRECUENCIA DE PALABRAS:

- palabras: ['sistema', 'sistemas', 'investigadores', 'publicaciones', 'investigación', 'factores', 'recomendador', 'éxito', 'base', 'importancia', 'procesos', 'recomendadores', 'desarrollo', 'parte', 'recomendación', 'autor', 'datos', 'análisis', 'revista']
- frecuencia: ['35', '32', '31', '26', '25', '24', '20', '19', '19', '19', '17', '16', '14', '14', '13', '13', '12', '12', '11']

b) Número de palabras:

- Con palabras de parada: 4362
- Sin palabras de parada: 1868

c) Numero de palabras de parada: 2494

d) Riqueza léxica: 0.42

B. DOC-2:

Finca_agroecologica_sostenible_de_la_Universidad_de_Granma.pdf

- Número de páginas: 3

a) FRECUENCIA DE PALABRAS:

- palabras: ['biodiversidad', 'agricultura', 'popular', 'plantas', 'especies', 'cultivos', 'desarrollo', 'finca', 'área', 'mujeres', 'centro', 'universidad', 'total', 'animales', 'agrícola', 'trabajo', 'familias', 'variedades', 'introducción', 'materia']
- Frecuencia: ['8', '8', '7', '7', '7', '7', '7', '6', '6', '6', '5', '5', '5', '5', '5', '5', '4', '4', '4', '4']

b) Número de palabras:

- Con palabras de parada: 1725
- Sin palabras de parada: 726

c) Numero de palabras de parada: 999
d) Riqueza léxica: 0.67

C. DOC-3:

Guia_virtual_interactiva_en_Android_a_través_de_codigos_QR_en_el_Museo_de_la_Escuela_Fiscal_Isidro_Ayora_del_Ecuador.pdf

- Número de páginas: 10

a) FRECUENCIA DE PALABRAS:
   - palabras: ['información', 'museo', 'aplicación', 'códigos', 'datos', 'código', 'figura', 'isidro', 'guía', 'virtual', 'museos', 'escuela', 'arte', 'desarrollo', 'usuarios', 'android', 'sistema', 'metodología', 'prueba']
   - frecuencia: ['44', '41', '37', '25', '22', '21', '20', '19', '17', '17', '15', '14', '14', '14', '14', '13', '13', '13', '13']
b) Número de palabras:
   - Con palabras de parada: 5566
   - Sin palabras de parada: 2410
c) Numero de palabras de parada: 3156
d) Riqueza léxica: 0.48

D. DOC-4:

Optimizacion_con_Colonia_de_Hormigas_para_la_Planificacion_optima_de_la_Fuerza_de_Trabajo.pdf

- Número de páginas: 7

a) FRECUENCIA DE PALABRAS:
   - palabras: ['trabajo', 'algoritmo', 'hormigas', 'universidad', 'problema', 'exploración', 'trabajador', 'optimización', 'número', 'valor', 'planificación', 'fuerza', 'solución', 'colonia', 'puede', 'cada', 'técnica']
   - frecuencia: ['27', '26', '18', '18', '17', '17', '17', '16', '14', '14', '13', '13', '13', '13', '13', '13', '12']
b) Número de palabras:
   - Con palabras de parada: 4687
   - Sin palabras de parada: 1791

c) Numero de palabras de parada: 2896
d) Riqueza léxica: 0.45

E. DOC-5:

Utopia_o_realidad_de_aplicaciones_informaticas_en_la_educacion_Caso_universidad_ecuatoriana.pdf

- Número de páginas: 19

a) FRECUENCIA DE PALABRAS:

- palabras: ['educación', 'sociedad', 'universidad', 'información', 'ecuatoriana', 'realidad', 'estudiantes', 'utilización', 'utopía', 'aplicaciones', 'caso', 'informáticas', 'revista', 'issn', 'componentes', 'nuevos', 'investigación', 'base']
- frecuencia: ['55', '50', '29', '29', '27', '26', '26', '25', '24', '21', '21', '20', '19', '19', '19', '15', '15', '15']

b) Número de palabras:

- Con palabras de parada: 5604
- Sin palabras de parada: 2294

c) Numero de palabras de parada: 3310
d) Riqueza léxica: 0.39

F. DISTANCE SIMILARITY DOC1-DOC2

a) DISTANCIA SIMILITUD :

- similarity: 0.133
- chebyshev: 0.274
- correlation: 1.136
- cosine: 0.867
- dice: 0.992
- euclidean: 1.317
- jaccard: 0.661
- minkowski: 0.436

A. DISTANCE SIMILARITY DOC1-DOC3

a) DISTANCIA SIMILITUD:

- similarity: 0.265
- chebyshev: 0.315
- correlation: 0.890
- cosine: 0.735
- dice: 0.985
- euclidean: 1.212
- jaccard: 0.703
- minkowski: 0.464

B. DISTANCE SIMILARITY DOC1-DOC4

b) DISTANCIA SIMILITUD:

- similarity: 0.173
- chebyshev: 0.288
- correlation: 1.029
- cosine: 0.827
- dice: 0.990
- euclidean: 1.286
- jaccard: 0.684
- minkowski: 0.450

C. DISTANCE SIMILARITY DOC1-DOC5

c) DISTANCIA SIMILITUD:

- similarity: 0.233
- chebyshev: 0.358
- correlation: 0.921
- cosine: 0.767
- dice: 0.985
- euclidean: 1.238
- jaccard: 0.719

- minkowski: 0.495

D. DISTANCE SIMILARITY DOC2-DOC3

d) DISTANCIA SIMILITUD:

- similarity: 0.110
- chebyshev: 0.338
- correlation: 1.130
- cosine: 0.890
- dice: 0.994
- euclidean: 1.334
- jaccard: 0.645
- minkowski: 0.469

E. DISTANCE SIMILARITY DOC2-DOC4

e) DISTANCIA SIMILITUD:

- similarity: 0.129
- chebyshev: 0.258
- correlation: 1.165
- cosine: 0.871
- dice: 0.993
- euclidean: 1.320
- jaccard: 0.659
- minkowski: 0.396

F. DISTANCE SIMILARITY DOC2-DOC5

f) DISTANCIA SIMILITUD:

- similarity: 0.130
- chebyshev: 0.375
- correlation: 1.106
- cosine: 0.870
- dice: 0.992

- euclidean: 1.319
- jaccard: 0.675
- minkowski: 0.478

G. DISTANCE SIMILARITY DOC3-DOC4

g) DISTANCIA SIMILITUD:

- similarity: 0.156
- chebyshev: 0.315
- correlation: 1.033
- cosine: 0.844
- dice: 0.991
- euclidean: 1.299
- jaccard: 0.705
- minkowski: 0.465

H. DISTANCE SIMILARITY DOC3-DOC5

h) DISTANCIA SIMILITUD:

- similarity: 0.249
- chebyshev: 0.359
- correlation: 0.894
- cosine: 0.751
- dice: 0.985
- euclidean: 1.226
- jaccard: 0.735
- minkowski: 0.500

I. DISTANCE SIMILARITY DOC4-DOC5

i) DISTANCIA SIMILITUD:

- similarity: 0.161
- chebyshev: 0.375
- correlation: 1.018

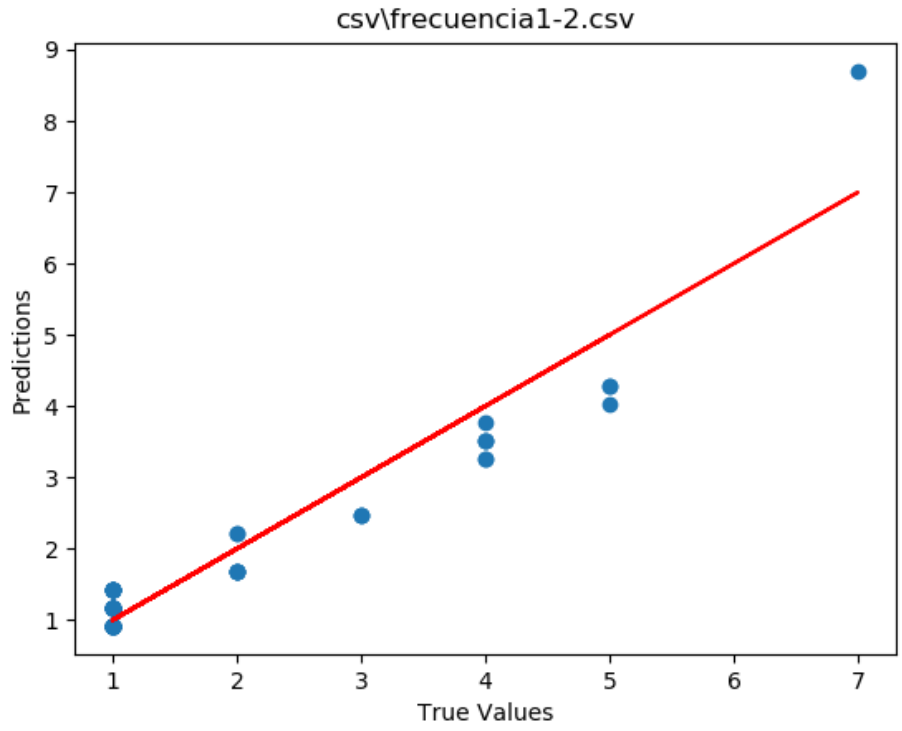
- cosine: 0.839
- dice: 0.990
- euclidean: 1.295
- jaccard: 0.678
- minkowski: 0.490

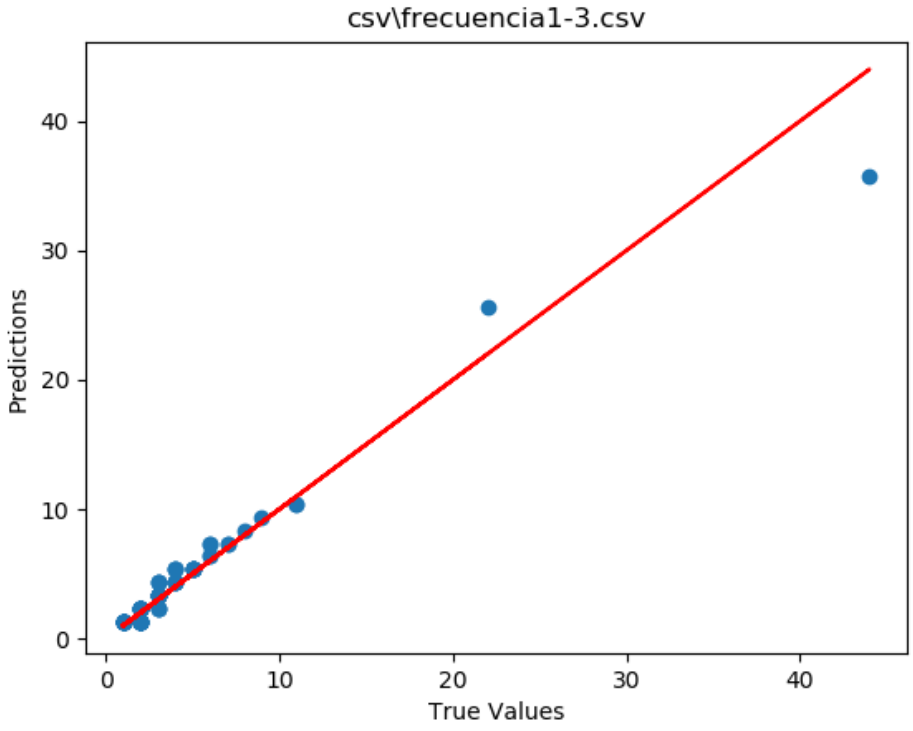
J. DISTANCE SIMILARITY DOC5-DOC5

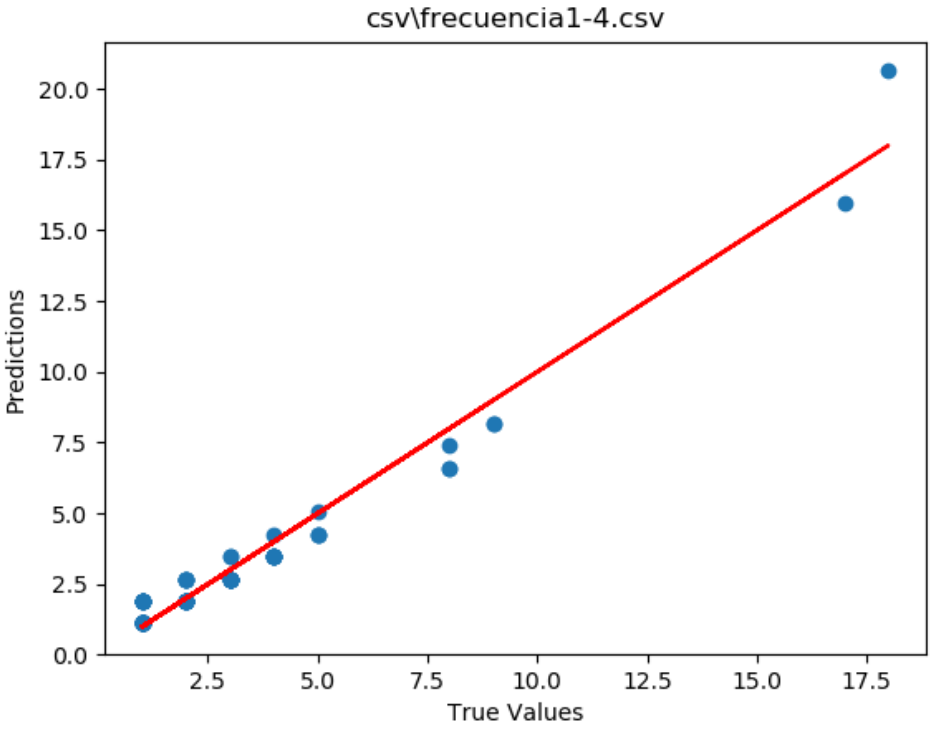
j) DISTANCIA SIMILITUD:

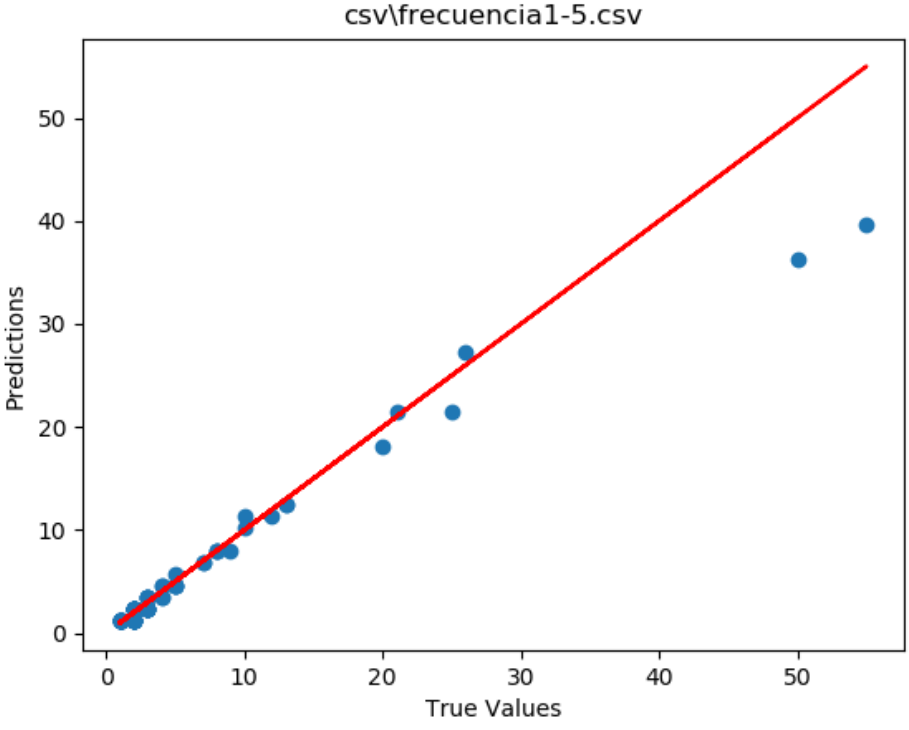
- similarity: 1.000
- chebyshev: 0.000
- correlation: 0.000
- cosine: 0.000
- dice: 0.936
- euclidean: 0.000
- jaccard: 1.000
- minkowski: 0.000

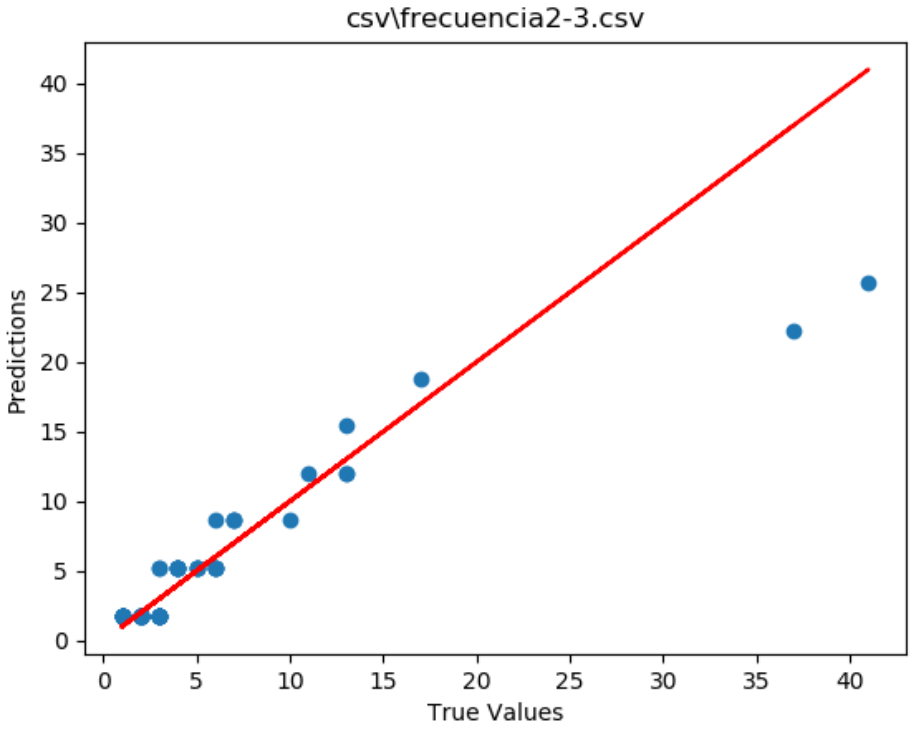
A-VALIDACIÓN CRUZADA: Gráficos generados a partir de los valores verdaderos y los de predicción usando un modelo de regresión lineal.

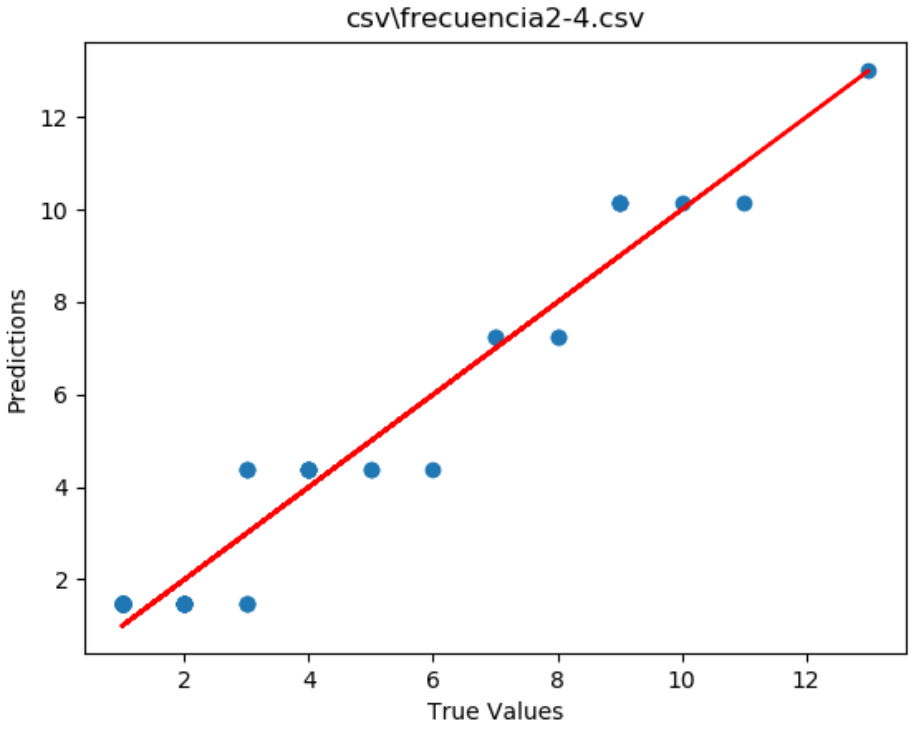


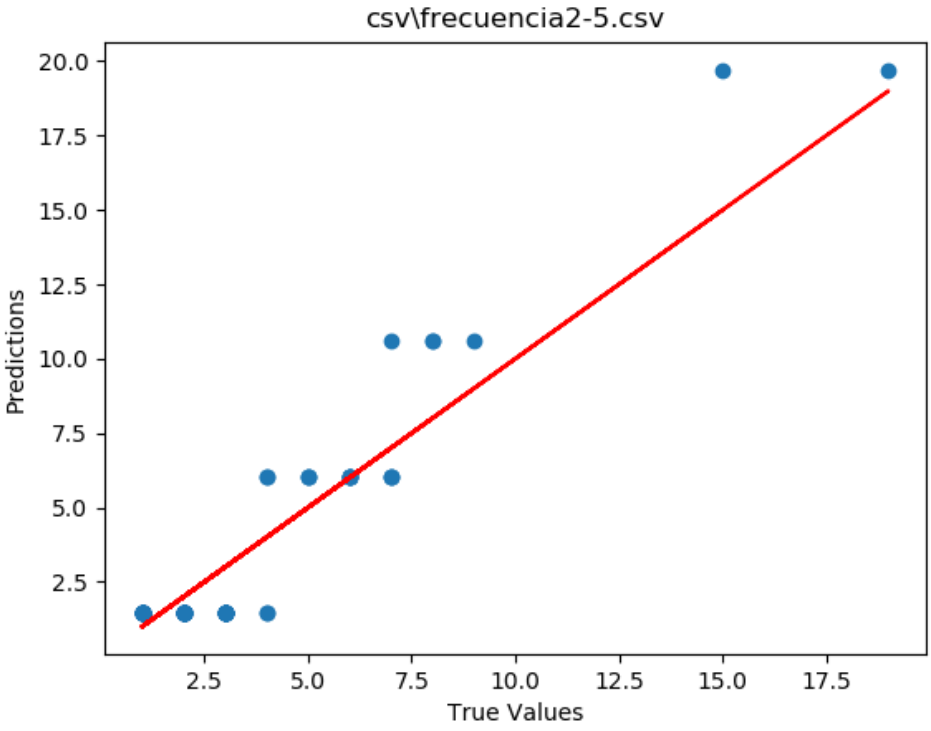


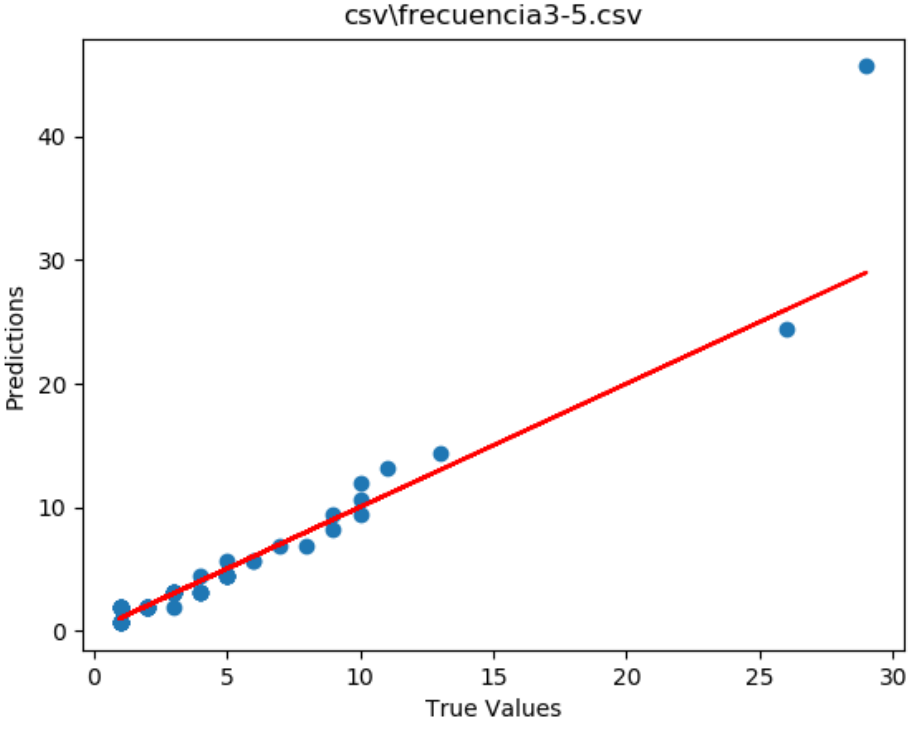

csv\frecuencia1-4.csv
Predictions
True Values
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

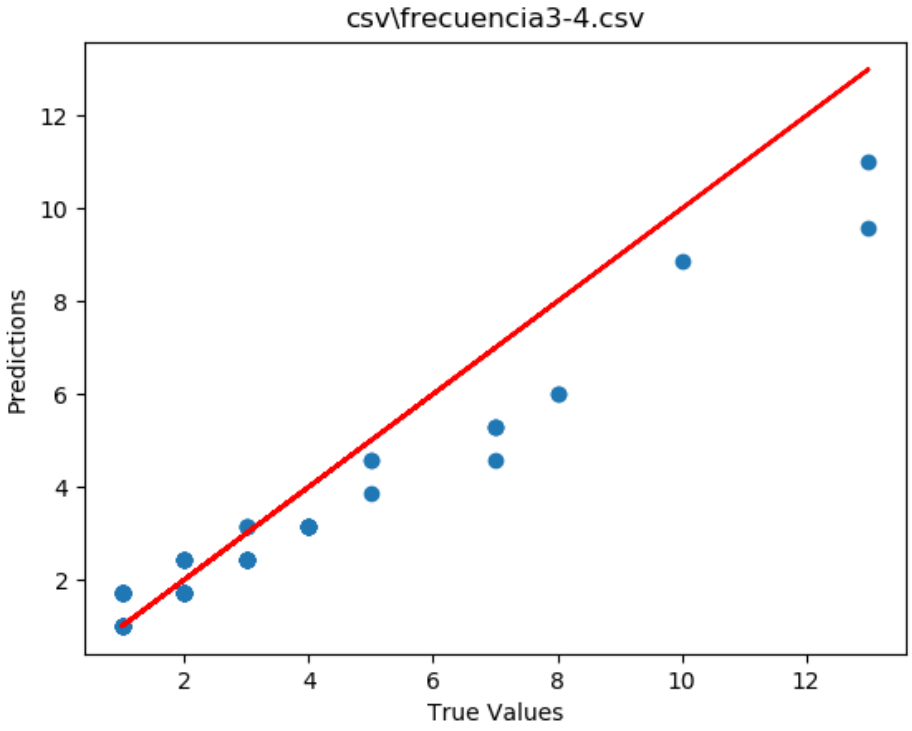

csv\frecuencia1-5.csv
Predictions
True Values
0
10
20
30
40
50

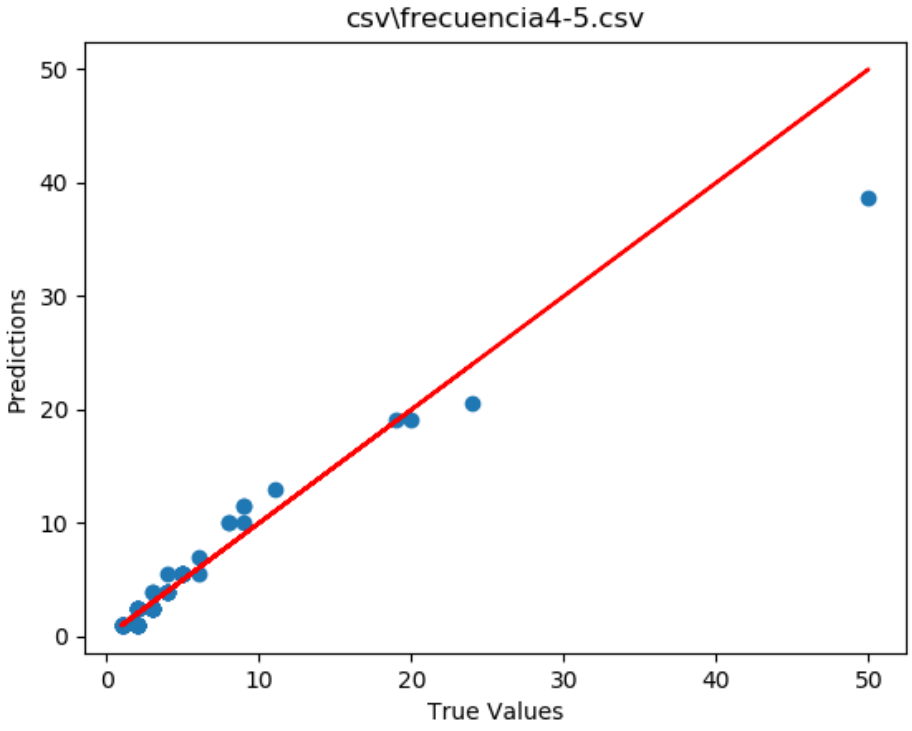

csv\frecuencia2-3.csv
Predictions
True Values
0
5
10
15
20
25
30
35
40


csv\frecuencia2-4.csv
Predictions
True Values
2
4
6
8
10
12


csv\frecuencia2-5.csv
Predictions
True Values
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
2.5
5.0
7.5
10.0
12.5
15.0
17.5


csv\frecuencia3-5.csv
Predictions
True Values
40
30
20
10
0
0
5
10
15
20
25
30


csv\frecuencia3-4.csv
Predictions
True Values
2
4
6
8
10
12


csv\frecuencia4-5.csv
Predictions
True Values
0
10
20
30
40
50

Anexo 3. Resultado de la aplicación de la metodología Scrum.

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. 1. **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.** 2. **El usuario selecciona la opción análisis de datos.** 3. **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.**	

Nº caso	**CU-003**
H.U	HU-003: Distancia y similtud
Nombre	obtener corpus
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: 1. **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.** 2. **El usuario selecciona el articulo y da click en ver detalle.** 3. **La plataforma muestra el detalle del artículo en un modal**	

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 graficas 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.**

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.**

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.**

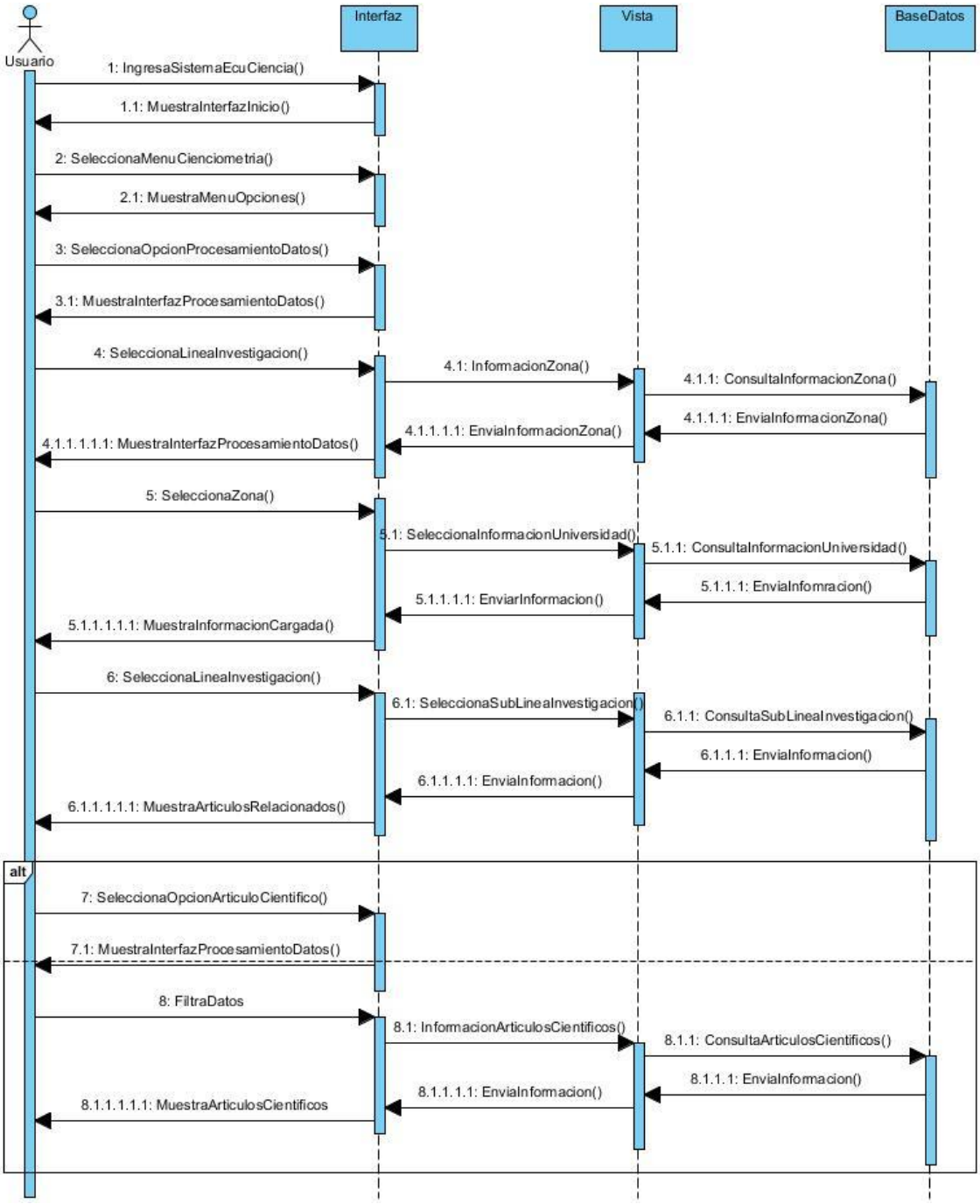
Flujo Normal:

1. **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.**
2. **El usuario selecciona en la opción ver análisis.**

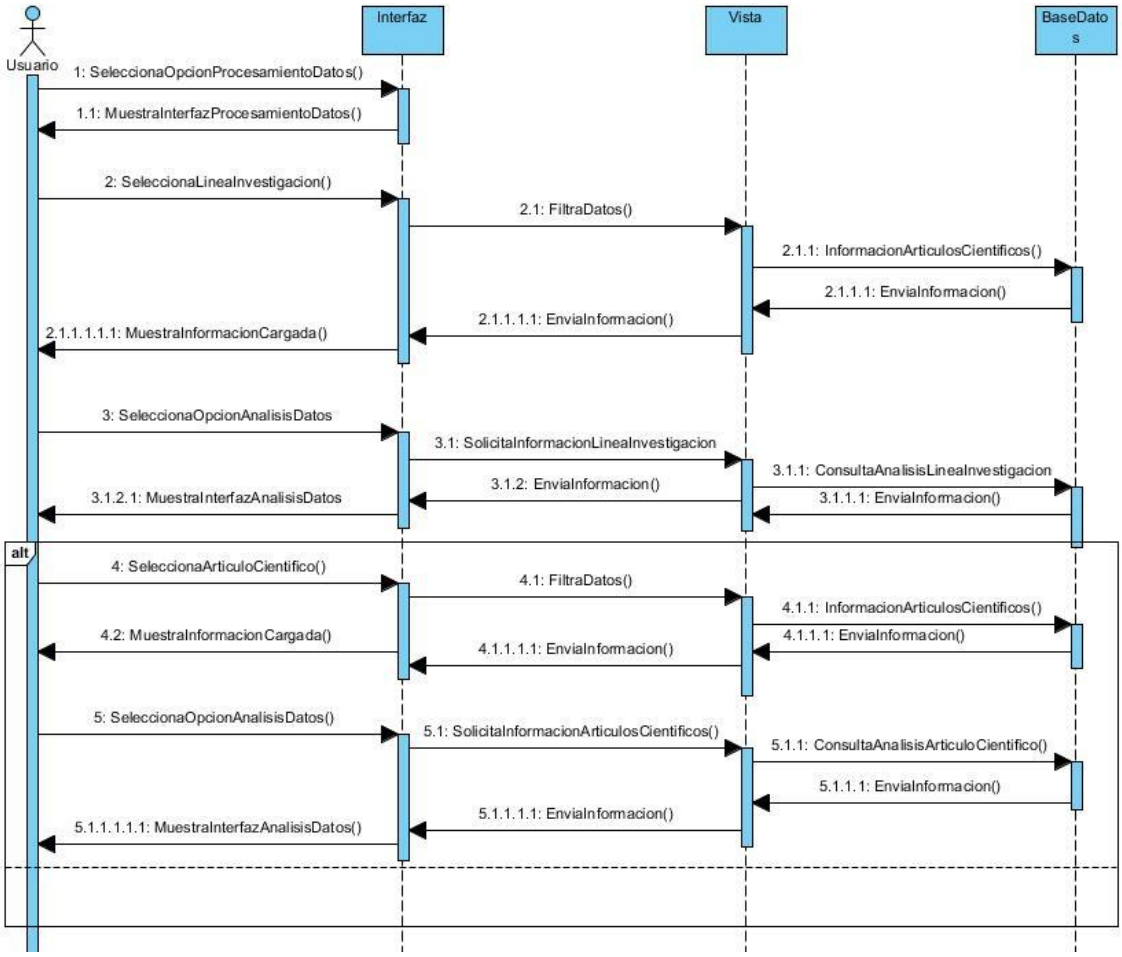
3. **La plataforma muestra el detalle del artículo en un modal con las opciones descargar articulo o descargar corpus.**
4. **El usuario da click y se descarga.**

DIAGRAMAS DE SECUENCIA

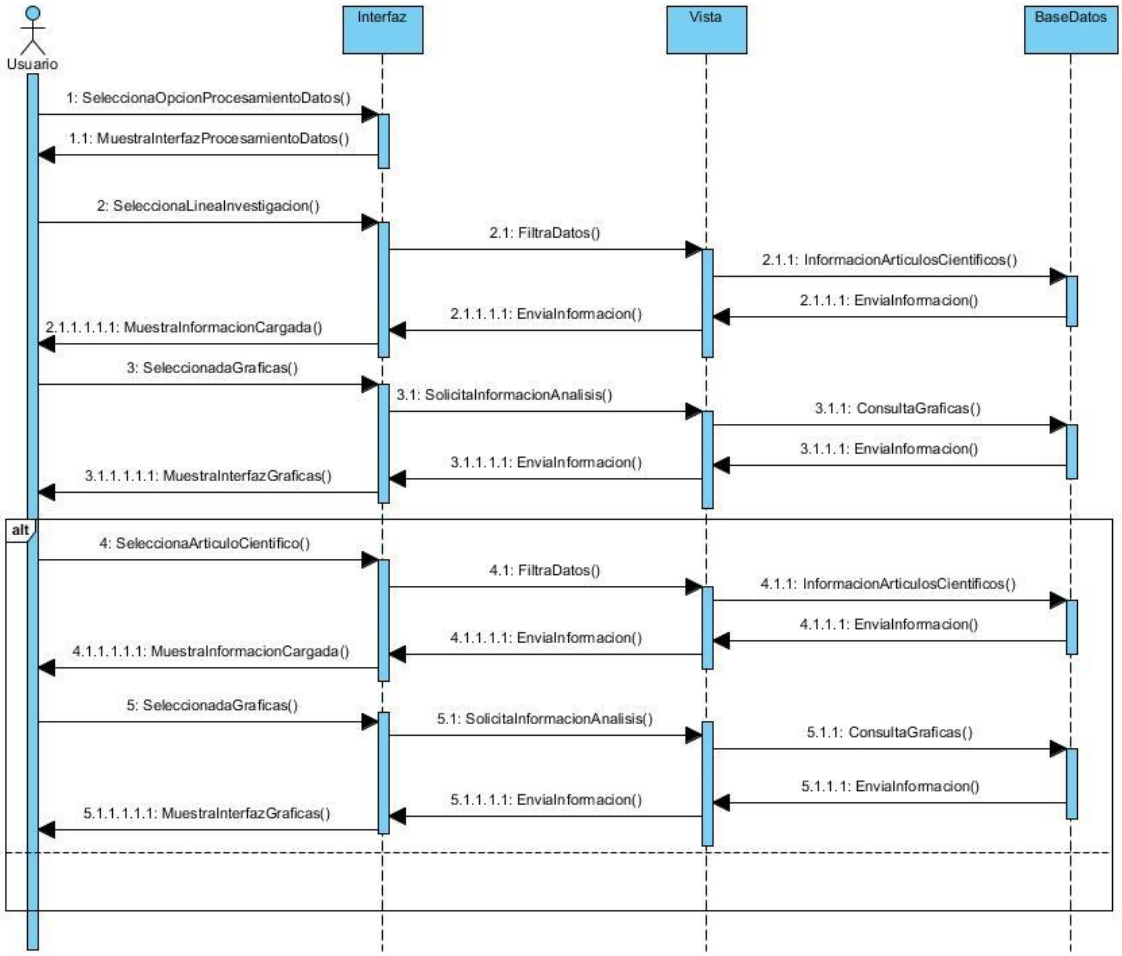
Filtrado de datos


Usuario
Interfaz
Vista
BaseDatos
1: IngresaSistemaEcuCiencia()
1.1: MuestraInterfazInicio()
2: SeleccionaMenuCienciometria()
2.1: MuestraMenuOpciones()
3: SeleccionaOpcionProcesamientoDatos()
3.1: MuestraInterfazProcesamientoDatos()
4: SeleccionaLineaInvestigacion()
4.1: InformacionZona()
4.1.1: ConsultaInformacionZona()
4.1.1.1: EnviaInformacionZona()
4.1.1.1.1: EnviaInformacionZona()
4.1.1.1.1.1: MuestraInterfazProcesamientoDatos()
5: SeleccionaZona()
5.1: SeleccionaInformacionUniversidad()
5.1.1: ConsultaInformacionUniversidad()
5.1.1.1: EnviaInfomracion()
5.1.1.1.1: EnviarInformacion()
5.1.1.1.1.1: MuestraInformacionCargada()
6: SeleccionaLineaInvestigacion()
6.1: SeleccionaSubLineaInvestigacion()
6.1.1: ConsultaSubLineaInvestigacion()
6.1.1.1: EnviaInformacion()
6.1.1.1.1: EnviaInformacion()
6.1.1.1.1.1: MuestraArticulosRelacionados()
alt
7: SeleccionaOpcionArticuloCientifico()
7.1: MuestraInterfazProcesamientoDatos()
8: FiltraDatos
8.1: InformacionArticulosCientificos()
8.1.1: ConsultaArticulosCientificos()
8.1.1.1: EnviaInformacion()
8.1.1.1.1: EnviaInformacion()
8.1.1.1.1.1: MuestraArticulosCientificos

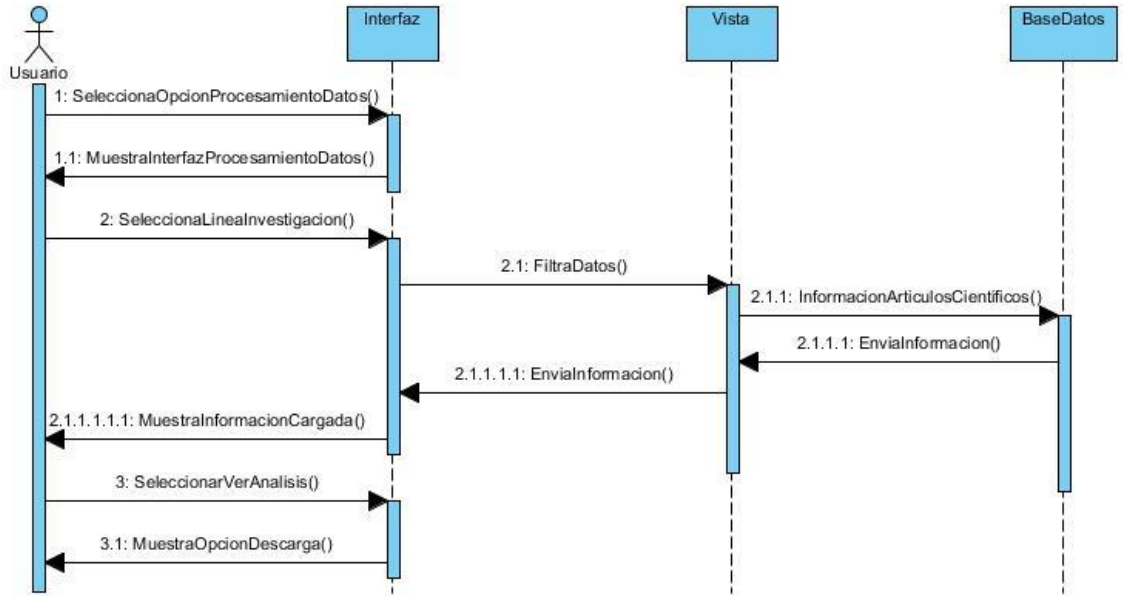
Obtener corpus



Distancia:


Usuario
Interfaz
Vista
BaseDatos
1: SeleccionaOpcionProcesamientoDatos()
1.1: MuestraInterfazProcesamientoDatos()
2: SeleccionaLineaInvestigacion()
2.1: FiltraDatos()
2.1.1: InformacionArticulosCientificos()
2.1.1.1: EnviaInformacion()
2.1.1.1.1: EnviaInformacion()
2.1.1.1.1.1: MuestraInformacionCargada()
3: SeleccionadaGraficas()
3.1: SolicitaInformacionAnalisis()
3.1.1: ConsultaGraficas()
3.1.1.1: EnviaInformacion()
3.1.1.1.1: EnviaInformacion()
3.1.1.1.1.1: MuestraInterfazGraficas()
alt
4: SeleccionaArticuloCientifico()
4.1: FiltraDatos()
4.1.1: InformacionArticulosCientificos()
4.1.1.1: EnviaInformacion()
4.1.1.1.1: EnviaInformacion()
4.1.1.1.1.1: MuestraInformacionCargada()
5: SeleccionadaGraficas()
5.1: SolicitaInformacionAnalisis()
5.1.1: ConsultaGraficas()
5.1.1.1: EnviaInformacion()
5.1.1.1.1: EnviaInformacion()
5.1.1.1.1.1: MuestraInterfazGraficas()

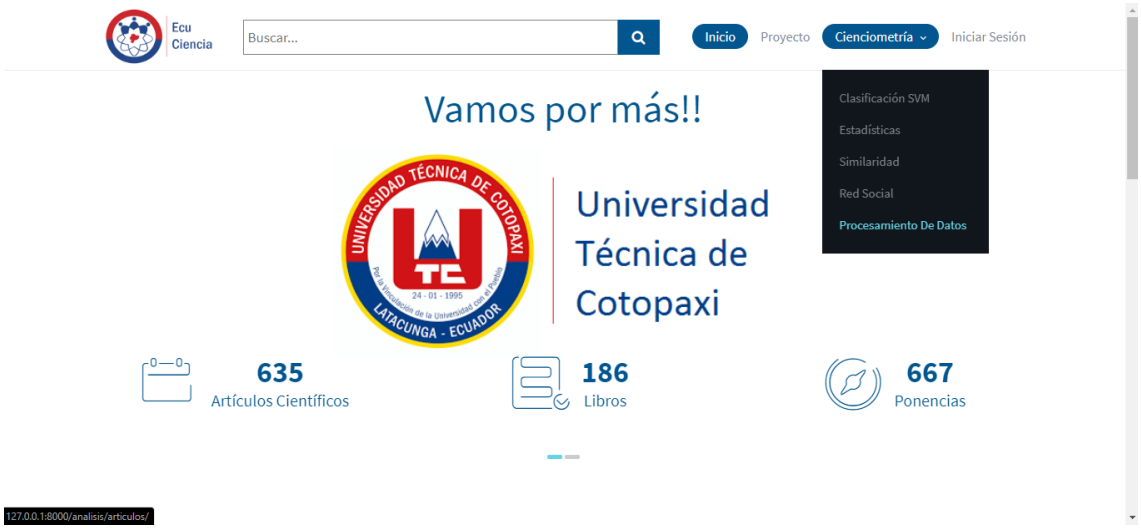
Descargar:


Usuario
Interfaz
Vista
BaseDatos
1: SeleccionaOpcionProcesamientoDatos()
1.1: MuestraInterfazProcesamientoDatos()
2: SeleccionaLineaInvestigacion()
2.1: FiltraDatos()
2.1.1: InformacionArticulosCientificos()
2.1.1.1: EnviaInformacion()
2.1.1.1.1: EnviaInformacion()
2.1.1.1.1.1: MuestraInformacionCargada()
3: SeleccionarVerAnalisis()
3.1: MuestraOpcionDescarga()

IMPLEMENTACIÓN DE LOS SPRINTS

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.

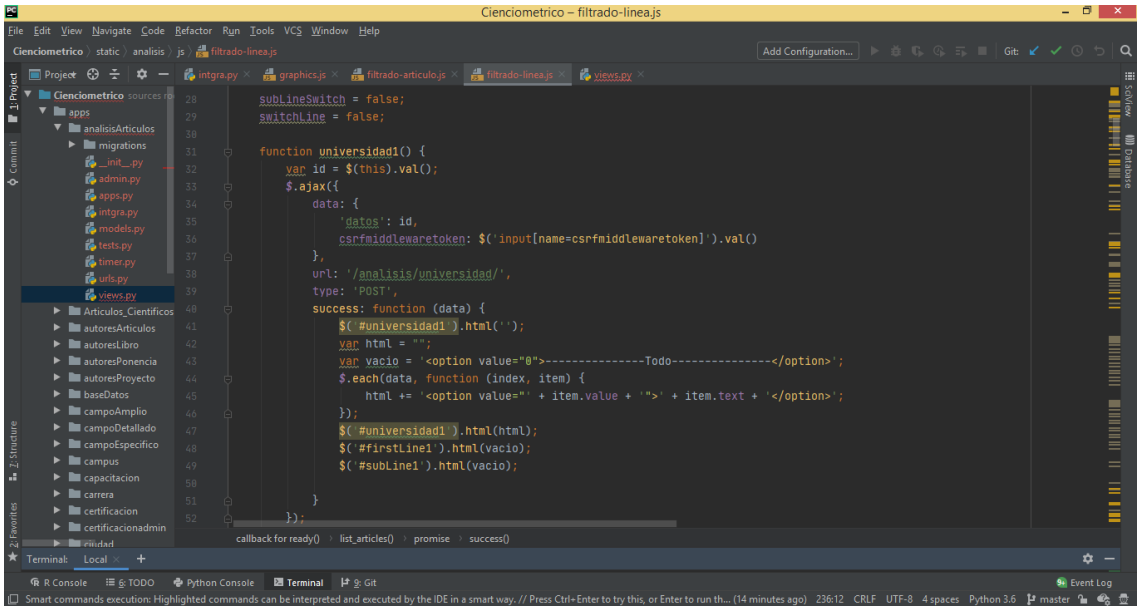
PANTALLA PRINCIPAL



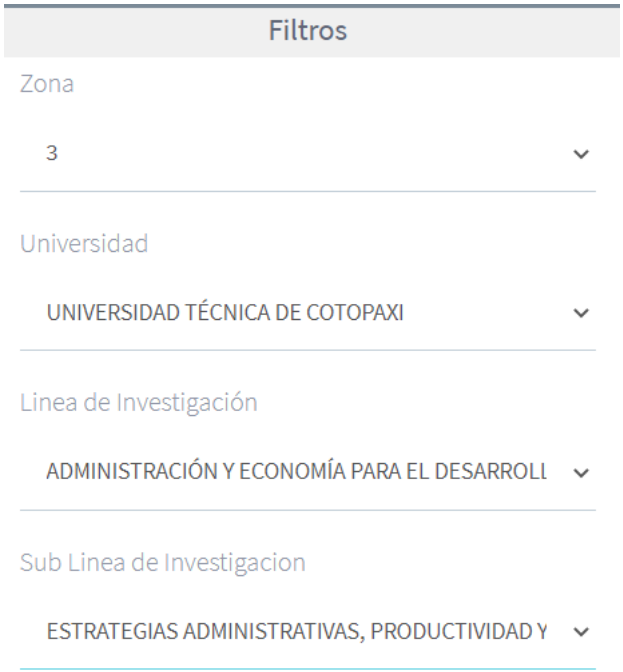
PANTALLA DEL MODULO PROCESAMIENTO DE DATOS

FILTRADO DE DATOS

Desarrollo del código del Sprint 1



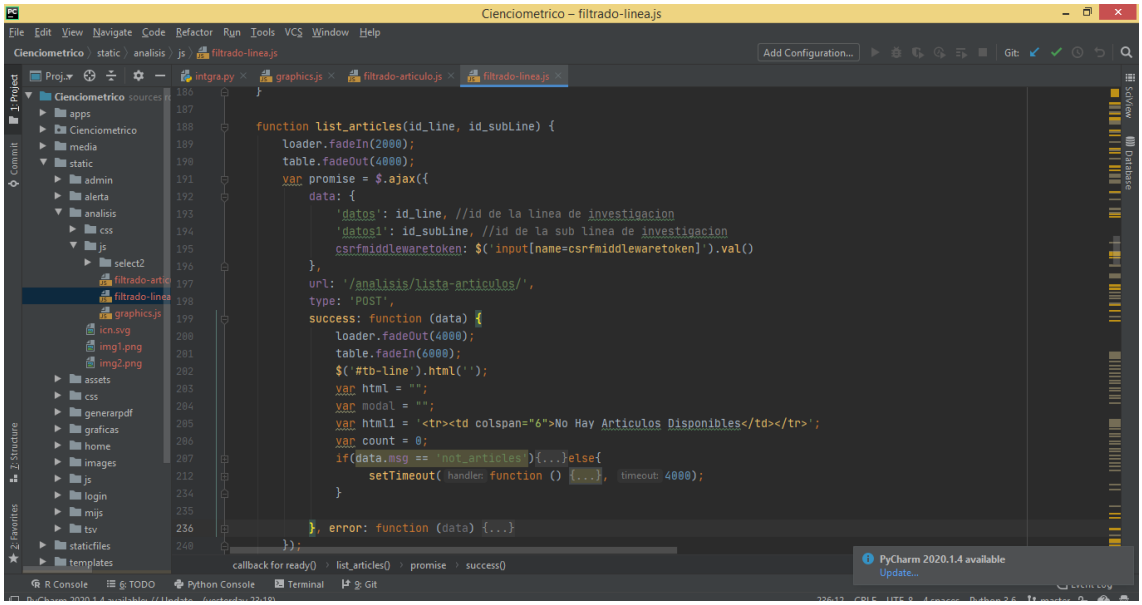
INTERFAZ GRAFICA DEL FILTRADO DE DATOS



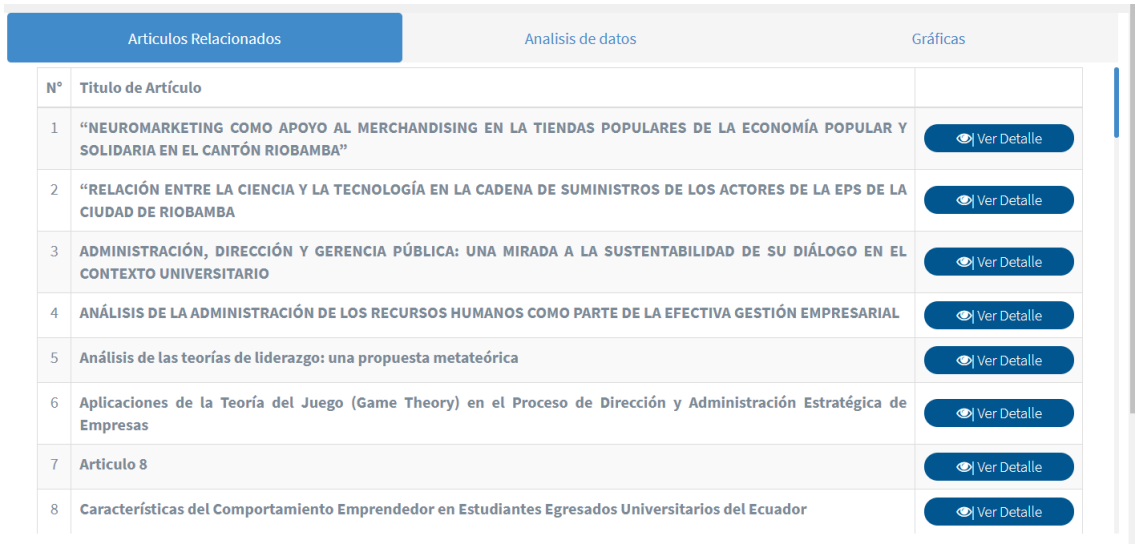
INTERFAZ GRAFICA POR LÍNEA DE INVESTIGACIÓN

INTERFAZ GRAFICA POR ARTICULO CIENTÍFICO

CARGAR ARTÍCULOS EN LA TABLA

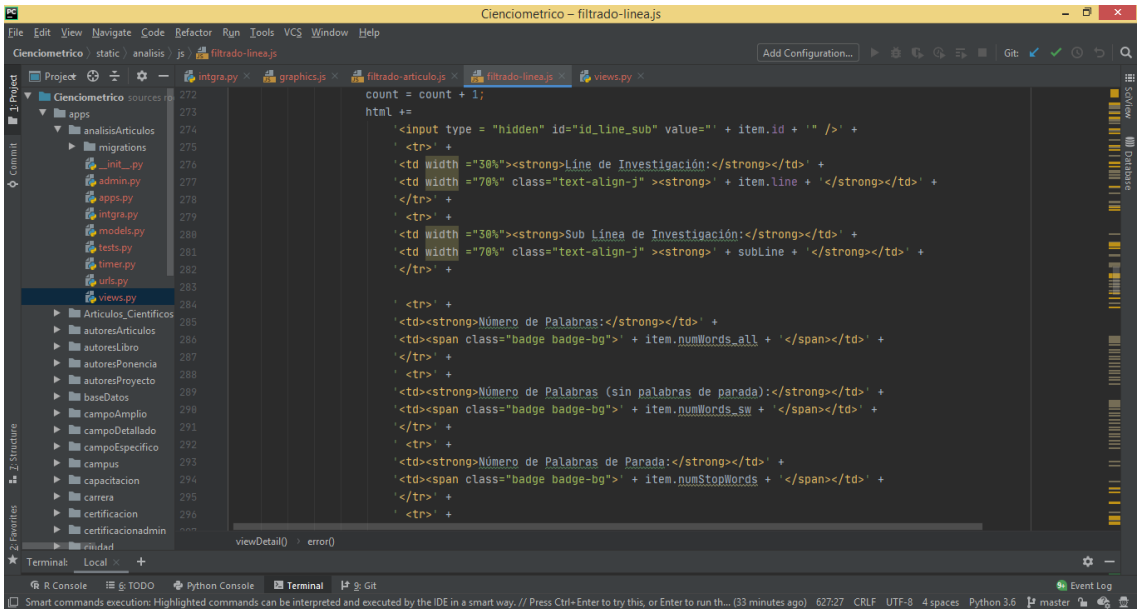


INTERFAZ GRAFICA



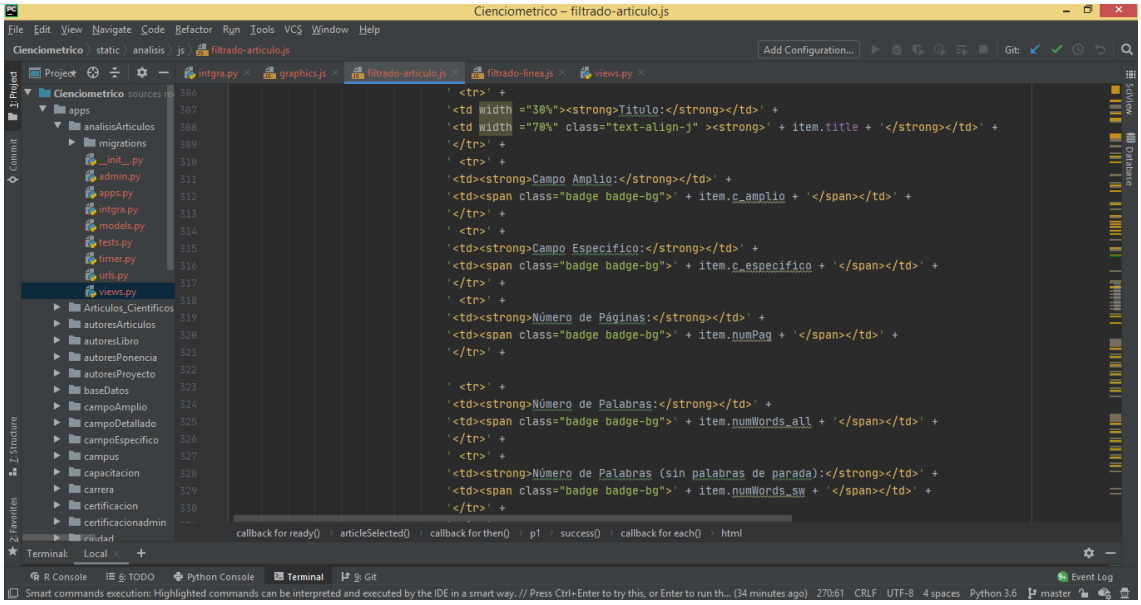
VISUALIZAR ANÁLISIS DEL CORPUS POR LÍNEA DE INVESTIGACIÓN

Desarrollo del código del Sprint 2



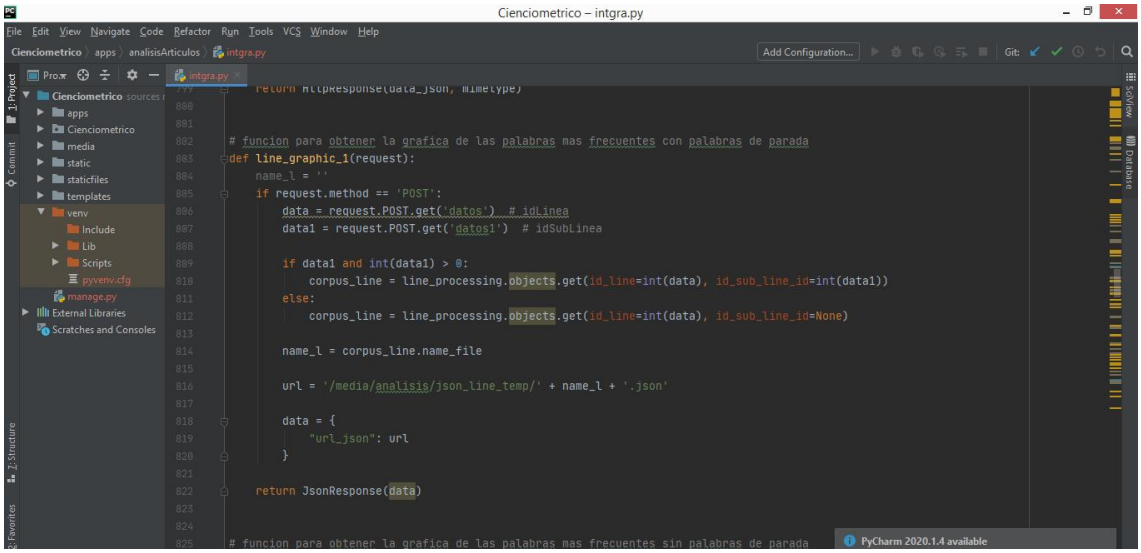
INTERFAZ GRAFICA

VER ANÁLISIS DEL CORPUS POR ARTICULO CIENTÍFICO

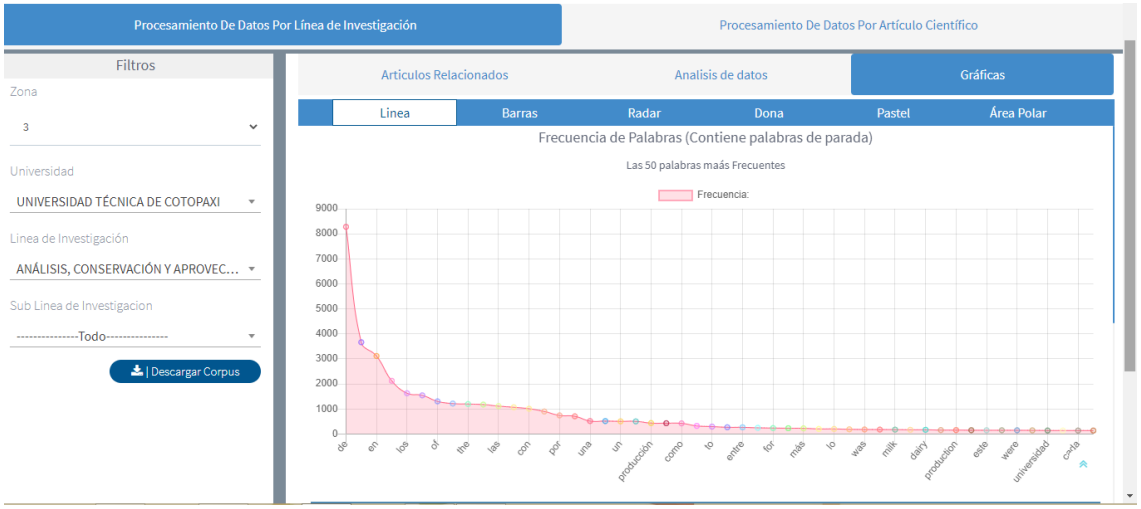


INTERFAZ GRAFICA

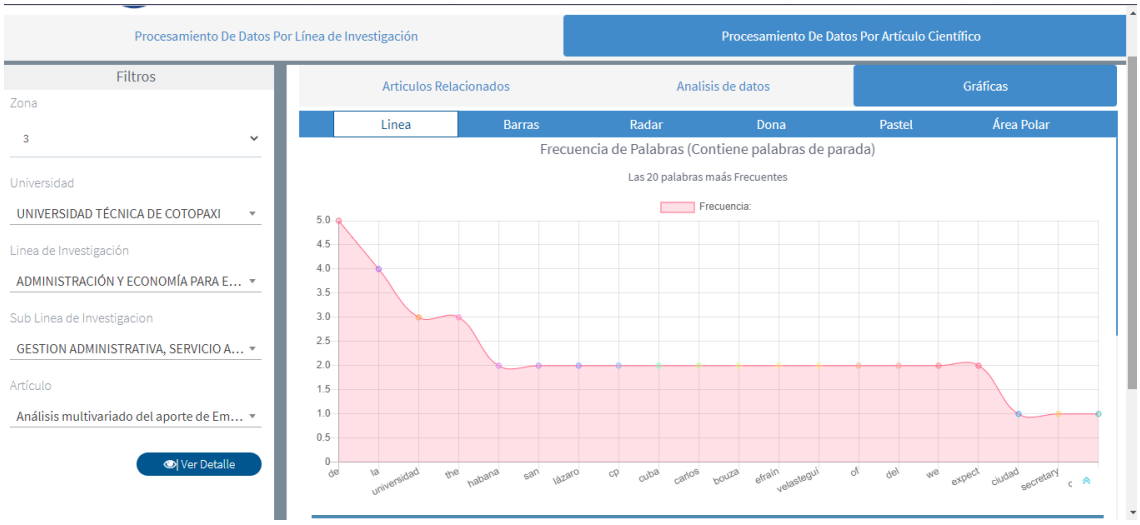
VER GRAFICAS con palabras de parada



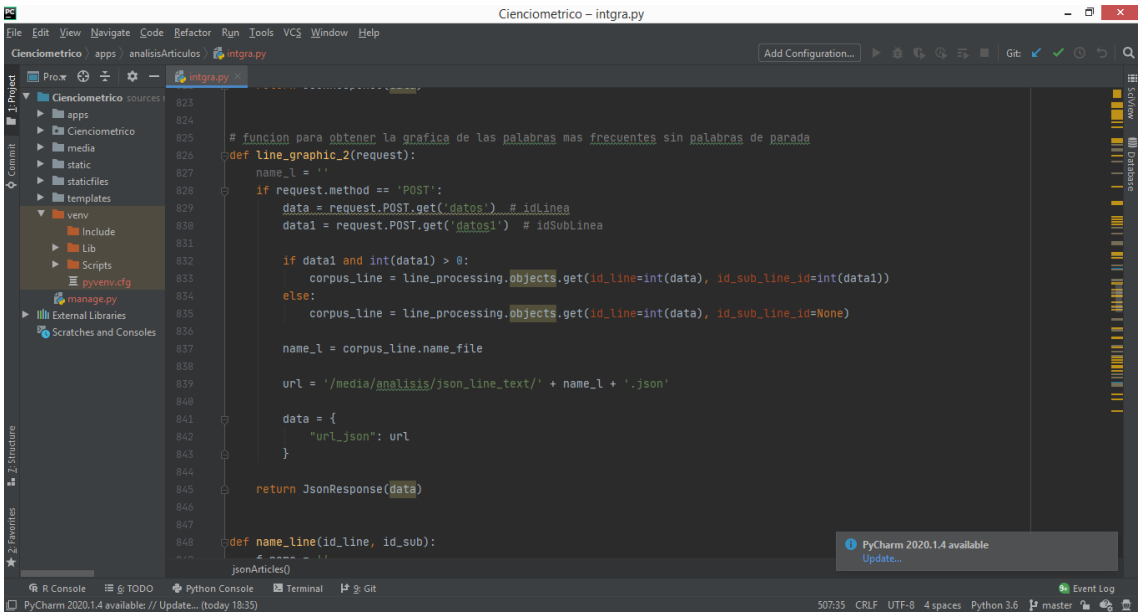
Gráfica línea por línea de investigación



Grafica línea por articulo científico



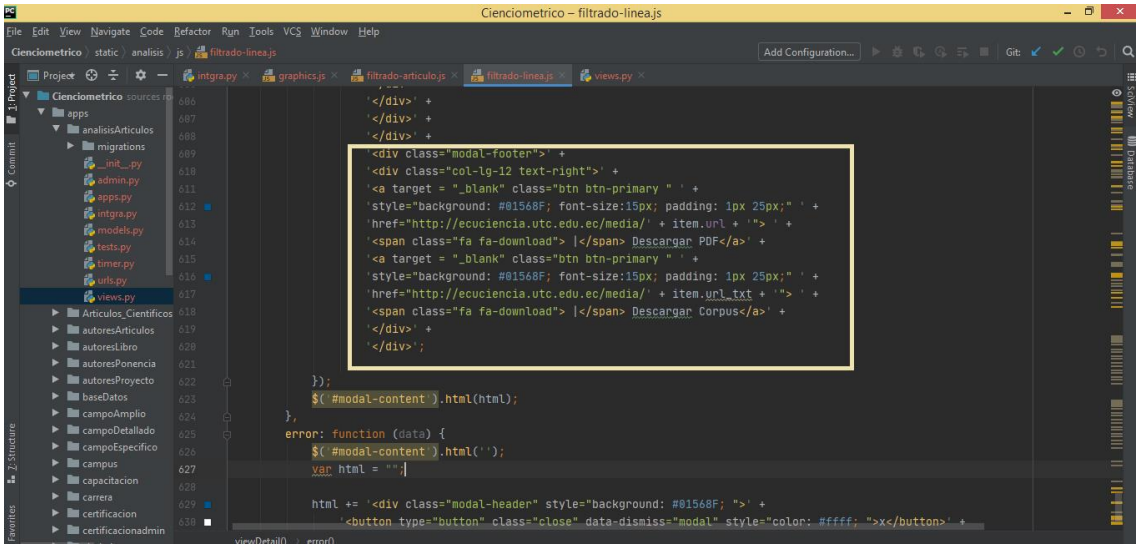
Ver graficas sin palabras de parada



Gráfica línea por línea de investigación

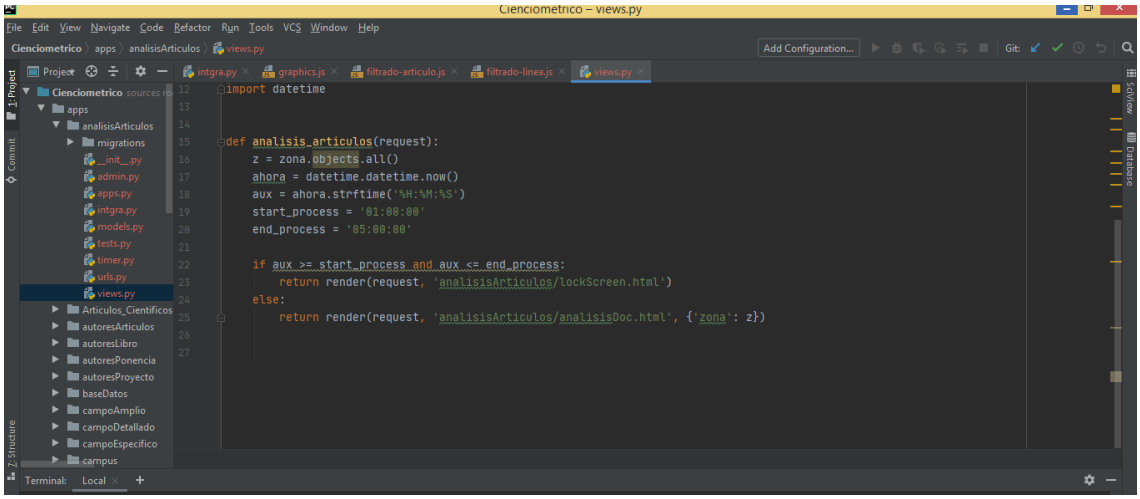
Gráfica línea por articulo científico

Ver detalle del articulo y descarga de corpus



INTERFAZ GRAFICA

Actualizar información (Pantalla de bloqueo)



INTERFAZ GRAFICA

CASOS DE PRUEBA

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

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

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

CP004:	Visualizar Graficas
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

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

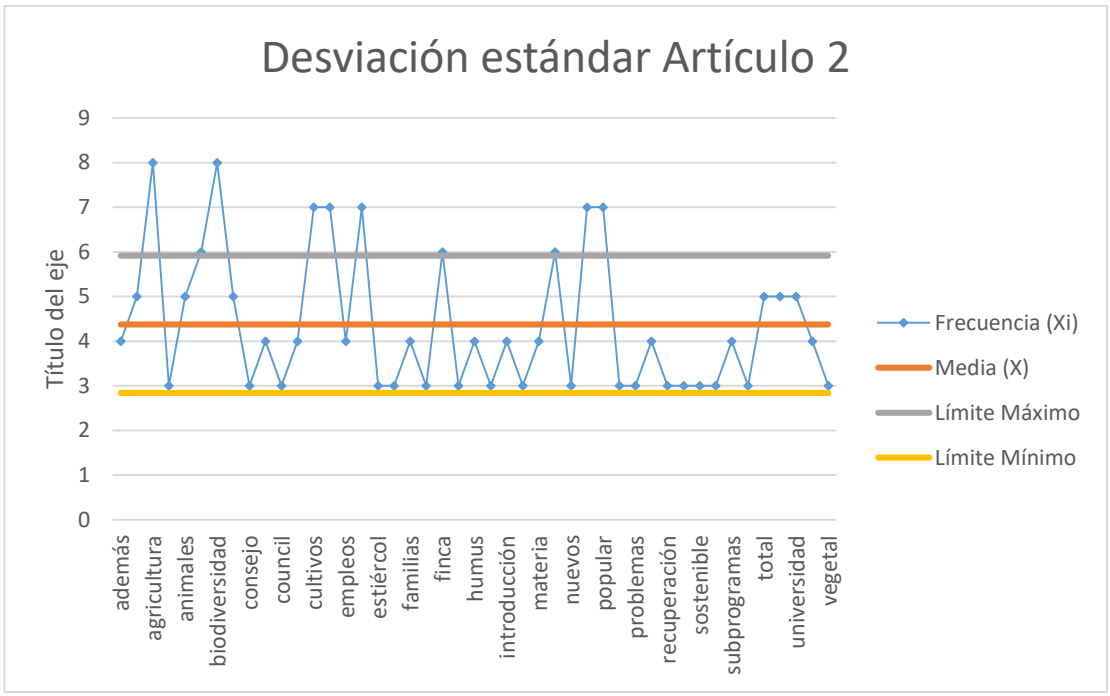
CP006:	Descargar Corpus
H.U:	006
Fecha:	25/07/2020
Responsable	Scrum Team
Descripción	Permite al usuario descargar el corpus.
Precondiciones:	El usuario debe estar en la plataforma EcuCiencia.
Resultado esperado 1:	Obtener la descarga del corpus en un archivo de texto plano por línea de investigación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

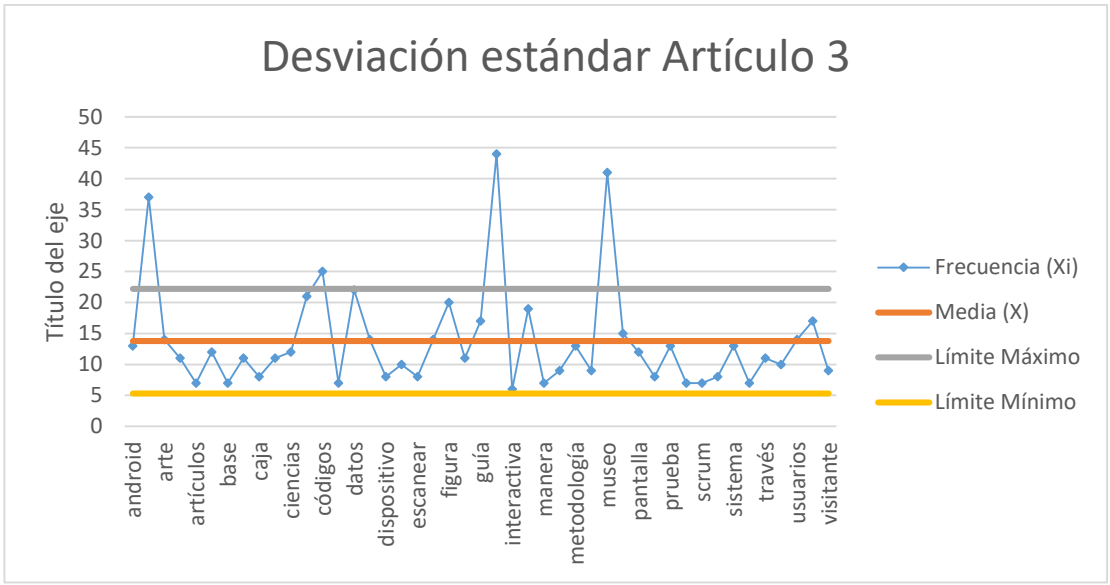
Anexo 4. Tabla que recoge los artículos de muestra con el valor de frecuencia, media, límite máximo y mínimo.

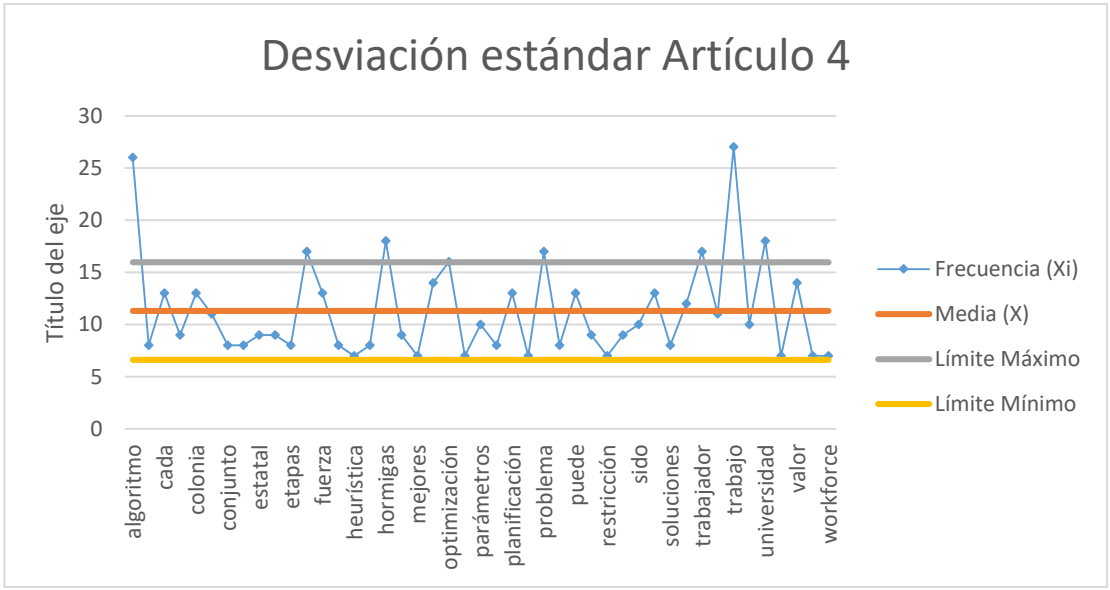
Artículo 1					Artículo 2					Artículo 3					Artículo 4					Artículo 5				
Palabras	Frecuencia (Xi)	Media (X)	Lím. Máx.	Lím. Mín.	Palabras	Frecuencia (Xi)	Media (X)	Lím. Máx.	Lím. Mín.	Palabras	Frecuencia (Xi)	Media (X)	Lím. Máx.	Lím. Mín.	Palabras	Frecuencia (Xi)	Media (X)	Lím. Máx.	Lím. Mín.	Palabras	Frecuencia (Xi)	Media (X)	Lím. Máx.	Lím. Mín.
análisis	12	13,27	20,5	6,1	además	4	4,38	5,92	2,84	android	13	13,8	22,2	5,29	algoritmo	26	11,3	16	6,61	acceso	12	16,1	26,2	6,02
área	10	13,27	20,5	6,1	agrícola	5	4,38	5,92	2,84	aplicación	37	13,8	22,2	5,29	algoritmos	8	11,3	16	6,61	actividades	8	16,1	26,2	6,02
autor	13	13,27	20,5	6,1	agricultura	8	4,38	5,92	2,84	arte	14	13,8	22,2	5,29	cada	13	11,3	16	6,61	acuerdo	10	16,1	26,2	6,02
base	19	13,27	20,5	6,1	agroecológica	3	4,38	5,92	2,84	artículo	11	13,8	22,2	5,29	cantidad	9	11,3	16	6,61	años	9	16,1	26,2	6,02
colaboración	7	13,27	20,5	6,1	animales	5	4,38	5,92	2,84	artículos	7	13,8	22,2	5,29	colonia	13	11,3	16	6,61	aplicaciones	21	16,1	26,2	6,02
conocer	7	13,27	20,5	6,1	área	6	4,38	5,92	2,84	ayora	12	13,8	22,2	5,29	colony	11	11,3	16	6,61	aprendizajes	9	16,1	26,2	6,02
datos	12	13,27	20,5	6,1	biodiversidad	8	4,38	5,92	2,84	base	7	13,8	22,2	5,29	conjunto	8	11,3	16	6,61	base	15	16,1	26,2	6,02
desarrollo	14	13,27	20,5	6,1	centro	5	4,38	5,92	2,84	cada	11	13,8	22,2	5,29	costo	8	11,3	16	6,61	caracterizació	11	16,1	26,2	6,02
determinar	7	13,27	20,5	6,1	consejo	3	4,38	5,92	2,84	caja	8	13,8	22,2	5,29	estatal	9	11,3	16	6,61	caso	21	16,1	26,2	6,02
diseño	9	13,27	20,5	6,1	consejos	4	4,38	5,92	2,84	capa	11	13,8	22,2	5,29	etapa	9	11,3	16	6,61	componentes	19	16,1	26,2	6,02
encuesta	10	13,27	20,5	6,1	council	3	4,38	5,92	2,84	ciencias	12	13,8	22,2	5,29	etapas	8	11,3	16	6,61	desarrollo	13	16,1	26,2	6,02
éxito	19	13,27	20,5	6,1	cuales	4	4,38	5,92	2,84	código	21	13,8	22,2	5,29	exploración	17	11,3	16	6,61	diferentes	11	16,1	26,2	6,02
factores	24	13,27	20,5	6,1	cultivos	7	4,38	5,92	2,84	códigos	25	13,8	22,2	5,29	fuerza	13	11,3	16	6,61	docentes	10	16,1	26,2	6,02
grupos	7	13,27	20,5	6,1	desarrollo	7	4,38	5,92	2,84	comunicació	7	13,8	22,2	5,29	función	8	11,3	16	6,61	ecuador	10	16,1	26,2	6,02
importancia	19	13,27	20,5	6,1	empleos	4	4,38	5,92	2,84	datos	22	13,8	22,2	5,29	heurística	7	11,3	16	6,61	ecuatoriana	27	16,1	26,2	6,02
importante	11	13,27	20,5	6,1	especies	7	4,38	5,92	2,84	desarrollo	14	13,8	22,2	5,29	horas	8	11,3	16	6,61	educación	55	16,1	26,2	6,02
información	8	13,27	20,5	6,1	estiércol	3	4,38	5,92	2,84	dispositivo	8	13,8	22,2	5,29	hormigas	18	11,3	16	6,61	empleo	9	16,1	26,2	6,02
investigación	25	13,27	20,5	6,1	extensionismo	3	4,38	5,92	2,84	ecuador	10	13,8	22,2	5,29	matriz	9	11,3	16	6,61	enseñanza	8	16,1	26,2	6,02
investigadores	31	13,27	20,5	6,1	familias	4	4,38	5,92	2,84	escanear	8	13,8	22,2	5,29	mejores	7	11,3	16	6,61	entornos	9	16,1	26,2	6,02
mediante	10	13,27	20,5	6,1	financiamient	3	4,38	5,92	2,84	escuela	14	13,8	22,2	5,29	número	14	11,3	16	6,61	estudiantes	26	16,1	26,2	6,02
neuronal	8	13,27	20,5	6,1	finca	6	4,38	5,92	2,84	figura	20	13,8	22,2	5,29	optimizació	16	11,3	16	6,61	evolución	9	16,1	26,2	6,02
parte	14	13,27	20,5	6,1	fuerza	3	4,38	5,92	2,84	gestionar	11	13,8	22,2	5,29	optimizatior	7	11,3	16	6,61	forma	8	16,1	26,2	6,02
predicción	7	13,27	20,5	6,1	humus	4	4,38	5,92	2,84	guía	17	13,8	22,2	5,29	parámetros	10	11,3	16	6,61	gráfico	12	16,1	26,2	6,02
proceso	8	13,27	20,5	6,1	incrementar	3	4,38	5,92	2,84	información	44	13,8	22,2	5,29	período	8	11,3	16	6,61	herramientas	11	16,1	26,2	6,02
procesos	17	13,27	20,5	6,1	introducción	4	4,38	5,92	2,84	interactiva	6	13,8	22,2	5,29	planificació	13	11,3	16	6,61	información	29	16,1	26,2	6,02
publicaciones	26	13,27	20,5	6,1	manejo	3	4,38	5,92	2,84	isidro	19	13,8	22,2	5,29	planning	7	11,3	16	6,61	informáticas	20	16,1	26,2	6,02
puede	10	13,27	20,5	6,1	materia	4	4,38	5,92	2,84	manera	7	13,8	22,2	5,29	problema	17	11,3	16	6,61	internet	8	16,1	26,2	6,02
realizar	11	13,27	20,5	6,1	mujeres	6	4,38	5,92	2,84	mediante	9	13,8	22,2	5,29	problemas	8	11,3	16	6,61	investigación	15	16,1	26,2	6,02
recomendación	13	13,27	20,5	6,1	nuevos	3	4,38	5,92	2,84	metodología	13	13,8	22,2	5,29	puede	13	11,3	16	6,61	issn	19	16,1	26,2	6,02

investigadores	31	13,27	20,5	6,1	familias	4	4,38	5,92	2,84	escanear	8	13,8	22,2	5,29	mejores	7	11,3	16	6,61	entornos	9	16,1	26,2	6,02
mediante	10	13,27	20,5	6,1	financiamient	3	4,38	5,92	2,84	escuela	14	13,8	22,2	5,29	número	14	11,3	16	6,61	estudiantes	26	16,1	26,2	6,02
neuronal	8	13,27	20,5	6,1	finca	6	4,38	5,92	2,84	figura	20	13,8	22,2	5,29	optimizació	16	11,3	16	6,61	evolución	9	16,1	26,2	6,02
parte	14	13,27	20,5	6,1	fuerza	3	4,38	5,92	2,84	gestionar	11	13,8	22,2	5,29	optimizatior	7	11,3	16	6,61	forma	8	16,1	26,2	6,02
predicción	7	13,27	20,5	6,1	humus	4	4,38	5,92	2,84	guía	17	13,8	22,2	5,29	parámetros	10	11,3	16	6,61	gráfico	12	16,1	26,2	6,02
proceso	8	13,27	20,5	6,1	incrementar	3	4,38	5,92	2,84	información	44	13,8	22,2	5,29	período	8	11,3	16	6,61	herramientas	11	16,1	26,2	6,02
procesos	17	13,27	20,5	6,1	introducción	4	4,38	5,92	2,84	interactiva	6	13,8	22,2	5,29	planificació	13	11,3	16	6,61	información	29	16,1	26,2	6,02
publicaciones	26	13,27	20,5	6,1	manejo	3	4,38	5,92	2,84	isidro	19	13,8	22,2	5,29	planning	7	11,3	16	6,61	informáticas	20	16,1	26,2	6,02
puede	10	13,27	20,5	6,1	materia	4	4,38	5,92	2,84	manera	7	13,8	22,2	5,29	problema	17	11,3	16	6,61	internet	8	16,1	26,2	6,02
realizar	11	13,27	20,5	6,1	mujeres	6	4,38	5,92	2,84	mediante	9	13,8	22,2	5,29	problemas	8	11,3	16	6,61	investigación	15	16,1	26,2	6,02
recomendación	13	13,27	20,5	6,1	nuevos	3	4,38	5,92	2,84	metodología	13	13,8	22,2	5,29	puede	13	11,3	16	6,61	issn	19	16,1	26,2	6,02
recomendador	20	13,27	20,5	6,1	plantas	7	4,38	5,92	2,84	móvil	9	13,8	22,2	5,29	realizar	9	11,3	16	6,61	mismos	10	16,1	26,2	6,02
recomendadore	16	13,27	20,5	6,1	popular	7	4,38	5,92	2,84	museo	41	13,8	22,2	5,29	restricción	7	11,3	16	6,61	nuevos	15	16,1	26,2	6,02
relevante	10	13,27	20,5	6,1	populares	3	4,38	5,92	2,84	museos	15	13,8	22,2	5,29	resultados	9	11,3	16	6,61	obtenido	9	16,1	26,2	6,02
requerimientos	9	13,27	20,5	6,1	problemas	3	4,38	5,92	2,84	pantalla	12	13,8	22,2	5,29	sido	10	11,3	16	6,61	prácticas	10	16,1	26,2	6,02
resultados	10	13,27	20,5	6,1	producción	4	4,38	5,92	2,84	proceso	8	13,8	22,2	5,29	solución	13	11,3	16	6,61	punto	10	16,1	26,2	6,02
retrieved	8	13,27	20,5	6,1	recuperación	3	4,38	5,92	2,84	prueba	13	13,8	22,2	5,29	soluciones	8	11,3	16	6,61	realidad	26	16,1	26,2	6,02
revista	11	13,27	20,5	6,1	recursos	3	4,38	5,92	2,84	pruebas	7	13,8	22,2	5,29	técnica	12	11,3	16	6,61	revista	19	16,1	26,2	6,02
sistema	35	13,27	20,5	6,1	sostenible	3	4,38	5,92	2,84	scrum	7	13,8	22,2	5,29	trabajador	17	11,3	16	6,61	sociedad	50	16,1	26,2	6,02
sistemas	32	13,27	20,5	6,1	species	3	4,38	5,92	2,84	servidor	8	13,8	22,2	5,29	trabajadores	11	11,3	16	6,61	superior	9	16,1	26,2	6,02
software	7	13,27	20,5	6,1	subprogramas	4	4,38	5,92	2,84	sistema	13	13,8	22,2	5,29	trabajo	27	11,3	16	6,61	tecnología	13	16,1	26,2	6,02
sugerencias	8	13,27	20,5	6,1	supported	3	4,38	5,92	2,84	software	7	13,8	22,2	5,29	trabajos	10	11,3	16	6,61	tecnologías	13	16,1	26,2	6,02
tabla	11	13,27	20,5	6,1	total	5	4,38	5,92	2,84	través	11	13,8	22,2	5,29	universidad	18	11,3	16	6,61	tecnológico	9	16,1	26,2	6,02
tecnologias	7	13,27	20,5	6,1	trabajo	5	4,38	5,92	2,84	usuario	10	13,8	22,2	5,29	utiliza	7	11,3	16	6,61	universidad	29	16,1	26,2	6,02
través	9	13,27	20,5	6,1	universidad	5	4,38	5,92	2,84	usuarios	14	13,8	22,2	5,29	valor	14	11,3	16	6,61	universidades	11	16,1	26,2	6,02
usuario	7	13,27	20,5	6,1	variedades	4	4,38	5,92	2,84	virtual	17	13,8	22,2	5,29	variante	7	11,3	16	6,61	utilización	25	16,1	26,2	6,02
variables	9	13,27	20,5	6,1	vegetal	3	4,38	5,92	2,84	visitante	9	13,8	22,2	5,29	workforce	7	11,3	16	6,61	utopía	24	16,1	26,2	6,02

Anexo 5. Gráficos de la desviación estándar

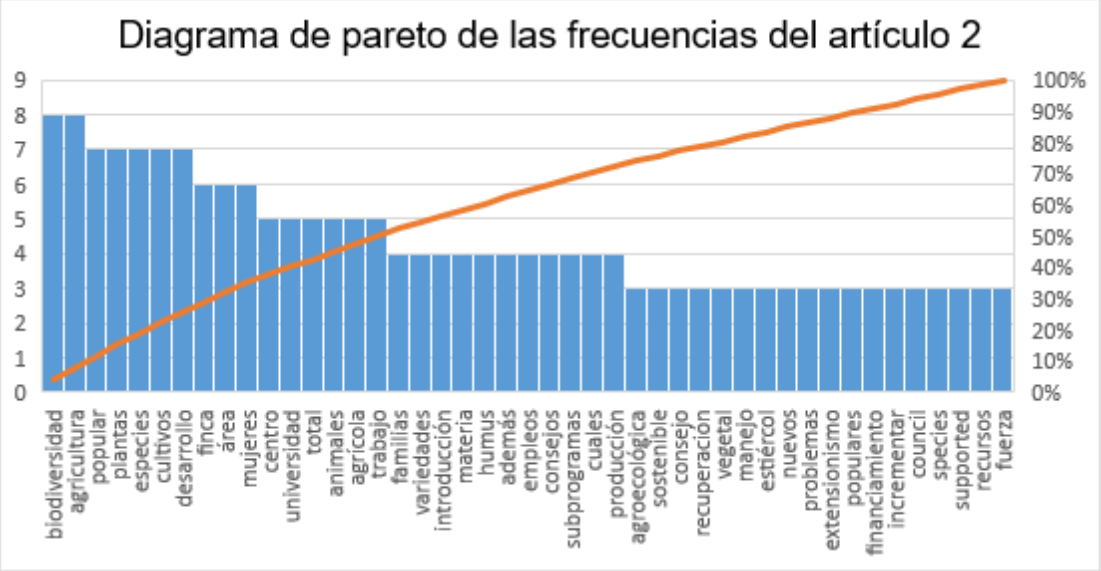


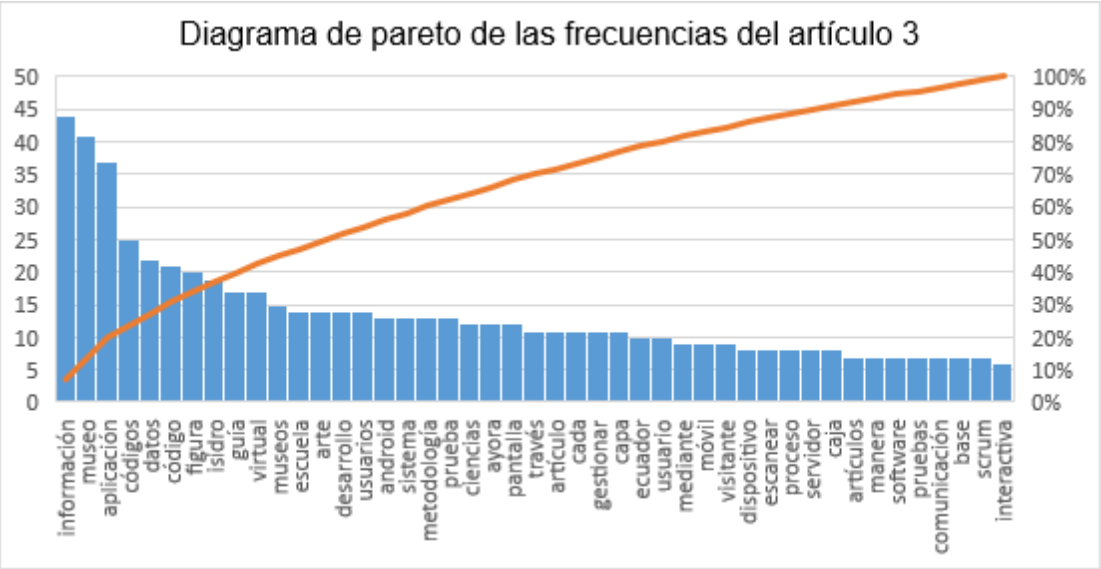



Desviación estándar Artículo 4
Título del eje
30
25
20
15
10
5
0
algoritmo
cada
colonia
conjunto
estatal
etapas
fuerza
heurística
hormigas
mejores
optimización
parámetros
planificación
problema
puede
restricción
sido
soluciones
trabajador
trabajo
universidad
valor
workforce
Frecuencia (Xi)
Media (X)
Límite Máximo
Límite Mínimo


Desviación estándar Artículo 5
Título del eje
60
50
40
30
20
10
0
acceso
acuerdo
aplicaciones
base
caso
desarrollo
docentes
ecuatoriana
empleo
entornos
evolución
gráfico
información
internet
issn
nuevos
prácticas
realidad
sociedad
tecnología
tecnológico
universidades
utopía
Frecuencia (Xi)
Media (X)
Límite Máximo
Límite Mínimo

Anexo 6. Gráficos de pareto de los artículos, representando el 20% de mayor peso en importancia y el 80 % del resto de las palabras en análisis





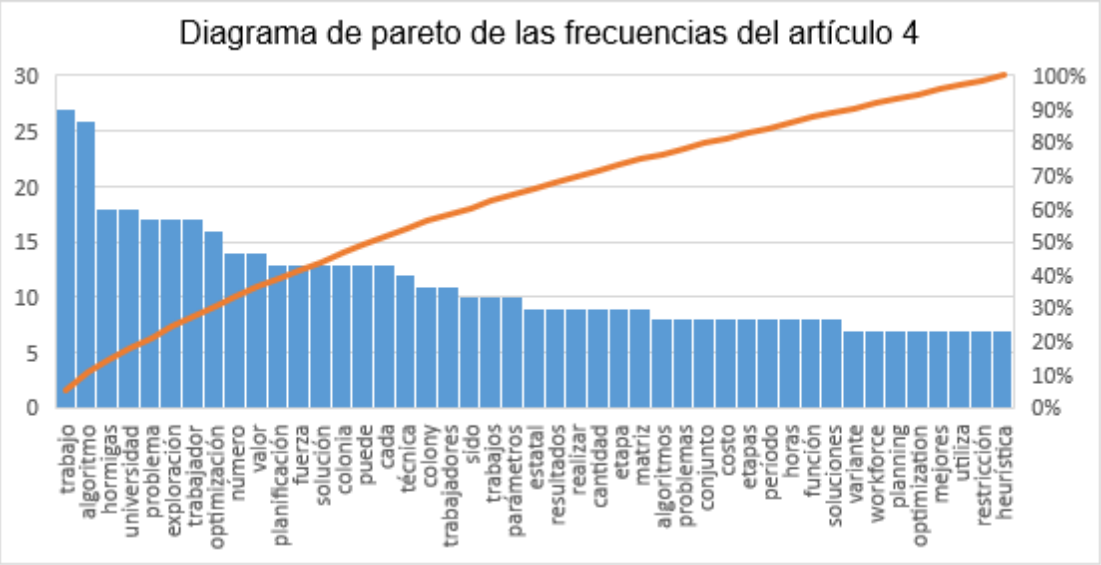


Diagrama de pareto de las frecuencias del artículo 4
30
25
20
15
10
5
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
trabajo
algoritmo
hormigas
universidad
problema
exploración
trabajador
optimización
número
valor
planificación
fuerza
solución
colonia
puede
cada
técnica
colony
trabajadores
sido
trabajos
parámetros
estatal
resultados
realizar
cantidad
etapa
matriz
algoritmos
problemas
conjunto
costo
etapas
periodo
horas
función
soluciones
variante
workforce
planning
optimization
mejores
utiliza
restricción
heurística

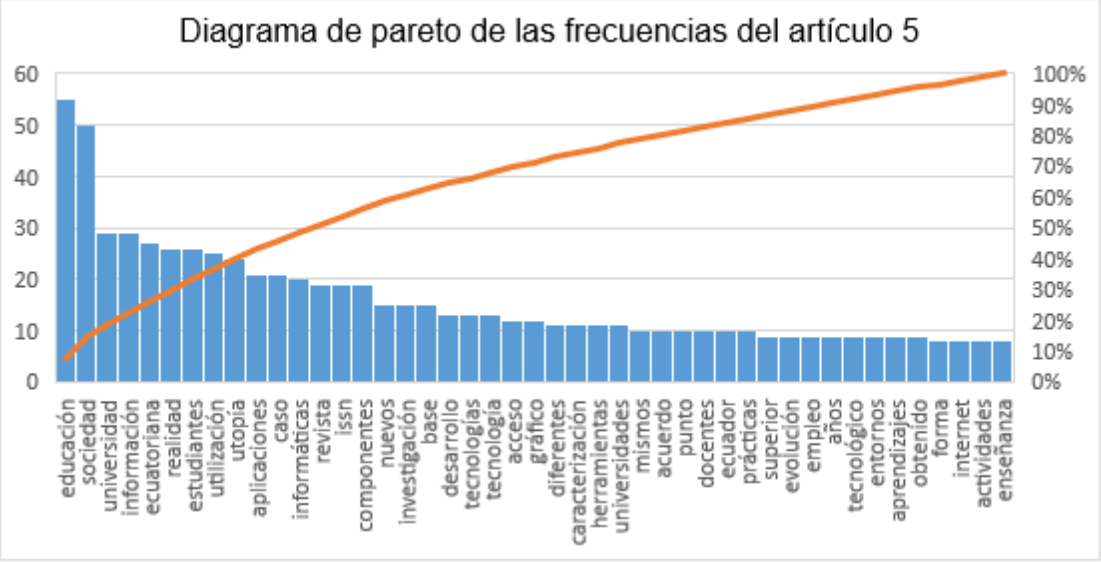


Diagrama de pareto de las frecuencias del artículo 5
60
50
40
30
20
10
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
educación
sociedad
universidad
información
ecuatoriana
realidad
estudiantes
utilización
utopía
aplicaciones
caso
informáticas
revista
issn
componentes
nuevos
investigación
base
desarrollo
tecnologías
tecnología
acceso
gráfico
diferentes
caracterización
herramientas
universidades
mismos
acuerdo
punto
docentes
ecuador
prácticas
superior
evolución
empleo
años
tecnológico
entornos
aprendizajes
obtenido
forma
internet
actividades
enseñanza