

<CÓDIGO>

Análisis de las redes sociales académicas de la Pontificia Universidad Javeriana

Frank Sebastián Franco Hernández

Miguel Ángel Gutiérrez Ibagué

Diana Marcela Herrán Giraldo

Luis Javier López González

PONTIFICIA UNIVERSIDAD JAVERIANA

FACULTAD DE INGENIERIA

MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

BOGOTÁ, D.C.

2023

<CÓDIGO>

Análisis de las redes sociales académicas de la Pontificia Universidad Ja-
veriana

Autor:

Frank Sebastián Franco Hernández

Miguel Ángel Gutiérrez Ibagué

Diana Marcela Herrán Giraldo

Luis Javier López González

MEMORIA DEL TRABAJO DE GRADO REALIZADO PARA CUMPLIR UNO
DE LOS REQUISITOS PARA OPTAR AL TÍTULO DE
MAGÍSTER EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Directora

Alexandra Pomares Quimbaya

Comité de Evaluación del Trabajo de Grado

<Nombres y Apellidos Completos del Jurado >

<Nombres y Apellidos Completos del Jurado >

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERIA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
Mayo, 2023

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Rector Magnífico

Luis Fernando Múnera Congote, S J.

Decano Facultad de Ingeniería

Ingeniero Lope Hugo Barrero Solano

Director Maestría en Ingeniería de Sistemas y Computación

Ingeniera Angela Carrillo Ramos

Director Departamento de Ingeniería de Sistemas

Ing. César Julio Bustacara Medina, Ph.D.

Artículo 23 de la Resolución No. 1 de Junio de 1946

“La Universidad no se hace responsable de los conceptos emitidos por sus alumnos en sus proyectos de grado. Sólo velará porque no se publique nada contrario al dogma y la moral católica y porque no contengan ataques o polémicas puramente personales. Antes bien, que se vean en ellos el anhelo de buscar la verdad y la Justicia”

AGRADECIMIENTOS

Primero, queremos agradecerles a quienes son los motores de nuestras vidas, padres, parejas e hijos, quienes nos han apoyado y siempre han estado con nosotros desde el inicio de este proceso. Igualmente, a todas aquellas personas con las que hemos compartido a lo largo de estos años de posgrado, amigos, profesores, y miembros de la facultad, quienes son personas que nos han enseñado, apoyado y escuchado en nuestro diario vivir académico.

Finalmente queremos agradecer especialmente a nuestra directora de tesis Alexandra Pomares Quimbaya, quien nos acompañó en este camino, brindándonos la confianza necesaria en este proceso, además de los consejos, espacios y experiencias únicas. Siendo así una colaboración especial que siempre agradeceremos.

Contenido

1. INTRODUCCIÓN.....	10
2. DESCRIPCIÓN GENERAL	12
2.1. PROBLEMÁTICA.....	12
2.2. OPORTUNIDAD	13
3. DESCRIPCIÓN DEL PROYECTO.....	14
3.1. OBJETIVO GENERAL	14
3.2. OBJETIVOS ESPECÍFICOS	14
3.3. FASES DE DESARROLLO	14
4. TRABAJOS RELACIONADOS	17
5. MARCO TEÓRICO.....	21
5.1. REDES SOCIALES.....	21
5.2 PROCESAMIENTO DE LENGUAJE NATURAL.....	22
6. DESARROLLO.....	23
6.2. ENTENDIMIENTO DEL NEGOCIO	23
<i>Evaluación de la situación</i>	<i>24</i>
<i>Herramientas.....</i>	<i>25</i>
<i>Técnicas de PLN.....</i>	<i>26</i>
<i>Restricciones</i>	<i>26</i>
<i>Objetivos de minería de datos</i>	<i>27</i>
6.3. ENTENDIMIENTO DE LOS DATOS	27
<i>Recolección de datos Iniciales</i>	<i>28</i>
<i>Descripción de los datos.</i>	<i>30</i>
<i>Exploración de los Datos</i>	<i>30</i>
6.4. PREPARACIÓN DE LOS DATOS	36
<i>Datos provenientes de SCOPUS</i>	<i>38</i>
<i>Construcción de los datos</i>	<i>40</i>
6.5. MODELADO	41
<i>Arquitectura SNAPUJ</i>	<i>42</i>
<i>Vista lógica.....</i>	<i>43</i>
<i>Vista de desarrollo</i>	<i>45</i>
<i>Vista física</i>	<i>46</i>

6.6. EVALUACIÓN	46
6.7. DESPLIEGUE	47
<i>Análisis exploratorio de los datos</i>	47
<i>Redes de investigadores</i>	52
<i>Detección de comunidades entre artículos</i>	56
<i>Análisis de palabras claves</i>	58
<i>Análisis de objetivos de desempeño sostenible versus artículos publicados</i>	60
<i>Visualización geográfica</i>	61
7. VALIDACIÓN	66
7.1. VALIDACIÓN EXPERIMENTAL	66
7.2 CUMPLIMIENTO DE REQUERIMIENTOS	67
7.3 PRUEBA DE USO	68
8. CONCLUSIONES Y TRABAJOS FUTUROS	69
9. REFERENCIAS	71
10. ANEXOS.....	74

ABSTRACT

The Pontifical Xaverian University has a need related to the consolidation of information sources associated with the researchers who are attached to the institution, which is not satisfied due to the difficulty of extracting information on research work. From this need found by the team, the tool called SNAPUJ was born, which will address the various problems related to obtaining detailed information on researchers, their fields of action and the working communities where they work.

RESUMEN

La Pontificia Universidad Javeriana posee una necesidad relacionada a la consolidación de las fuentes de información asociadas a los investigadores que están adscritos a la institución, la cual no está satisfecha debido a la dificultad de extraer la información referente a los trabajos de investigación. De esta necesidad encontrada por el equipo, nace la herramienta llamada SNAPUJ, que abordará las distintas problemáticas relacionadas a la obtención de información detallada de los investigadores, sus campos de actuación y las comunidades de trabajo donde se desempeñan.

RESUMEN EJECUTIVO

Desde la Vicerrectoría de Investigación de la Pontificia Universidad Javeriana se detectó la necesidad de poder tener una forma de tener un consolidado de todas las fuentes de información donde se encuentre asociado un investigador adscrito a la Universidad, porque actualmente, están disgregados tanto en las bases de datos académicas (como por ejemplo Clarivate Web of Science o Elsevier Scopus) como en los documentos de información de proyecto propios de la institución.

De esta necesidad manifestada por la Universidad, del equipo surge la herramienta conocida como “SNAPUJ”, una aplicación web que se alimentará de las bases de datos académicas y de los proyectos con el fin de emitir métricas como el índice h, que indica la productividad de las citas que tiene un autor determinado, junto con la exploración de otra información relevante de datos de los investigadores, poder hallar comunidades de investigadores que se desconocían de antemano al evaluar el contenido de las palabras clave de sus obras (y también el análisis a secas de palabras clave de las mismas), la relación de estas obras con los Objetivos de Desarrollo Sostenible (ODS) y como último, pero no menos importante, la obtención de la red de colaboración de un autor, que relaciona al autor con sus colaboradores más inmediatos.

Se espera que con la implementación de SNAPUJ en el proceso de la Vicerrectoría, puedan tener una mejor trazabilidad de las redes de colaboración de los investigadores adscritos a la institución, un mejor conocimiento del impacto de las investigaciones que se estén realizando o ya se hayan realizado (en cuanto a cantidad de citas) y, en general, tener una mirada más amplia y objetiva en cuanto a la producción intelectual (en el ámbito investigativo) que pueda tener la Universidad mediante sus investigadores.

1. INTRODUCCIÓN

Este documento presenta el desarrollo y resultados obtenidos para el trabajo de grado titulado “Análisis de las redes sociales académicas de la PUJ”. Idea que surgió a partir de evidenciar la necesidad de la Pontificia Universidad Javeriana de consolidar fuentes de datos asociadas a los investigadores adscritos a la misma. Las instituciones académicas tienen a su disposición redes sociales académicas que apoyan los procesos de publicación de sus investigadores adscritos. Sin embargo, hoy en día es difícil para las instituciones poder realizar un seguimiento adecuado de las publicaciones actuales de sus investigadores adscritos, debido a las múltiples fuentes, por lo cual, resulta complejo poder obtener métricas, conocer sobre qué temas se está trabajando actualmente, el impacto de las investigaciones realizadas, entre otros.

Gestionar la información científica de forma oportuna y eficiente, es un proceso primordial, ya que es la base de cualquier investigación. Es por esto, que para instituciones que están en constante producción de este tipo de material, (como lo es el caso de la PUJ) les resulta clave tener información de forma centralizada proveniente de diferentes fuentes como bases de datos académicas externas o propias. De esta manera identificar los investigadores que han contribuido de forma activa en el desarrollo investigativo de la universidad, como también la exploración de sus colaboraciones y comunidades de trabajo, se convierte en un factor determinante para fomentar la colaboración y el intercambio de conocimientos en la comunidad académica.

En la actualidad, existen diversos repositorios de información que permiten la identificación de autores y la búsqueda de información científica, como es el caso de Scopus y Web Of Science (WOS), dos de las bases de datos más importantes y utilizadas en el ámbito académico [1], [2], [3]. Sin embargo, estas herramientas tienen limitaciones en cuanto a la integración de información de los investigadores dentro de una misma universidad y la identificación de sus colaboradores y comunidades de trabajo.

En este trabajo de grado se propone el desarrollo de una herramienta que permita a la Universidad obtener información detallada sobre los investigadores relacionados a un campo específico, así como conocer sus coautores, comunidades de trabajo e impacto obtenido. Se utilizó la información disponible en bases de datos de Scopus y WOS, como también la información que

tiene disponible la PUJ con relación a los proyectos de investigación, para ofrecer una visión unificada y detallada de la actividad científica de los investigadores de la PUJ.

2. DESCRIPCIÓN GENERAL

2.1. Problemática

Actualmente, la Pontificia Universidad Javeriana, en sus sedes de Bogotá y Cali, cuenta con 129 grupos de investigación, de los cuales, 87 de ellos se encuentran dentro de las dos categorías más altas en la comunidad científica en Colombia según Minciencias [4], con un 47% en la categoría A1 y el 21% en la categoría A. La producción científica generada por la PUJ debe estar presente en los principales sitios digitales como, plataformas sociales académicas y revistas indexadas.

Dada la gran cantidad de grupos de investigación presentes dentro de la Vicerrectoría de Investigación de la PUJ, se hace necesario tener información consolidada que permita conocer datos relevantes acerca de los aportes de investigadores que han realizado su aporte científico, así como sus redes de colaboración. Se evidencian varias necesidades que requieren ser consolidadas para facilitar y apoyar las funciones correspondientes en la vicerrectoría. A continuación, se listan las más relevantes:

A través de una serie de entrevistas con integrantes de la Vicerrectoría de Investigación de la PUJ, se identificaron las siguientes necesidades:

- 1) Integrar las diversas fuentes de datos relacionadas con investigadores de la PUJ.
- 2) Desambiguar el nombre de los autores en las publicaciones.
- 3) Conocer la red de colaboración entre investigadores.
- 4) Identificar las universidades con las que la PUJ ha colaborado.
- 5) Extraer información relevante asociada a:
 - a. Identificar los proyectos de investigación en los cuales la PUJ cuenta con investigadores con proyectos en el estado Aprobado.
 - b. Clasificar los temas de investigación en los que sus investigadores cuentan con proyectos en el estado Aprobado.
 - c. Clasificar los investigadores por facultades
 - d. Clasificar las investigaciones por palabras clave
 - e. Realizar segmentaciones por comunidades y regiones

- f. Contar con la trazabilidad de sus investigadores a través del tiempo
- g. Identificar fortalezas (más investigaciones) y debilidades (menos investigaciones) en temas de investigación en la PUJ.
- h. Obtener índice H.

El problema planteado desde la Vicerrectoría es que a la fecha la PUJ no cuenta con una herramienta que permita integrar y consolidar fuentes de datos de redes sociales académicas de diversas fuentes, lo cual, ha dificultado caracterizar la investigación en sentido amplio en la PUJ.

2.2. Oportunidad

Como alternativa de solución a la problemática planteada, se establece un sistema en la cual se pueden consolidar y analizar las fuentes de datos de Scopus, WOS que tienen relación con el nombre de afiliación “Javeriana”, como también, se pueden consolidar y analizar los proyectos de investigación que tiene la PUJ en sus bases de datos internas.

3. DESCRIPCIÓN DEL PROYECTO

3.1. Objetivo general

Desarrollar e implementar un sistema de análisis de datos de investigación académica utilizando herramientas y técnicas de analítica de redes para identificar patrones y tendencias en los proyectos y productos de investigación de la PUJ, con el fin de proporcionar información valiosa para la toma de decisiones estratégicas y mejorar la calidad de la investigación académica.

3.2. Objetivos específicos

- Identificar los trabajos y herramientas relevantes de analítica de redes investigación y producción académica.
- Implementar un sistema de análisis de datos de investigación académica.
- Desarrollar una prueba de concepto del sistema construido usando los proyectos y productos de investigación de la PUJ.

3.3. Fases de desarrollo

Este proyecto utilizó una de las metodologías más usadas para describir proyectos de minería de datos. Se trata de CRISP-DM (The cross-industry standard process for data mining) [5]. En su implementación no solo se puede deducir información de datos en general sino también de forma especializada. Está diseñada para adaptarse a procesos de investigación de ingeniería aportando las bases necesarias para preparar los datos que posteriormente serán puestos en producción; no obstante, esta metodología no ofrece un manejo adecuado de calidad de datos para esto se debe proponer una técnica que permita exponer los datos con la mejor calidad posible y que estén alineadas a las necesidades del negocio. Las fases metodológicas permiten recorrer las etapas del ciclo de vida de los proyectos para el análisis de datos, se compone de 6 fases que se ejecutan de forma secuencial e iterativa, lo que permite una mejora continua pues hay una retroalimentación y puesta en marcha de nuevas soluciones que son resultado de hallazgos en fases anteriores.

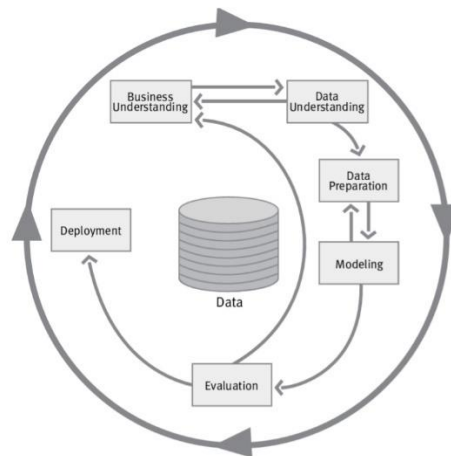


Ilustración 1 Fases del modelo de referencia CRISP-DM [6]

A continuación, se describen las fases metodológicas a utilizar:

Entendimiento del negocio: Esta fase inicial se centra en comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial, para luego convertir este conocimiento en una definición del problema de minería de datos y un plan preliminar diseñado para lograr los objetivos.

Entendimiento de los datos: La fase de comprensión de los datos comienza con una recopilación inicial de datos y continúa con actividades para familiarizarse con los datos, identificar problemas de calidad de datos, descubrir las primeras ideas sobre los datos o detectar subconjuntos interesantes para formular hipótesis sobre información oculta.

Preparación de los datos: La fase de preparación de datos abarca todas las actividades para construir el conjunto de datos final (los datos que se utilizarán en la(s) herramienta(s) de modelado) a partir de los datos en bruto iniciales. Las tareas de preparación de datos se realizan varias veces y no en un orden prescrito. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y limpieza de los datos para las herramientas de modelado.

Modelado: En esta fase, se seleccionan y aplican diversas técnicas de modelado, y se calibran sus parámetros para obtener los valores óptimos. Por lo general, existen varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requisitos específicos

sobre la forma de los datos. Por lo tanto, a menudo es necesario volver a la fase de preparación de datos.

Evaluación: En esta etapa del proyecto, se ha construido un modelo (o modelos) que parece tener una alta calidad desde la perspectiva del análisis de datos. Antes de pasar a la implementación es importante evaluar el modelo y revisar los pasos ejecutados para construirlo, para asegurarse de que cumpla con los objetivos empresariales. Un objetivo clave es determinar si hay algún problema empresarial importante que no se haya considerado suficientemente. Al final de esta fase, se debe tomar una decisión sobre el uso de los resultados de la minería de datos.

Despliegue: La creación del modelo generalmente no es el final del proyecto, se debe realizar la entrega al cliente. Es importante que el cliente comprenda de antemano qué acciones deben llevarse a cabo para realmente aprovechar los modelos creados.

Para llevar a cabo cada una de las fases relacionadas anteriormente, se deben tener en cuenta la tareas o subsecciones que propone la metodología para lograr obtener mejores resultados.

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives Background Business Objectives Business Success Criteria Assess Situation Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits Determine Data Mining Goals Data Mining Goals Data Mining Success Criteria Produce Project Plan Project Plan Initial Assessment of Tools and Techniques	Collect Initial Data Initial Data Collection Report Describe Data Data Description Report Explore Data Data Exploration Report Verify Data Quality Data Quality Report	Select Data Rationale for Inclusion/Exclusion Clean Data Data Cleaning Report Construct Data Derived Attributes Generated Records Integrate Data Merged Data Format Data Reformatted Data Dataset Dataset Description	Select Modeling Techniques Modeling Technique Modeling Assumptions Generate Test Design Test Design Build Model Parameter Settings Models Model Descriptions Assess Model Model Assessment Revised Parameter Settings	Evaluate Results Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models Review Process Review of Process Determine Next Steps List of Possible Actions Decision	Plan Deployment Deployment Plan Plan Monitoring and Maintenance Monitoring and Maintenance Plan Produce Final Report Final Report Final Presentation Review Project Experience Documentation

Ilustración 2 Actividades de las fases del modelo CRISP-DM. [6]

4. TRABAJOS RELACIONADOS

Antes de proceder a construir una solución alrededor de la problemática detectada, se realizó un análisis de las tecnologías y artículos identificados en la literatura que tienen relación con el análisis de redes sociales académicas y las herramientas que se han usado para su entendimiento. Este capítulo presenta las tecnologías identificadas y el siguiente los artículos.

De acuerdo con [7], Cytoscape es software de código abierto con el cual se pueden modelar e integrar redes biomoleculares con grandes bases de datos de proteínas-proteínas, proteínas-DNA e interacciones genéticas de humanos y modelos de organismos.

A partir de esta base y buscando expandir el análisis de proteínas o genes a análisis entre colegas o instituciones, en [8] definen que las redes que representan conexiones entre los individuos puede ser una herramienta valiosa. Por tal motivo, crean una aplicación llamada Social Network Cytoscape la cual es capaz de crear automáticamente un resumen visual de individuos conectados. El objetivo de esta aplicación es crear resúmenes visuales de individuos conectados por enlaces de coautorías en la academia, en función a los datos bibliográficos de Scopus, In-Cites y PubMed. El resultado de la red de coautorías puede ser visualizado y analizado para entender mejor la red de colaboración de investigadores o para comunicar el grado de colaboración y productividad de una publicación en un grupo de investigadores. También puede ser útil como herramienta de investigación para identificar aspectos relevantes de investigación, investigadores y artículos en el área de interés.

Posteriormente [9], afirman que Cytoscape es una de las herramientas de análisis y visualización de redes de biología; sin embargo, tiene con algunas restricciones al momento de crear flujos de trabajo reproducibles y escalables. Para resolver esta situación, proponen integrarla con sistemas de flujos de trabajo altamente productivos como por ejemplo Python/R en Jupyter/RStudio, donde exponen alrededor de 270 funciones claves y 34 aplicaciones de Cytoscape.

En [10], proponen realizar una comparación sistémica de citaciones de Google Scholar, Web of Science y Scopus, teniendo en cuenta 2'448.005 citaciones de 2299 artículos en inglés altamente citados en 252 categorías publicados en 2006, encontrando que el 46,9% de las

citaciones fueron encontradas en las 3 bases de datos, el 36.9% solo se encontraban en Google Scholar, el 3.4% solo se encontraban en Scopus y el 1% solo se encontraban en Web of Science.

En cuanto a la detección de comunidades en redes, en [11] realizan una revisión multidisciplinaria de los métodos y sus respectivas aplicaciones en varios dominios de la vida real, como también de las ventajas y desventajas de las diversas aproximaciones que ofrecen dichos métodos. Así mismo, realizan comparaciones frente al desempeño de los algoritmos, permitiendo de esta manera, una rápida familiarización con los aspectos más importantes de esta temática. En cuanto a los algoritmos analizados, estos corresponden a dos grandes clasificaciones, la primera, algoritmos de detección de comunidad separadas. La cual, a su vez, se ramifican en tres ramas, (1) los algoritmos tradicionales, como lo son segmentación particional, segmentación jerárquica, segmentación espectral y grafos de particionamiento, (2) los algoritmos basados en modularidad y (3) los algoritmos dinámicos. La segunda clase son los algoritmos de detección de comunidades sobre posicionadas, pasando por detección difusa, métodos basados en inferencia estadística, y métodos de reducción de dimensionalidad como lo son la factorización de matrices no negativas y el método de componentes principales. Dentro de sus aplicaciones menciona temas relacionados con redes sociales en línea, redes de comunicaciones, comercio electrónico, en academia y en cienciometría, en sistemas biológicos y de la salud, en economía, detección de fraude, predicción de enlaces, refactorización de paquetes de software y redes de detección de anomalías. En [12] recopilan estudios previos con relación a la problemática de establecer la trazabilidad de la evolución de las redes sociales dinámicas a través del tiempo.

En [3] definen que el concepto de red social académica es creado en el contexto del grande volumen de datos escolares, refiriéndose a la compleja red que se puede formar con entidades académicas y sus relaciones. En su artículo proponen realizar una revisión de los antecedentes, el estado actual y las tendencias de las redes sociales académicas, como también un análisis de los modelos, sus principales métricas, sus propiedades y la disponibilidad de las herramientas para el análisis académico. Como aspectos relevantes destacan las características básicas de las ontologías semánticas académicas, los principales conjuntos de datos disponibles y la información de las mejores herramientas para el análisis de redes sociales académicas, entre las cuales se destacan CiNetExplorer, NetworkX, Pajek y VOSviewer, entre otras. En cuanto a las

principales técnicas para la minería de datos relaciona 4 grupos principales asociados a medidas de similitud, aprendizaje estadístico relacional, minería de grafos y aprendizaje de máquina.

En [13], presentan pybliometrics, el cual ofrece unas clases para interactuar con diversas APIs de Scopus para extraer información como nombres de los autores, sus afiliaciones, resúmenes de sus artículos, entre otros.

En [14], presentan ScientoPy, una herramienta con interfaz gráfica con la cual se pueden importar y unir los conjuntos de datos de WOS y Scopus, encontrar y remover los documentos duplicados y generar un reporte con el resumen de las actividades de preprocesamiento, extraer el índice H de los temas analizados, extraer la país y la institución de la afiliación del autor, generar los top por autor, país o institución basados en el primer autor del documento o en todos los autores del documento, generar la tendencia de temas basados en el top de la tasa promedio de crecimiento y diversa gráficas de visualización, entre otras funcionalidades.

De acuerdo con [15], realizan un análisis bibliométrico sobre el aprendizaje adaptativo en la educación utilizando las herramientas de CiteSpace y VOSviewer y teniendo en cuenta artículos indexados en WOS desde el año 2000 al 2022. CiteSpace fue utilizado para realizar el análisis de la evolución, medir la similitud de unidades de datos a través de un enfoque de teoría de conjuntos relativo a la estandarización de los datos para presentar cronológicamente la evolución y el cambio de tendencias en la investigación. VOSviewer fue utilizado para realizar el análisis estadístico de países y autores, identificar los investigadores principales en la temática y análisis de segmentación en las palabras claves.

Gracias a los artículos analizados, pudimos tener una base de conocimiento estructurada y especializada en los temas que abarcaremos en nuestro trabajo, ayudando así a construir desde el conocimiento científico, una propuesta tecnológica que busca satisfacer inicialmente los requerimientos de la vicerrectoría, impactando las decisiones que se puedan tomar, para que se pueda orientar mejor procesos como, la producción científica, desarrollo de proyectos por palabras claves, encontrar investigadores que han venido realizando sus investigaciones enfocadas en algún tema en particular por ejemplo, proyectos de investigación orientados a los ODS (Objetivos de Desarrollo Sostenible), detección de comunidades e índice H. Se espera que con la propuesta que entregaremos en este trabajo, se puedan realizar análisis a partir de

las funcionalidades que se implementen, de esta manera dar cumplimiento a los objetivos estratégicos de la vicerrectoría de investigación y de paso impactando los objetivos estratégicos de la PUJ.

5. MARCO TEÓRICO

En esta sección se expone la base de conocimiento investigada para el desarrollo de este trabajo.

5.1. Redes Sociales

Según [16] una red social es definida como una red de interacciones o relaciones, donde los nodos consisten en actores y los vínculos consisten en las relaciones o interacciones entre estos actores. Teniendo en cuenta este concepto podríamos decir que las aplicaciones como Facebook o Instagram no son redes sociales, sino que son plataformas que permiten que redes sociales que comúnmente ya existen entre nodos previamente vinculados puedan ser materializados de forma digital, permitiendo su expansión y su conexión con nuevos nodos. Generalmente, las redes sociales formadas digitalmente a través de este tipo de plataformas son habitualmente redes de gran escala ya que acorde a [17] las redes que contienen centros o nodos importantes, parecieran tener enlaces ilimitados y ningún nodo es típico de otros, debido a que las redes de gran escala obedecen a la ley de potencia se puede determinar que entre mayor sea el grado de distribución de un nodo (centro) o, en otras palabras, el número de vínculos que este nodo tenga, mayor será la probabilidad de que otro nodo se conecte a él. Trasladando este análisis a la plataforma Instagram se podría decir que una persona que tenga un alto número de seguidores tiene más posibilidades que personas que aún no la siguen la decidan seguir que otra que tenga menos seguidores; sin embargo, esto no garantiza que el peso de los vínculos que posee la persona con mayor número de seguidores sea mayor que la que tiene menos seguidores, acorde a [18] se argumenta que las personas con las que tenemos vínculos débiles (en el caso de los "conocidos" en lugar de amigos cercanos o familiares) son en realidad más importantes para nuestras oportunidades laborales que nuestros amigos cercanos.

Ahora, si extrapolamos los análisis anteriores al análisis de redes sociales académicas, encontramos que las técnicas de medición se hacen basadas en otros atributos que nos ofrece este tipo de red y su estructura, pero la base de la formación de la red es la misma. Por ejemplo, cuando conocemos la cantidad de coautorías relacionadas a un investigador en especial de la PUJ, nos permite conocer que dicho investigador es un centro importante en la red. De esta

forma las personas cuyo interés sea conocer quién es el investigador con mayores colaboraciones, pueden tomar decisiones relevantes para el dominio de la investigación.

5.2 Procesamiento de Lenguaje Natural

Hoy en día, vemos cómo tareas comunes son ejecutadas por las personas y a su vez apalancadas con la ayuda de la Inteligencia Artificial. En muchas ocasiones de forma inconsciente, cada vez estamos más conectados a las IA por la forma como nos ayudan y solucionan de manera eficiente algunas de nuestras necesidades. Sin embargo, no siempre el resultado es lo que se espera, pues han sido entrenadas con información que en ocasiones no contempla todo el contexto.

Una de las técnicas de IA que permite que la interacción entre el ser humano y las máquinas se pueda establecer de forma orgánica, es la de NLP (Natural Language Processing) o en español Procesamiento de Lenguaje Natural. Esta habilidad de los programas computacionales les permite a las máquinas interpretar, manipular y entender el lenguaje humano tal como se habla o se escribe, según describen tanto [19] como [20].

Entre muchas de las capacidades que ofrece NLP están: el análisis de sentimientos, la traducción automática, los asistentes virtuales, entre otros. Se encuentra que el procesamiento de textos es el más relevante, pues es la base de las demás capacidades, que incluyen extracción y clasificación de textos, traducción de máquina y generación de lenguaje natural [19]. Este permite obtener información estructurada que es determinante para la toma de decisiones de forma oportuna, en tiempos significativamente menores a lo que podría generar una persona. Para lograr esto, se deben realizar una serie de pasos que involucran desde hacer una limpieza de los datos hasta utilizar técnicas de aprendizaje profundo. En el proyecto se trabajará con NLP, y se utilizarán diferentes herramientas para procesar los textos, clasificarlos y extraer la información relevante, entre muchas otras cosas.

6. DESARROLLO

6.2. Entendimiento del negocio

La producción científica generada por las universidades debe ser parte de su objetivo estratégico y muchas veces implícito en su objetivo social, gracias a esto, se logra mejorar la calidad educativa y los procesos de formación de sus estudiantes. Adicionalmente, se logra entregar nuevo conocimiento a la sociedad que permita tener avances científicos, aportar a procesos de innovación y al desarrollo local o regional en las diferentes áreas del conocimiento.

Para la Pontificia Universidad Javeriana, la investigación es una actividad estratégica y prioritaria para dar cumplimiento a sus propósitos fundamentales. Desde la Vicerrectoría de Investigación y por medio de una de las áreas de competencia: la Dirección de Investigación, se ofrece acompañamiento a las unidades académicas en el desarrollo de sus procesos de investigación. Entre las principales funciones que realiza la dirección se encuentran las siguientes:

- Acompañar la formulación de proyectos presentados a convocatorias externas.
- Apoyar la gestión administrativa de proyectos de investigación en alianza con entidades externas.
- Gestionar los sistemas de información y registro de los proyectos de investigación.
- Coordinar el fortalecimiento de capacidades para la investigación (integridad científica y ética de la investigación, alianzas estratégicas institucionales para la investigación).
- Apoyar los procesos de internacionalización de la investigación de la Universidad.
- Apoyar a las unidades académicas en el fortalecimiento de los grupos de investigación.
- Promover la divulgación del conocimiento generado en la investigación.
- Apoyar con recursos de cienciometría los procesos de acreditación institucional y de programas académicos.
- Apoyar las actividades de los semilleros de investigación.
- Apoyar la gestión administrativa de los proyectos de innovación y creación artística.
- Gestionar a nivel institucional la normatividad regulatoria en materia de investigación (trámites y permisos ambientales, acceso a recursos genéticos). [21]

Para el desarrollo de estas funciones, la dirección debe desarrollar actividades que demandan obtener información oportuna y veraz, para tomar decisiones que impactan las necesidades de la Universidad. Es por esto por lo que han surgido necesidades que buscan ser satisfechas, desde el cuerpo académico, con esto, lograr obtener la información de forma sistemática y organizada.

Evaluación de la situación

Para entender la situación actual de la dirección de investigación, se realizó una entrevista semiestructurada como técnica de investigación. Esta se encuentra en el Anexo 1. (Acta entrevista vicerrectora de investigación) al final de este documento; Por medio de ella obtuvimos: datos cualitativos y cuantitativos, comprensión del contexto, experiencia individual, necesidades y limitaciones.

A lo largo del desarrollo del proyecto se obtuvo información de diferentes *stakeholders*, como también personas vinculadas a las fuentes a procesar. En el caso de Scopus, se realizaron varias sesiones donde se nos orientó sobre el uso de la API de la plataforma, como también, con funcionarios de la PUJ, que nos provisionaron la información de los datos estructurados y no estructurados de los proyectos de investigación.

A partir del proceso de investigación realizado previamente, se definen de manera concreta los requerimientos asociados al sistema, y por tanto las funcionalidades esperadas de la misma. Las actividades realizadas para la identificación de requerimientos fueron las siguientes:

1. Definición de casos de uso: Se establecen los casos de uso del sistema, teniendo en cuenta la información recolectada, incluyendo la investigación de trabajos relacionados y entrevista con stakeholders. Anexo 3 casos de uso
2. Definición de requerimientos funcionales: A partir de los casos de uso establecidos se formalizan una versión inicial de los requerimientos funcionales del sistema. Anexo 4
3. Refinamiento de requerimientos funcionales: Posterior a la definición inicial de los requerimientos, se realizó una revisión asociada a los mismos con el fin de refinarlos y empezar la construcción de historias de usuario.

Herramientas

Inventario de fuentes

A continuación, se relacionan los datos provenientes de las diferentes fuentes con los que se trabajó durante el transcurso del proyecto.

- Dataset Scopus: se obtuvo por medio de una consulta usando la API de Scopus. Se obtienen datos como autores, nombre de artículo, *keywords*, entre otros.
Formato: CSV.
- Dataset WOS: Se obtuvo por medio de una consulta generada de forma manual desde la herramienta.
Formato: TXT.
- Dataset de los datos de Proyectos de investigación
Formato: PDF, Word 2007 (DOCX) y Word 97 (DOC)

En la siguiente fase de entendimiento de los datos se ampliará la información de los datasets.

Inventario de recursos tecnológicos

A continuación, se relacionan las herramientas utilizadas para llevar a cabo lo requerido por el proyecto.

- **Python:** Lenguaje de programación de alto nivel utilizado para la analítica de datos, dada su gran eficiencia y facilidad de aprendizaje, multiplataforma y cuenta con una amplia comunidad de desarrolladores que actualizan y mejoran sus librerías constantemente.
Librerías usadas:
 - NLTK [22]
 - aspose.words
 - texttract
- **Pybliometrics:** Librería de Python diseñada para acceder y extraer de la base de datos de Scopus información importante para la producción científica, como autores, citas, afiliaciones, palabras claves, entre otros. Esto se hace a través de un API Key generado en <http://dev.elsevier.com/myapikey.html>. El uso de esta librería facilita la reproducibilidad

de proyectos de investigación y mejora la integridad de los datos para los investigadores que utilizan datos de Scopus. [13]

- **MongoDB:** Base de datos NoSql (Not Only SQL), diseñada para almacenar grandes volúmenes de datos, gracias a su flexibilidad al no requerir que los datos se almacenen en un esquema o estructura determinada, hace que sea ideal para el uso de aplicaciones móviles o web que requieran obtener información de forma eficiente.
- **ScientoPy:** ScientoPy es una herramienta de análisis cientométrico de código abierto basada en Python que se utiliza para analizar y visualizar datos bibliométricos y de investigación científica. Esta herramienta permite importar datos de Web of Science (WoS) y Scopus, filtrar publicaciones por tipo de documento, encontrar y eliminar duplicados, extraer el índice H para los temas analizados, así como la información sobre país e institución desde las afiliaciones de los autores. [14]
- **Streamlit:** Librería de Python que permite crear aplicaciones web para análisis de datos y machine learning de una forma sencilla y rápida. [23]

Técnicas de PLN

- Algoritmo de clasificación: Se realiza la desambiguación de nombres a través del uso de algoritmos que permitan encontrar los mejores resultados con porcentaje de precisión tolerable con base a los recursos provenientes en las diferentes corporas analizadas.
Se hace uso de diferentes librerías de Python que permiten el análisis de los datos, uso de modelos y técnicas de evaluación que se presentaran más adelante en el capítulo del modelo.
- K-means: Para la detección de comunidades, se tuvo presente lo expuesto en [11], y se toma como referencia la técnica de K-means y el análisis de componentes principales para realizar las segmentaciones.

Restricciones

- Las fuentes de datos que se recibirán vendrán en formatos CSV, DOCX, TXT y PDF, se realizará análisis netamente de texto

- Las fuentes externas serán únicamente de SCOPUS y WOS.

Objetivos de minería de datos

- Extraer de forma expedita la información conducente a poder perfilar las comunidades de investigación.
- Caracterizar el impacto de las investigaciones desarrolladas por la PUJ por medio de algoritmos de clasificación.

6.3. Entendimiento de los datos

En esta fase se busca entender los datos que se tienen disponibles, los cuales serán el insumo para resolver los problemas del negocio, se determinará si son suficientes para atender las necesidades del negocio. Se analizará su estructura, si estos vienen agregados o si en cambio vienen desagregados y se deben convertir para manejar datos consolidados, se estudiará su naturaleza y caracterización, de tal forma que esto nos permita entender mejor el contexto de la investigación de la PUJ a través de los datos.

Se seleccionaron los datos que para la vicerrectoría resultan más relevantes, no se tuvieron en cuenta aquellos que no se encuentran dentro del alcance de este proyecto, como puede ser información que esté presente en libros, actas de conferencias, tesis o informes técnicos.

Conforme a lo hablado con la vicerrectora de investigación, se determinó que se analizará información proveniente únicamente de los proyectos de investigación de la PUJ y de artículos provenientes de Wos y Scopus. Se sugirió que en la información que se obtiene del campo texto que contiene el *abstract* de los artículos, se podrían encontrar palabras claves como potencial de impacto que ayudarán a identificar y categorizar los temas, caracterizar temas como los ODS, apropiación social, comunidades e innovación.

Así mismo, se representarán los conceptos lingüísticos que se obtienen después de la clasificación y preprocesamiento de los datos.

Recolección de datos Iniciales

Se contó con tres conjuntos de datos recolectados; el primero son los datos extraídos de SCOPUS mediante el API ofrecido por Elsevier, el segundo son los datos extraídos manualmente de WOS y el tercero son los datos entregados por la PUJ con relación a los proyectos de investigación. Para cada conjunto de datos, se realizó un análisis exploratorio y una verificación de calidad de los datos.

El proceso de extracción de los datos de SCOPUS, se realizó a través del uso de la herramienta Pybliometrics, donde se generó una consulta que permitió acceder a los elementos cuya afiliación estaba asociada con la Pontificia Universidad Javeriana Bogotá, Cali y Medicina. Para su utilización se tuvieron en cuenta los siguientes parámetros:

Término 1: ‘AF-ID (“Universidad Javeriana Facultad de Medicina” 60077399) AND SUBJAREA ({})’

Término 2: ‘AF-ID (“Pontificia Universidad Javeriana” 60033545) AND SUBJAREA ({})’

Término 3: ‘AF-ID (“Pontificia Universidad Javeriana Cali” 60087099) AND SUBJAREA ({})’

Por cada afiliación se realizaron las consultas en relación con artículos en los siguientes temas:

Temas: Agricultura, Artes, Bioquímica, Negocios, Ingeniería Química, Química, Ciencias de la computación, Ciencias de decisión, Odontología, Ciencias de la tierra, Economía, Energía, Ingeniería, Ciencias del medio ambiente, Profesiones de la salud, Inmunología, Ciencias de materiales, Matemáticas, Medicina, Neurociencias, Enfermería, Farmacología, Física, Psicología, Ciencias sociales, Veterinaria y Multidisciplinario.

Posteriormente, instanciando la clase ScopusSearch se iteró por cada afiliación y por cada tema para extraer los datos requeridos y se almacenan los resultados en MongoDB, a continuación, se ilustra el pseudocódigo utilizado:

Términos = [Término 1, Término 2, Término 3]

Temas = ['agri', 'arts', 'bioc', 'busi', 'ceng', 'chem', 'comp', 'deci', 'dent', 'eart', 'econ', 'ener', 'engi', 'envi', 'heal', 'immu', 'mate', 'math', 'medi', 'neur', 'nurs', 'phar', 'phys', 'psyc', 'soci', 'vete', 'mult']

for Ter in Términos:

for Tem in Temas:

s = ScopusSearch(t,format(sub),subscriber = True, view='COMPLETE')

Generar archivo CSV()

Almacenar resultados en MongoDB() Se creó una base de datos con dos colecciones, una para almacenar los datos extraídos de Scopus a través de pybliometrics y otra para almacenar los datos extraídos manualmente de WOS.

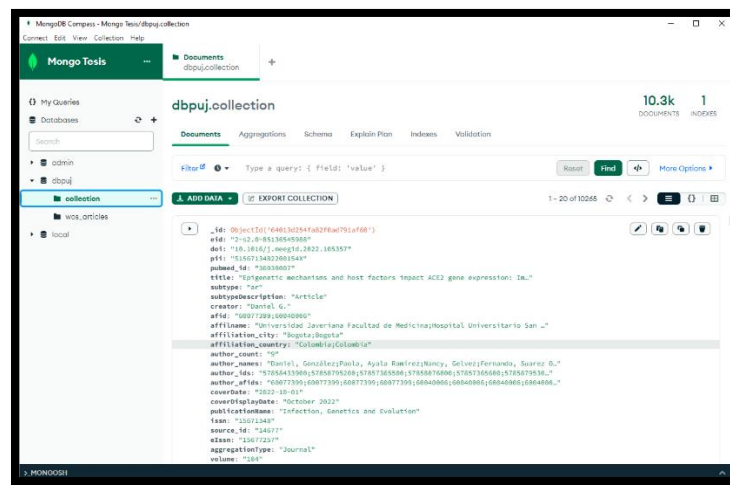


Ilustración 3. Colección de documentos de SCOPUS

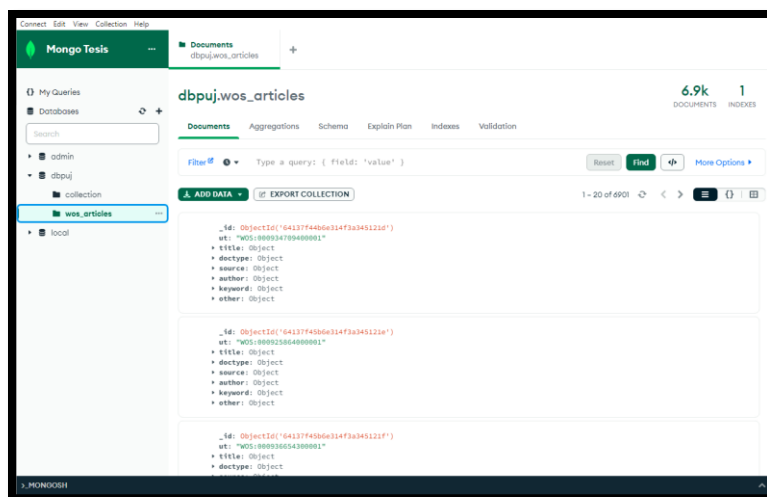


Ilustración 4. Colección de documentos de WOS

Descripción de los datos.

En el anexo 2 se da una descripción de variables de Scopus recolectadas. Scopus descargó treinta y seis variables, con un total de 10382 registros. Como se verá en la exploración de los datos, los datos descargados desde Scopus son de alta calidad al no tener gran cantidad de columnas en valores NaN (nulos).

Exploración de los Datos

Datos de SCOPUS

Del reporte de exploración de los datos extraídos de MongoDB se detectaron 10382 observaciones y 36 variables, dentro de las cuales se encuentran el país, la ciudad, el nombre y el ID de la afiliación, nombres, ID y palabras claves de los autores, título de la publicación, entre otros. En el Anexo 1, se encuentra la descripción de las variables. En cuanto a la verificación de la calidad de los datos se identificaron en las variables que 31 son tipo object, 3 tipo

int64 y 2 de tipo float64. En la Ilustración Ilustración 5 se observan la cantidad de datos faltantes detectados para cada una de las variables.

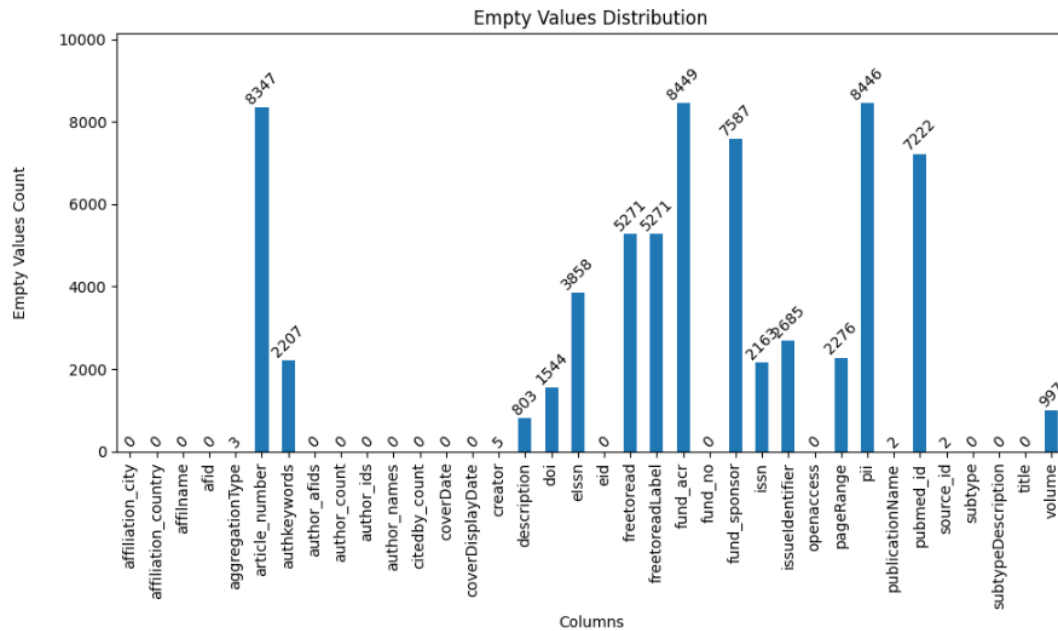


Ilustración 5 Cantidad de datos faltantes por variable

Entendimiento de los datos de WOS

Actualmente, la consulta directa de los datos de WOS mediante una API, emite un objeto JSON que cuenta con una estructura semejante a la mostrada en la siguiente ilustración:

```
articles[0]

{'author': {'authors': ['Gregorio-Chaviano, Orlando',
                        'Marin-Florez, Alexander',
                        'Lopez-Mesa, Evony Katherine',
                        'Lopez-Cordoba, Maria Angelica',
                        'Gomez, Maximino Lopez',
                        'Zamora, Maria-Consuelo'],
            'book_authors': None,
            'book_group_authors': None},
 'doctype': {'doctype': ['Article']},
 'keyword': {'keywords': ['Impact factor',
                           'Emerging Sources Citation Index',
                           'Journal Citation Reports',
                           'Latin American Scientific journals',
                           'Bibliometric indicators']},
 'other': {'contributor_researcher_id_names': None,
            'contributor_researcher_id_researcher_ids': None,
            'identifier_article_no': ['ARTN e91382'],
            'identifier_doi': ['10.5007/1518-2924.2023.e91382'],
            'identifier_eissn': ['1518-2924'],
            'identifier_ids': ['D7GC6'],
            'identifier_isbn': None,
            'identifier_issn': None,
            'identifier_xref_doi': None,
            'researcher_id_disclaimer': ['ResearcherID data provided by '
                                         'Clarivate Analytics']},
 'source': {'book_series_title': None,
            'issue': None,
            'pages': [''],
            'published_biblio_date': None,
            'published_biblio_year': ['2023'],
            'source_title': ['ENCUNTROS BIBLI-REVISTA ELETRONICA DE '
                             'BIBLIOTECONOMIA E CIENCIA DA INFORMACAO'],
            'special_issue': None,
            'volume': ['28']},
 'title': {'title': ['Effect of Emerging Sources Citation Index journal '
                     'citations on impact factor values']},
 'ut': 'WOS:000970363900001'}
```

Ilustración 6: Información que se extrae directamente de WOS para los artículos de investigación.

Existen siete grandes tipos de datos que se pueden encontrar:

1. Authors, que contienen tres tipos de autores: un listado de personas naturales que publicaron un artículo, bajo la propiedad 'authors', 'book_authors', que son autores, pero de libro y 'book_group_authors', que alude a los autores colectivos.
2. Doctype se refiere al o a los tipos de producción referenciadas por este artículo, conteniendo uno de los siguientes 12 tipos de documento, obtenidos mediante una consulta directa hacia los artículos descargados desde la API:


```
▶ typeart=[]  
  for article in articles:  
    for doctype in article.doctype.doctype:  
      typeart.append(doctype)  
  arttype= set(typeart)  
  arttype  
  
↳ {'Article',  
   'Biographical-Item',  
   'Book Chapter',  
   'Book Review',  
   'Correction',  
   'Data Paper',  
   'Early Access',  
   'Editorial Material',  
   'Letter',  
   'Meeting Abstract',  
   'News Item',  
   'Proceedings Paper',  
   'Reprint',  
   'Review',  
   'preprint'}
```

Ilustración 7 Tipos de documento existentes para el material encontrado en WOS, relacionado con la Pontificia Universidad Javeriana.

3. *Keyword*: hacen referencia al conjunto de palabras clave de cada material.
4. *Other*: hace referencia a los identificadores de cada uno de los materiales: los nombres de los investigadores contribuyentes, el identificador del artículo, el DOI, el ISSN, el propio ID dentro de WOS, ISBN e ISSN (en caso de tener), referencias cruzadas a otros DOI y un disclaimer de investigador
5. *Source*: hace referencia al documento mayor en el cual está incluido el material referenciado en esta fuente de datos, donde el dato más importante de este objeto es el *source_title*, que es el título de esta fuente mayor.
6. *Title* hace referencia al título del material bibliográfico encontrado.
7. UT es el identificador de WOS para cada material bibliográfico.

Al no haber surgido información suficiente desde este conjunto de objetos descargado desde la API de WOS, se optó por volcar de forma manual todos aquellos documentos que se pudieran identificar con la Pontificia Universidad Javeriana. Los archivos descargados contienen un máximo de 500 entradas bibliográficas (fue el límite que permitió descargar WOS) y están

separados por el carácter tabulador. Una vez se procesó el conjunto de archivos, con el dataset actual se poseen 6514 entradas procesadas con 71 atributos: 57 de tipo object, 13 flotantes y uno solo que es un entero largo. Gran parte de los atributos tiene una gran cantidad de valores nulos (si no es que toda la columna tiene valores nulos únicamente). La distribución de los valores nulos a través del conjunto de datos está en la siguiente ilustración:

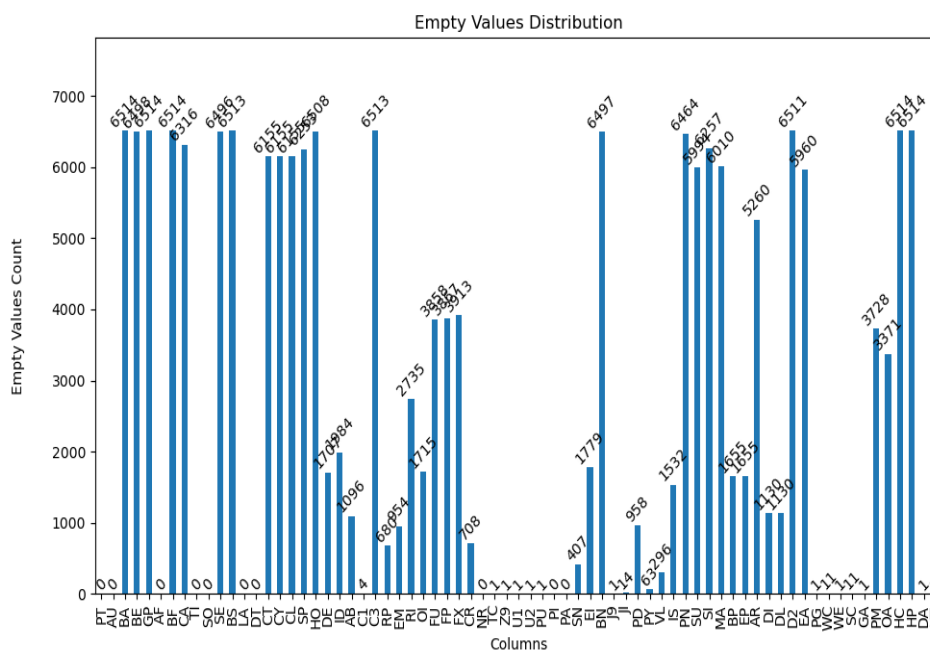


Ilustración 8 Distribución de los valores nulos en el dataset descargado manualmente desde WOS.

Entendimiento de los datos de Proyectos de investigación

Los proyectos de investigación vienen en tres formatos de archivo distintos: PDF, Word 2007 (DOCX) y Word 97 (DOC). Se trata de información, por definición, no estructurada, debido a que viene en un archivo binario. Originalmente, a modo de prueba, se suministró por parte de la Dirección de Investigación, un inventario de 20 archivos, 6 PDFs, 2 DOCs y los restantes 12 son DOCX.





















Nombre	Fecha de modificación	Tipo	Tamaño
 00008204.pdf	22/07/2022 11:21	Documento Adob...	341 KB
 00008389.pdf	22/07/2022 11:23	Documento Adob...	347 KB
 00008390.pdf	22/07/2022 11:23	Documento Adob...	592 KB
 00008769.pdf	22/07/2022 11:23	Documento Adob...	741 KB
 00008993.pdf	22/07/2022 11:24	Documento Adob...	1.074 KB
 00009584.pdf	22/07/2022 11:24	Documento Adob...	739 KB
 00008092.docx	21/04/2023 22:03	Documento de Mi...	1.344 KB
 00008244.docx	21/04/2023 22:02	Documento de Mi...	44 KB
 20018.docx	21/04/2023 22:02	Documento de Mi...	17 KB
 20097.docx	21/04/2023 22:02	Documento de Mi...	14 KB
 20101.docx	21/04/2023 22:02	Documento de Mi...	15 KB
 20106.docx	21/04/2023 22:02	Documento de Mi...	18 KB
 20292.docx	21/04/2023 22:02	Documento de Mi...	15 KB
 20300.docx	21/04/2023 22:02	Documento de Mi...	16 KB
 20316.docx	21/04/2023 22:02	Documento de Mi...	17 KB
 20319.docx	21/04/2023 22:02	Documento de Mi...	14 KB
 20439.docx	21/04/2023 22:02	Documento de Mi...	11 KB
 20519.docx	21/04/2023 22:03	Documento de Mi...	14 KB
 00007278.doc	22/07/2022 11:14	Documento de Mi...	274 KB
 00008145.doc	22/07/2022 11:18	Documento de Mi...	1.731 KB

Ilustración 9: Los 20 documentos entregados para efectos de la prueba.

El archivo más grande es 00008145.doc, con 1731 KB de espacio en disco y el más pequeño es 20439.docx, con 11 KB de espacio en disco.

Los proyectos que se guardan en PDF contienen, en principio, varias tablas, en las cuales se almacena el título del proyecto, convocatoria, duración y los miembros del equipo investigador, con sus distintos datos que los acreditan como tales, incluyendo la institución asociada, último título y su año de obtención, publicaciones y si pertenece o no a un grupo de investigación que esté clasificado por COLCIENCIAS. Luego de dichas tablas, se brinda un resumen del proyecto, por qué se hace la propuesta, los objetivos, la metodología a usar por la investigación, los resultados esperados, los entregables y la bibliografía. Algunos de los proyectos incluyen presupuesto.

Los archivos en formato DOCX empiezan a divergir en cuanto a su estructura global, pero contienen informaciones semejantes a las ya vistas anteriormente en los archivos PDF (en algunos casos no se incluyen tablas, sino que dichos proyectos expresan en texto toda la información que mencionarán), y de la misma forma pasa con los archivos DOC (que en ocasiones puede ir en tablas la información y en otras ocasiones puede ir en texto claro).

6.4. Preparación de los datos

Con relación a las fuentes de datos de SCOPUS y WOS, el principal problema atendido fue la desambiguación de nombres, el cual, se logró identificar los respectivos números de identificación de autores en cada fuente, para su respectiva homologación. Posteriormente, se utilizó la herramienta ScientoPy para realizar la remoción de duplicados.

A través de la interfaz gráfica ofrecida por ScientoPy, se ubican en la carpeta de datos de entrada los archivos en formato .csv de Scopus y en formato .txt de WOS. La versión utilizada es la 2.1.1 y se encuentra disponible en el siguiente repositorio de Github https://github.com/jpruiz84/ScientoPy/releases/download/v2.1.1/ScientoPy_v2.1.1.zip

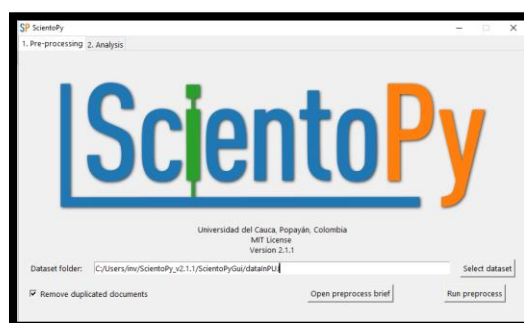


Ilustración 10. Selección de carpeta con archivos de entrada

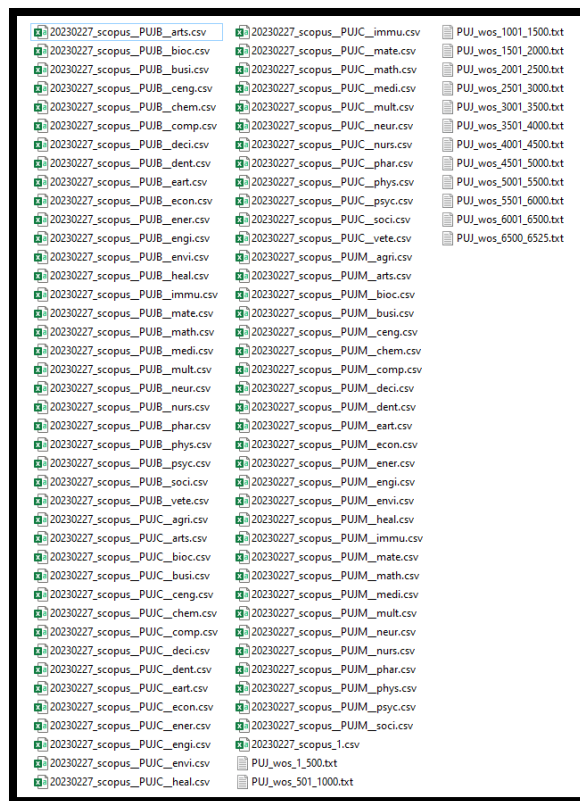


Ilustración 11. Archivos de entrada de SCOPUS y WOS

Una vez se seleccionan los archivos de entrada, se ejecuta la rutina de preprocesamiento, que cuenta con filtrado de tipo de documento, unificación de nombres de campos, normalización de nombres de autores, remover duplicados, cantidad de citas, instituciones y países de los artículos. A continuación, se observa el resumen generado por ScientoPy, donde se detecta el cargue de 22020 artículos, omite 2035 por ser tipos de documento como capítulos de libros, cartas, notas, libros, reportes de erratas, resúmenes de reuniones, correcciones, entre otros, dejando de esta manera 19985 artículos. Identifica que el 72.9% son artículos de SCOPUS y el restante son de WOS y remueve 9653 artículos duplicados, dejando de esta manera 10332 artículos, el 48.4% de SCOPUS y el restante de WOS.

```
[Info,Number,Percentage,Source,Conference Paper,Article,Review,Proceedings Paper,Article in Press,Total
**** Original data ****
Loaded papers,22020,
Omitted papers by document type,2035,9.2%,
Total papers after omitted papers removed,19985,
Loaded papers from WoS,5412,27.1%,
Loaded papers from Scopus,14573,72.9%,
,,,WoS,"0, 0.0%","4908, 24.6%","504, 2.5%","0, 0.0%","0, 0.0%","5412, 27.1%"
,,,Scopus,"1870, 9.4%","11499, 57.5%","1204, 6.0%","0, 0.0%","0, 0.0%","14573, 72.9%"
,,,,,
Duplicated removal results:
Duplicated papers found,9653,48.3%,
Removed duplicated papers from WoS,83,1.5%,
Removed duplicated papers from Scopus,9570,65.7%,
Duplicated documents with different cited by,4862,50.4%,
Total papers after rem. dupl.,10332,
Papers from WoS,5329,51.6%,
Papers from Scopus,5003,48.4%,
,,,,,
Statics after duplication removal filter,
,,,WoS,"0, 0.0%","4830, 46.7%","499, 4.8%","0, 0.0%","0, 0.0%","5329, 51.6%"
,,,Scopus,"985, 9.5%","3632, 35.2%","386, 3.7%","0, 0.0%","0, 0.0%","5003, 48.4%"
```

Ilustración 12. Resumen de preprocesamiento de ScientoPy

Adicionalmente, en la carpeta de datos preprocesados, genera un archivo .csv con los nuevos datos preprocesados y otro con el resumen.

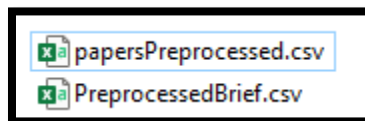


Ilustración 13. Resultados del preprocesamiento de ScientoPy

Datos provenientes de SCOPUS

Para resolver el problema de desambiguación de nombres se trabajó en primera instancia con las variables IDs de autores y Nombres de autores, estas variables contienen las listas de IDs y de nombres, tal y como se observa en la Ilustración Ilustración 14 (a). Posteriormente en (b), se crea una tabla con la relación uno a uno entre el ID y su correspondiente nombre. En (c), se remueven los IDs duplicados, definiendo de esta manera una tabla de homologación, donde cada ID tiene un único nombre. Finalmente, en (d), se observa el mapeo realizado en la variable Nombre de autores, con su correspondiente ID y su correspondiente nombre.

	author_ids	author_names
0	54898017800;57209011866;54988830000;5785634200...	Corredor-Santamaría, W.;Calderón-Delgado, I. C...
1	58189358100;57212998692;55373619400;6601953528	Archbold, George;Parra, Carlos;Carrillo, Henry...
2	56479997600;57419419300;58160053900;58160276800	Aguilar-Garavito, Mauricio;Cortina-Segarra, Jo...
3	58094091400;57212489782;58094183500	Aponte Amaya, Flor Marina;Escobar-Vargas, Jorg...
4	56645757100;56499871900;37047886600;6602898341...	Barreto, Elisa;Lim, Marisa C.W.;Rojas, Danny;D...

(a)

	ID	Name
0	54898017800	Corredor-Santamaría, W.
1	57209011866	Calderón-Delgado, I. C.
2	54988830000	Arbeli, Z.
3	57856342000	Navas, J. M.
4	8585122500	Velasco-Santamaría, Y. M.
...
65912	7006560368	Hunter, F. R.
65913	7003974229	Dravid, A. R.
65914	7005282783	Di Perri, Raoul
65915	24745703800	Morillo, A.
65916	7004180494	Himwich, H. E.

65917 rows × 2 columns

(b)

	author_names
0	Corredor-Santamaría, W.;Calderón-Delgado, I. C...
1	Archbold, George;Parra, Carlos;Carrillo, Henry...
2	Aguilar-Garavito, Mauricio;Cortina-Segarra, Jo...
3	Aponte Amaya, Flor Marina;Escobar-Vargas, Jorg...
4	Barreto, Elisa;Lim, Marisa C.W.;Rojas, Danny;D...

(d)

	ID	Name
0	54898017800	Corredor-Santamaría, W.
1	57209011866	Calderón-Delgado, I. C.
2	54988830000	Arbeli, Z.
3	57856342000	Navas, J. M.
4	8585122500	Velasco-Santamaría, Y. M.
...
65911	6506495592	Ospina, Bertha
65913	7003974229	Dravid, A. R.
65914	7005282783	Di Perri, Raoul
65915	24745703800	Morillo, A.
65916	7004180494	Himwich, H. E.

30753 rows × 2 columns

(c)

Ilustración 14. Desambiguación de nombres en Scopus.

A manera de ejemplo ilustrativo, se muestra a continuación un caso del ID 24832026300 antes de desambiguar con 6 opciones de diferentes nombres y después de ser desambiguado solo con una opción de nombre.

Name	
Pomares-Quimbaya, Alexandra	24
Quimbaya, Alexandra Pomares	21
Pomares, Alexandra	12
Pomares Q, Alexandra	1
Pomares Quimbaya, Alexandra	1
Pomares-Quimbaya, A.	1
dtype: int64	

Antes

Desambiguación



Name	
Quimbaya, Alexandra Pomares	1
dtype: int64	

Después

Ilustración 15. Ejemplo de desambiguación para el ID 24832026300.

Construcción de los datos

Datos provenientes de WOS

Debido a las limitaciones del API de WoS, se usan datos obtenidos a través de la descarga manual de documento y procesados posteriormente. Como campos principales se tienen en cuenta los autores del artículo, el id de los mismos, el abstract de los artículos, sus títulos y sus palabras clave.

Datos provenientes de los proyectos de investigación

Al haberse suministrado estos documentos en un solo paquete enviado por correo electrónico, se introdujeron estos documentos en una carpeta de Google Drive para después ser procesados en un notebook de Google Colab, utilizando las librerías de Python `aspose.words` (para los documentos DOCX), `PyPDF2` (para los PDF) y `textract` (para los DOC), iterando por todos los archivos de la carpeta de Google Drive utilizando el módulo nativo `os` de Python. Cada uno de los archivos se descargó hacia un archivo de texto plano donde en la medida de los límites técnicos de cada librería, se descargó todo el contenido de los archivos.

Con los DOCX se trabaja de forma sencilla, abriendo el documento y guardando su contenido en un archivo plano. De los PDF se extrae cada página y se guarda el texto contenido en ellas, añadiendo un salto de línea en cada ocasión. En el caso de los DOC, se extrae directamente el texto (importante aclarar que solo saca el texto, las tablas las salta) y lo almacena en el archivo de texto plano.





















 00007278.txt	28/04/2023 18:35	Documento de te...	81 KB
 00008092.txt	28/04/2023 18:37	Documento de te...	17 KB
 00008145.txt	28/04/2023 18:35	Documento de te...	24 KB
 00008204.txt	28/04/2023 18:35	Documento de te...	30 KB
 00008244.txt	28/04/2023 18:36	Documento de te...	17 KB
 00008389.txt	28/04/2023 18:35	Documento de te...	19 KB
 00008390.txt	28/04/2023 18:35	Documento de te...	26 KB
 00008769.txt	28/04/2023 18:35	Documento de te...	25 KB
 00008993.txt	28/04/2023 18:35	Documento de te...	57 KB
 00009584.txt	28/04/2023 18:36	Documento de te...	56 KB
 20018.txt	28/04/2023 18:36	Documento de te...	17 KB
 20097.txt	28/04/2023 18:37	Documento de te...	17 KB
 20101.txt	28/04/2023 18:36	Documento de te...	17 KB
 20106.txt	28/04/2023 18:36	Documento de te...	17 KB
 20292.txt	28/04/2023 18:37	Documento de te...	17 KB
 20300.txt	28/04/2023 18:37	Documento de te...	17 KB
 20316.txt	28/04/2023 18:36	Documento de te...	17 KB
 20319.txt	28/04/2023 18:37	Documento de te...	17 KB
 20439.txt	28/04/2023 18:37	Documento de te...	17 KB
 20519.txt	28/04/2023 18:37	Documento de te...	17 KB

Ilustración 16: Archivos de texto plano obtenidos gracias al procesamiento de los archivos originales.

Nótese que los archivos TXT provenientes de los DOCX mostrados en la Ilustración 9 suelen ser de mayor tamaño que sus contrapartes (aunque por las limitaciones conocidas de Aspose.Words en su versión gratuita es posible que no haya descargado completamente uno o varios de los archivos, aparte de poner marcas de agua a comienzo y final del texto).

En un segundo paso, ya teniendo los archivos de texto plano, se procedió a obtener tres piezas de información concretas de los documentos, que también están en los documentos de Scopus o Web of Science: título, autores y resumen. Al ser tan diferentes los archivos en su estructura, se realizó verificación manual de patrones dentro del mismo texto para poder, posteriormente, extraer de manera automática la información necesaria para realizar los análisis posteriores.

6.5. Modelado

Para esta fase se crearon diversos subconjuntos de datos que pudieran dar respuesta a las necesidades identificadas en la fase de entendimiento del negocio, con los cuales se puedan plantear segmentaciones, rankings y frecuencias, entre otros aspectos. Considerando que los modelos preexistentes por sí solos no solucionan todos los requerimientos, se diseñó un sistema que integra diferentes componentes.

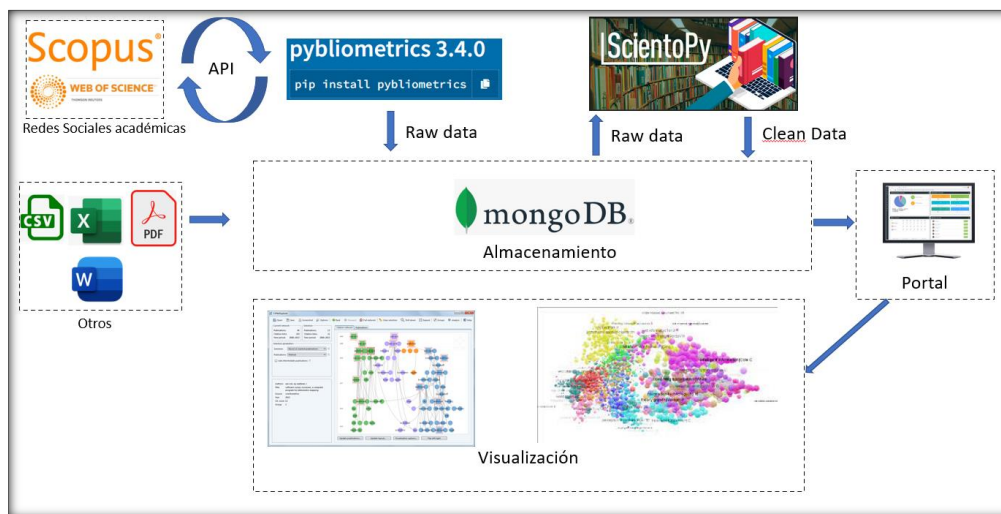


Ilustración 17 Diagrama del funcionamiento de SNAPUJ

El sistema de análisis de datos de investigación académica propuesto lo titulamos SNAPUJ (Social Networks Analisys Pontificia Universidad Javeriana). Este sistema integra la información obtenida por las fuentes externas (Scopus/Web of Science) con las fuentes de la Universidad, mediante distintos tipos de procesamiento. El destino último de la información preprocesada es un almacenamiento en MongoDB, que se carga en un portal web, que permite visualizar los distintos indicadores de interés respecto de los investigadores, sus redes de colaboración y las comunidades que a partir de sus palabras claves se puedan generar. Esto permite dar alcance a los objetivos propuestos en este trabajo cubriendo las necesidades de análisis de la vicerrectoría de investigación.

Arquitectura SNAPUJ

El diseño de este sistema se basa en el modelo de vistas arquitecturales 4+1 propuesto por Kruchten [24]. Este modelo utiliza como entrada principal la vista de casos de uso y/o escenarios, que se corresponde con los casos de uso presentados en la sección anterior. Además, en el modelo se define la arquitectura y el modelo de datos que se utilizarán en el desarrollo del sistema.

Para definir la arquitectura del sistema, se utiliza un modelo en tres capas que cuenta con alta interactividad con los usuarios. Para ello, se definen las siguientes vistas:

1. **Vista lógica:** La vista lógica se define utilizando la notación Entity Boundary Control (EBC), con el estilo de cliente-servidor y combinado con un modelo en 3 capas.
2. **Vista de desarrollo:** A partir de la vista lógica, se desarrollan diagramas de paquetes y componentes que son trazables y abarcan todos los casos de uso descritos en la sección 5.2.
3. **Vista física:** Al establecer los componentes del sistema, es posible construir el diagrama de despliegue para obtener una visualización física del sistema.

A continuación, se presentan los diagramas propuestos previamente:

Vista lógica

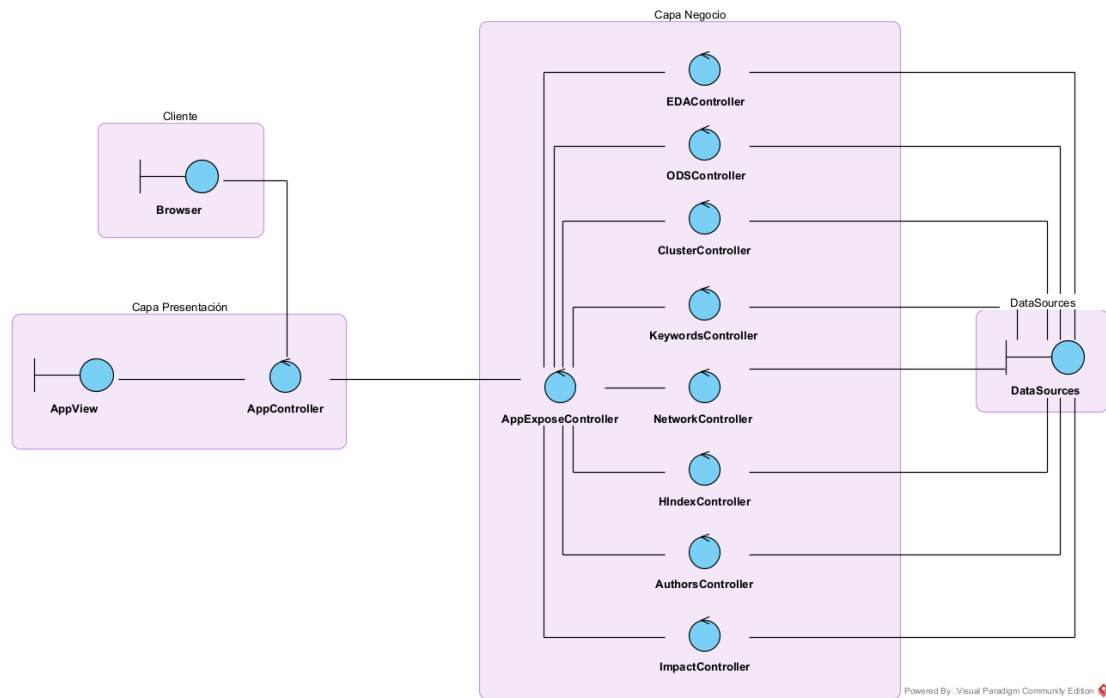


Ilustración 18 Diagrama EBC - Vista lógica

En la Ilustración 18, se puede observar el diagrama EBC (Entity Boundary Control) propuesto para el sistema, el cual se encuentra distribuido en las siguientes capas:

- **Cliente:** El cliente es el punto de entrada al sistema. Para este caso particular, se refiere al navegador web que será utilizado para hacer uso de las funcionalidades del sistema.
- **Capa de Presentación:** La capa de presentación, en este caso particular, contiene una única vista que será dinámica, por lo tanto, un único boundary y un único controller. Esta capa se enfoca en la interacción con el usuario, expone funcionalidades tales como el análisis exploratorio de datos, la revisión de objetivos de desarrollo, entre otras. Adicionalmente, para la obtención de la información cuenta con una comunicación directa con la capa de negocio.
- **Capa de Negocio:** La capa de negocio es la encargada de realizar todo el procesamiento de los datos y del cálculo y generación de los indicadores y métricas propuestas. Existe un controlador por cada una de las funcionalidades expuestas. Cada uno de dichos controladores cuentan con la conexión hacia las fuentes de datos, bien sean bases de datos o archivos.
- **DataSources:** Esta capa corresponde a las fuentes de datos que proveen al sistema con la información necesaria para la generación de las métricas propuestas.

Vista de desarrollo

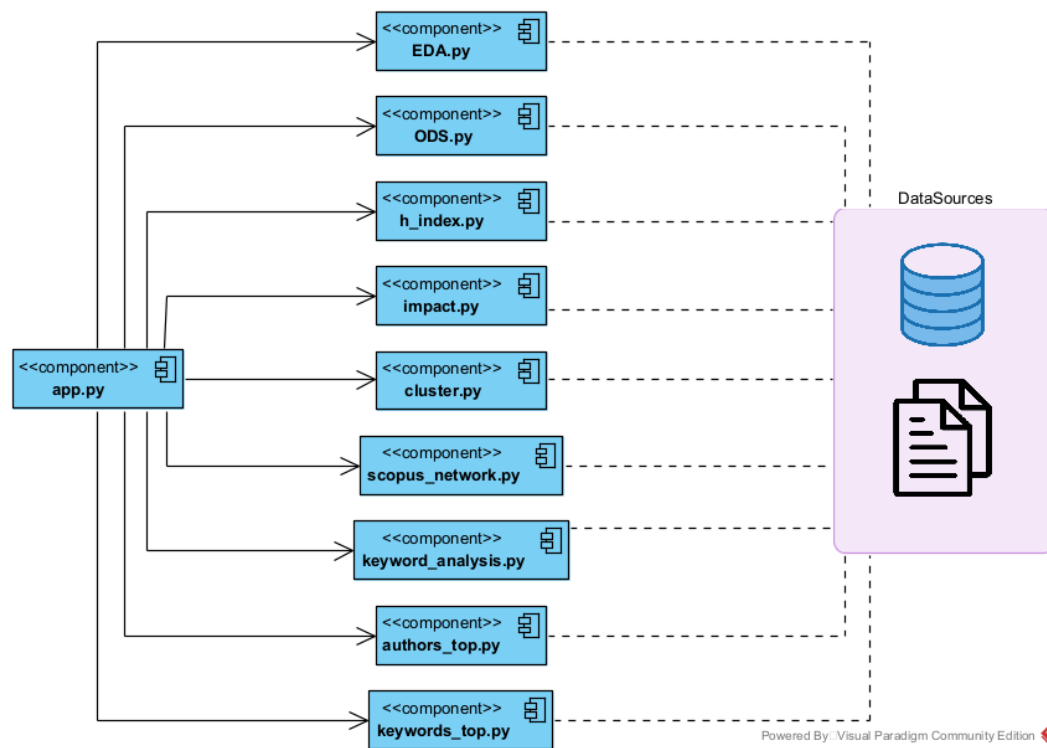


Ilustración 19 Diagrama de Componentes - Vista de desarrollo

En la Ilustración 19, se observa el diagrama de componentes, el cual se encuentra compuesto por 3 niveles, descritos a continuación:

- **Nivel de presentación:** Este nivel se encuentra conformado por el componente app.py que se encarga de recibir las interacciones con el usuario y procesarlas para ser enviadas al componente correspondiente.
- **Nivel de negocio:** El nivel de negocio contiene los componentes Python relacionados a cada una de las métricas propuestas para ser generadas por el sistema, así mismo proporciona la escalabilidad y mantenibilidad necesaria en caso de agregar nuevas métricas.
- **Fuentes de datos:** El último nivel corresponde a las diferentes fuentes de datos aceptadas por el sistema para ser procesadas, siendo estas bases de datos o archivos.

Vista física

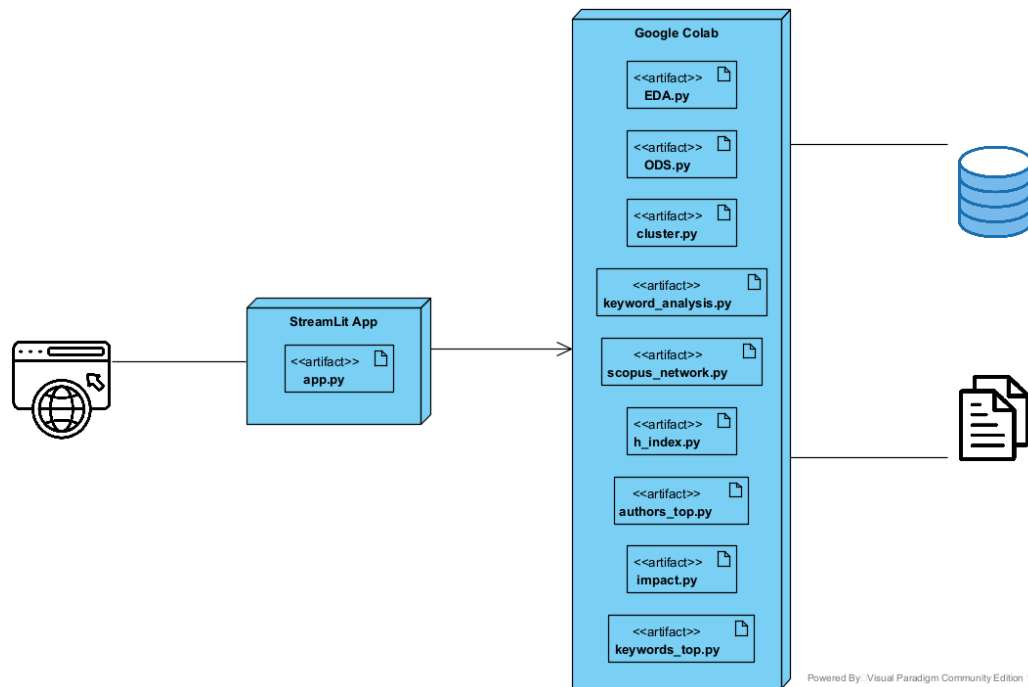


Ilustración 20 Diagrama de despliegue - Vista física

El diagrama de despliegue se encuentra en la Ilustración 20, el cual, se encuentra compuesto por 2 *Tiers* principales, los cuales son los siguientes:

- **StreamLit App:** Este Tier contiene la presentación del sistema a través del app.py, así mismo, es el encargado de establecer la conexión con la capa de “backend” localizada en Google Colab, para poder obtener la información relacionada a las métricas que busca obtener el sistema.
- **Google Colab:** El Tier de Google Colab contiene los componentes de negocio del sistema, expone un artefacto para cada uno de los filtros, y, cada uno de ellos cuenta con una conexión directa hacia las diferentes fuentes de datos.

6.6. Evaluación

Se realizaron sesiones de reto con los principales involucrados de la Vicerrectoría para alinear las expectativas frente a las necesidades planteadas.

6.7. Despliegue

Se creó una aplicación web llamada SNAPUJ en la cual se pueden acceder a las funcionalidades desarrolladas para realizar análisis, visualizaciones e identificar aspectos relevantes y de impacto para la vicerrectoría de investigación.

A través de la aplicación desarrollada con la librería Streamlit, se desplegaron las siguientes funcionalidades, las cuales se pueden acceder a través de un enlace web y la URL externa asociada, a continuación, se comentarán sus características principales:

Selecciona una opción

- ☒ Exploración de datos de Scopus
- ☐ Exploración de datos de WoS y Scopus
- ☐ Exploración de Proyectos de Investigación
- ☐ Redes de Investigadores
- ☐ Red de afiliaciones de la PUJ
- ☐ Segmentación de artículos
- ☐ Análisis de palabras claves
- ☐ Relación de ODS con artículos
- ☐ Índice H de citaciones
- ☐ Top de autores con más publicaciones
- ☐ Visualización geográfica por países
- ☐ Visualización geográfica por ciudades de Colombia

Redes de investigación de la PUJ

Conoce aspectos relevantes de los datos de investigadores de Scopus

Datos Previos

	affiliation_city	affiliation_country
0	;Bogota;Madrid	Colombia;Colombia;Spain
1	Bogota;Ghent;San Andres Island;Medellin	Colombia;Belgium;Colombia;Colombia
2	Bogota;Bogota;Alicante	Colombia;Colombia;Spain
3	Bogota;Bogota	Colombia;Colombia
4	Cali;Copenhagen;Goiania;Stony Brook;Prague Praha;Birmensdorf	Colombia;Denmark;Brazil;United States;Cz

Número de observaciones: 9848

Número de variables: 36

Ilustración 21. Aplicación Web SNAPUJ

Análisis exploratorio de los datos

Con base en los datos consolidados a través de ScientoPy, se realizó un análisis exploratorio de los datos unificados, revisando sus primeras cinco observaciones, la cantidad total de observaciones y variables contenidas en el conjunto de datos, un resumen estadístico de todas las variables y la cantidad de valores faltantes por variable, como también el top 10 de las variables que por su nombre se ven relevantes. En las siguientes ilustraciones se observan los resultados obtenidos a través de SNAPUJ para esta funcionalidad.

A continuación se puede observar el conjunto de datos, se cuenta con un deslizador vertical y otro horizontal para navegar sobre los datos y conocer la manera en la cual son recibidos en primera instancia.

Análisis Exploratorio de Datos		
Datos Previos		
	Authors	Title
0	Orlik, Y., Moreno, A., Velazquez, M.C.	10th anniversary of the J
1	Hoyos, J.A., Bermeo, A.M.P., Nino, J.F.P.	10-years experience in th
2	Moreno-Fuquen, R., Loaiza, A.E., Diaz-Velandia, J., Kennedy, A.R., Morrison, C.A.	1-Benzylpiperidin-4-one
3	Morales, A., Puerta, R., Rommel, S., Monroy, I.T.	1 Gb/s chaotic encoded v
4	Pulido, N., Guevara-Morales, J.M., Rodriguez-Lopez, A., Pulido, A., Diaz, J., Edrada-Ebo	1H-nuclear magnetic res

Número de observaciones: 10194

Número de variables: 42

Ilustración 22. Visualización de datos previos y cantidad de observaciones y variables

También se pueden consultar valores estadísticos como la cantidad de variables, cantidad de variables únicas, la observación que más veces aparece en dicha variable, entre otras. Para las variables numéricas se pueden consultar sus valores como la media, varianza, máximo, mínimo y otros percentiles.

Resumen estadístico:										
	Authors	Title	Year	Source title	Volume	Issue	Art. No.	Page start	Page end	Page co
count	10194	10194	10,194	10194	9321	7797	1853	7992	7967	5345
unique	9311	10183	None	4084	663	354	1731	1812	1836	84
top	Orticochea	Programa	None	UNIVERSITAS	13	1	9	1	41	9
freq	17	2	None	183	233	1640	7	179	35	447

Ilustración 23. Resúmenes estadísticos de datos consolidados

De forma general, se pueden consultar la cantidad de valores faltantes por variable, lo cual nos permite tener un entendimiento de la completitud del conjunto de datos y de esta manera empezar a inferir que variables están con todos sus registros al 100% y cuáles no.

Valores faltantes:

	Column	Missing Count	Missing Percent
0	Authors	0	0.00%
1	Title	0	0.00%
2	Year	0	0.00%
3	Source title	0	0.00%
4	Volume	873	8.56%
5	Issue	2,397	23.51%
6	Art. No.	8,341	81.82%
7	Page start	2,202	21.60%
8	Page end	2,227	21.85%
9	Page count	4,849	47.57%

Ilustración 24 Valores de datos faltantes por variable en cantidad y porcentaje

Una variable de interés es la afiliación, donde se encontró que las diversas versiones que se tienen de la Pontificia Universidad Javeriana y también se consultó el top 10 de las afiliaciones más frecuentes en el conjunto de datos.

	Affiliations
Pontificia Universidad Javeriana, Cali, Colombia	20
Pontificia Univ Javeriana, Bogota, Colombia	18
Dept. Cir., Univ. Javeriana, Bogota, Colombia	8
Dept. Cir., Fac. Med., Univ. Javeriana, Bogota, Colombia	8
Pontificia Univ Javeriana, Fac Psicol, Bogota, Colombia	7
Fac. Med., Univ. Javeriana, Bogota, Colombia	7
Pontificia Univ Javeriana, Cali, Colombia	6
Pontificia Universidad Javeriana Cali, Colombia	5
Univ Javeriana, Bogota, Colombia	5
Dept. Med. Int., Fac. Med., Univ. Javeriana, Bogota, Colombia	5

Ilustración 25 Top 10 de las afiliaciones más frecuentes.

También se puede destacar el crecimiento que se ha tenido en la generación de publicaciones a través de los años, pasando de 456 en el año 2014 a 1159 en el año 2021.

	Year
2,021	1,159
2,020	1,077
2,019	939
2,018	850
2,017	821
2,016	669
2,022	665
2,015	543
2,013	467
2,014	456

Ilustración 26. Cantidad de publicaciones por año

En cuanto a la variable país, se detectó que en una sola observación se almacenan los países de los diversos autores de cada publicación, encontrando el top 10 de las observaciones más frecuentes.

	country
Colombia	2,910
Colombia;United States	301
United States;Colombia	233
Colombia;Spain	210
Spain;Colombia	165
Colombia;France	81
Colombia;United Kingdom	65
Brazil;Colombia	61
Colombia;Chile	60
Mexico;Colombia	58

Ilustración 27. Top 10 de publicaciones por país.

Análisis Exploratorio de Datos del API de WoS

Para Web of Science, inicialmente se planteaba utilizar el API provista, para este caso particular, woS lite. A continuación, se presenta un perfilamiento de los datos obtenidos a través de dicha API.

Datos Previos

	_ut	_title	_doctype
0	WOS:000971687400019	{'title': ['Design of a wearable device for respiratory rate monitoring and ']	{'doctype':
1	WOS:000967236900001	{'title': ['Novel Use of Feminization Laryngoplasty']}]	{'doctype':
2	WOS:000972043400001	{'title': ['Conservation at the edge: connectivity and opportunities from ']	{'doctype':
3	WOS:000975984900001	{'title': ['High cost drugs in Latin America: access and barriers']}]	{'doctype':
4	WOS:000862858100001	{'title': ['Impacts of pastures and forestry plantations on herpetofauna: A ']	{'doctype':

Número de observaciones: 7066

Número de variables: 8

Ilustración 28 Datos previos obtenidos de vos lite

A su vez, consultamos algunos análisis estadísticos asociados a los datos obtenidos. Encontrando que, los datos obtenidos no tienen valores numéricos en este caso.

Resumen estadístico:

	_ut	_title	_doctype	_source	_author	_keyword	_other	discriminator
count	7066	7066	7066	7066	7066	7066	7066	0
unique	7066	7031	19	6729	6736	5221	6890	None
top	WOS:000971687400019	{'title': ['Design of a wearable device for respiratory rate monitoring and ']	{'doctype':	{'book_se	{'authors	{'keywords	{'contri	None
freq	1	7	5211	17	12	1834	9	None

Ilustración 29 Resumen estadístico de los datos obtenidos de vos

Por último, evaluando los datos faltantes, se encuentra que, los datos no tienen datos faltantes en la mayoría de sus columnas, a excepción de la columna “discrimination”, tal como se puede evidenciar a continuación.

Valores faltantes:

	Columna	Conteo de Faltantes	Porcentaje de Faltantes
0	_ut	0	0.00%
1	_title	0	0.00%
2	_doctype	0	0.00%
3	_source	0	0.00%
4	_author	0	0.00%
5	_keyword	0	0.00%
6	_other	0	0.00%
7	discriminato	7,066	100.00%

Ilustración 30 Valores faltantes - Datos obtenidos de wos lite

Finalmente, y teniendo en cuenta la información obtenida, se decide no utilizar el API de WoS debido a que no es posible obtener la información requerida para realizar un análisis adecuado de los datos obtenidos. Por lo cual, se decide realizar la descarga manual de dichos artículos.

Redes de investigadores

Utilizando los datos desambiguados de Scopus, se procedió a crear la red de investigadores de la PUJ, para lo cual utilizo el preprocesamiento comentado en la sección de preparación de los datos. Posteriormente se realizó la construcción de todos los posibles nodos con sus respectivos enlaces, para contar con esta información para la construcción de la red por el autor que se quiera consultar. Dada la gran cantidad de investigadores y en algunos casos su gran cantidad de enlaces se optó por construir redes que permitan conocer a los coautores del autor consultado, como también, los coautores de estos coautores, lo que se denomina, la red de coautores a primer y segundo nivel. A continuación, se observa el resultado de la red al

seleccionar el autor “Calderón-Delgado I. C.” y su relación de coautores a primer y segundo nivel los coautores de estos coautores.

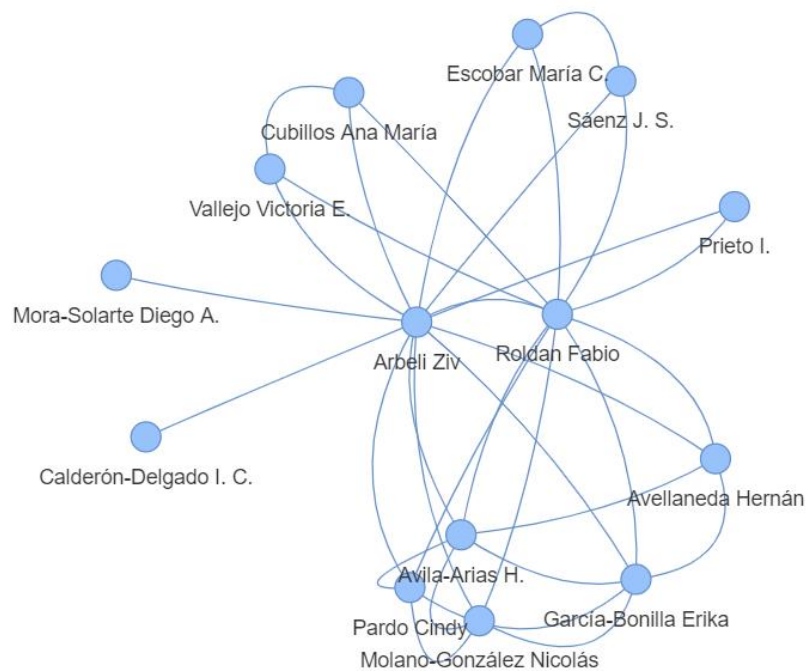


Ilustración 31. Red por autor en primer y segundo nivel

Autores a nivel 1:

```
▼ [  
  0 : "Arbeli Ziv"  
]
```

Autores a nivel 2:

```
▼ [  
  0 : "Roldan Fabio"  
  1 : "Avellaneda Hernán"  
  2 : "Calderón-Delgado I. C."  
  3 : "Sáenz J. S."  
  4 : "Mora-Solarte Diego A."  
  5 : "Prieto I."  
  6 : "Molano-González Nicolás"  
  7 : "Avila-Arias H."  
  8 : "Escobar María C."  
  9 : "Cubillos Ana María"  
 10 : "García-Bonilla Erika"  
 11 : "Vallejo Victoria E."  
 12 : "Pardo Cindy"  
]
```

Ilustración 32. Lista de autores a nivel 1 y 2 relacionados con el autor García Juan Carlos

Detección de comunidades entre artículos

A través de la detección de la importancia estadística de las palabras en los resúmenes de los documentos y la identificación de la cantidad óptima de clústeres a seleccionar a través del método Elbow y el análisis de silueta, se entrega una funcionalidad que le permite al usuario ajustar la cantidad de clúster que desea analizar y a través de la descomposición de componentes principales, se crean los grupos o comunidades existentes entre los artículos analizados.

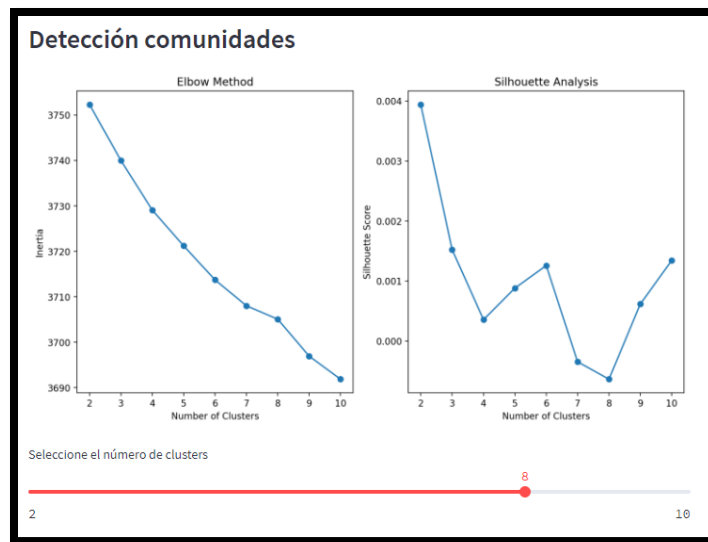


Ilustración 33 Análisis de codo (izquierda) y de silueta (derecha)

A continuación, se observan los dos componentes principales al seleccionar 8 clústeres y sus respectivas agrupaciones de artículos con sus respectivos autores en forma tabular.

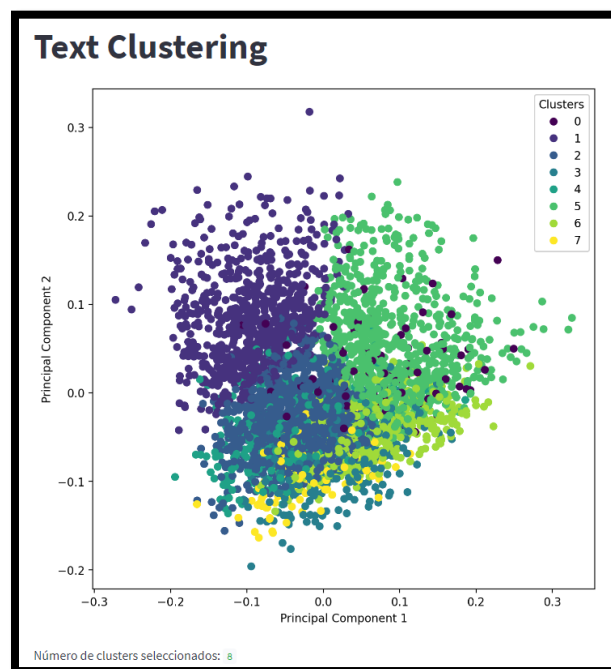


Ilustración 34. Visualización de componentes principales con 8 clústeres

Clusters:

	Autores	Cluster
6	Bautista-Molano, W., Saldarriaga-Rivera, L.M., Junca-Ramirez, A., Fernandez-Aldana, .	1
8	Saldarriaga-Rivera, L.M., Bautista-Molano, W., Junca-Ramirez, A., Fernandez-Aldana, .	1
11	Carpeta, S., Pineda, T., Martinez, M.C., Osorio, G., Porras-Hurtado, G.L., Rojas, J., Zarah	1
17	Barreto-Duran, E., Mejia-Cruz, C.C., Jaramillo-Garcia, L.F., Leal-Garcia, E., Barreto-Prie	4
21	Canon-Ramirez, L.E., Prieto-Sandoval, V.	5
22	Baez, E., Guio-Vega, G.P., Echeverria, V., Sandoval-Rueda, D.A., Barreto, G.E.	4
28	Bogoya, J.M., Vargas, A., Cuate, O., Schutze, O.	6
29	Siqueira, J.R., ter Horst, E., Molina, G., Losada, M., Mateus, M.A.	2
31	Tellez-Beltran, D., Gonzalez-Munoz, A., Baron-Cifuentes, V., Pradilla-Gomez, J.M.	2
32	Garcia, P.K., Vargas, D.C., Contreras, K., Gonzalez, C., Rodriguez, M.P., Bermudez, L.E.	1

Ilustración 35. Segmentación de grupos de artículos con sus respectivos autores de forma tabular.

Análisis de palabras claves

Se entregó una funcionalidad para buscar la(s) palabra(s) clave de interés y generar las respectivas nubes de palabras donde se encontraron al interior de algún título, resumen y/o palabras claves. Adicionalmente, se generó una lista detallada de los datos relevantes de los artículos relacionados.

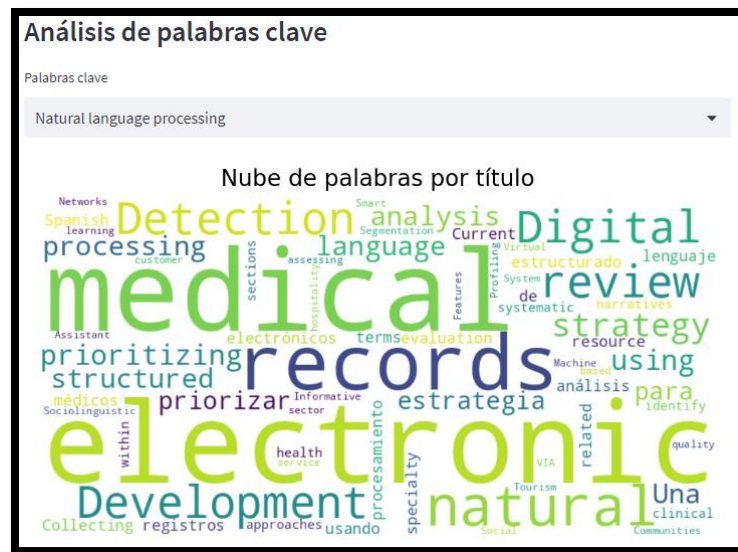


Ilustración 36. Nube de palabras de títulos relacionados con las palabras clave "Natural language processing"

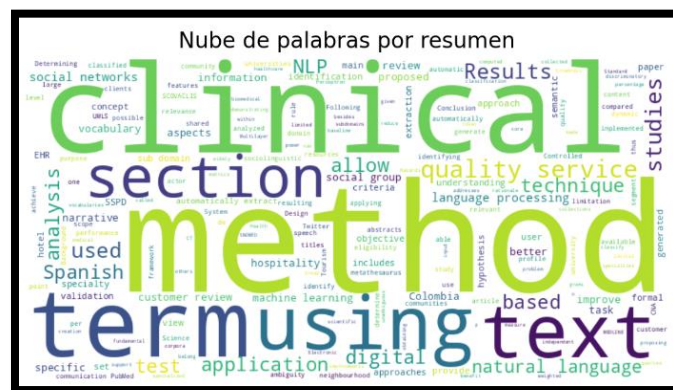


Ilustración 37. Nube de palabras de resúmenes relacionados con las palabras clave "Natural language processing"

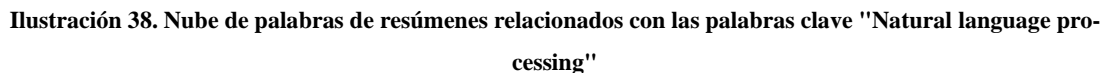


Ilustración 39. Lista detallada con los datos de los artículos identificados

Para esta funcionalidad se contaba con la información de los Objetivos de Desarrollo Sostenibles (ODS) en idioma español y los resúmenes de los artículos en inglés. Por lo cual fue necesario traducir los ODS a idioma inglés para encontrar la métrica del coseno de similitud entre los resúmenes y cada ODS. Posteriormente, se generó una lista consolidando por cada

ODS en idioma español sus tres artículos con mayor similitud. A continuación, se observa el resultado de la funcionalidad entregada.

¿Cuál objetivo quieres revisar?

Selecciona un objetivo de desarrollo sostenible

1 . 1 De aquí a 2030 erradicar para todas las personas y en todo el mundo la pobreza extrema (a... ▼

	Authors	Title
1,782	Gomez-Restrepo, C., Rincon, C.J., Medina-Rico, M.	Chronic diseases in the p
5,806	Aguado-Quintero, L.F., Osorio-Mejia, A.M., Ahumada-Castro, J.R., Riascos-Correa, G.I.	Measuring poverty from t
9,073	Mallarino, C.U., Marin, J.J.	The frontiers of poverty i

Ilustración 40. Relación entre ODS y el top 3 de artículos con mayor similitud.

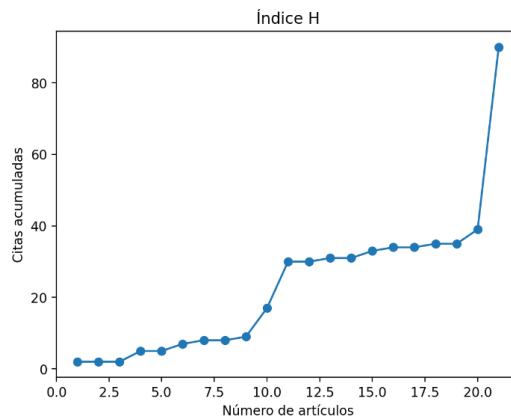
Visualización de Índice H de citaciones

Para esta funcionalidad se realiza un conteo de la cantidad de citaciones de cada uno de los artículos para un autor seleccionado, a su vez, presenta el valor calculado de índice H para el autor seleccionado, el cual se define como el valor más alto de H tal que el autor ha publicado al menos H artículos que han recibido al menos H citas cada uno. A continuación, el resultado de esta funcionalidad.

Revisemos el índice de citaciones

Seleccione el autor

Pomares-Quimbaya, A



El índice H para el autor Pomares-Quimbaya, A es 12

Ilustración 41 Visualización de índice H de citaciones.

Top de autores con más publicaciones

A través de la identificación de todos los autores que han realizado publicaciones tanto en Scopus como en WoS, se realiza un conteo de publicaciones por autor, y se ofrece la posibilidad de conocer los autores que más publicaciones han realizado, ofreciendo la posibilidad de hacer una selección de la cantidad de autores deseados en el top. A continuación, se puede ver el output de esta funcionalidad.

Revisemos el top de autores con más publicaciones

Seleccione cuantos autores quiere ver

5

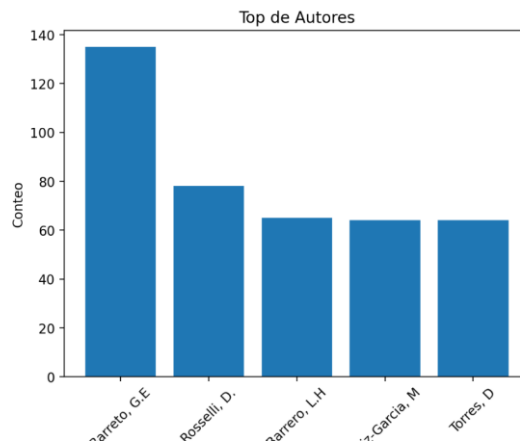


Ilustración 42 Visualización top de autores con más publicaciones

Visualización geográfica

Luego de hacer un tratamiento de preprocesamiento a la variable country, se logró realizar el conteo de publicaciones por cada país, donde encontramos que el país Colombia aparece en 6183 publicaciones, seguido de España con 792 y Chile con 473. A continuación, se observa el resultado de esta funcionalidad.

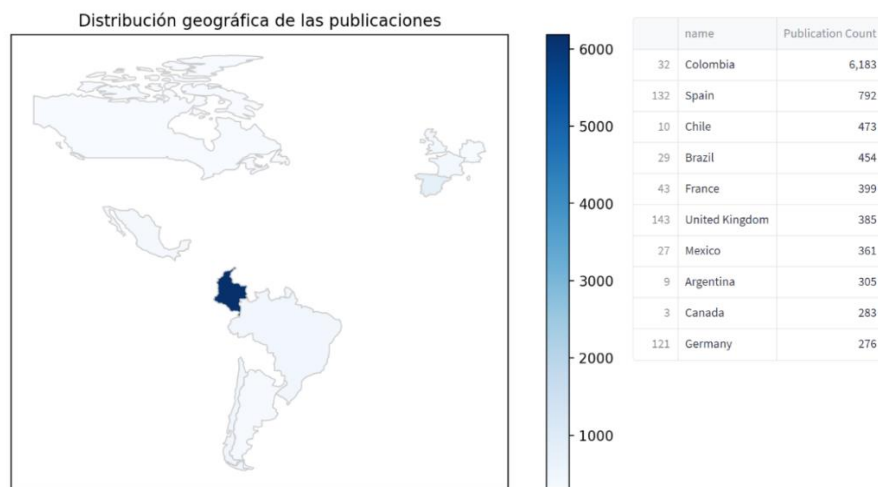


Ilustración 43. Visualización geográfica de las publicaciones.

Red de afiliaciones de la PUJ

Para el análisis de esta funcionalidad se utilizaron los datos de la fuente Scopus, debido a que esta fuente permitió identificar los registros de la PUJ gracias a su ID de afiliación, se realizó un preprocesamiento de los datos similar al de la desambiguación de autores, pero en este caso se realizó con las afiliaciones es decir los nombres de las universidades y su ID.

En este análisis se evidencio como desde la fuente de Scopus, la universidad está registrada con diferentes códigos de afiliación lo que ocasiona que la información este segregada en las fuentes originales. Según análisis realizado, se estima que existen alrededor de 9 registros con el nombre de la PUJ pero con diferente código de afiliación.

Para la elaboración del grafo, inicialmente se pensó en generar una red egocentrada, con esto ver aunque no siendo el objetivo de este tipo de redes las posibles relaciones entre sus alter, pero cuando se tomó el nodo que corresponde al nombre 'Pontificia Universidad Javeriana' donde uno de sus ID el 60033545 por ser el que cuenta con mayor número de registros, se encontraron un total de 8039 registros y al calcular los vínculos directos de este nodo con sus vecinos arrojó un valor de 6010, lo cual hace que esta red sea imposible de interpretar por medio de una visualización, por lo que se optó en graficar una red centralizada con otro número de ID con menor cantidad de vínculos.

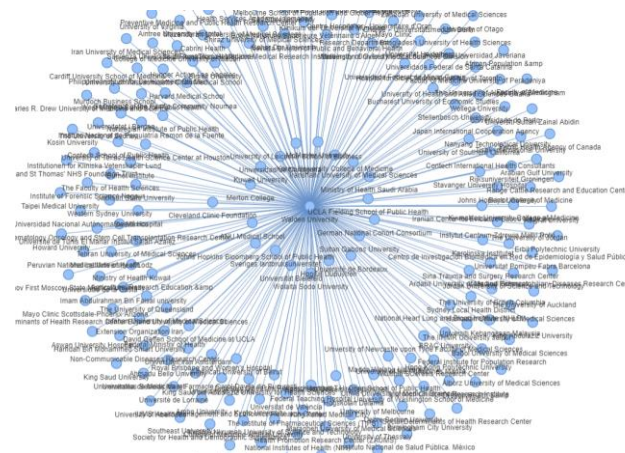


Ilustración 444. Red de universidades con las que la PUJ ha colaborado.

Esta funcionalidad permite visualizar la red de universidades vinculadas con la PUJ, como también exportar a un archivo csv con esta información.

universidades vinculadas

```

▶ [ 0 - 100 ]
▶ [ 100 - 200 ]
▼ [ 200 - 300 ]
200 : "Charité - Universitätsmedizin Berlin"
201 : "Université de Lorraine"
202 : "Työterveyslaitos"
203 : "Western Sydney University"
204 : "Klinikum der Universität München"
205 : "Zahedan University of Medical Sciences"
206 : "Kermanshah University of Medical Sciences"
207 : "Erbil Polytechnic University"
208 : "Hong Kong Polytechnic University"
209 : "The Institute of Pharmaceutical Sciences (TIPS)"
210 : "Universidad Nacional Autónoma de México"

```

Exportar a CSV

Se ha exportado la lista de nodos diferentes a VínculosUniversidades.csv

Ilustración 455. Red de afiliaciones de la PUJ.

7. VALIDACIÓN

7.1. Validación experimental

Para llevar a cabo la validación experimental de nuestro sistema, diseñamos y realizamos un experimento siguiendo los siguientes pasos:

- **Diseño del experimento:** Establecimos un grupo de control y un grupo de tratamiento. En el grupo de control, los participantes utilizaron un enfoque convencional, mientras que en el grupo de tratamiento utilizaron nuestro sistema. La variable independiente fue la aplicación del sistema, y las variables dependientes incluyeron varias métricas específicas. Estas métricas incluyeron el tiempo que tardaban los participantes en encontrar un artículo relevante de su interés, la precisión en la búsqueda (evaluando si los artículos cumplían o no con las expectativas iniciales del usuario) y la cantidad de clics necesarios para obtener información relevante en nuestro sistema.
- **Recopilación de datos:** Obtenemos los datos originales de Scopus, WoS y algunos archivos PDF y Word como nuestra muestra de estudio. Utilizamos criterios específicos para seleccionar los datos relevantes y asegurar la representatividad de la muestra.
- **Procesamiento de datos:** Utilizamos la herramienta ScientoPy para procesar los datos recopilados. Aplicamos técnicas específicas de análisis y extracción de información para obtener resultados cuantitativos y cualitativos relevantes.
- **Generación de gráficos:** Utilizamos herramientas como VOSViewer, Citenet y SNAPUJ para generar gráficos basados en los datos procesados. Nos enfocamos en representar visualmente las relaciones entre las entidades y analizar las tendencias emergentes.
- **Comparación y análisis de resultados:** Realizamos una comparación exhaustiva de los resultados obtenidos entre el grupo de control y el grupo de tratamiento.

Establecimos criterios y métricas específicas para evaluar la precisión y consistencia de las métricas obtenidas. Se llevaron a cabo pruebas estadísticas y análisis de significancia para respaldar nuestras conclusiones.

Estos pasos nos permitieron validar la consistencia y precisión de nuestro sistema. Los resultados obtenidos fueron consistentes con nuestras expectativas y demostraron que nuestro sistema supera al enfoque convencional en términos de las métricas evaluadas. Esta validación experimental proporciona un respaldo sólido a los resultados obtenidos en nuestra investigación.

7.2 Cumplimiento de requerimientos

En el anexo 4 se especifican todos y cada uno de los requerimientos funcionales de SNAPUJ. En esta sección se evaluará el cumplimiento de dichos requerimientos como parte de la validación que se ha de hacer sobre el sistema. Por simplicidad, se mencionará en la siguiente tabla el código del requerimiento y se menciona si cumplió o no cumplió.

Código	Cumplimiento
R1	Se cumple al existir una funcionalidad que evalúa el impacto de cada autor por cada área.
R2	Se cumple porque cada opción de SNAPUJ cumple una funcionalidad específica. No hay una sola que haga todo.
R3	Cada funcionalidad que lo permite tiene filtros (por ejemplo, por autores u obras). Se cumple.
R4	Se cumple luego de los procesamientos previos de la información.
R5	Cumplió mediante la funcionalidad de la métrica del índice h.
R6	Cumplió en la funcionalidad de la detección de comunidades entre artículos.
R7	Cumplió cuando se consulta cada autor a la hora de efectuar su red de colaboración.
R8	Cada tabla permite ser exportada fácilmente a formatos legibles por el usuario, por lo tanto, se cumple.
R10	Relacionado con R7.
R11	Relacionado con R5

7.3 Prueba de Uso

Se realiza un acercamiento a la aplicación, con un estudiante de la maestría de Analítica para la inteligencia de negocios, que, aunque no es el usuario final inmediato, si consideramos importante, conocer las opiniones sin sesgo alguno a partir solo de la experiencia con el uso de la herramienta.

A través de la prueba de uso, se validaron las funcionalidades que se lograron implementar en el sistema SNAPUJ, con esto, se buscó obtener una retroalimentación de las funcionalidades y usabilidad de la aplicación, lo que nos permitirá saber que tantas necesidades fueron abarcadas con el desarrollo del sistema, como también, encontrar en este estudiante un usuario en potencia, que pueda encontrar en esta herramienta un apoyo, para una futura producción científica.

Esta prueba de uso, se realizó por medio de una sesión en Teams, en esta se realizaron las preguntas del cuestionario Anexo 4. Cuestionario prueba de uso, donde se registran las respuestas obtenidas mediante la exploración de las funcionalidades.

Con base a la información obtenida, se puede evidenciar que se logro cumplir con las necesidades de la vicerrectoría, sin estas ser notificadas al estudiante previamente. Si bien, una sola prueba de uso no es suficiente para evaluar el impacto de la aplicación, al menos, nos permite conocer que tan cerca estuvimos de dar respuesta a los requerimientos establecidos en este trabajo de grado.

8. CONCLUSIONES Y TRABAJOS FUTUROS

En esta memoria se especificó todo el proceso de desarrollo del sistema SNAPUJ, desde su especificación hasta los requerimientos que debería tener para su correcto funcionamiento. Se detalló el diseño del producto, se demostró, mediante el proceso investigativo, por qué SNAPUJ es la herramienta adecuada para la necesidad de nuestro cliente y se justificó la elaboración de esta herramienta de cara a la necesidad descrita anteriormente. SNAPUJ permitirá un perfilamiento sencillo del material bibliográfico presente en las distintas fuentes, generar las redes de colaboración entre investigadores, analizar los Objetivos de Desarrollo Sostenible contra los títulos de los materiales, analizar palabras clave, detectar posibles comunidades de investigación entre los distintos materiales bibliográficos analizados, tener visualización geográfica de los autores y contar tanto con el índice h de citas como con un top de autores con más publicaciones.

Para el trabajo futuro se especificaron tres ítems, los cuales son: adición de indicadores nuevos al sistema, generación de cuadros de mando adicionales y generación de reportería adicional, incluyendo exportaciones de los análisis a diversos formatos, incluyendo Excel. Un trabajo futuro adicional, detectado en el proceso de despliegue, Streamlit debe desplegarse fuera de Colab.

En cuanto a las necesidades comunicadas por la vicerrectoría, requerimientos como comparar la producción científica sobre algún área de conocimiento realizada por la PUJ con la de otras universidades, como identificar características del investigador como la edad, el sexo y su profesión, quedan como requerimientos que podrían ser subsanados en un trabajo futuro.

Actualmente, el portal web contiene doce opciones: el análisis exploratorio completo de la información cargada, la visualización de redes (que incluye la generación del grafo y los coautores y coautores de los coautores obtenidos), la detección de comunidades (que genera clústeres según los autores), el análisis de palabras clave y el análisis de Objetivos de Desarrollo Sostenible (ODS), la obtención del índice h de un autor, visualización geográfica y el top de más publicaciones. El trabajo futuro relacionado está en añadir nuevas funcionalidades, aprovechando la arquitectura modular que posee la aplicación, inclusive, aprovechando la

flexibilidad que permite Streamlit, la aplicación puede migrar de una aplicación de única página (SPA, en sus siglas en inglés), hacia una aplicación que contenga varias páginas web, si el desempeño de la original se ve afectada por la potencial sobrecarga de datos que pueda llegarse a dar.

SNAPUJ actualmente genera sus textos únicamente en español. Como trabajo futuro, podrían implementarse librerías que ayuden con el trabajo de internacionalización, como Gettext, con el fin de que la aplicación sea accesible no solamente a usuarios hispanohablantes sino a personas que se expresen en otros idiomas.

9. REFERENCIAS

- [1] C. Schock, J. Dumler y F. Doepper, «Data Acquisition and Preparation – Enabling Data Analytics Projects within Production,» *Procedia CIRP*, vol. 104, pp. 636-640, 2021.
- [2] A. Almuhanha, W. M. S. Yafooz y A. Alsaeedi, «An Interactive Scholarly Collaborative Network Based on Academic Relationships and Research Collaborations,» *Applied Sciences*, vol. 12, n° 2, p. 915, 2022.
- [3] X. Kong, Y. Shi, S. Yu, J. Liu y F. Xia, «Academic social networks: Modeling, analysis, mining and applications,» *Journal of Network and Computer Applications*, vol. 132, pp. 86-103, 2019.
- [4] Ministerio de Ciencia, Tecnología e Innovación, «La Ciencia en Cifras - Grupos de Investigación Reconocidos,» [En línea]. Available: <https://minciencias.gov.co/la-ciencia-en-cifras/grupos>. [Último acceso: 18 Mayo 2023].
- [5] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer y R. Wirth, CRISP-DM 1.0 Step-by-step data mining guide, SPSS Inc., 2000.
- [6] M. Teunis y J.-W. Lankhaar, «Chapter 3 CRISP-DM as a guide for Data Mining and EDA,» 16 Febrero 2022. [En línea]. Available: https://rstudio-connect.hu.nl/redamoi_test/crisp-dm-as-a-guide-for-data-mining-and-eda.html. [Último acceso: 18 Mayo 2023].
- [7] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski y T. Ideker, «Cytoscape: A software Environment for integrated models of biomolecular interaction networks,» *Genome Research*, vol. 13, n° 11, pp. 2498-2504, 2013.
- [8] V. Kofia, R. Isserlin, A. M. Buchan y G. D. Bader, «Social Network: a Cytoscape app for visualizing co-publication networks,» *F1000Res*, vol. 4, p. 481, Agosto 2015.
- [9] D. Otasek, J. H. Morris, J. Bouças, A. R. Pico y B. Demchak, «Cytoscape Automation: empowering workflow-based network analysis,» *Genome Biology*, vol. 20, n° 1, 2019.

-
- [10] A. Martín Martín, E. Orduña Malea, M. Thelwall y E. Delgado López-Cózar, «Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories,» *Journal of Informetrics*, vol. 12, n° 4, pp. 1160-1177, 2018.
- [11] M. A. Javed, M. S. Younis, S. Latif, Q. Junaid y A. Baig, «Community detection in networks: A multidisciplinary review,» *Journal of Network and Computer Applications*, vol. 108, pp. 87-111, 2018.
- [12] N. Dakiche, F. Benbouzid-Si Tayeb, Y. Slimani y K. Benatchba, «Tracking community evolution in social networks: A survey,» *Information Processing & Management*, vol. 56, n° 3, pp. 1084-1102, 2019.
- [13] M. E. Rose y J. R. Kitchin, «pybliometrics: Scriptable bibliometrics using a Python interface to Scopus,» *SoftwareX*, vol. 10, 2019.
- [14] J. Ruiz Rosero, G. Ramírez González y J. Viveros Delgado, «Software survey: ScientoPy, a scientometric tool for topics trend analysis in scientific publications,» *Scientometrics*, vol. 121, n° 2, pp. 1165-1188, 2019.
- [15] Y. Jing, L. Zhao, K. Zhu, H. Wang, C. Wang y Q. Xia, «Research Landscape of Adaptive Learning in Education: A Bibliometric Study on Research Publications from 2000 to 2022,» *Sustainability*, vol. 15, n° 4, p. 3115, 2023.
- [16] C. C. Aggarwal, «An introduction to social network data analytics,» de *Social Network Data Analytics*, New York, Springer, 2011, pp. 1-14.
- [17] A.-L. Barabási y E. Bonabeau, «Scale-Free Networks,» *Scientific American*, vol. 288, n° 5, pp. 60-69, 2003.
- [18] M. S. Granovetter, «The Strength of Weak Ties,» *American Journal of Sociology*, vol. 78, n° 6, pp. 1360-1380, 1973.
- [19] B. Lutkevich y E. Burns, «What is natural language processing? - An Introduction To NLP,» Techtarget.com, [En línea]. Available: <https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP>. [Último acceso: 18 Mayo 2023].
- [20] Amazon Web Services, «¿Qué es NLP?,» [En línea]. Available: <https://aws.amazon.com/es/what-is/nlp/>. [Último acceso: 18 Mayo 2023].

- [21] Pontificia Universidad Javeriana, «Vicerrectoría de Investigación,» [En línea]. Available: <https://www.javeriana.edu.co/vicerrectoria-de-investigacion>. [Último acceso: 15 Mayo 2023].
- [22] S. Bird, E. Loper y E. Klein, *Natural Language Processing with Python*, O'Reilly Media Inc., 2009.
- [23] streamlit.io, «Streamlit · A faster way to build and share data apps,» [En línea]. Available: <https://streamlit.io/>. [Último acceso: 18 Mayo 2023].
- [24] P. Kruchten, «Architectural Blueprints—The “4+1” View Model of Software Architecture,» *IEEE Software*, vol. 12, n° 6, pp. 42-50, 1995.
- [25] F. Carrillo Carrillo y H. Alcalde Heras, «Modes of innovation in an emerging economy: a firm-level analysis from Mexico,» *Innovation: Organization & Management*, vol. 22, n° 3, pp. 334-352, 2020.

10. ANEXOS

[Anexo 1] Acta entrevista vicerrectora de investigación.

[Anexo 2] Descripción de variables de Scopus.

[Anexo 3] Casos de uso

[Anexo 4] Requerimientos.

[Anexo 5] Cuestionario prueba de uso

Anexo 1. Acta entrevista vicerrectora de investigación.

En la sesión se busca entender la necesidad del negocio, conocer que preocupaciones y/o limitaciones se tienen, de esta forma fijar el alcance del proyecto de trabajo de grado. A continuación, se relacionan las personas que hicieron parte de la sesión y sus roles dentro de la misma.

Astrid Liliana Sanchez Mejia. - vicerrectora de investigación.

Rol: Entrevistada

Alexandra Pomares Quimbaya. – directora de investigación.

Rol: Entrevistada y participante

Luis Javier López González – Estudiante maestría en ingeniería de sistemas y computación.

Rol: Entrevistador

Frank Sebastián Franco Hernández – Estudiante maestría en ingeniería de sistemas y computación.

Rol: Entrevistador

Miguel Ángel Gutiérrez Ibagué– Estudiante maestría en ingeniería de sistemas y computación.

Rol: Entrevistador

Diana Marcela Herrán Giraldo– Estudiante maestría en ingeniería de sistemas y computación.

Rol: Entrevistadora

¿Cuánto tiempo llevas trabajando en la universidad?

Empecé a trabajar en la universidad desde el año 2003.

¿Qué cargo tienes actualmente?

Soy la vicerrectora de investigación del departamento de vicerrectoría de investigación.

¿Cuánto tiempo llevas en este cargo actualmente?

Desde abril del 2022 pasé a la vicerrectoría de investigación, antes estaba en la facultad de ciencias jurídicas, como profesora de la facultad.

¿Qué responsabilidades tienes alrededor de este cargo?

Liderar los procesos que están relacionados con el fomento de la investigación, la innovación, la creación artística y el emprendimiento en la universidad.

¿Para el desempeño de este rol tienes experiencia en el uso de redes sociales académicas? ¿Usas alguna en particular? ¿Te apoyas en alguna en particular para realizar alguna tarea que quieres ejercer?

Si, Me apoyo en redes sociales académicas especializadas como Academia, y sistemas de revistas indexadas como Scopus y WoS, tengo la experiencia de haber realizado publicaciones, antes de llegar a la vicerrectoría, dirigí un doctorado en ciencias jurídicas; Para lograr su posicionamiento y proceso de acreditación, hicimos un análisis sobre la producción de investigaciones por parte de los profesores y la presencia de ellos en ciertos sitios, Scopus fue uno de los sitios pero también lo fue Google Scholar, ya que este último a nivel de ciencias es más importante. Me apoyé en esos datos obtenidos para generar un análisis para procesos de acreditación y para hacer la planeación de la investigación de la facultad de ciencias jurídicas.

¿Cuándo hacías este análisis tuviste algún inconveniente con la data obtenida? ¿Al momento de descargarla o en su almacenamiento?

No, en general, en scopus y WoS se podían extraer de manera sencilla, pero de Google scholar era menos sencillo, había una persona encargada de descargar los datos desde estas fuentes y los organizaba en la base de datos, también había un asistente que se encargaba de organizar la información que venían desde CvLAC (registros de los investigadores) y GrupLAC (registros de grupo de investigación) que se encuentra en la plataforma ScienTI.

¿La descarga de los datos se realizaba de forma manual o por medio de alguna herramienta de forma automática? ¿Usas algún tipo de métrica que te permita hacer tu análisis?

Entiendo que había algunos que se podían extraer directamente de alguna de estas bases, pero había otros que exigían hacerlo de forma manual. Nosotros usamos dos variables de análisis frente a los textos, una era número de citaciones y la otra, el índice H (Es el resultado del cruce del número de publicaciones y el número de las citas), la información la podíamos sacar directamente de Scopus o Google Scholar, en este último aparecía listado de las publicaciones y cuantas citaciones tiene, lo que no puedes verificar son las auto citaciones.

¿Actualmente en la universidad, cuentas con alguna herramienta que te permita monitorear estos indicadores?

No, actualmente no. Se ha venido haciendo de forma manual ya que el número de profesores en el doctorado eran de 18 a 20, ahora, ya en la vicerrectoría es cuando estamos hablando de 1800 profesores por lo cual se vuelve necesaria una herramienta que nos permita obtener esa información y así poder generar algún reporte.

¿Qué esperarías obtener de una herramienta que consolide esta información?

Lo que esperamos es tener los datos disponibles de manera sistemática que nos permitan fundamentar las decisiones y políticas que se definan en la PUJ, es importante tener los datos agregados de la universidad, pero también los datos desagregados como facultad, departamento, incluso individual por profesor, con esto poder realizar un monitoreo desde la perspectiva de cada una de las unidades.

¿Qué características principales te gustaría que se conocieran de los investigadores?

La edad, el sexo y si es posible su profesión.

¿Deseas conocer quién es la persona que cito al investigador, de que país y universidad es?

Si, Esto permite ver la red académica y con quienes están interactuando nuestras profesoras y profesores, con esto podremos saber con qué universidades y países tenemos cooperación.

¿Te gustaría poder visualizar métricas sobre quiénes son los investigadores o facultades con mayor cantidad de publicaciones?

Más que estos dos que mencionas, nos gustaría identificar por área de conocimiento y hacer las comparaciones de esa misma área de conocimiento entre las facultades, posteriormente, comparar con la producción del área de conocimiento de otras universidades, para saber que tan productivos son nuestros profesores.

Otro análisis en el que se podría pensar es por los temas de las investigaciones, más allá que el área de conocimiento, para saber cuáles son los temas más recurrentes y cuáles son las áreas en las que más se está investigando y escribiendo en la PUJ.

¿Cuál es el tipo de impacto que deseas medir con base a los temas encontrados en los artículos?

Por medio de los *abstract* se podría obtener palabras claves como potencial de impacto que ayudaran a identificar y categorizar los temas, caracterizar temas como los ODS, apropiación social, comunidades e innovación. Es importante conocer datos demográficos como municipios, departamento o regiones del país, para análisis de perspectivas territoriales, como también información de la facultad, departamento y grupo de investigación que realiza la investigación.

¿Qué tipo de reportes te gustaría generar en función a este conjunto de datos?

Reportes visuales, pero también que se pueda descargar las bases de datos en formatos como csv, que me permita posteriormente hacer análisis adicionales.

¿Qué esperas realizar con los datos obtenidos?

La idea es compartirlos con otros tomadores de decisiones en la PUJ, pero también generar piezas de comunicación y divulgación con los datos para crear visibilidad y también para la toma de decisiones.

¿Con que frecuencia usarías la herramienta a desarrollar?

Mínimo una vez al día.

Anexo 2. Descripción de variables de Scopus.

Variable	Descripción
affiliation_city	Ciudad de la afiliación
affiliation_country	País de la afiliación
Affilname	Nombre de la afiliación
Afid	ID de la afiliación
aggregationType	Tipo de agregación
article_number	Número del artículo
authkeywords	Palabras claves del autor
author_afids	ID de las afiliaciones del autor
author_count	Cantidad de autores
author_ids	IDs de los autores
author_names	Nombres de los autores
citedby_count	Cantidad de citas
coverDate	Fecha de publicación
coverDisplayDate	Fecha de publicación (visualización)
Creator	Creador
description	Descripción
Doi	DOI

eIssn	ISSN Electrónico
Eid	Eid
freetoread	Acceso gratuito
freetoreadLabel	Etiqueta de acceso gratuito
fund_acr	Acrónico del fondo de financiamiento
fund_no	Número del fondo de financiamiento
fund_sponsor	Patrocinador del fondo de financiamiento
Issn	ISSN
issueIdentifier	Identificador del número de la publicación
Openaccess	Acceso abierto
pageRange	Rango de páginas
Pii	Pii
publicationName	Nombre de la publicación
pubmed_id	ID de pubmed
source_id	ID de la Fuente
Subtype	Subtipo
subtypeDescription	Descripción del subtipo
Title	Título
Volumen	Volumen

Anexo 3 Casos de uso

Có- digo	Caso de uso
C1	Consultar área de conocimiento más investigada por facultad
C2	Visualizar información desagregada
C3	Filtrar la información de las consultas realizadas
C4	Consultar información no duplicada después de los filtros realizados
C5	Conocer el impacto de los investigadores adscritos a la universidad
C6	Conocer con que comunidades trabaja una facultad y/o un investigador
C7	Conocer los diferentes investigadores con nombres desambiguados
C8	Exportar resultados de las consultas a Excel
C9	Actualizar la información consultada al menos una vez al mes
C10	Consultar las redes de colaboración por grupos
C11	Generar índices de números de citaciones, índice H, entre otros

Anexo 4 Requerimientos funcionales

Có- digo	Requerimiento funcional
R1	El sistema debe permitir la consulta del área de conocimiento más investigada por facultad, para que los usuarios puedan conocer las áreas de investigación más activas de cada facultad.
R2	El sistema debe permitir la visualización de información desagregada, para que los usuarios puedan obtener detalles específicos y relevantes de la información consultada.
R3	El sistema debe permitir realizar filtros a las consultas, para que los usuarios puedan obtener información específica y relevante.
R4	El sistema debe permitir consultar información no duplicada después de los filtros realizados, para evitar la duplicación de información en las consultas.
R5	El sistema debe permitir conocer el impacto de los investigadores adscritos a la universidad, para que los usuarios puedan conocer la producción científica y la relevancia de los trabajos de investigación realizados por los investigadores de la institución.

R6	El sistema debe permitir conocer con qué comunidades trabaja una facultad y/o un investigador, para que los usuarios puedan conocer las colaboraciones de cada facultad o investigador con otras comunidades.
R7	El sistema debe permitir a los usuarios conocer los diferentes investigadores adscritos a la universidad, donde cada investigador esté identificado de forma única y desambiguada, para que los usuarios puedan consultar la información de los investigadores de forma precisa.
R8	El sistema debe permitir a los usuarios exportar los resultados de las consultas realizadas a un archivo Excel.
R9	El sistema debe actualizar la información consultada al menos una vez al mes, para que los usuarios puedan tener acceso a información actualizada y precisa.
R10	El sistema debe permitir a los usuarios consultar las redes de colaboración entre diferentes grupos de investigación, para que puedan conocer las colaboraciones existentes entre ellos.
R11	El sistema debe permitir a los usuarios generar índices como número de citas e índice H para cada investigador, para que puedan conocer la relevancia de los trabajos de investigación realizados por los investigadores de la institución.

Anexo 5. Cuestionario prueba de uso.

A continuación, se indican los comentarios recibidos por parte del estudiante por funcionalidad.

Cuestionario resuelto por el estudiante:

Juan Sebastián Quiroga Bernal, quien actualmente cursa

Programa: maestría de analítica para la inteligencia de negocios.

Semestre: Tercer semestre de la

Correo: jsebastianquirogab@javeriana.edu.co

Funcionalidades:

A continuación, se indican las preguntas generalizadas que se aplicarán en cada funcionalidad:

1. ¿Es relevante la información que visualizas?
2. ¿La información es clara o necesita más descripción?
3. ¿Crees que esta funcionalidad es importante incluirla en la aplicación?
4. ¿Qué cambios sugieres para esta funcionalidad?

Exploración de datos de Scopus

1. Rta:// la información de datos previos no es clara, se debe tener algún insight que permita entender lo que se esta mostrando, aunque puede ser información enriquecedero para el usuario administrativo final, la veo desconectada, por otro lado, el análisis que muestra de variables relevantes me parece importante.
2. Rta:// si requiere más descripción.
3. Rta:// Si se eliminan los datos previos creo que quedaría mas claro.
4. Rta:// Agregar *Story Telling* y cambiar el nombre de la funcionalidad.

Exploración de datos de WoS y Scopus

1. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior.
2. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior.
3. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior.
4. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior.

Exploración de Proyectos de Investigación

1. Rta:// Si, es relevante.
2. Rta:// Requiere una descripción de los que se muestra en la funcionalidad.
3. Rta:// Si, veo que puede ser útil para las personas que van a realizar una investigación.
4. Rta:// Solo Añadir la descripción.

Redes de Investigadores

1. Rta://Si, bastante.

2. Rta://Adicionar un hipervínculo que permita conocer un poco más de la visualización por grafos, puede que un usuario no sepa cómo interpretar la información.
3. Rta://Si, es el mayor valor agregado, que he visto hasta ahora.
4. Rta://No, está bien elaborada.

Red de afiliaciones de la PUJ

1. Rta:// si, es relevante.
2. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior.
3. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior.
4. Rta:// Sugiero incluir en una sola funcionalidad la visualización de redes de investigadores y red de afiliaciones de la PUJ.

Segmentación de artículos

1. Rta://si, es relevante.
2. Rta://Esta información es bastante clara, las nubes de palabras te dicen todo.
3. Rta://Si, da información de lo que más se habla en la universidad.
4. Rta://No, ningún cambio.

Análisis de palabras claves

1. Rta:// Si, bastante relevante, al igual que los grafos es lo más relevante.
2. Rta:// Es clara.
3. Rta:// Si, de las más importantes.
4. Rta:// Ninguna, me gusta mucho.

Relación de ODS con artículos

1. Rta:// si, es muy relevante.
2. Rta:// es bastante clara.
3. Rta:// mucho.
4. Rta:// ninguno, es muy intuitiva.

Índice H de citaciones

1. Rta://si, muy relevante.
2. Rta://si, es clara, aunque realizando una consulta de un autor, múltiples nombres aparecieron.
3. Rta://Si. Da información con enfoque organizacional.
4. Rta://Ninguno.

Top de autores con más publicaciones

1. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior.
2. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior.
3. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior.
4. Rta:// Agregar más filtros, por facultad o área de conocimiento

Visualización geográfica por países

1. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior
2. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior
3. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior
4. Rta:// No, es bastante intuitivo, información muy útil.

Visualización geográfica por ciudades de Colombia

1. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior
2. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior
3. Rta:// Aplicar la respuesta de la funcionalidad inmediatamente anterior
4. Rta:// Se puede integrar con la funcionalidad anterior.