UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

MAESTRÍA EN INFORMÁTICA

TESIS

EXTRACCIÓN AUTOMÁTICA DE METADATOS PARA LA ADMINISTRACIÓN DEL REPOSITORIO INSTITUCIONAL DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO

PRESENTADA POR:

ALAIN PAUL HERRERA URTIAGA

PARA OPTAR EL GRADO ACADÉMICO DE:

MAGISTER SCIENTIAE EN INFORMÁTICA
MENCIÓN EN GERENCIA DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES

PUNO, PERÚ

2022

DEDICATORIA

Al ser más supremo del universo, por concederme vida y sabiduría. Dios nuestro señor todo poderoso.

Con respeto y admiración a mi madre y a mi mamita:
Victoria Urtiaga Chambi y María Concepción Chambi Ticona por su invalorable amor y sacrificio; y su constante apoyo incondicional en mi formación profesional y humana.

A mis hermanos, hermanas por el aliento constante en mi formación profesional.

A mis compañeros de trabajo con los cuales compartimos conocimientos y experiencias, lo que constituye un aliento y ánimo para la realización del presente trabajo

Alain Paul.

AGRADECIMIENTOS

A los docentes de la Facultad de Ingeniería Estadística e Informática de la Universidad Nacional del Altiplano por compartir sus conocimientos con sus estudiantes y contribuir en la formación profesional, por absolver cada uno de mis dudas, por su paciencia y calma en las sesiones de aprendizaje, mi cariño, respeto y admiración por cada una de ellos.

Un agradecimiento muy grande también a mi asesor Dr. Charles Ignacio Mendoza Mollocondo por sus sugerencias y aporte en la mejora de este trabajo de investigación.

A los miembros del jurado, por sus sugerencias y aportes en la mejora de este trabajo de investigación.

Un agradecimiento muy grande al Dr. Vladimiro Ibañez Quispe y al Dr. Bernabe Canqui Flores, quienes brindaron su apoyo en la realización de este trabajo de investigación.

ÍNDICE GENERAL

Puno, 22 de julio de 2022

ÁREA: Ingeniería del Software
TEMA: Extracción automática de metadatos para la administración del Repositorio Institucional de la Universidad Nacional del Altiplano
LÍNEA: Desarrollo del Software

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Pág.

ÍNDICE DE ANEXOS

Pág.

RESUMEN

Los Repositorios Institucionales permiten organizar y preservar la producción científica de una Institución, la presente investigación tiene como finalidad optimizar la extracción de metadatos y publicación de documentos de investigación procesos fundamentales para la administración de Repositorios Institucionales que requieren de tiempo, mediante la implementación del software "E-MeRI", cuya población se compone por 1518 documentos de investigación. Para el desarrollo del sistema se utilizó la programación por capas y para el contraste de la hipótesis se utilizó prueba t para muestras relacionadas. Con respecto a la extracción automática se elaboró un algoritmo mediante técnicas de procesamiento de lenguaje natural, al cual se determinó la complejidad algorítmica lineal O(n) y demostró ser eficiente en comparación a otras herramientas extractoras. A la misma vez se determinó el nivel de precisión entre 96% y 99% de resultados correctos en base a las métricas *Precisión y Recall*. De la diferencia del tiempo de extracción, el sistema logra reducir en 5 minutos y 21 segundos por documento y permitió extraer en un minuto 4 documentos. Se concluye que la extracción automática de metadatos y la publicación de documentos de investigación mejoran la administración del Repositorio Institucional de la Universidad Nacional del Altiplano, reduciendo el tiempo de extracción y publicación de forma significativa con un valor $p\,(0.000) < \alpha = 0.05$, además la evaluación del software basado en la norma ISO 25000 obtuvo un valor de 8.93 de calidad total, logrando un nivel cumple con los requisitos y un grado muy satisfactorio.

Palabras clave: Algoritmos, Herramienta, Extracción automática, Metadatos Procesamiento de lenguaje natural.

ABSTRACT

The purpose of this research is to optimize the extraction of metadata and publication of research documents, fundamental processes for the administration of Institutional Repositories that require time, through the implementation of the "E-MeRI" software, whose population is composed of 1,518 research documents. For the development of the system, layered programming was used and for the contrast of the hypothesis, t-test for related samples was used. With respect to automatic extraction, an algorithm was developed using natural language processing techniques, whose linear algorithmic complexity O(n) was determined and proved to be efficient in comparison with other extraction tools. At the same time, the level of accuracy was determined between 96% and 99% of correct results based on the metrics Precision and Recall. From the difference in extraction time, the system managed to reduce by 5 minutes and 21 seconds per document and allowed to extract 4 documents in one minute. It is concluded that the automatic extraction of metadata and the publication of research documents improve the administration of the Institutional Repository of the National University of the Altiplano, reducing the extraction and publication time significantly with a p-value (0.000)< α=0.05, also the evaluation of the software based on the ISO 25000 standard obtained a value of 8.93 of total quality, achieving a level meets the requirements and a very satisfactory grade.

Keywords: Algorithms, Automatic extraction, Metadata, Natural language processing, Tool.

INTRODUCCIÓN

La disponibilidad de grandes depósitos de documentos electrónicos, accesibles y diversos desde la web están aumentando rápidamente. Gran cantidad de esta información está en forma de texto no estructurado, lo que dificulta la consulta de la información, pero en la actualidad hay formas más estructuradas de acceder a la información. En este sentido emergen cuatro conceptos fundamentales: objetos de aprendizaje, metadatos, estándares y repositorios institucionales. Los usos de los metadatos se han popularizado debido a la aparición de los recursos digitales y recursos de información electrónica. La definición de los metadatos más común es "datos sobre datos", sin embargo, según (Bermudez, 2010), define a los metadatos en la siguiente forma: Metadata (Pronunciado de igual forma) es data que describe otra data.

Los Repositorios institucionales son archivos electrónicos de la producción científica de una institución, la cual se almacena en un formato digital, donde se permite la búsqueda y la recuperación para difundir los recursos académicos y científicos de una institución. Los Repositorios son medios que nos permiten importar, identificar, almacenar, preservar, recuperar y exportar un conjunto de objetos digitales, usualmente desde una página web (Bustos-Gonzalez & Porcel, 2007).

El Repositorio Institucional de la UNA Puno tiene como objetivo almacenar toda la producción científica y académica, a través de la administración en la cual están involucrados la publicación de documentos de investigación, para lo que es necesario el proceso de extracción de metadatos, el cual se encarga de obtener los atributos que identifican a cada documento (Senso & Rosa Piñero, 2003). Este proceso se realiza manualmente por el personal encargado, lo cual requiere de tiempo y pueden demorar dependiendo de la cantidad de documentos de investigación a los cuales serán extraídos los metadatos. Conforme a lo mencionado, esta investigación aborda la temática de extracción de metadatos relacionándose con la línea de investigación: tecnologías de información.

Para dar solución a esta problemática, en esta investigación se presenta la optimización del proceso de extracción de metadatos y publicación de documentos de investigación que permita mejorar la administración del Repositorio Institucional mediante el desarrollo de un software haciendo uso del procesamiento de lenguaje natural, en el cual fueron aplicados a 380 documentos de investigación que han sido publicados en el 2021. Por

otro lado, para la implementación del software se utiliza el lenguaje de programación Python y el framework Flask, herramientas de código abierto, por lo que reduce los costos en el desarrollo del software.

Por lo expuesto, esta investigación está orientada a la extracción automática de metadatos de los documentos de investigación que son presentados y publicados en el Repositorio Institucional de la Universidad Nacional del Altiplano, debido a que son procesos fundamentales en la administración de repositorios haciendo énfasis en documentos de tesis de pregrado y posgrado

La investigación está dividida por cuatro capítulos, que se exponen a continuación: Capítulo I, en este capítulo se detallan, los conceptos sobre la extracción de metadatos, así como el uso del procesamiento de lenguaje natural. Capítulo II, en este capítulo se identifica el problema, justificación y objetivos de estudio por los cuales se va a desarrollar la investigación. Capítulo III, en este capítulo se describe la metodología de la investigación. Capítulo IV. se describe los resultados y discusión. Los resultados se presentan por objetivos específicos, a la misma vez se da la interpretación de la información contenida en tablas, figuras y algoritmos, demostrando la aceptación o rechazo de la hipótesis mediante la prueba de hipótesis, la conclusión se desarrolló de acuerdo a los objetivos específicos y se redactó en párrafos finalmente las recomendaciones se presentaron en relación a los objetivos específicos, teniendo en cuenta las orientaciones y medidas a realizarse

CAPÍTULO I

REVISIÓN DE LITERATURA

1.1 Marco Teórico

1.1.1 Repositorio Institucional

Los repositorios son sitios en donde se almacena y resguarda información de forma centralizada y son accedidos principalmente desde redes informáticas o de internet (Rivera, 2009). La definición anterior es un poco general, por lo tanto, se puede detallar más con la propuesta de Barrueco & Garcia (2009) quienes definen los repositorios institucionales digitales como: Aquellos servicios prestados por las universidades, al conjunto de la comunidad, para recopilar, administrar, difundir y preservar la producción documental digital generada en la institución, cualquiera que sea su tipología, a través de la creación de una colección digital organizada, abierta e interoperable a través del protocolo OAI-PMH, para garantizar un aumento de la visibilidad e impacto de la misma.

En esta misma conceptualización, Chazarra, Requena, & Valverde (2010) indican que un repositorio digital es una herramienta tecnológica que hace uso de internet para facilitar el acceso a los contenidos desde cualquier ubicación del usuario, como señalan textualmente: Un repositorio de contenidos digitales es un sistema que hace uso de Internet, que sirve para almacenar y controlar la información guardada en los contenidos digitales y que facilita el acceso de sus usuarios a estos contenidos, generalmente desde cualquier lugar del mundo.

De Giusti, Lira, Oviedo, Villarreal, & Texier, (2012) coinciden con la definición brindada por Rivera (2009) y señalan que los repositorios institucionales digitales deben ser interoperables con otros repositorios semejantes, tal como se evidencia en la siguiente cita: Una infraestructura web capaz de brindar un conjunto de servicios

a una comunidad, destinados a recopilar, gestionar, difundir y preservar contenidos a través de una colección organizada y accesible en abierto que debe estar provista de facilidades que le permiten interoperar con otros repositorios similares.

1.1.2 Repositorios Institucionales en Perú

El Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto Ley Nº 30035, "ofrece acceso abierto al patrimonio intelectual resultado de la producción en materia de ciencia, tecnología e innovación realizada en entidades del sector público o con financiamiento del estado" (Congreso De La República Del Perú, 2013).

La Ley que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de acceso abierto, aprobada en marzo del año 2013, concede obligatoriedad en nuestro país para publicar los resultados de todas las investigaciones científicas financiadas con fondos públicos, los cuales deben ponerse a disposición en repositorios digitales de acceso abierto (Congreso de la República del Perú, 2013). La dación de esta normativa ha posicionado al Perú como el segundo país de América Latina seguido de Argentina en elevar una legislación nacional sobre este tema. Como parte de la implementación de la ley CONCYTEC (Consejo nacional de ciencia, tecnología e innovación tecnológica) ha sido responsable de la coordinación e integración de toda la documentación científica nacional en el Repositorio Nacional Digital denominado ALICIA (Acceso libre a información científica para la innovación), cuya plataforma empezó a funcionar en abril de 2014.

El objetivo del Repositorio Nacional ALICIA es conformar una red interoperable de repositorios institucionales, a partir del establecimiento de políticas, estándares y protocolos para el intercambio de información comunes a todos los integrantes de la Red, esta actividad recae en el CONCYTEC, organismo gubernamental que implementa, integra, estandariza, almacena, preserva y gestiona la información; además establece las políticas para su seguridad y sostenibilidad, dispone estándares de interoperabilidad y promueve el uso y aprovechamiento de dicha información (ALICIA CONCYTEC, 2021). El Repositorio Nacional ALICIA tiene como meta albergar en formato digital toda la producción científica - tecnológica del país, la cual permite la búsqueda y recuperación de la información científica en forma libre y sin restricciones dentro o fuera del país.

1.1.3 Software para repositorios

Los repositorios institucionales digitales se basan en la administración y el almacenamiento de contenidos digitales; además de proporcionar acceso a los usuarios, estos pueden ser de tipo abierto o protegido, para lograr lo anterior, es necesario un software que permita gestionar todas las acciones requeridas por las instituciones, como, por ejemplo: almacenar y descargar documentos, gestionar contenidos, entre otros, como se observa en la figura 1, DSpace es el software desarrollado bajo licencia como el software libre más utilizado a nivel mundial, sin fines de lucro, con el objetivo del uso en el ámbito académico, para instituciones, organizaciones comerciales y repositorios de acceso abierto, en resumen, permite un acceso fácil y abierto a todo tipo de contenido digital (OpenDoar, 2021).

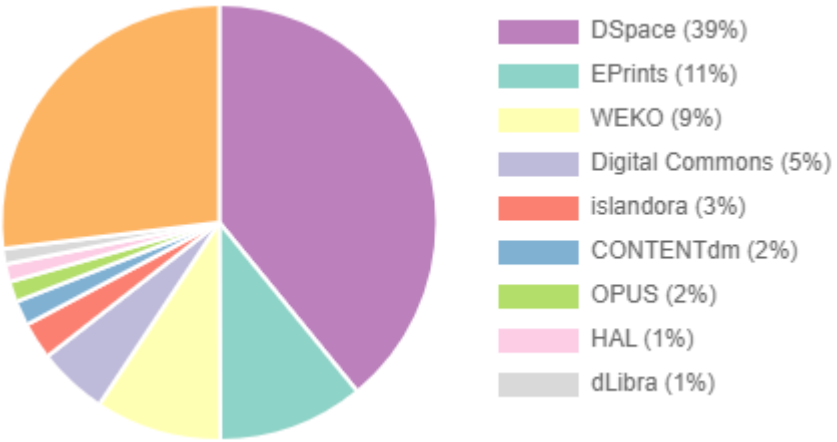


Figura 1. Software más utilizado por los repositorios institucionales

Fuente: (OpenDoar, 2021)

1.1.4 Administración de Repositorios Institucionales

Según Nkiko, Bolu, & Michael-Onuoha (2014) manifiesta que "categorizar, clasificar documentos, conversión de PDF y la carga son los procesos involucrados en la administración de un Repositorio Institucional (RI)", indicaron además que las principales actividades rutinarias de administración se centran en la extracción de metadatos y publicación de documentos. Los repositorios institucionales pueden servir como una empresa académica para la institución de educación superior, expandiéndose y creciendo con el tiempo para servir no solo a la comunidad de la

institución local, sino también a las partes interesadas dentro de la comunidad en general. Sin embargo, Lynch (2005) habían advertido que los RI pueden fallar con el tiempo, principalmente debido a: (i) falta de financiamiento, (ii) administración incompetente y probablemente (iii) problemas técnicos. A lo largo de los años, los investigadores han puesto énfasis en el estado de implementación, desarrollo, experiencia y desafíos e identificación de factores críticos de éxito. Greene (2010) había observado que muy pocos estudios se habían centrado en investigar las actividades involucradas en el manejo de los RI. Por otro lado, Nabe (2012) menciona que los repositorios institucionales bien administrados podrían ayudar a la institución a comercializar su investigación y tener un mejor impacto visual a través de los procesos relacionados a la clasificación, extracción de metadatos y publicación de documentos.

1.1.5 Protocolo OAI-PMH

Para una recuperación y búsqueda de información existen diversos mecanismos para generar y obtener metadatos del internet, el más usado es el protocolo OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting) que puede implementarse en cualquier sistema que requiriese la comunicación de metadatos. El protocolo OAI-PMH es un mecanismo sin barreras para la interoperabilidad entre repositorios y adhiere al esquema de metadatos Dublin Core sin calificar (DC). (Testa & Degiorgi, 2013). En la cosecha o harvesting existen dos clases de participantes:

a. Proveedores de datos: Son aquellos repositorios que exponen metadatos estructurados de los objetos que poseen.

b. Proveedores de servicios: Son los repositorios que cosechan esos metadatos y le otorgan al usuario servicios de valor añadido, tales como sistemas de búsquedas e identificación, alertas y estadísticas de uso e impacto.

El protocolo OAI-PMH, para comunicarse utiliza transacciones HTTP (Hiper Text Transfer Protocol) para la transferencia de contenidos Web, la cual se basa en un modelo cliente/servidor que transmite preguntas y respuestas entre un proveedor de datos y un proveedor de servicios (The OAI Executive, 2015). OAI-PMH es el protocolo preferido por las redes de repositorios institucionales más importantes por su simplicidad, adaptabilidad y también porque tiene las funciones básicas necesarias

para interoperar. Mediante el uso de verbos o acciones permite obtener: la identificación del repositorio, lista de conjuntos de datos, identificadores de objetos, registros en distintos formatos, registros actualizados desde una fecha dada y posibilita, además, combinar las distintas acciones para obtener distintos conjuntos de datos. (Pérez Velandia & Felipe Silva, 2007).

1.1.6 Metadatos

La palabra metadato en su primer significado que se le dio y es el que en la actualidad es el más frecuente, fue la de "dato sobre datos", debido a que proporcionan la información mínima necesaria para identificar un recurso. Según Bermudez (2010) define a los metadatos en la siguiente forma: Metadata (Pronunciado de igual forma) es data que describe otra data, es información que describe el contenido un archivo u objeto. Por ejemplo, una imagen digitalizada de una orden de compra es la data, la descripción de este documento, como lo es el número de la orden de compra, dirección física, nombre a quien va dirigido, fecha todo esto sería la metadata (Martínez Arellano, 2017), las páginas de web pueden incluir metadata (conocido como meta tags) los cuales describen el contenido de una página. La metadata se utiliza frecuentemente para indexar información en una base de datos para localizar fácilmente un documento, archivo u objeto

La National Information Standards Organization (NISO) dice que "referirse en estos momentos a metadatos es hacerlo con el fin de significar las estructuras de los archivos, conjuntos de datos u otra entidad de información que aseguran la descripción técnica que se necesita para representar las partes del objeto digital" (NISO, 2019).

Desde el punto de vista de las Ciencias de la Documentación y la Información, Pavani (2009) define que "los metadatos son un conjunto de atributos de catalogación de los documentos, que permiten su identificación sin tener que ejecutarlos", sobre esta definición se debe aclarar que, en este contexto, ejecutar, se refiere a la apertura del documento digital para conocer su contenido. El registro de metadatos realizado correctamente permite conocer el contenido de un documento digital sin tener que verlo o leerlo por completo.

1.1.7 Estándar de metadatos Dublin Core

El esquema de metadatos más usado a nivel mundial es Dublin Core. Fue desarrollado en 1995 por (DCMI, 1995) este estándar procura ser un conjunto básico de elementos de metadatos para facilitar la recuperación de objetos. El conjunto de elementos de Dublin Core está compuesto por 15 elementos de metadatos divididos en tres grupos:

- Contenido: Agrupa los metadatos utilizados para describir el contenido del recurso como son: título, tema, descripción, fuente, idioma, relación y cobertura.
- Propiedad intelectual: Agrupa los metadatos utilizados para identificar el autor, editor, colaborador y derechos.
- Instanciación: Agrupa los metadatos utilizados para identificar a esa instancia propiamente dicha, como fecha, tipo, formato y identificador en caso que pueda existir réplicas del recurso.

En este estándar todos los elementos son opcionales, lo que permite usar algunos de los elementos para tener descripciones simples.

Tabla 1

Elementos del estándar de metadatos Dublin Core

Nombre	Etiqueta	Definición
Title/Titulo	dc.title	Nombre del objeto
Source/Fuente	dc.source	Objetos, impresos o electrónicos a partir de los cuales se derivó este objeto
Language/Lenguaje	dc.language	Idioma del contenido intelectual
Relation/Relación	dc.relation	Relación del presente objeto con otros
Coverage/Cobertura	dc.coverage	Características espaciales y/o temporales del contenido intelectual del objeto.
Description/Descripción	dc.description	Descripción textual del contenido del objeto
Subject/Tema	dc.subject	Tópico tratado por el trabajo
Creator/Creador	dc.creator	Persona responsable del contenido intelectual del trabajo

Publisher/Editor	dc.publisher	Entidad responsable de poner a su disposición el objeto en su formato actual
Rights/Derechos	dc.rights	Información sobre los derechos del autor del objeto, para que el usuario final conozca sus condiciones de uso y acceso
Contributor/Contribuidor	dc.contributor	Personas tales como editores o correctores que hicieron contribuciones intelectuales significativas a los objetos
Date/Fecha	dc.date	Fecha asociada con la creación o publicación del objeto
Format/Formato	dc.format	Forma física o digital en que se presenta el objeto
Identifier/Identificador	dc.identifier	Cadena o número utilizado para identificar el objeto de manera única
Type/Tipo	dc.type	Categoría del objeto tal como, por ejemplo, tesis, libros, artículos, etc

Fuente: Dublin Core metadata Initiative (DCMI, 2018)

1.1.8 Extracción de metadatos

La extracción de metadatos se ha convertido en un problema abierto y de difícil solución, esto se debe a variedad de tipos de recursos, los diferentes formatos de archivos utilizados y falta o diversidad de estructura en los mismos. Este problema ha sido parcialmente abordado en (Pire, Deco, Casali, & Espinase, 2011) donde se analizaron algunos sistemas dedicados a la extracción automática de metadatos educativos de objetos de aprendizaje, que aunque son relevantes, algunas no están implementadas o no están disponibles como herramientas libres. Esto, sumado a la estandarización de metadatos y el auge de los repositorios institucionales y del acceso abierto al conocimiento que se vislumbró y describió anteriormente, da como resultado el fundamento necesario para comprender la verdadera importancia del desarrollo de nuevos algoritmos para la extracción automática en repositorios institucionales. (Pinilla Gómez, 2017) en base esto también se menciona que hay tres aspectos importantes que se deben tener en cuenta al momento de seleccionar o diseñar un sistema de extracción automática de metadatos:

- Los tipos de archivo que serán procesados (como, html, txt, pdf, doc, etc.).
- Los metadatos a ser extraídos, los cuales serán considerados enfocándose en especial atención en aquellos a nivel educativo.
- Las técnicas y recursos utilizados para realizar la extracción, como el uso de herramientas para el procesamiento de lenguaje natural (NLP), ontologías entre otras.

Herramientas utilizadas en la extracción de metadatos

Existen diferentes propuestas existentes para la extracción automática de metadatos, muchas de estas propuestas son relevantes, además algunas no están implementadas o no están disponibles para su uso libre.

SAFEX (System for Automatic eXtraction of E-learning object Features)

Sistema creado por The Center on Communication Studies (Univ. Palermo, Italia), que automáticamente extrae indicadores didácticos de cualquier página Web. (Alfano, Lenzitti, & Visalli, 2007)

AlchemyAPI

AlchemyAPI se encuentra en (http://www.alchemyapi.com/, es una plataforma de minería de texto la cual proporciona un conjunto de herramientas que permiten el análisis semántico utilizando técnicas de procesamiento de lenguaje natural y machine learning, más precisamente algoritmos de deep learning (Deng & Yu, 2014). Proporciona un conjunto de servicios que permiten el análisis de forma automática de documentos de texto, esta herramienta presenta varios servicios a través de su RESTful API (http://www.alchemyapi.com/api/calling.html), entre ellos se encuentran:

- Identificación del autor
- Identificación de entidades
- Generación de palabras calves
- Categorización del contenido
- Identificación del idioma

AlchemyAPI presenta su versión gratuita, el servicio tiene una limitación de 1000 consultas por dia y un lımite por consulta de 150 kbs.

KEA Automatic Keyphrase Extraction

KEA (http://community.nzdl.org/kea/), es la implementación en JAVA del algoritmo KEA desarrollado por Witten *et al.* (1999) donde presenta la herramienta la cual extrae automáticamente frases claves del texto completo sobre el documento a analizar, el conjunto de todas las frases seleccionadas en un documento en el cual se identifican haciendo uso del procesamiento léxico rudimentario. A la misma vez utiliza técnicas de machine-learning para generar un clasificador que determina que frases candidatas son las que deben ser asignadas como frases clave. Esta herramienta puede ser utilizada localmente y se necesita una fase previa de entrenamiento.

ParsCit

Es una herramienta de código abierto que permite el análisis de referencias bibliográficas. ParsCit realiza el análisis examinando cada referencia e identificando cada campo que lo compone. Consta de dos tareas para la extracción de referencias, el preprocesado y el postprocesado para el procesado ParsCit utiliza métodos heurísticos para convertir el documento en formato PDF a texto plano, luego en el procesado utiliza CRF++, implementación del método de aprendizaje automático CRF, para obtener cada uno de los tokens que componen la referencia (Cartic, Ramakrishnan; Abhishek, Patnia; Eduard, Hovy;Gully, 2012). Esta herramienta puede ser utilizada tanto como servicio web o como una aplicación independiente.

Mr Dlib

Mr Dlib ubicada en (http://mr-dlib.org/) es una biblioteca digital que facilita el acceso a una gran cantidad de artículos de texto completo y sus metadatos en formato XML y JSON mediante un servicio web RESTful, en su etapa beta de desarrollo, sus funcionalidades son utilizadas por terceros y permite extraer Tıtulo y Autores (Beel et al., 2011).

1.1.9 DSpace

DSpace es un software de código abierto desarrollado por el Massachusetts Institute of Technology (MIT) y los laboratorios Hewlett Packard (HP) para gestionar repositorios de ficheros digitales, facilitando su depósito, el objetivo de DSpace es coleccionar y organizar la producción intelectual, organizándola en comunidades, asignándole metadatos y permitiendo su difusión a recolectores o indexadoras. Estas características han hecho que junto con EPrints, sea uno de los softwares más

preferidos por las instituciones académicas para gestionar el repositorio donde los investigadores depositan sus publicaciones y materiales de búsqueda con el objetivo de darles una mayor visibilidad. (Rodríguez Gairín & Sulé, 2008)

1.1.10 Técnicas de Procesamiento de Lenguaje Natural (NLP)

El lenguaje natural es aquel que utilizan los seres humanos para comunicarse unos con otros. A su vez, el procesamiento del lenguaje natural (PLN o NLP por sus siglas en inglés Natural Language Processing) se encarga del procesamiento computacional del lenguaje natural y sobre su aplicación para dar solución a problemas de ingeniería. El procesamiento del lenguaje natural involucra una transformación a una representación formal, manipula esta representación y por último, si es necesario, lleva los resultados nuevamente a lenguaje natural (Hernández & Gómez, 2013)

1.1.10.1 Aplicaciones del procesamiento de lenguaje natural

Las aplicaciones del NLP son muy variadas, ya que su alcance es muy extenso, algunas de las aplicaciones son traducción automática, recuperación de la información, extracción de la información y resúmenes, reconocimiento de voz, análisis de sentimientos, chat bots, traductores automáticos, los correctores de estilo y ortografía de los procesadores de texto, entre otros. (Vásquez, Huerta, Quispe, & Huayna, 2009)

1.1.10.2 Componentes del procesamiento de lenguaje natural

El NLP tiene cuatro componentes básicos de análisis que facilitan y optimizan las tareas de procesamiento y extracción de información. La ejecución de estos componentes normalmente es secuencial, aunque no todos los análisis siguen esa secuencia, sino que depende del objetivo de aplicación. (Pinilla Gómez, 2017).

Análisis morfológico o léxico: Es el análisis interno de las palabras que forman oraciones para extraer lemas, rasgos flexivos, unidades léxica compuestas. Es indispensable para la información básica: categoría sintáctica y significado léxico.

Análisis sintáctico: Determina si una secuencia de componentes léxicos o tokens cumplen una determinada estructura gramatical, siguiendo las reglas gramaticales del lenguaje analizado. Esto se logra haciendo uso de los resultados obtenidos del análisis léxico y el etiquetado morfológico.

Análisis semántico: Determina si una secuencia de componentes léxicos o tokens forman una sentencia bien construida, coherente y con sentido, haciendo uso del árbol sintáctico generado por el analizador sintáctico.

Análisis pragmático: Analiza la relación entre las palabras y el contexto donde son utilizadas, evaluando la influencia del mismo en su significado e interpretación. Este análisis se lleva a cabo haciendo uso de los resultados del análisis semántico.

1.1.10.3 Arquitectura de un sistema de NLP

La arquitectura de un sistema de procesamiento de lenguaje natural se sustenta en una definición del lenguaje natural (LN) por niveles estos son: fonológico, morfológico, sintáctico, semántico y pragmático.

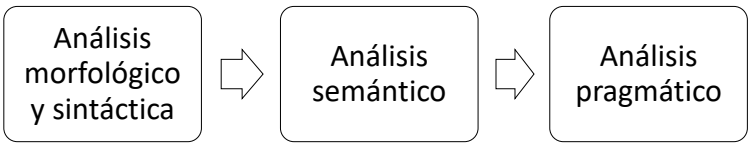


Figura 2. Arquitectura de un Sistema de Procesamiento de Lenguaje Natural

Fuente: (Vásquez et al., 2009)

La arquitectura del sistema de NLP muestra como la computadora interpreta y analiza las oraciones que le sean proporcionadas. La explicación de este sistema es:

1. El usuario expresa a la computadora que es lo que quiere hacer.
2. La computadora analiza las oraciones proporcionadas, en el sentido morfológico y sintáctico, lo que significa, si las frases contienen palabras compuestas por morfemas y si la estructura de las oraciones es

correcta. En este proceso el analizador lexicográfico y el analizador sintáctico juegan un papel importante. El primero denominado escaner se encarga de identificar los componentes léxicos definidos a priori, el segundo denominado parser se encarga de verificar si cumple un orden gramatical entre los elementos identificados por el scanner.

3. La siguiente etapa es analizar las oraciones semánticamente, es decir que es lo que significa de cada oración, luego se debe asignar el significado de estas oraciones a expresiones lógicas (verdadero o falso).
4. Terminada la etapa anterior, ahora se puede realizar el análisis pragmático de la instrucción, es decir después de ser analizada cada oración, ahora se deben de analizar todas juntas, teniendo en cuenta la situación de cada oración, analizando las oraciones anteriores, cuando se completa este paso, la computadora ya sabe lo que va a hacer, es decir, ya tiene la expresión final.
5. Cuando se obtiene la expresión final, el siguiente paso es la ejecución de esta, para obtener así el resultado y poder proporcionarlo al usuario.

1.1.10.4 Tokenizacion

La tokenizacion es el proceso de dividir un documento de texto en distintos componentes, descartando ciertos caracteres como espacios en blancos, saltos de líneas entre otros. Un token es una cadena de caracteres (palabra o un signo de puntuación) que contiene algún significado en el contexto de un texto, como ejemplo el texto "El procesamiento de texto" tiene 4 tokens: ('El', 'procesamiento', 'de', 'texto'). (Webster & Kit, 1992)

1.1.10.5 Segmentación

La segmentación consiste en separar el texto en fragmentos que pueden tratarse de forma independiente. La manera más usual de segmentación es dividir el texto en párrafos u oraciones

1.1.11 Expresiones Regulares

Una expresión regular sirve como un descriptor de un lenguaje. También es una herramienta para describir patrones de texto. Formalmente, el objetivo de las expresiones regulares es representar todos los posibles lenguajes definidos sobre un alfabeto, en base a una serie de lenguajes primitivos, y operadores de composición

(Billhardt, 2007). Las expresiones regulares permiten mostrar lenguajes regulares, tanto por su capacidad de especificación mediante un número reducido de operadores como por sus aplicaciones prácticas en la construcción de analizadores léxicos. Todo lenguaje finito es regular, y por el teorema de Kleene, todo lenguaje regular se puede representar mediante autómatas de estado finito. En la teoría de lenguaje formal, según la jerarquía de Chomsky (2011) está clasificado como gramáticas de tipo 3, el subconjunto con menor expresividad.

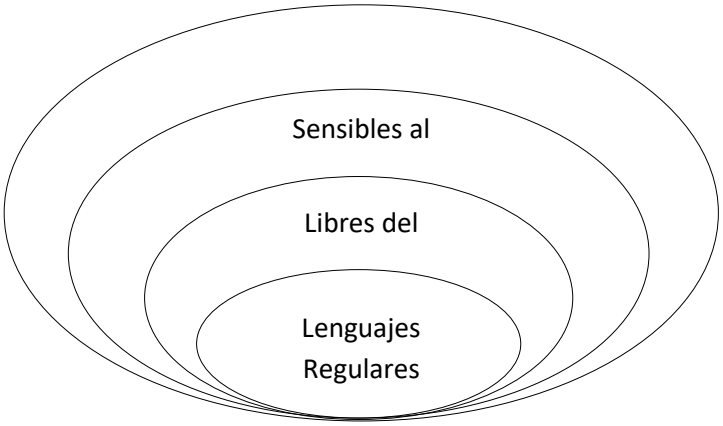


Figura 3. Jerarquía de Chomsky

Fuente: (Chomsky, 1956)

La expresión regular es una forma para describir los lenguajes regulares, formalmente definido como:

- Λ (cadena vacía) y ϕ (lenguaje vacío) son regulares.
- Para cada símbolo a del alfabeto Σ, a es regular
- Si $r_1 + r_2$ son regulares, también lo son sus respectivos concatenación y unión: $(r_1 + r_2), (r_1.r_2)$.
- Si r es regular, $(r *)$ también lo es.
- Todas las expresiones sobre Σ obtenidas aplicando los pasos anteriores.

Mencionado esto, una expresión regular es una herramienta potente para encontrar patrones en textos, en la actualidad está integrado en la mayoría de lenguajes de programación con extensiones que incluso le da la capacidad de expresar lenguajes no regulares.

1.1.12 Modulo re

En Python la operación con expresión regular está incluido en el módulo re.py, donde se puede crear una cadena de caracteres para el matching con combinaciones de letras del alfabeto, dígitos o símbolos especiales que representan conjuntos de los cuales se puede ver en la siguiente figura:

Character	Description	Example
\A	Returns a match if the specified characters are at the beginning of the string	"\AThe"
\b	Returns a match where the specified characters are at the beginning or at the end of a word (the "r" in the beginning is making sure that the string is being treated as a "raw string")	r"\bain" r"ain\b"
\B	Returns a match where the specified characters are present, but NOT at the beginning (or at the end) of a word (the "r" in the beginning is making sure that the string is being treated as a "raw string")	r"\Bain" r"ain\B"
\d	Returns a match where the string contains digits (numbers from 0-9)	"\d"
\D	Returns a match where the string DOES NOT contain digits	"\D"
\s	Returns a match where the string contains a white space character	"\s"
\S	Returns a match where the string DOES NOT contain a white space character	"\S"
\w	Returns a match where the string contains any word characters (characters from a to Z, digits from 0-9, and the underscore _ character)	"\w"
\W	Returns a match where the string DOES NOT contain any word characters	"\W"
\Z	Returns a match if the specified characters are at the end of the string	"Spain\Z"

Figura 4. Símbolos del Módulo re de Python

Fuente: (W3schools, 2021)

También se tiene la posibilidad de crear patrones de búsqueda más avanzado mediante los siguientes operadores:

Character	Description	Example
[]	A set of characters	"[a-m]"
\	Signals a special sequence (can also be used to escape special characters)	"\d"
.	Any character (except newline character)	"he..o"
^	Starts with	"^hello"
$	Ends with	"planet$"
*	Zero or more occurrences	"he.*o"
+	One or more occurrences	"he.+o"
?	Zero or one occurrences	"he.?o"
{}	Exactly the specified number of occurrences	"he{2}o"
|	Either or	"falls|stays"
()	Capture and group	

Figura 5. Meta caracteres del Módulo re

Fuente: (W3schools, 2021)

En conclusión, la búsqueda se realiza con los métodos re.search o re.match sobre una expresión definida, donde se puede etiquetar y clasificar fragmentos de texto que han hecho match mediante la directiva (?P <etiqueta> …). Por ejemplo con la expresión (Sr. o Sra.) (?P<name> [A-Z] [a-z]*) podemos encontrar nombres de personas, que es una secuencia seguida de Señor o Señora empezando por una letra mayúscula y luego un número indefinido de minúsculas. Si se ha encontrado algún match, entonces el objeto devuelto por la búsqueda permite recuperar texto accediéndolo como un diccionario, mediante la etiqueta que se había definido: matchText = res.group (“name”), (Pan, 2020)

1.1.13 Stop-words

Las stop-words son palabras que no tienen significado notorio en un sistema de recuperación. Son una parte del lenguaje natural, son generalmente palabras de alta frecuencia que no proporcionan ninguna información adicional. La eliminación de stop words reduce la dimensionalidad del espacio de términos. Las stop-words o también conocidas como palabras vacías, son palabras en documento de texto como artículos, preposiciones, pro-sustantivos, entre otros, que no da el significado a los documentos, no se miden como palabras clave en aplicaciones de minería de texto (Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J. & Nithya, 2015)

Las Stop Words o palabras vacías tienen una función gramatical, pero no aportan nada al contenido del documento, es decir, tienen un rol más sintáctico que semántico. Mediante la eliminación de tales términos se puede contribuir a mejorar la recuperación de la información, ya que es poco probable que los usuarios realicen búsqueda de documentos por esas palabras.

Los estudios correspondientes a este fenómeno fueron iniciados por Hans Peter Luhn en 1958 con su investigación sobre el índice KWIC (Keyword in Context), (Luhn, 2014) en su investigación presenta una técnica de indexación que organiza las palabras según su consideración como claves para la recuperación o no de la información, teniendo en cuenta el contexto del documento. Este proceso derivó en la definición del término "palabra vacía" para referirse a aquellas con un bajo poder discriminatorio y representativo del contenido del documento. Los análisis estadísticos realizados por Luhn demostraron que la indexación es un proceso más rápido cuando se prescinde de tales términos, contribuyendo al ahorro de espacio necesario para el almacenamiento de la información. A la misma vez se demostró que entre un 30% y un 50% de las palabras de un texto corresponden a tal categoría.

Una de las técnicas más utilizadas para remover palabras vacías consiste en revisar si la palabra está dentro de un listado que contiene las palabras vacías más comunes, como son adverbios, conjunciones, preposiciones, entre otras. Sin embargo, cada colección de documentos es única. Por lo tanto, es razonable tener un listado de palabras vacías diferente para diferentes colecciones, con el fin de maximizar el rendimiento de un algoritmo de recuperación de la información. No obstante, la técnica de eliminación de palabras vacías ha mejorado, debido a la introducción de técnicas que tienen en cuenta el significado de tales palabras cuando están acompañadas de sustantivos, en casos en los que no pueden ser separadas o eliminadas por conformar una denominación propia, así como por pérdidas en el significado semántico de un sintagma, frase o palabra (Ochando, 2013).

1.1.14 Stemming

El proceso de stemming o lematizacion, consiste en remover los sufijos en palabras que pertenezcan a la misma familita semántica, es decir, reducir las palabras a sus elementos mínimos con significado, las raíces de las palabras, si bien el objetivo principal es el reducir las diferentes formas lingüísticas de una palabra a una forma

común y así facilitar el acceso a la información durante el posterior proceso de búsqueda, paralelamente se está reduciendo el número de términos diferentes del sistema, lo que permite a su vez una segunda reducción de los recursos de almacenamientos requeridos (Halascy, 2006).

Las implementaciones de estos algoritmos están vinculados a los idiomas y por esto es que se encuentran más implementaciones para palabras en inglés y muy pocas para el español. Por lo que, tanto en inglés como en español y en cualquier otro idioma, un término puede ser reducido a su común denominador, permitiendo la recuperación de todos los documentos cuyas palabras tengan la misma raíz común, como, por ejemplo: catálogo, catálogos, catalogación, catalogador, catalogar, catalogando, catalogado, catalogándonos todos los términos derivan en tal caso de "catalog", haciendo posible que la recuperación sea completa en más de 8 supuestos distintos (Peinado Rodriguez, 2003). No obstante, no siempre esta técnica permite resolver perfectamente todas las consultas que un usuario pueda plantear.

1.1.15 POS Tagging

Pos Tagging (Part-of-speech tagging) sirve para dividir el texto en varios elementos gramaticales para su posterior análisis, resuelve el problema de asignar a cada palabra de una frase, la función que cumple en esa frase, es decir, se trata de un etiquetado gramatical. El etiquetado de palabras es la tarea de asignar apropiadamente la categoría sintáctica a las palabras de un texto, parte de la problemática del etiquetado es la ambigüedad en la asignación de la categoría sintáctica apropiada a ciertas palabras del corpus, que pueden tener más de una etiqueta debido al contexto en el que se encuentren. La entrada para el proceso de POS Tagging es un texto más un conjunto de etiquetas y la salida del proceso es un texto etiquetado, asignado a cada palabra una única etiqueta (France & Allen, 1997). Esta sencilla técnica, puede ser llevada a cabo por diferentes métodos que tienen en cuenta el contexto local, el POS Tagging es de interés para muchas aplicaciones, como, por ejemplo: reconocimiento y generación de palabras, acceso a bases de datos textuales, análisis sintáctico, recuperación de información entre otras.

1.1.16 Reconocimiento de entidades nombras (NER)

El reconocimiento de entidades nombradas (NER por sus siglas en ingles), también conocido como extracción de entidades, permite localizar y clasificar entidades

dentro de un texto, de acuerdo a categorías tales como nombres de personas, organizaciones, ubicaciones, cantidades, expresiones de tiempo, valores monetarios entre otros. Como por ejemplo “Perú”, “Banco de crédito del Perú”. Una entidad nombrada también puede ser entendida como aquella que puede ser referenciada con un nombre propio (Lin, 2011).

La tarea de reconocimiento de entidades nombradas comúnmente denominada como NER consiste en identificar las entidades del texto y clasificarlas en un conjunto predefinido de tipos, tales como persona, organización y ubicación (Aggarwal & Zhai, 2012). En la siguiente figura se presenta un ejemplo de etiquetado NER. La oración de ejemplo es: “Backus, La Iberica y BCP permanecen más de un siglo en el mercado local”. Donde ORG hace referencia a la entidad “Organización”, AMOUNT a la entidad “Cantidad” y NE indica que la palabra no es una entidad.

Backus	,	La	Ibérica	y	BCP	permanecen	más
ORG	NE	ORG		NE	ORG	NE	NE
de	un	siglo	en	el	mercado	local	
NE	AMOUNT		NE	NE	NE	NE	NE

Figura 6. Oración de ejemplo de etiquetado NER

Fuente: (Copara Zea, 2017)

1.1.17 Complejidad algorítmica

La complejidad del algoritmo evalúa la obtención del conteo de operaciones, realizadas por un algoritmo dado como un tamaño de entrada datos en la función (Kuznetsov & Obiedkov, 2010). Para explicarlo de una manera más simple, la complejidad algorítmica es una aproximación del número de pasos necesarios para ejecutar un algoritmo.

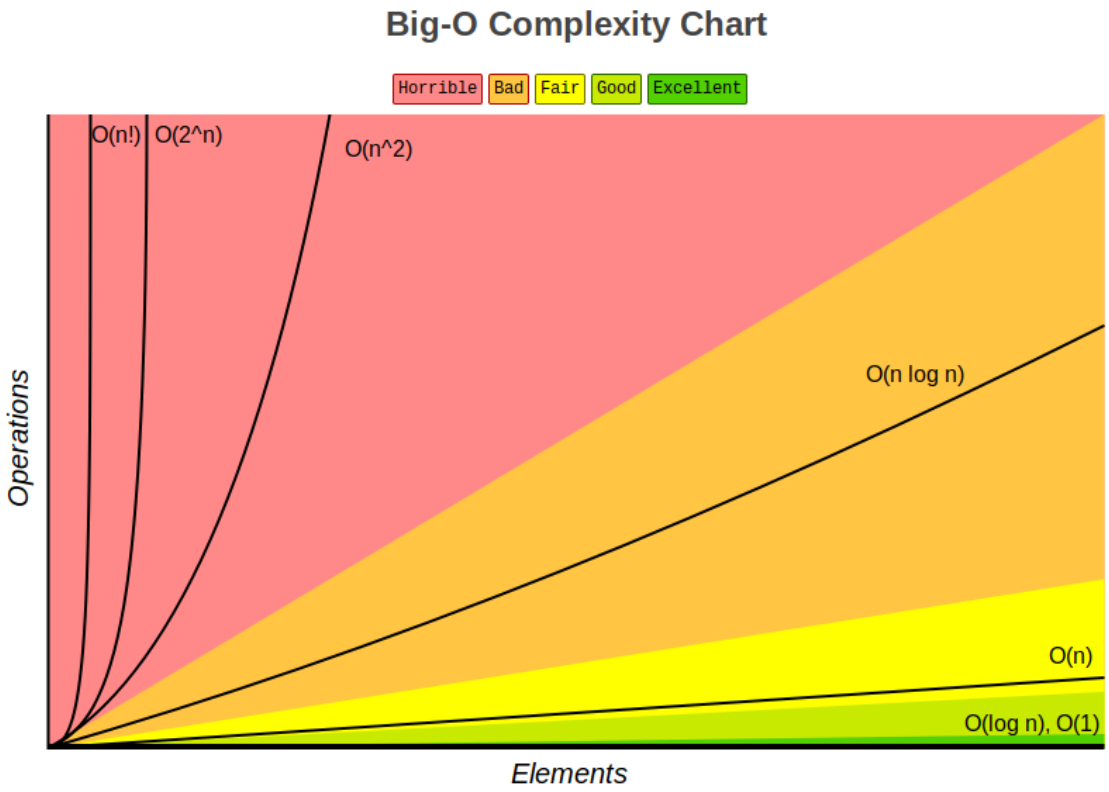


Figura 7. Complejidad Big-O

Fuente: (Sanz de Diego, 2021)

Notación Big-O:

La notación Big O consiste en el análisis de capacidades científicas y graduales. Es importante representar la complejidad algorítmica de cada solución. Hay dos tipos de métricas para analizar la complejidad del algoritmo. Complejidad basada en la eficiencia algorítmica y la otra es la complejidad en la estructura del algoritmo (Singh & Tiwari, 2015). La Notación Big O es utilizada para identificar la base de la función en sus tasas de crecimiento (Burnim, Jukevar, & Sen, 2009). La longitud de la función de entrada en La notación Big O generalmente se refiere a los tres aspectos. Hay los peores casos, caso promedio y mejor caso. Este escenario ayuda a los desarrolladores de algoritmos a predecir el comportamiento de sus algoritmos (Y. Han & Thorup, 2002). También pueden resolver cuál de los múltiples algoritmos se debe utilizar.

Samsudin (2020) menciona los tipos de complejidad algorítmicas:

- O(1) Constante: La cantidad de input que reciba no afecta al resultado, es decir siempre demorara el mismo tiempo.
- O(n) Lineal: la complejidad crecerá de forma proporcional a medida que se incremente el input.

- O(log n) Logarítmica: La función crecerá de forma logarítmica con respecto al input. Es decir que en un inicio el incremento será rápido, pero luego se estabilizara.
- O(n log n) Log lineal: El algoritmo crecerá de forma logarítmica pero junto con una constante.
- O(n²) Polinomial: Se incrementa de forma cuadrática.
- O(2^n) Exponencial: Crecerá de forma exponencial, lo que implica que la carga es muy alta. Para nada recomendable en ningún caso, solo para análisis conceptual.
- O(n!) Factorial: Se incrementa de forma factorial, por lo que al igual que el exponencial su carga es muy alta, por lo que jamás utilizar algoritmos de este tipo.

1.1.18 Modelo de calidad ISO/IEC 25000

La norma ISO/IEC 2500 presenta como objetivo principal: "guiar el desarrollo de productos de software mediante la especificación de requisitos y evaluación de características de calidad", la cual se encuentra dividida en cinco partes como se ve en la figura (ISO 25000, 2022).

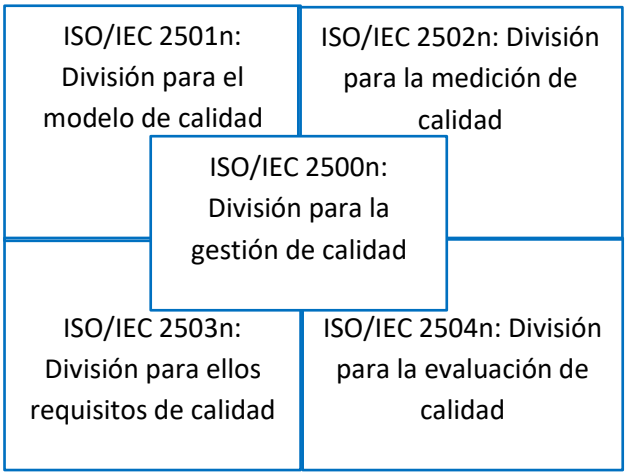


Figura 8. División de la norma ISO/IEC 25000

1.1.18.1 La ISO/IEC 2500n División de gestión de calidad

Las normas que se encuentra en este apartado definen todos los términos, modelos y definiciones referenciados por las otras normas del entorno 25000. En la actualidad esta división se encuentra conformada por:

- **ISO/IEC 25000:** Guía de SQuaRE, contiene el modelo de la arquitectura SQuaRE.
- **ISO/IEC 25001:** Establece los requisitos y orientaciones para gestionar la evaluación y especificación de los requisitos del producto software.

1.1.18.2 La ISO/IEC 2501n División de modelo de calidad

Las normas de este apartado muestran los modelos de calidad especificados los cuales incluyen las características para la calidad externa, interna y en uso del producto de software, las cuales estas divididas por:

- **ISO/IEC 25010:** Determina las características de calidad del producto software que se pueden evaluar, dentro de estas características se presentan: funcionalidad, portabilidad, compatibilidad, rendimiento, fiabilidad, seguridad, mantenibilidad y usabilidad.
- **ISO/IEC 25012:** Esta norma determina un modelo general para la calidad de los datos, los cuales son aplicados a datos que se encuentran almacenados de forma estructurada y forman parte de un sistema de información.

1.1.18.3 La ISO/IEC 2502n Divisiones de mediciones de calidad

Las normas que se encuentran en esta división incluyen un modelo de referencia de calidad del producto de software, definiciones matemáticas para las mediciones de calidad y una guía práctica para su aplicación. En la actualidad está dividida por:

- **ISO/IEC 25020:** Presenta una explicación introductoria y un modelo de referencia común a los elementos de medición de calidad. A la misma vez facilita una guía con la finalidad de que los usuarios seleccionen, desarrollen y apliquen medidas propuestas por las normas ISO.
- **ISO/IEC 25021:** Esta norma define y especifica un conjunto recomendado de métricas base las cuales son derivadas con el fin de ser usadas en el transcurso de todo el ciclo de vida del desarrollo del software.

- **ISO/IEC 25022:** La presente norma determina específicamente cuales son las métricas para realizar la medición sobre la calidad en uso del producto.
- **ISO/IEC 25023:** Define exactamente las métricas para realizar los sistemas de software y la medición de la calidad de productos.
- **ISO/IEC 25024:** Determina las métricas para realizar la medición de la calidad de datos.

1.1.18.4 La ISO/IEC 2503n División de requisitos de calidad

Las normas que se encuentran en este apartado contribuyen a la especificación de requisitos de calidad, las cuales esta conformadas por:

- **ISO/IEC 25030:** Las normas ubicadas en esta división permiten la especificar los requisitos de calidad, los cuales pueden ser utilizados en el proceso de especificación sobre los requisitos de calidad para un producto software, el cual va a ser desarrollado o como entrada para un proceso de evaluación.

1.1.18.5 La ISO/IEC 2504n División de evaluación de calidad

Dentro de este apartado se incluye normas que proporcionan requisitos, recomendaciones y guías para llevar a cabo el proceso de evaluación del producto software, las que se encuentran dividas por:

- **ISO/IEC 25040:** Esta norma define el proceso de evaluación de la calidad del producto software, el cual se encuentra formado por cinco actividades: Establecer cuáles son los requisitos, especificar la evaluación, diseñar la evaluación, ejecutar la evaluación y concluir la evaluación.
- **ISO/IEC 25041:** Especifica los requisitos y recomendaciones para la implementación práctica de la evaluación del producto software desde el punto de vista de los desarrolladores, evaluadores y de los usuarios finales.

1.1.19 Calidad de software

Según Pressman (2022) menciona la "concordancia con los requisitos funcionales y de rendimiento explícitamente establecidos con los estándares de desarrollo plenamente documentados y con las características implícitas que se espera de todo software desarrollado profesionalmente", con base en los requisitos funcionales y no funcionales identificados en la etapa de análisis del sistema, insumo principal para implementar dichos requisitos con los atributos mínimos de calidad, fomentando la aplicación de procesos estandarizados y criterios necesarios en cada una de sus etapas, así se fomenta que el avance en el ciclo de vida del software minimice el riesgo de fracaso del proyecto.

1.2 Antecedentes

Los estudios previos que guardan relación con este trabajo de investigación son los siguientes:

Choudhury *et al.* (2021), realizaron un estudio en el cual proponen un modelo de campo aleatorio que combina características visuales para la extracción de metadatos diseñados para documentos exclusivamente digitales, como documentos escaneados. El modelo propuesto contiene un corpus de 500 portadas, el modelo propuesto logro una medida de 81,3% y 96% en la extracción de siete metadatos extraídos correctamente.

Adam & Kiran (2021), presentan un estudio donde profundizan el conocimiento de cómo el personal implicado en los repositorios institucionales en sus respectivas instituciones, identifica los procesos fundamentales que están detrás de la administración de sus repositorios institucionales, también explora los principales retos en el contexto de la administración de los RI.

Heinrichs *et al.* (2021), describe un caso de uso que utilizan la extracción de metadatos interoperables de entidades de datos para anotar automáticamente sus datos, sin embargo surge el problema de que estos valores de metadatos no son reutilizables debido a su falta de interoperabilidad, por lo tanto este estudio presenta una solución para mapear los valores de metadatos definidos con metadatos interoperables extrayéndolos primero utilizando un algoritmo de extracción de metadatos general y luego proponiendo un método para el mapeo. El método esta validado contra el caso de uso y muestra resultados prometedores para otros.

Hasan *et al.* (2020), proponen un modelo heurístico para extraer metadatos de las portadas escaneadas mediante la aplicación de expresiones regulares para cada campo, capturando patrones para siete campos de metadatos como: título, autores, año, grado académico, programas, instituciones y asesores. El método se evalúa en un conjunto de datos, logrando una precisión de extracción de hasta el 97% en los campos de los archivos de texto, el método propuesto plantea una base sólida para los métodos basados en el aprendizaje automático.

Pan (2020), manifiesta que la extracción se hace posible mediante técnicas de procesamiento de lenguaje natural, así como clasificación de entidades combinado con la búsqueda de patrones sobre el texto lematizado mediante expresión regular. Estos datos, permitieron un posterior análisis cuantitativo sobre cómo ha evolucionado el litigio en España, quienes son los actores políticos involucrados; cuales son los problemas de políticas públicas más recurridas, y hasta qué punto existe una mayor "ligitiosidad". Permitiendo a la sociedad en su conjunto conocer con más precisión el papel de tribunales y su impacto en las políticas.

Iturbe Herrera *et al.* (2019), afirma que el uso de técnicas de lenguaje natural y la información tipográfica del texto en el documento para la extracción de metadatos, tales como: título, autores, editorial y fecha de publicación. Los resultados obtenidos en la evaluación con documentos digitales no estructurados indican el potencial del enfoque propuesto, que es capaz de producir buenos resultados en la extracción de metadatos.

Flores *et al.* (2017), realizaron un componente para la extracción automática de metadatos bibliográficos desde corpus textuales en formato PDF, En este artículo se describe un componente web para la extracción automática de metadatos bibliográficos, para validar si el componente de extracción de metadatos reduce el tiempo de extracción se realiza un diseño experimental a partir de un caso de estudio. Además de validar el componente a través del diseño experimental se le aplican un conjunto de pruebas de calidad.

Tkaczyk (2017), propone un algoritmo automático, preciso y flexible para extraer una amplia gama de metadatos directamente de artículos científicos en forma digital nativa. La información extraída incluye metadatos básicos del documento, texto completo estructurado y sección de bibliografía. Diseñado como una solución universal, el algoritmo propuesto puede manejar una gran variedad de diseños de publicación con alta

precisión y, por lo tanto, es muy adecuado para analizar colecciones de documentos heterogéneos.

Júnior Grossi (2016), compara la capacidad de extraer metadatos de algunas de las herramientas preseleccionadas (Cermine, CiteSeer, CrossRef y ParsCit) utilizando un experimento empírico con un conjunto de artículos para esto. Por lo tanto, en base a los resultados presentados, podemos identificar o apoyar cada una de las otras herramientas no relacionadas con su capacidad para extraer metadatos adecuadamente. Superé la herramienta CrossRef, todo con más resultados que el 60%, tragando el 86.83% de Cermine. Además, se evidenciaron por la fragilidad principal de estos componentes; puntos donde se necesitarán ajustes; Para quienes, se obtiene un gran éxito. Esta calificación se calcula en función de los resultados obtenidos en la extracción de metadatos de la selección de artículos realizados.

Casali *et al.* (2015), propone facilitar al usuario el autoarchivo de sus Objetos Digitales Educativos en un repositorio institucional, para esto, se modifica el flujo de carga estándar de la plataforma DSpace, proponiendo un nuevo flujo para el depósito de objetos de modo que pueda integrarse en este proceso un extractor de metadatos. Los metadatos extraídos automáticamente son luego validados por el usuario en el proceso de descripción del objeto, el extractor fue desarrollado un prototipo en JAVA.

Miranda *et al.* (2015), describe un posible enfoque metodológico para automatizar esta actividad mediante la extracción de metadatos directamente de los archivos que configuran el objeto de aprendizaje en sí. El enfoque propuesto intenta reunir la teoría de la información, los modelos de aprendizaje, el análisis estadístico y la heurística ad hoc para extraer un amplio conjunto de campos de los metadatos. Los resultados del experimento son particularmente alentadores para pensar en este enfoque como una solución para admitir repositorios de objetos de aprendizaje y otras plataformas que tienen necesidades para administrar un amplio almacenamiento de contenido y una gran cantidad de usuarios con diversas características personales, dispositivos para la interacción.

Tkaczyk *et al.* (2014), presentaron CERMINE - Automatic extraction of metadata and references from scientific literature, es un sistema integral de código abierto para extraer metadatos y referencias bibliográficas analizadas de artículos científicos en forma digital nativa. El sistema se basa en un flujo de trabajo modular, cuya arquitectura permite la

capacitación y evaluación en un solo paso, permite modificaciones y reemplazos sin esfuerzo de componentes individuales y simplifica la expansión de la arquitectura.

Pinilla *et al.* (2014), ofrece una perspectiva sobre el estado actual de las investigaciones acerca de extracción automática de metadatos, estableciendo las bases para futuras investigaciones en el marco concreto de objetos de aprendizaje en repositorios institucionales de acceso abierto, en el estudio se evalúan diferentes propuestas actuales de extracción automática de metadatos, a la luz del cumplimiento de estándares y algunos aspectos relevantes para el análisis y diseño de sistemas de extracción automática.

Kern *et al.* (2012), presenta el algoritmo TeamBeam analiza un artículo científico y extrae metadatos estructurados, como el título, el nombre de la revista y el resumen, así como información sobre los autores del artículo (por ejemplo, nombres, direcciones de correo electrónico, afiliaciones). La entrada del algoritmo es un conjunto de bloques generados a partir del texto del artículo luego se aplica un algoritmo de clasificación, que tiene en cuenta la secuencia de entrada, en dos fases consecutivas. En la evaluación del algoritmo, su desempeño se compara con dos heurísticas y tres sistemas de extracción de metadatos existentes. Se utilizan tres conjuntos de datos diferentes con características variables para evaluar la calidad de los resultados de la extracción. TeamBeam se desempeña bien bajo las pruebas y se compara favorablemente con los enfoques existentes.

Granitzer *et al.* (2011), manifiesta que el uso de campos aleatorios condicionales y máquinas de vectores de soporte, implementados en dos sistemas de última generación del mundo real, a saber, ParsCit y Mendeley Desktop, para extraer automáticamente metadatos bibliográficos. Comparamos la precisión de los sistemas en dos conjuntos de datos del mundo real recién creados recopilados de los repositorios Mendeley y Linked-Open-Data.

Kovačević *et al.* (2011), realizaron un estudio con el objetivo de desarrollar un sistema para la extracción automática de metadatos de documentos científicos en formato PDF para el sistema de información para monitorear la actividad de investigación científica de la Universidad de Novi Sad (CRIS UNS). Obteniendo como resultados sobre el análisis de los metadatos extraídos de estas publicaciones mostró que el rendimiento del sistema para los datos no vistos anteriormente está de acuerdo con el obtenido por la validación cruzada de ocho clasificadores SVM separados.

Pire *et al.* (2011), realizaron un estudio donde se analizan, por un lado, la importancia de los metadatos de los objetos de aprendizaje con el fin de poder utilizarlos en un sistema de recomendación automática personalizada. Por otro lado, se explora el estado del arte de las técnicas de extracción automática de metadatos, y se analizan y comparan diferentes sistemas de extracción.

Cui y Chen (2010), propuso un Modelo de Markov Oculto (HMM) mejorado para extraer metadatos en la literatura académica. Se ha creado un conjunto de datos que incluye 458 publicaciones de las conferencias VLDB, que contiene la característica visual de los bloques de texto. Los experimentos mostraron que la precisión de extracción es superior a la de cualquier trabajo existente.

Cortez *et al.* (2009), realiza un método novedoso para extraer componentes (por ejemplo, nombres de autores, títulos de artículos, lugares, números de página) de citas bibliográficas. Este método no se basa en patrones que codifiquen delimitadores específicos utilizados en un estilo de cita particular. Los resultados de estos experimentos muestran niveles de precisión y recuperación superiores al 94% para todos los campos, y una extracción perfecta para la gran mayoría de las citas analizadas.

Marinai (2009), realizo un estudio donde se analiza sobre el uso de técnicas de análisis de documentos para la extracción de metadatos de documentos PDF. Se describe un paquete diseñado para extraer metadatos básicos de estos documentos. Además, se utiliza la información recopilada de una base de datos de citas ampliamente conocida (DBLP) para ayudar a la herramienta en la difícil tarea de identificación del autor. El sistema se prueba en algunas colecciones de papel seleccionadas de actas de conferencias recientes.

Flynn *et al.* (2007), se realizó un estudio que tuvo como objetivo desarrollar y demostrar un nuevo sistema para extraer metadatos. Primero, se examinó un documento en un intento de reconocerlo como una instancia de un diseño de documento conocido. Luego, se aplicó una plantilla, una descripción escrita de cómo asociar bloques de texto en el diseño con campos de metadatos, para extraer los metadatos. La extracción se valida después del procesamiento posterior para evaluar la calidad de la extracción y, si es necesario, para marcar extracciones no confiables para el reconocimiento humano.

Alfano *et al.* (2007), realizaron un estudio con el objetivo de desarrollar un sistema, llamado SAXEF (Sistema para la extracción automática de características de objetos de

aprendizaje), que es capaz de extraer automáticamente los indicadores didácticos una especie de ADN de cualquier página web o grupo de páginas que se encuentra en Internet y le permite al maestro evaluar fácilmente si esa página con sus contenidos es de su interés web.

Flynn *et al.* (2007), describen los procedimientos para desarrollar un conjunto de herramientas y un proceso para la extracción automatizada de metadatos de colecciones de documentos grandes, diversas y en evolución. El proceso automatizado desarrollado permite que muchos más documentos estén disponibles en línea de lo que hubiera sido posible debido a limitaciones de tiempo y costos. Se describe la arquitectura, implementación e ilustramos la efectividad del conjunto de herramientas al proporcionar resultados experimentales en dos colecciones principales DTIC (Centro de Información Técnica de Defensa) y NASA (Administración Nacional de Aeronáutica y del Espacio).

Han et al. (2003), describe un método basado en la clasificación de Support Vector Machine para la extracción de metadatos de la parte del encabezado de los trabajos de investigación y muestra que supera a otros métodos de aprendizaje automático en la misma tarea. La extracción adicional de metadatos se realiza buscando los mejores límites de fragmentos de cada línea. Se descubrió que la búsqueda y el uso de los patrones estructurales de los datos y el agrupamiento de palabras basado en el dominio pueden mejorar el rendimiento de extracción de metadatos.

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

2.1 Identificación del problema

La Web es actualmente uno de los recursos educativos más importantes, donde los distintos usuarios académicos tienen acceso a una gran cantidad de información a su disposición. Para el proceso de búsqueda las personas usan motores de búsqueda, como Google, DuckDuckGo, Bing, Ask entre otros, donde la mayoría de los casos, no retornan la información deseada o en algunos casos devuelven una gran cantidad de enlaces. Ahora existen otras formas más estructuradas de acceder a la información, en base a esto se muestran aparecen conceptos fundamentales: objetos de aprendizaje, metadatos, estándares y repositorios institucionales.

Los Repositorios Digitales son medios que nos permiten gestionar, almacenar, preservar, difundir facilitar el acceso a los objetos digitales que este alberga (Polanco-Cortés, 2016). Las herramientas más utilizadas en software libre para implementar un repositorio son las siguientes: DSpace, E-prints, Fedora, LUCENE, entre otros. Estas herramientas facilitan a las instituciones públicas y privadas, la instalación, implementación y mantenimiento de repositorios digitales, con el objetivo de organizar y administrar la información de forma digital. Los repositorios institucionales, organizan y almacenan la producción científica o académica resultado de la actividad docente e investigadora de de una o varias instituciones, recuperando, preservando, difundiendo y dando acceso abierto a los recursos digitales depositados en ellos. En la actualidad son las universidades o institutos de investigación los que, en general, administran este tipo de repositorios, y constituyen una herramienta clave de sus políticas científicas y académicas, además de una pieza de apoyo fundamental para la enseñanza y la investigación (Paradelo Luque, 2009).

El Repositorio Institucional de la Universidad Nacional del Altiplano Puno, almacena la producción de trabajos de investigación científica realizados por los estudiantes de pregrado y posgrado, permitiendo una búsqueda más acotada para la recuperación y reutilización de estos recursos digitales. Estos objetos se almacenan utilizando metadatos descriptivos que proporcionan información adicional sobre los documentos de investigación. La información que se almacena en estos metadatos es primordial para una mejor recuperación de los mismos y se convierte en un aspecto importante en la administración de repositorios. Existen distintos estándares de metadatos tales como DublinCore y IEEE LOM 4, que utilizan distintas categorías para el caso peruano las recolecciones de los metadatos son validados según las directrices de DRIVER y el esquema de metadatos utilizado es Dublín Core 2 (ALICIA CONCYTEC, 2021).

No obstante, es necesario investigar las fuerzas impulsoras en la administración de los Repositorios Institucionales, es útil identificar y comprender los procesos clave que desempeñan un papel importante en la administración de un Repositorio Institucional (RI). Los metadatos que son extraídos de los documentos de investigación y la información cargada sobre los mismos en el Repositorio Institucional de la UNA - PUNO, vienen a ser los procesos impulsores detrás de la administración de un RI, que en su mayoría son de baja calidad o incompleta. Esto se debe a que la carga de metadatos, suele ser una tarea tediosa, la cual requiere de tiempo y muchas veces el personal encargado de llenar los mismos lo hacen erróneamente. Para facilitar la carga de metadatos en el repositorio el presente estudio tiene como propósito optimizar el proceso de extracción y publicación de documentos mediante el desarrollo de un software haciendo uso del procesamiento de lenguaje natural que permita mejorar la administración del Repositorio Institucional.

2.2 Enunciado del problema.

Teniendo en cuenta lo expresando anteriormente nos planteamos la siguiente premisa:

¿De qué manera la extracción automática de metadatos mejora la administración del Repositorio Institucional de la Universidad Nacional del Altiplano?

A la misma vez las siguientes interrogantes específicas:

¿En qué medida el procesamiento de lenguaje natural contribuye al desarrollo del algoritmo para la extracción automática de metadatos en los documentos de investigación del Repositorio Institucional Universidad Nacional del Altiplano?

¿En qué medida la precisión del algoritmo permite la extracción automática de metadatos de los documentos de investigación?

¿Cuál es la diferencia del tiempo de ejecución en el proceso de extracción de metadatos antes y después de la implementación del software?

2.3 Justificación

Los medios electrónicos y digitales actualmente son indispensables en la vida cotidiana, esto se debe a la gran facilidad de manipulación y movilidad que ofrecen estos medios, a diferencia de los objetos físicos como los libros, revistas entre otros. Actualmente es posible acceder a una gran cantidad de información digital de libros o revistas que se encuentra disponibles físicamente, es aquí donde intervienen las diferentes maneras de obtención de información como los repositorios institucionales o bibliotecas digitales

Durante los últimos años los repositorios institucionales han cobrado importancia en la sociedad académica y científica, porque representan una fuente de información digital especializada, organizada y accesible para los investigadores de distintas áreas (Texier, De Giusti, Oviedo, Villarreal, & Lira, 2012). La gran cantidad de trabajos de investigación almacenadas en el Repositorio Institucional de la Universidad Nacional del Altiplano, hoy en día es enorme y está en constante crecimiento. Un repositorio institucional moderno y completamente funcional para proporcionar servicios de alta calidad requiere de una buena administración a través de un acceso no solo a las fuentes de los documentos almacenados, sino también a sus metadatos como título, autores, palabras clave, resúmenes, fechas, temas entre otros.

La extracción de metadatos se encarga de obtener los atributos o etiquetas que identifican a cada documento de investigación. Estos metadatos permiten la recuperación, búsqueda, autenticación y evaluación de un recurso dentro del repositorio institucional. La información confiable de estos metadatos es crucial para la organización de datos puesto que este proceso es necesario para la administración de un RI, desafortunadamente, el repositorio institucional no cuenta con alguna herramienta que permita el acceso rápido a metadatos de calidad y confiables. En tales casos, el repositorio necesita un método

confiable para extraer metadatos y referencias de los diferentes documentos de investigación disponibles que son publicados en el repositorio institucional de la UNA – PUNO.

2.4 Objetivos

2.4.1 Objetivo general

Optimizar la extracción de metadatos y publicación de documentos de investigación para la administración del Repositorio Institucional de la Universidad Nacional del Altiplano.

2.4.2 Objetivos específicos

- Desarrollar un algoritmo para la extracción automática de metadatos basada en el procesamiento de lenguaje natural (NLP).
- Implementar y determinar el nivel de precisión del algoritmo de extracción automática de metadatos basada en procesamiento de lenguaje natural (NLP).
- Evaluar la diferencia del tiempo de extracción de metadatos antes y después de la implementación del software.

2.5 Hipótesis

2.5.1 Hipótesis General

La extracción automática de metadatos mejora la administración del Repositorio Institucional de la Universidad Nacional del Altiplano

2.5.2 Hipótesis especificas

- El procesamiento de lenguaje natural contribuye al desarrollo del algoritmo para la extracción automática de metadatos en los documentos de investigación del Repositorio Institucional UNA- Puno
- El algoritmo basado en procesamiento de lenguaje natural logra un nivel de precisión adecuado para la extracción automática de metadatos de los documentos de investigación.
- La implementación del software mejora el tiempo en el proceso de extracción de metadatos de los documentos de investigación

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1 Lugar de estudio

La Universidad Nacional del Altiplano de Puno, es una universidad pública ubicada en la ciudad de Puno, Perú. Fue creada por Ley Nº 406 el 29 de agosto de 1856, firmada por Don Ramón Castilla y Marquesado, con la denominación de Universidad de Puno, para la enseñanza de la Teología, Jurisprudencia, Medicina, Filosofía y Letras, Matemáticas y Ciencias Naturales, actualmente, cuenta con 20 facultades, 36 escuelas profesionales, una Escuela de Posgrado, 8 centros experimentales y 6 centros de servicios; albergando a más de 16 mil estudiantes, mil 200 docentes y 800 trabajadores administrativos. (UNAP, 2021)

3.2 Población

La población de estudio está conformada por todos los documentos de investigación publicados en el Repositorio Institucional de la Universidad Nacional del Altiplano del 2021.

Tabla 2

Población de estudio

NIVEL	TOTAL
PREGRADO	1181
POSGRADO	337
TOTAL	**1518**

Fuente: Repositorio Institucional de la Universidad Nacional del Altiplano

3.3 Muestra

El diseño de muestra que se utilizó en el trabajo de investigación fue de tipo no probabilístico, utilizando el muestreo por conveniencia, este tipo de muestreo se caracteriza por obtener muestras accesibles representativas, por lo que, se consideró como muestra los documentos de investigación que fueron publicados en las colecciones de pregrado y posgrado en el Repositorio Institucional durante el periodo de julio y diciembre del 2021, esto con el fin de obtener un mejor aprendizaje automático, cumpliendo los siguientes criterios.

Criterios de Inclusión

- Documentos de investigación en tipo de contenido de tesis de pregrado, tesis de maestría y tesis de doctorado
- Documentos de investigación en tipo de obra tesis conducente al grado académico
- Documentos de investigación en formato PDF y Word

Criterios de Exclusión

- Documentos de investigación en tipo de contenido de monografía, reporte, libro, revisión, conferencia, documentos internos, documentos técnicos entre otros.
- Documento de investigación en tipo de obra de trabajos de investigación, trabajo de suficiencia profesional y trabajo académico
- Documentos de investigación en formato PDF no editable o formato grafico (archivos escaneados)

Criterios de eliminación

- Documentos de investigación que no cumplieron con el formato para ser publicados en el Repositorio Institucional

Según a los criterios mencionados el tamaño de muestra que se usó figura en la siguiente tabla.

Tabla 3

Muestra de estudio

NIVEL	TOTAL
PREGRADO	266
POSGRADO	114
TOTAL	**380**

Fuente: Equipo de trabajo

3.4 Método de investigación

3.4.1 Tipo de investigación

El tipo de Investigación al que corresponde el presente trabajo de investigación es el aplicado (Arias, 2012), por lo que la extracción automática de metadatos basada en procesamiento de lenguaje natural busca mejorar la administración del Repositorio Institucional de la Universidad Nacional del Altiplano Puno.

Según el enfoque de la investigación: es una investigación explicativa porque describe las causas de los fenómenos que están en estudio (Hernández Sampieri, Fernández Collado, Baptista Lucio, Mendoza Torres, & Méndez Valencia, 2014).

3.4.2 Diseño de investigación

El diseño de investigación que se utilizo es el diseño experimental y el nivel es cuasi experimental con pre prueba y post prueba con un solo grupo, el esquema del diseño se muestra a continuación:

Tabla 4

Diseño pre prueba y post prueba con un solo grupo

Aplicación de pre-test o medición inicial	Aplicación del estímulo o tratamiento	Aplicación del post-test o medición final
G O1	X	O2

Fuente: (Arias, 2012)

Donde:

G: Personal administrativo encargado del Repositorio Institucional

O1: Administración del Repositorio Institucional antes de la extracción automática de metadatos mediante el software basado en procesamiento de lenguaje natural.

X: Software para extracción automática de metadatos y publicación de documentos

O2: Administración del Repositorio Institucional después de la extracción automática de metadatos mediante el software basado en procesamiento de lenguaje natural

Al finalizar se establecen las diferencias entre O1 y O2 para determinar si la extracción automática de metadatos y publicación de documentos mejoran la administración del Repositorio Institucional.

3.4.3 Método de tratamiento de datos

Para el tratamiento de datos se usó la prueba t de Student es un estadístico paramétrico que es usado para comparar la media de dos muestras relacionadas y determinar si existen diferencias entre ellas. Esta prueba es usada cuando el investigador quiere comparar 2 grupos con variables cuantitativas continuas y con distribución normal, dicho de otra manera es la comparación de promedios entre 2 grupos (Flores-Ruiz, Miranda-Novales & Villasís-Keever, 2017).

i. Planteamiento de hipótesis

$H_0 : \mu_x = \mu_y$, con la implementación del software de extracción automática de metadatos, no se reduce el tiempo de publicación de documentos de investigación en la administración del Repositorio Institucional

$H_a : \mu_x \neq \mu_y$, con la implementación del software de extracción automática de metadatos, si se reduce el tiempo de publicación de documentos de investigación en la administración del Repositorio Institucional

ii. Nivel de significancia

Se usó un nivel de significancia del 5%, es decir $\alpha = 0.05$

iii. Regla de decisión

Si $p \geq 0.05$, entonces se acepta la H_0 y se rechaza la H_a

Si $p < 0.05$, entonces se rechaza la H_0 y se acepta la H_a

iv. Conclusión

Dependiendo del resultado de la regla de decisión, se dará una interpretación acerca de los datos analizados

3.5 Descripción detallada de métodos por objetivos específicos

Para el cumplimiento de los objetivos específicos utilizaremos los siguientes métodos:

3.5.1 Metodología para el desarrollo del algoritmo de extracción automática de metadatos

En Repositorio Institucional de la Universidad Nacional del Altiplano hace uso de la plataforma de DSpace que utiliza colecciones para agrupar los documentos de investigación que se encuentra dentro de una comunidad y subcomunidad que son creadas por el administrador. Para enviar un documento al repositorio se siguen procesos que se divide en una serie de pasos. En estos pasos se selecciona una comunidad Pregrado/Posgrado, selección de subcomunidad para Áreas/ Posgrados, selección de subcomunidad para las escuelas profesionales/Programas de posgrado, se crea una colección, se describe el objeto mediante los metadatos, se suben los archivos que lo componen, se aceptan las licencias y se realiza una revisión previa a que se complete el deposito. En la figura N° 9 se muestra el diagrama de flujo del proceso mencionado.

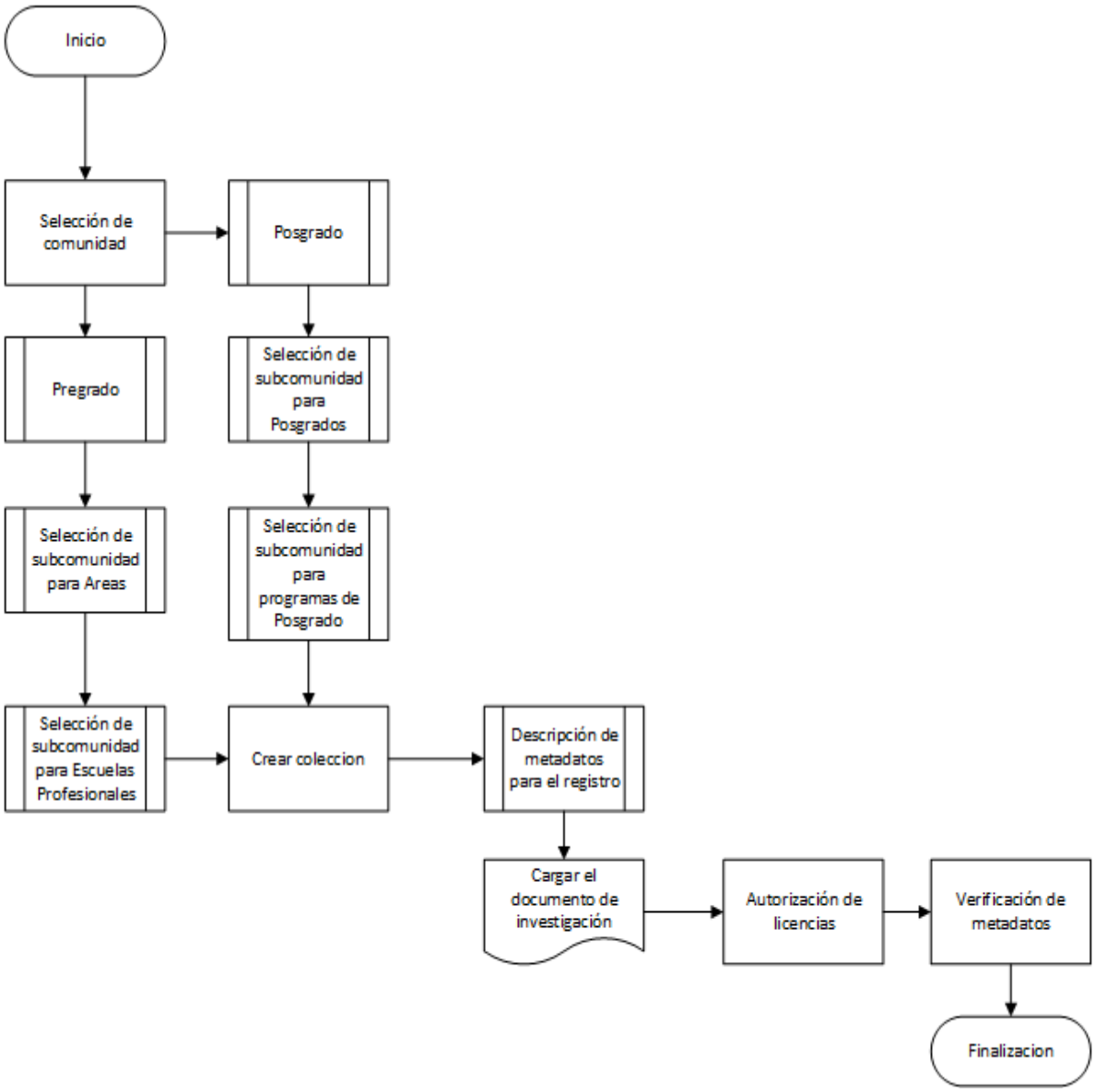


Figura 9. Diagrama de flujo para envió de documentos al Repositorio

Se desarrolló el algoritmo para la extracción automática de metadatos de los documentos de investigación que son publicados en el Repositorio basada en procesamiento de lenguaje natural (NLP), mediante el lenguaje de programación de Python, el cual contiene librerías que permiten resolver problemas relacionados con el (NLP). Para la implementación del desarrollo del algoritmo de extracción automática de metadatos, se reestructuro el proceso de envió de los documentos de investigación al Repositorio, modificando y reordenando los pasos de publicación los cuales permitirán mejorar la administración del Repositorio Institucional.

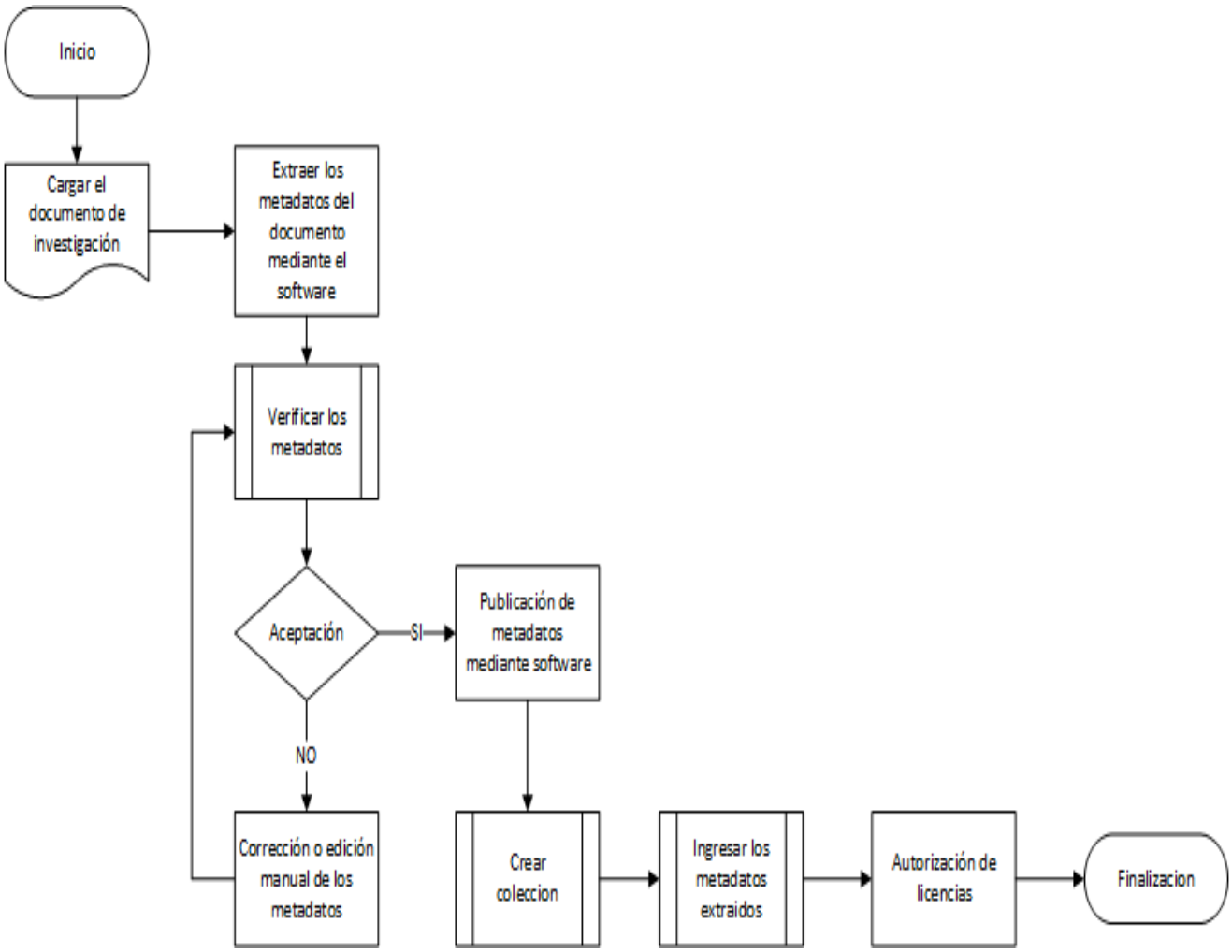


Figura 10. Diagrama de flujo para la implementación del algoritmo

Las principales modificaciones que se realizó en el flujo para enviar documentos al Repositorio son:

1. **Cargar el documento de investigación:** Este paso ahora se realiza antes de llenar los formularios de descripción del objeto debido a que el software implementado se encarga de la extracción automática de metadatos para el completado de los mismos.
2. **Verificación de los metadatos:** En este paso mediante el algoritmo de extracción automática de metadatos, se acepta los metadatos para su validación en los campos correspondientes del formulario o se corrige los metadatos que no fueron extraídos correctamente este proceso se realiza mediante la edición manual de metadatos por parte del personal encargado de publicar los documentos de investigación.

3. **Publicación de los documentos:** Ahora en el paso de descripción de metadatos para el registro del documento que pertenece al proceso de publicación se realiza de forma automática mediante el software al igual que la autorización de licencias necesarias para la publicación.

3.5.2 Nivel de precisión del algoritmo en la extracción automática de metadatos

Las métricas de evaluación permiten medir el rendimiento de un modelo de aprendizaje automático y son de gran importancia para los problemas de clasificación donde se busca diferenciar distintos algoritmos Machine y Deep learning con la finalidad de facilitar la elección del mejor algoritmo dependiendo del objetivo de investigación.

Para medir la precisión del algoritmo de extracción automática de metadatos en los documentos de investigación se utilizó las siguientes métricas (Scikit-Learn, 2022)

Precisión: Es el porcentaje de clasificaciones correctas de nuestro modelo clasificador dentro de las predicciones positivas. Esta métrica permite saber cuántos metadatos extraídos de los documentos de investigación mediante el algoritmo de extracción automática basado en procesamiento de lenguaje natural son realmente los correctos o positivos, y se calcula mediante la siguiente ecuación.

$$Precision: \frac{TP}{TP + FP}$$

Donde:

TP: (True Positive) Verdaderos positivos, son los valores que el algoritmo clasifica como positivos y que realmente son positivos

FP: (False Positive) Falsos positivos, son los valores que el algoritmo clasifica como positivo cuando realmente son negativos

Recall o cobertura: Es una medida que permite conocer la proporción de casos positivos que fueron correctamente clasificados. Esta métrica permite saber cuántos metadatos extraídos de los documentos de investigación mediante el algoritmo de extracción automática basado en procesamiento de lenguaje natural son valores

positivos y fueron correctamente clasificados, se calcula mediante la siguiente ecuación:

$$Recall = \frac{TP}{TP + FN}$$

Donde:

TP: (True Positive) Verdaderos positivos, son los valores que el algoritmo clasifica como positivos y que realmente son positivos

FN: (False Negative) Falsos negativos, son los valores que el algoritmo clasifica como negativo cuando realmente son positivos.

3.5.3 Evaluación de la diferencia del tiempo de extracción de metadatos antes y después del desarrollo del software

En esta etapa se puso a prueba el tiempo del algoritmo para la extracción de los metadatos en comparación con el tiempo que demora el personal encargado de la publicación de los documentos de investigación. Para medir el tiempo se procedió a registrar y analizar el tiempo en minutos que toma al personal encargado en extraer los metadatos sobre una cantidad de documentos de investigación en 4 tareas secuenciales.

- El procesamiento de extraer metadatos en 50 documentos de investigación (G1)
- El procesamiento de extraer metadatos en 150 documentos de investigación (G2)
- El procesamiento de extraer metadatos en 300 documentos de investigación (G3)
- El procesamiento de extraer metadatos en 380 documentos de investigación (G4)

Para la aplicación de estas tareas se utilizó el siguiente diseño experimental

Tabla 5

Diseño experimental propuesto

Grupo	Aplicación de pre-test	Estimulo o tratamiento	Aplicación del pos-test
Gi	OAi	X	ODi

Fuente: Equipo de trabajo

Donde:

Gi: Tarea “i” que realiza el personal administrativo encargado de la publicación de “n” documentos de investigación en el Repositorio Institucional

OAi: Tiempo en horas que demora el personal administrativo encargado de la publicación de “n” documentos de investigación antes del desarrollo del algoritmo de extracción automática de metadatos basada en procesamiento de lenguaje natural

X: Implementación del algoritmo de extracción automática de metadatos para la publicación de documentos de investigación en el Repositorio Institucional

ODi: Tiempo en horas que demora el personal administrativo encargado de la publicación de “n” documentos de investigación después del desarrollo del algoritmo de extracción automática de metadatos basada en procesamiento de lenguaje natural.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1 Resultados conforme al objetivo específico 1

Para extraer los metadatos de forma automática necesarios para la publicación de documentos, primero se identificó cuáles son las comunidades y subcomunidades que son usados para publicar documentos de investigación en el Repositorio Institucional de la Universidad Nacional del Altiplano

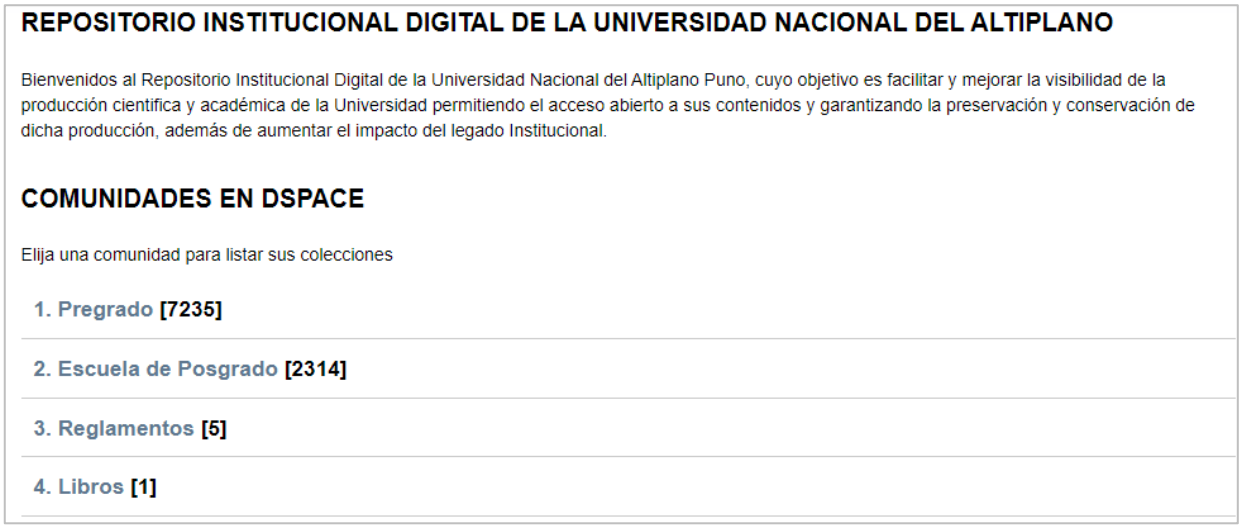


Figura 11. Comunidades del Repositorio Institucional UNAP

Fuente: (VRI UNAP, 2022)

Como se observa en la figura 11 existen cuatro comunidades (Pregrado, Escuela de Posgrado, Reglamentos y Libros) de estas comunidades solo dos son las más usadas para la publicación de documentos Pregrado y Escuela de Posgrado

Figura 12. Subcomunidades en la comunidad de Pregrado

En la comunidad de pregrado existen cuatro subcomunidades Ciencias Biomédicas, Ciencias de la Ingeniería, Ciencias Económicas y Empresariales y Ciencias Sociales las mismas que son consideradas en Áreas. En estas subcomunidades se publican los documentos de investigación (tesis, tesina, artículos) de las Facultades y Escuelas profesionales de la Universidad Nacional del Altiplano.

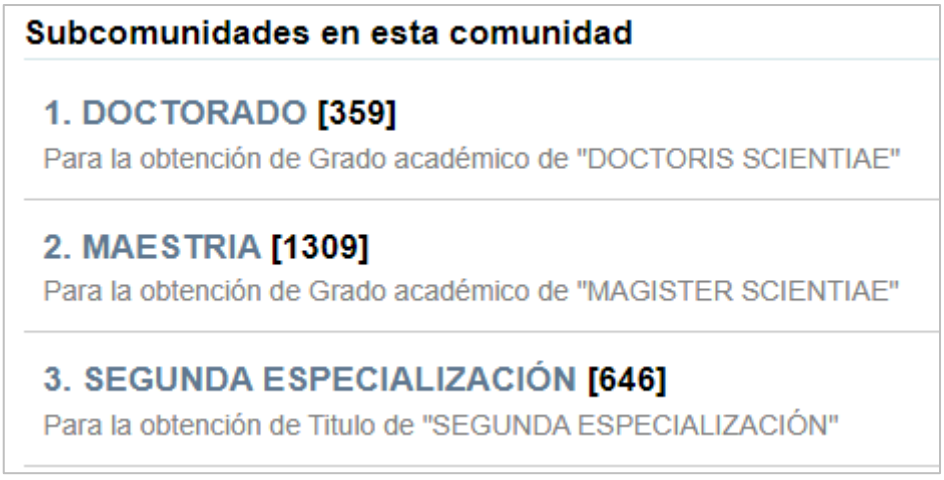


Figura 13. Subcomunidades de la comunidad Escuela de Posgrado

En la comunidad Escuela de Posgrado existe tres subcomunidades doctorado, maestría y segunda especialización, en estas subcomunidades se publican los documentos de investigación (tesis, artículos, informe de experiencia laboral) de los programas de maestrías, doctorados y segunda especialidades en la escuela de posgrado

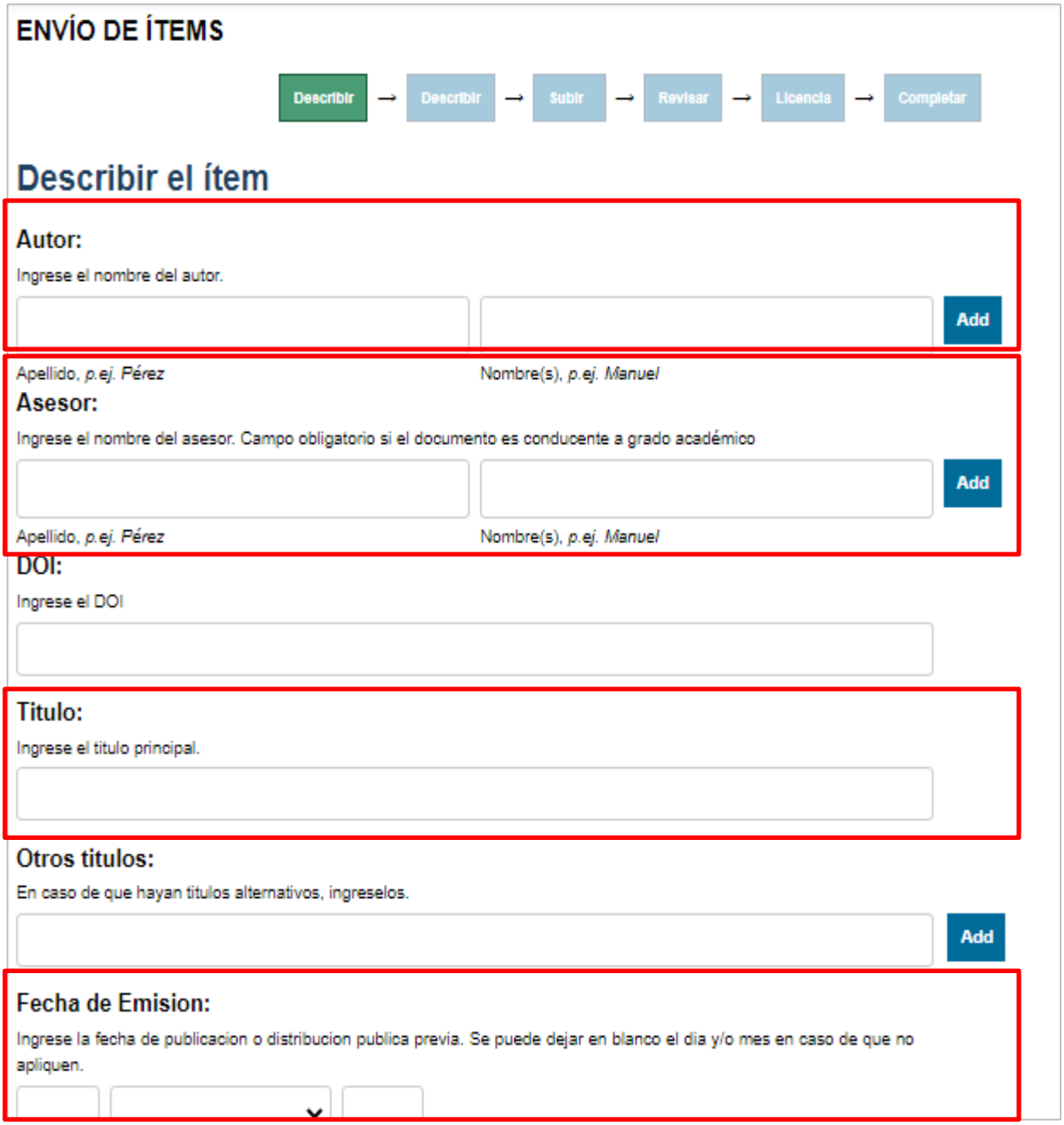


Figura 14. Formulario de envió para la publicación de documentos de investigación en DSpace

De la figura 14 se analizaron los metadatos a extraer para el desarrollo del algoritmo de extracción automática, por lo que los metadatos que serán extraídos por el algoritmo son los que se muestran en la siguiente tabla.

Tabla 6

Metadatos a ser extraídos por el algoritmo

Nombre	Metadato	Descripción
Titulo	dc.title	Nombre del objeto
Autor	dc.contributor.author	Persona responsable de la creación del contenido
Fecha de publicación	dc.date.issued	Fecha de publicación del documento de investigación
Tipo de publicación	dc.type	Tipo de documento a publicar (tesis de bachiller, tesis maestría o tesis doctorado)
Institución que otorga el grado académico	thesis.degree.grantor	Nombre completo de la institución educativa y el departamento o área responsable (Facultades)
Disciplina del campo de conocimiento	thesis.degree.discipline	Escuela profesional, nombre de la maestría o Doctorado
Denominación	thesis.degree.name	Es la denominación asociada al grado académico, tal como aparece en la tesis al obtener el título o grado académico
Nivel de educación	thesis.degree.level	Nivel de educación a obtener el grado o título (Título profesional, Maestría o Doctorado)
Área o Tema	dc.subject	Tema, áreas o líneas de investigación del documento a registrar
Resumen	dc.description.abstract	Resumen del documento a registrar

Fuente: (ALICIA CONCYTEC, 2017)

4.1.1 Diseño

A continuación, se describe el diseño del desarrollo del software de extracción automática de metadatos para la publicación de documentos de investigación en el Repositorio Institucional “E-MeRI”, mediante los diagramas UML

4.1.1.1 Diagrama de caso de uso para extracción de metadatos y publicación de los documentos

Los diagramas de caso de uso es una manera ilustrativa de mostrar toda la funcionalidad del software, en la figura siguiente se puede ver todo lo que el usuario puede realizar.

Figura 15. Diagrama casos de uso para extraer metadatos y publicar documentos

En la figura 15 muestra el caso de uso donde el actor operador puede realizar las siguientes acciones:

- Extracción automática de metadatos: El operador deberá de seleccionar un documento de investigación a extraer metadatos, el proceso se

realiza mediante el algoritmo de extracción automática, que se encarga de extraer los metadatos del documento, una vez extraídos los metadatos se debe de seleccionar el tipo de contenido (tesis de pregrado, tesis de maestría o tesis de doctorado), el tipo de obra (tesis, trabajo de investigación o trabajo de suficiencia profesional) y el nivel de educación (bachiller, título profesional, titulo de segunda especialidad, maestría o doctorado). Estos metadatos no han sido extraídos mediante el algoritmo de extracción debido a que no están disponibles en el documento de investigación, posterior a la visualización de los metadatos extraídos se muestran editables en caso de que algún metadato no se extrajo correctamente y finalmente almacenarlos en la base de datos

- Publicación de documentos: una vez que tengamos los metadatos extraídos y almacenados en la base de datos se debe de verificar, editar o eliminar los metadatos que no fueron extraídos correctamente para finalmente publicar en el Repositorio Institucional.
- Configurar: el operador deberá de registrar sus datos personales en el cual también incluye los datos de acceso al DSpace debido a que es necesario para la publicación automática de los documentos de investigación.

4.1.1.2 Arquitectura de software

Para la arquitectura del software se utilizó la programación por capas (PROP), en la cual se describe de la siguiente manera:

- Capa de presentación: Es la capa que ve el operador en cual se encarga del manejo de la interfaz gráfica, así como la tarea de presentar al operador la capa de dominio de una manera más entendible y la captura de los datos de entrada que realiza el operador los cuales son enviados a la capa de dominio.
- Capa de dominio: También conocida como la capa de lógica de negocio, en esta capa se muestra toda la información y funcionalidad del software, aquí está ubicado el algoritmo de extracción automática que se encarga de extraer los metadatos de los documentos de

investigación el cual está conectado a la capa de datos donde se realiza el almacenamiento de los metadatos extraídos y enviándolos a la capa de presentación haciendo posible la comunicación entre ambas capas. A la misma vez se muestra la publicación de documentos el cual se encarga de la publicación de los documentos procesados mediante el software.

- Capa de datos: Esta capa es la encargada de gestionar toda la información almacenada como también las acciones de guardar, modificar o eliminar los metadatos extraídos de los documentos de investigación. Las mismas que están compuestas por controladores de bases de datos que se comunican con la capa de dominio.

Esta arquitectura planteada tiene la ventaja de que los módulos trabajen de forma independiente, lo cual implica que pueden ser desarrolladas paralelamente logrando una mejor eficiencia de trabajo grupal, una ventaja adicional es también la reutilización del código, lo que permite que si existen cambios en una capa no afecta a las otras capas, por lo que solo es necesario realizar modificaciones al módulo a editar.

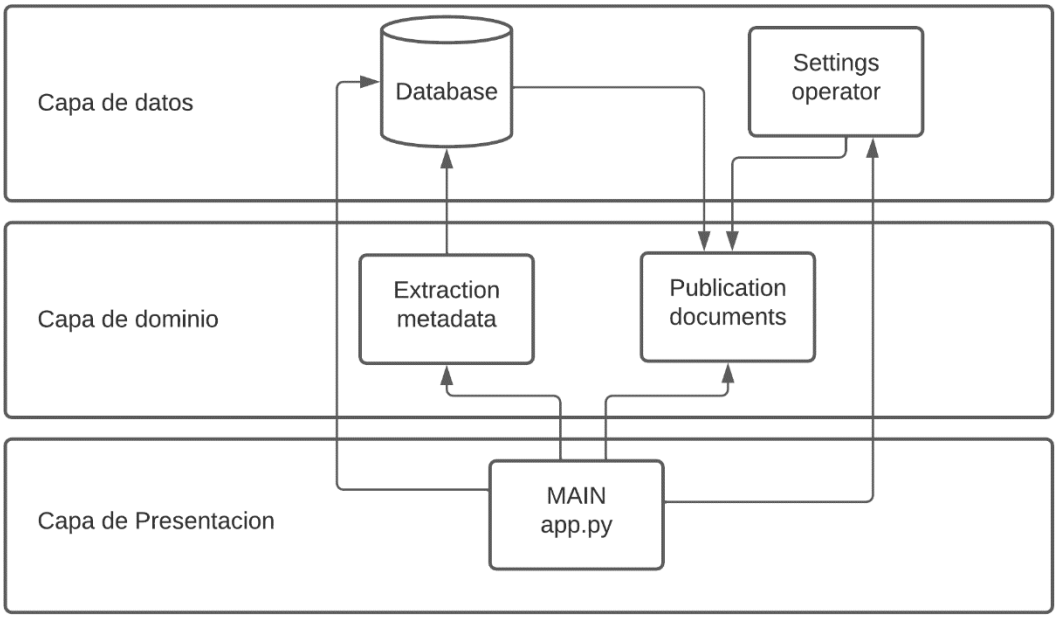


Figura 16. Arquitectura del software

4.1.2 Desarrollo del software

A continuación, se detallará la implementación de los módulos del software

4.1.2.1 Algoritmo para la extracción automática de metadatos

Para realizar la extracción automática de metadatos se utilizó el lenguaje de programación Python, debido a que cuenta con librerías que permiten el análisis lingüístico de documentos y el procesamiento de lenguaje natural. La plataforma usada para el análisis del texto fue, Natural Language Toolkit (NLTK).

Se inició el algoritmo cargando el documento de investigación al que se extraerá los metadatos el cual es recibido en formato PDF, para realizar el análisis del documento es necesario que este en formato de texto de manera que se procedió a convertir de PDF a texto para lo cual se usó la librería PyPDF2 que permite el manejo de PDFs como extraer información de un documento, dividir documentos por páginas, fusionar varios documentos en uno solo entre otras acciones.

Con la obtención del texto posteriormente se procedió con la limpieza esto significa convertir el texto en bruto en una lista de palabras y guardarlo de nuevo, para lo cual se dividió el documento por espacios en blanco, nuevas líneas, tabulaciones y saltos de líneas. Luego se eliminó la puntuación a cada palabra utilizando expresiones regulares para seleccionar los caracteres de puntuación con una constante denominada string.punctuation y reemplazándolas con nada. Para un mejor procesamiento del texto se hizo la conversión de mayúsculas a minúsculas de cada palabra del documento.

Una vez normalizado el texto, se utilizó NLTK para dividir cadenas en tokens con la función word_tokenize. Con el documento tokenizado se hizo más fácil el procesamiento del texto filtrando los tokens de interés para la extracción de metadatos del documento

```
#*****************************************************************************************************
#*              Module to extract metadata from documents                                            *
#*****************************************************************************************************
mydocument=open('tesis_a_extraer.pdf',mode="rb")
converttotext=pdftotext.PDF(mydocument)
# split words by space
words = converttotext.split()
#prepare regex for character filtering
re_punc = re.compile('[%s]' % re.escape(string.punctuation))
#convert to lowercase
words = [word.lower() for word in words]
#remove punctuation from each word
stripped = [re_punc.sub('', w) for w in words]
text=stripped
tokens=nltk.word_tokenize(text)
textDocument=nltk.Text(tokens)
#Metadata to extract
```

Figura 17. Código para la conversión y normalización del documento

La función getDescriptionUri() de la figura 18 recibe como parámetro el documento en consulta procesado y normalizado en tokens, su función es extraer el metadato tipo de obra (dc.type), iterando sobre todas las palabras del documento y comparando con la expresión regular para obtener el tipo de obra tesis, tesina, articulo o informe, debido a que son los documentos de investigación que son publicados en el Repositorio Institucional de la Universidad Nacional del Altiplano.

```
#Function to extract metadata Description Uri (Tipo de obra)
def getDescriptionUri(textDocument):

  for w in textDocument:
    match=re.search("tesis|tesina|articulo|informe",w)
    if (match):
      _metadataDescriptionUri=w
  if (_metadataDescriptionUri==""):
    _metadataDescriptionUri="Tipo de obra no encontrada"

  return _metadataDescriptionUri
```

Figura 18. Código para extraer el metadato tipo de obra (dc.type)

La función getTitle() como se muestra en la figura 19 recibe como parámetro el documento procesado y el metadato tipo de obra como tesis, tesina, articulo o informe. Su función es extraer el metadato título (dc.title), debido a que el

título del documento puede contener diferentes caracteres, se realizó la tokenizacion del documento con expresiones regulares. Después se itera sobre una lista que contiene un diccionario con las palabras clave que son incluidas en el título de un documento de investigación, estas se encuentran almacenadas en la base de datos, luego se extrae los tokens reconocidos para el metadato del título y se destokeniza para unir los tokens que conforman el título del documento de investigación.

```
#Function to extract metadata Title (Titulo)
def getTitle(textDocument,_metadataDescriptionUri):
  #connection to the database to use the dictionaries
  sql="SELECT palabra FROM diccionario;"
  conn=mysql.connect()
  cursor=conn.cursor()
  cursor.execute(sql)
  wordsTitle=cursor.fetchall()
  conn.commit()
  pattern = r'''(?x)(?:[A-Z]\.)+| \w+(?:-\w+)*| \$?\d+(?:\.\d+)?%?|'''

  for t in nltk.regexp_tokenize(textDocument,pattern):
    for w in wordsTitle:
      match=re.search(w,t)
      if (match):
        _Title=w

  tokenizer = RegexpTokenizer(r'(_Title(.*?)_metadataDescriptionUri)')
  text=TreebankWordDetokenizer().detokenize(textDocument)
  _metadataTitle = tokenizer.tokenize(text)
  if (_metadataTitle==""):
      _metadataTitle="Titulo no encontrado, se sugiere ingresar nuevo termino al diccionario"

  return _metadataTitle
```

Figura 19. Código para extraer el metadato título (dc.title)

La función getFirstLastName(), ver figura 20 recibe como parámetro el documento procesado y normalizado en tokens, su función es extraer el metadato nombre del autor (dc.contributor.author), iterando sobre todas las palabras del documento y comparando con la expresión regular para obtener el nombre del autor, luego se extrae los tokens reconocidos para el metadato nombre y se destokeniza para unir los tokens que conforman el nombre completo del autor. Debido a que el metadato autor del documento es extraído con nombres y apellidos juntos es necesario separarlos para la publicación del documento por lo que se usa la función splitName(), encargada de obtener el nombre del autor y apellido del autor por separado.

```
#Function to extract metadata author first name and last name(Nombres y apellidos del autor)
def getFirstLastName(textDocument):
  for w in textDocument:
    match=re.search("presentada|presentado|por|bach.|bach",w)
    if (match):
      AuthorName=w

  tokenizer = RegexpTokenizer(r'(_AuthorName(.*?)_para')
  text=TreebankWordDetokenizer().detokenize(textDocument)
  _metadataAuthorName = tokenizer.tokenize(text)
  if (_metadataAuthorName==""):
      _metadataAuthorName="Nombre del autor no encontrado"

  firstName=splitName(_metadataAuthoName)[0]
  lastName=splitName(_metadataAuthorName)[1]

  return [firstName,lastName]
```

Figura 20. Código para extraer el metadato nombre del autor (dc.contributor.author)

La función getDegreeName() de la figura 21 acepta como parámetro el texto del documento normalizado, su función es extraer el metadato denominación del grado académico (thesis.degree.name) iterando sobre todas las palabras del documento y comparando con la expresión regular para obtener el nombre de la denominación del grado académico, después se extrae los tokens reconocidos para el metadato y se destokeniza para unir los tokens que conforman la denominación del grado académico.

```
#Function to extract metadata degree name (denominacion del grado academico)
def getDegreename(textDocument):
  for w in textDocument:
    match=re.search("para|optar|profesional|grado|academico",w)
    if (match):
      DegreeName=w

  tokenizer = RegexpTokenizer(r'(_DegreeName(.*?)_place')
  text=TreebankWordDetokenizer().detokenize(textDocument)
  _metadataDegreeName = tokenizer.tokenize(text)
  if (_metadataDegreeName==""):
      _metadataDegreeName="Denominacion no encontrada"

  return _metadataDegreeName
```

Figura 21. Código para extraer el metadato denominación (thesis.degree.name)

De la figura 22 la función getDegreeDiscipline() recibe como parámetro el texto del documento tokenizado y el metadato título, su función es extraer el

metadato disciplina del campo de conocimiento (thesis.degree.discipline) que viene a ser el nombre la escuela profesional, maestría o doctorado. La función itera sobre todas las palabras del documento y comparando con la expresión regular para obtener la disciplina del campo de conocimiento, luego se extrae los tokens reconocidos para el metadato y se destokeniza para unir los tokens que conforman el metadato.

```
#Function to extract metadata degree discipline (escuela)
def getDegreeDiscipline(textDocument,_metadataTitle):
  for w in textDocument:
    match=re.search("escuela",w)
    if (match):
      escuela=w

  tokenizer = RegexpTokenizer(r'(escuela(.*?)_metadataTitle')
  text=TreebankWordDetokenizer().detokenize(textDocument)
  _metadataDegreeDiscipline = tokenizer.tokenize(text)
  if (_metadataDegreeDiscipline==""):
      _metadataDegreeDiscipline="Escuela profesional no encontrada"

  return _metadataDegreeDiscipline
```

Figura 22. Código para extraer el metadato disciplina del campo de conocimiento (thesis.degree.discipline)

Como se muestra en la figura 23 la función getDegreeGrantor() recibe como parámetro el texto del documento normalizado y el metadato de la disciplina del campo de conocimiento, su función es extraer el metadato de la institución que otorga el grado académico (thesis.degree.grantor) que es el nombre completo de la facultad. Esta función se encarga de iterar sobre todas las palabras del documento comparando con la expresión regular para obtener el metadato, luego se extrae los tokens reconocidos para la facultad con la clase RegexpTokenizer y se destokeniza con la clase TreebankWordTokenizer para finalmente unir los tokens.

```
#Function to extract metadata Degree Grantor (Facultad)
def getDegreeGrantor(textDocument,_metadataDegreeDiscipline):
  for w in textDocument:
    match=re.search("facultad",w)
    if (match):
      facultad=w
  tokenizer = RegexpTokenizer(r'(facultad(.*?)_metadataDegreeDiscipline')
  text=TreebankWordDetokenizer().detokenize(textDocument)
  _metadataDegreeGrantor= tokenizer.tokenize(text)
  if (_metadataDegreeGrantor==""):
      _metadataDegreeGrantor="Facultad no encontrada"

  return _metadataDegreeGrantor
```

Figura 23. Código extraer el metadato disciplina del campo de conocimiento (thesis.degree.grantor)

La función getDate() de la figura 24 acepta como parámetro el texto del documento tokenizado, su función es extraer el metadato fecha de publicación (dc.date.issued), iterando sobre todas las palabras del documento comparando con la expresión regular para obtener el metadato, luego se extrae los tokens reconocidos para obtener la fecha con la clase RegexpTokenizer y se destokeniza con la clase TreebankWordTokenizer para finalmente unir los tokens. Debido a que el metadato fecha de publicación se extrae en formato de fecha larga es necesario separarlos en dia, mes y año para la publicación del documento por lo que se usó una lista para convertir la fecha en un formato estándar de dd-mm-yyyy.

```
#Function to extract metadata Degree Date (Fecha)
def getDate(textDocument):
  for w in textDocument:
    match=re.search("fecha|fecha sustentacion|enero|febrero|marzo|abril|mayo|junio|julio|agosto|setiembre|octubre|noviembe|dic
    if (match):
      date=w
  tokenizer = RegexpTokenizer(r'(fecha(.*?)_year')
  text=TreebankWordDetokenizer().detokenize(textDocument)
  _metadataDate= tokenizer.tokenize(text)
  _date=_metadataDate.slipt()
  dicMonths=[['enero','febrero','marzo','abril','mayo','junio','julio','agosto','setiembre','octubre','noviembre','diciembre']
  for month in range(len(dicMonths)):
    for col in range(len(dicMonths[_date[1]])):
      if (dicMonths[month][col])==_date[2]:
        _month=dicMonths[2][col]
  _day=_date[1]
  _year=_date[4]
  if (_metadataDate==""):
      _metadataDate="Fecha no encontrada"

  return _year+"-"+_month+"-"+_day
```

Figura 24. Código para extraer el metadato de fecha de publicación (dc.date.issued)

De la figura 25 la función getSubject() recibe como parámetro el texto del documento normalizado, su función es extraer el metadato tema o área del documento a publicar (dc.subject). La función itera sobre todas las palabras del documento y comparando con la expresión regular para el tema, área o línea de investigación, luego se extrae los tokens reconocidos para el metadato con la clase RegexpTokenizer y se destokeniza con la clase TreebankWordTokenizer para finalmente unir los tokens.

```
#Function to extract metadata subjet (Area|Tema)
def getSubject(textDocument):
  for w in textDocument:
    match=re.search("area|tema|linea",w)
    if (match):
      subject=w
  tokenizer = RegexpTokenizer(r'(subject(.*?)[^0-9a-zA-Z]')
  text=TreebankWordDetokenizer().detokenize(textDocument)
  _metadataSubject = tokenizer.tokenize(text)
  if (_metadataSubject==""):
      _metadataSubject=subject +" no encontrado."

  return _metadataSubject
```

Figura 25. Código para extraer el metadato tema o área (dc.subject)

Como se muestra en la figura 26 la función getAbstract() acepta como parámetro el texto del documento tokenizado, su función es extraer el metadato resumen del documento (dc.abstract), Esta función se encarga de iterar sobre todas las palabras del documento a través del patrón y comparando con la expresión regular para obtener el metadato, luego se extrae los tokens reconocidos para el resumen con la clase RegexpTokenizer y se destokeniza con la clase TreebankWordTokenizer para finalmente unir los tokens.

```
#Function to extract metadata absctract (Resumen)
def getAbsctract(textDocument):

  pattern = r'''resumen|abstract|(?x)(?:[A-Z]\.)+| \w+(?:-\w+)*| \$?\d+(?:\.\d+)?%?|'''

  for t in nltk.regexp_tokenize(textDocument,pattern):
    match=re.search(w,t)
    if(match):
        _absctract=w
  tokenizer = RegexpTokenizer(r'(abastract(.*?)[^0-9a-zA-Z]')
  text=TreebankWordDetokenizer().detokenize(textDocument)
  _metadataAbstract = tokenizer.tokenize(text)
  if (_metadataAbstract==""):
      _metadataAbstract="Resumen no encontrado"

  return _metadataAbstract
```

Figura 26. Código para extraer el metadato resumen (dc.abstract)

A fin de evaluar la eficiencia del algoritmo para la extracción automática de metadatos se realizó la comparación con las herramientas extractoras GROBID, Keyphrase extraction algorithm (KEA), Mr.Dlib y ParsCit debido a que estas herramientas permiten la integración con otros proyectos de desarrollo como por ejemplo repositorios institucionales, bibliotecas digitales entre otros, a través de módulos o una aplicación independiente que permita cargar documentos. A partir de aquí para efectos de prueba se realizó la extracción de metadatos sobre 50 documentos del Repositorio Institucional de la Universidad Nacional del Altiplano con las herramientas mencionadas.

Las pruebas que se efectuaron tienen como finalidad evaluar los resultados obtenidos al realizar la extracción de los metadatos: nombre completo del autor, titulo, fecha de publicación, tipo de documento, institución que otorga el grado académico, escuela profesional nombre de la maestría o doctorado, denominación, nivel de educación, área, tema y resumen. A la mima vez se procedió a registrar y analizar el tiempo promedio en minutos de respuesta para extraer los metadatos y la cantidad de metadatos extraídos.

Tabla 7

Comparación en cantidad y tiempo de los metadatos extraídos con otras herramientas extractoras

Metadato	Herramientas						Algoritmo de extracción automática	
	KEA		Mr.Dlib		Parscit			
	Cant.	Tiempo promedio en min	Cant.	Tiempo promedio en min	Cant.	Tiempo promedio en min	Cant.	Tiempo promedio en min
Autor	30		24		25		50	
Titulo	27		33		31		50	
Fecha de publicación	-		-		-		46	
Tipo de documento	-		-		-		49	
Facultad, Maestría o Doctorado	-	00:39:49	-	00:31:65	-	00:38:74	50	00:12:88
Denominación	-		-		-		49	
Nivel de educación	-		-		-		50	
Área	-		-		-		44	
Tema	-		-		-		39	
Resumen o Abstract	25		11		23		50	

Fuente: Equipo de trabajo

En la tabla 7, se puede visualizar que las herramientas KEA, Mr.Dlib y Parscit extraen parcialmente los metadatos, los cuales son autor, título y resumen a diferencia del algoritmo que extrae todos los metadatos. En cuanto al tiempo de extracción, a las herramientas les toma más tiempo en comparación en el algoritmo de extracción automática de metadatos propuesto en esta investigación.

Complejidad algorítmica

Se realizó el análisis de la complejidad del algoritmo sobre el proceso de extracción automática de metadatos mediante la notación Big O

```
def extractMetadata(document):
  mydocument=open(document,mode="rb")       #(1)
  converttotext=pdftotext.PDF(mydocument)   #(1)
  # split words by space
  words = converttotext.split()             #(1)
  #prepare regex for character filtering
  re_punc = re.compile('[%s]' % re.escape(string.punctuation))  #(1)
  #convert to lowercase
  words = [word.lower() for word in words]  #(1)
  #remove punctuation from each word
  stripped = [re_punc.sub('', w) for w in words]  #(1)
  text=stripped     #(1)
  tokens=nltk.word_tokenize(text) #(1)
  textDocument=nltk.Text(tokens)  #(1)
  #Metadata to extract
  _metadataName=getFirstLastName(textDocument)  #O(n)
  _metadataTitle=getTitle(textDocument)   #O(n)
  _metadataDegreeName=getDegreeName(textDocument) #O(n)
  _metadataDegreeGrantor=getDegreeGrantor(textDocument)   #O(n)
  _metadataDegreeDiscipline=getDegreeDiscipline(textDocument)   #O(n)
  _metadataDescriptionUri=getDescriptionUri(textDocument)   #O(n)
  _metadataDate=getDate(textDocument)   #O(n)
  _metadataSubject=getSubject(textDocument)   #O(n)
  _metadataAbstract=getAbstract(textDocument)   #O(n)
```

Figura 27. Complejidad algorítmica para la extracción automática de metadatos

De la figura 27, para la extracción automática de metadatos se utiliza un módulo principal mediante una función el cual recibe como parámetro los documentos en formato PDF. Se inicia con la lectura del documento, donde se hace uso del procesamiento de lenguaje natural para convertir el texto en una lista de palabras que reciben un proceso con el fin de normalizar el texto del documento, estas líneas de código se ejecutan una sola vez lo que equivale a O(1). Luego el texto normalizado es enviado a las distintas funciones que tienen como objetivo extraer los distintos metadatos del documento, cada una de estas funciones contienen ciclos for para iterar en los tokens dentro del texto normalizado, el cual realizara un matching, por lo que las líneas de códigos dentro de las funciones iteran n veces, lo equivalen a O(n), para cada una de las funciones.

$$O(G) = 8 + 9(O(n))$$

$$O(G) = O(n)$$

Realizando la suma de todo lo calculado, simplificándolo y obteniendo el valor más representativo, se determina que la complejidad del algoritmo para la extracción automática de metadatos en los documentos de investigación es lineal O(n). Por lo que el tiempo de ejecución del algoritmo dependerá de la cantidad de tokens que se obtuvieron del texto normalizado mediante el procesamiento de lenguaje natural de cada documento.

4.1.2.2 Algoritmo para la publicación de documentos

Para realizar la publicación de documentos de investigación en el Repositorio Institucional mediante la plataforma DSpace se utilizó la herramienta Mechanize debido a que permite automatizar las interacciones con los sitios web. El algoritmo de automatización del publicado inicio abriendo la página web del Repositorio Institucional de la Universidad Nacional del Altiplano (ver figura 28), luego con los accesos se ingresó a la cuenta del personal encargado de la publicación de documentos. Para crear la colección en la subcomunidad correspondiente a la escuela profesional, maestria o doctorado del documento se usó la librería BeautifulSoup para iterar sobre los y comprobar si se trata de un que contiene la subcomunidad donde se publicara el documento posteriormente se selecciona la colección para él envió de ítems.

```
#*************************************************************************************
#*        MODULE TO PUBLISH DOCUMENTS                                                *
#*************************************************************************************
#

br = mechanize.Browser()
br.set_handle_robots(False)
br.open(_page)
#Login
br.select_form
posting = _urlcolecction
r = br.open(posting)
#Create colecction
br.form[_name]= _nameDocument
req = requests.get(_urlcolecction)
soup = BeautifulSoup(req.text, "lxml")
colecctions=soup.find_all('div',attrs={'class':'artifact-title'})
colecctionDocument=colecctions('span').getText()
```

Figura 28. Código para el ingreso al DSpace y creación de colección

Para él envió del documento a publicar en la colección (ver figura 29), se seleccionó el formulario mediante la función browser.select_form, luego para el rellenado de los campos se ingresó los metadatos extraídos del documento:

apellido del autor, nombre del autor, titulo, fecha de publicación, nombre de la institución, denominación del estudiante, nombre completo de la institución y facultad, disciplina del campo de conocimiento los cuales se muestran en la figura 30. Los metadatos tipo de contenido, idioma, derechos, licencia y el nombre completo de la institución se ingresan mediante declaraciones de variables definidas en el algoritmo de publicación. En cuanto a los metadatos apellido del asesor, nombre del asesor, tipo de contenido y nivel de educación se rellenan mediante selección en la aplicación del software. Después se usó br.submit() para ir al siguiente formulario del envió

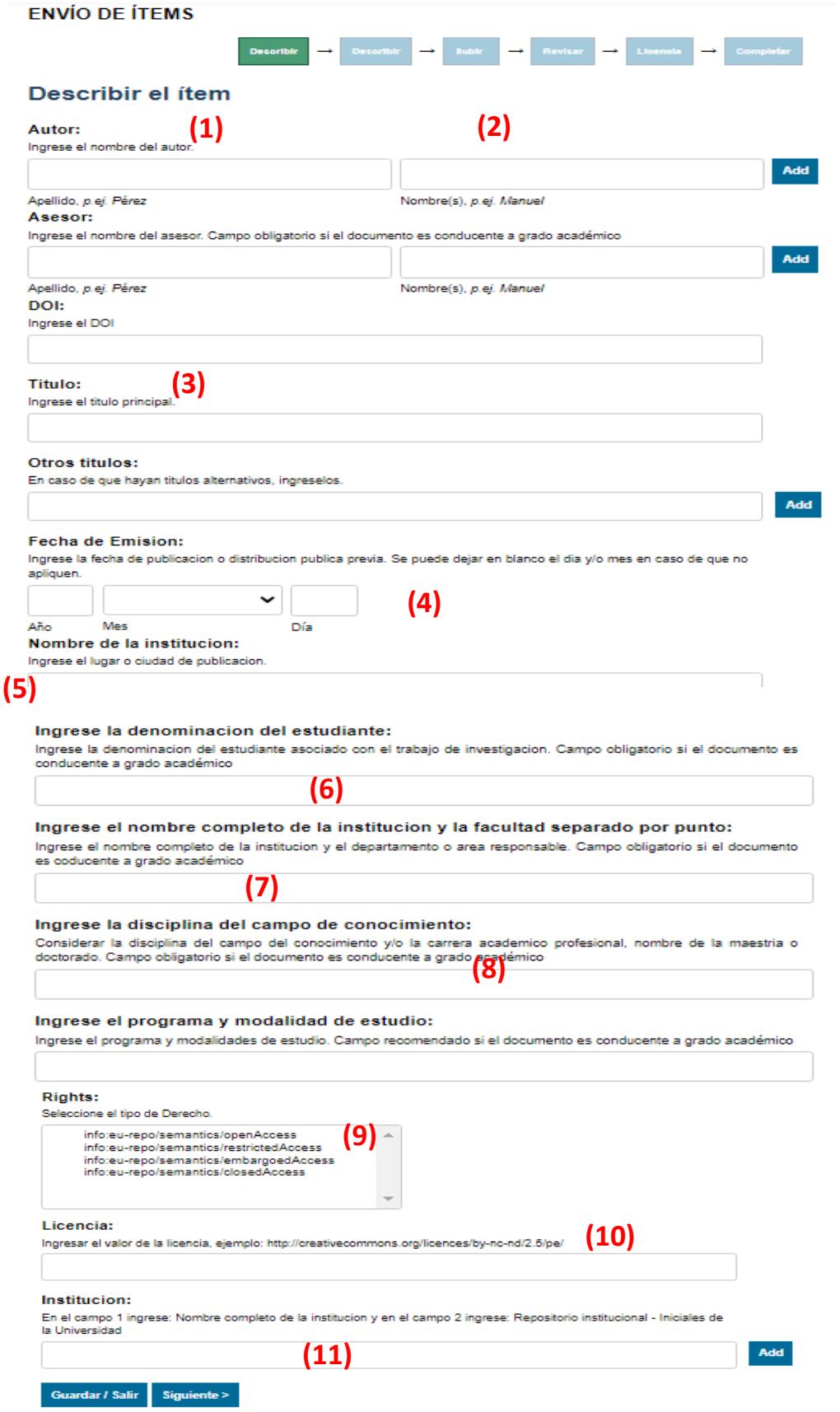
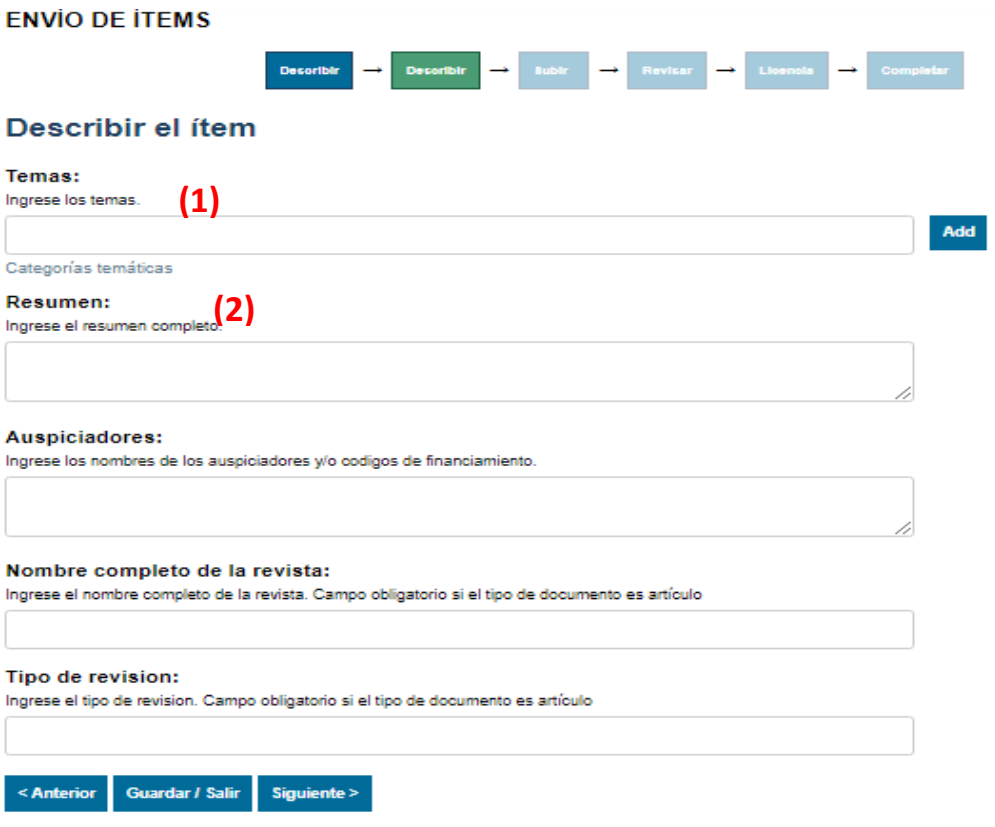


Figura 29. Formulario para él envió de ítems a una colección

```
#publication of document in the collection
br.select_form
br.form[_lastAuthor]= _metadataLastname                #(1)
br.form[_firstAuthor]=_metadataFirstname               #(2)
br.form[_title]=_metadataTitle                         #(3)
br.form[_date]=_metadataDate                           #(4)
br.form[_publisher]=_nameUniversidad                   #(5)
br.form[_degreeName]=_metadataDegreeName               #(6)
br.form[_degreeGrantor]=_metadataDegreeGrantor         #(7)
br.form[_degreeDiscipline]=_metadataDegreeDiscipline   #(8)
br.form[_rights]=_metadataRights                       #(9)
br.form[_rightsUri]=_metadataRightsUri                 #(10)
br.form[_source]=_nameRepository                       #(11)
br.submit(submit)
```

Figura 30. Código para el rellenado de campos en el formulario

En el paso siguiente del envió (ver figura 31), se rellenó los campos del formulario con los metadatos extraídos: área o tema y resumen o abstract del documento




```
#pick item description
br.select_form
br.form[_subject]=_metadataSubject        #(1)
br.form[_abstract]=metadataAbstract       #(2)
```

Figura 31. Código para el rellenado de los metadatos tema y resumen

Para el siguiente paso del envió (ver figura 32), se selecciona el documento de investigación a publicar, luego con br.submit() pasamos a la revisión del envió posteriormente aceptamos las licencias para finalmente publicar el documento de investigación en la colección correspondiente a la subcomunidad.

```
#pick document
br.add_file(open(_documentName'rb'))
br.submit("submit_next")
#Verification submission items
br.submit()
#Licenses
br.form[_decisions]=[(_accept or decline)]
br.submit()
```

Figura 32. Código para la selección del documento y aceptación de licencias

4.2 Resultados conforme al objetivo específico 2

Para la implementación del software, se utilizó Flask que es un microframework escrito y desarrollado en Python para crear aplicaciones como páginas web dinámicas, APIs entre otras bajo el patrón Modelo-Vista-Controlador (MVC), además permite desarrollar aplicaciones de una manera rápida y ágil debido a que se puede instalar extensiones o complementos acorde al tipo de proyecto a desarrollar.

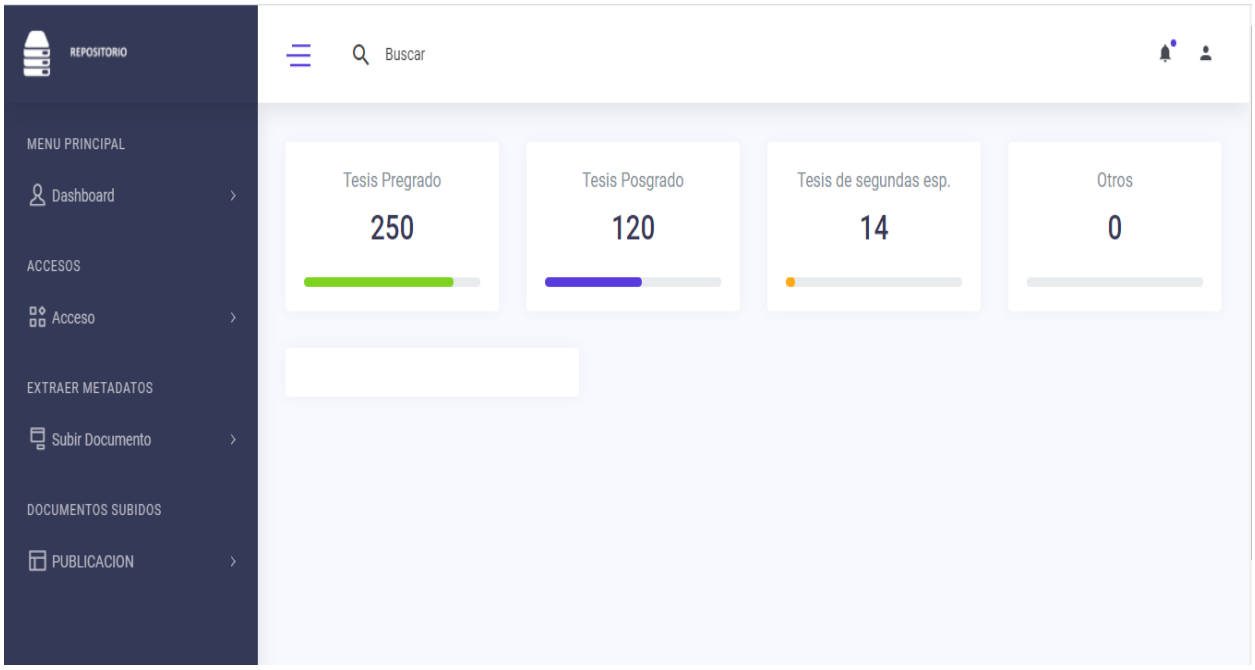


Figura 33. Ventana principal del software

En la figura 33 se muestra la ventana principal de la aplicación, el cual contiene un menú principal donde se encuentra el dashboard, perfiles de usuarios, el módulo de extracción de metadatos y el módulo de publicación de documentos de investigación donde figura todos los documentos extraídos hasta el momento.

Dentro del módulo de extracción de metadatos (ver figura 34), se muestra la opción subir documentos donde se selecciona el documento de investigación al cual se le aplicara la extracción automática de metadatos.

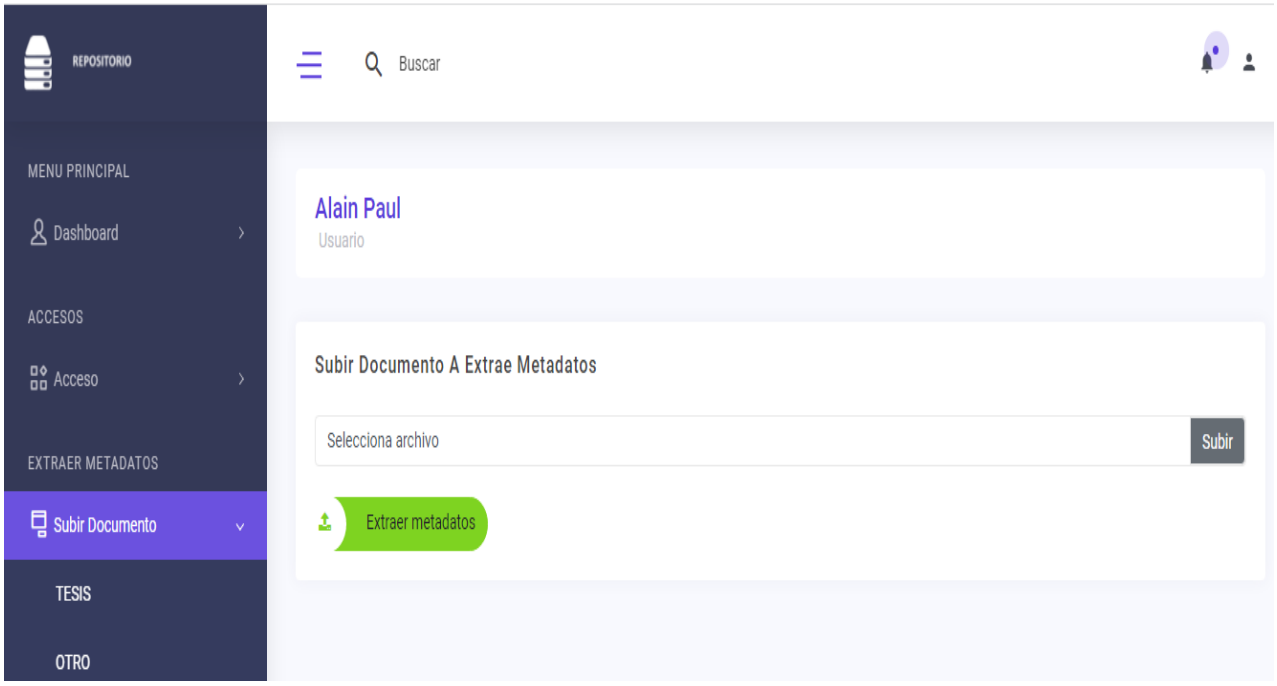


Figura 34. Vista del módulo de extracción automática de metadatos

Luego de seleccionar y enviar el documento de investigación al cual se extraerá los metadatos de forma automática, se visualiza los metadatos extraídos en la siguiente vista (ver figura 35).

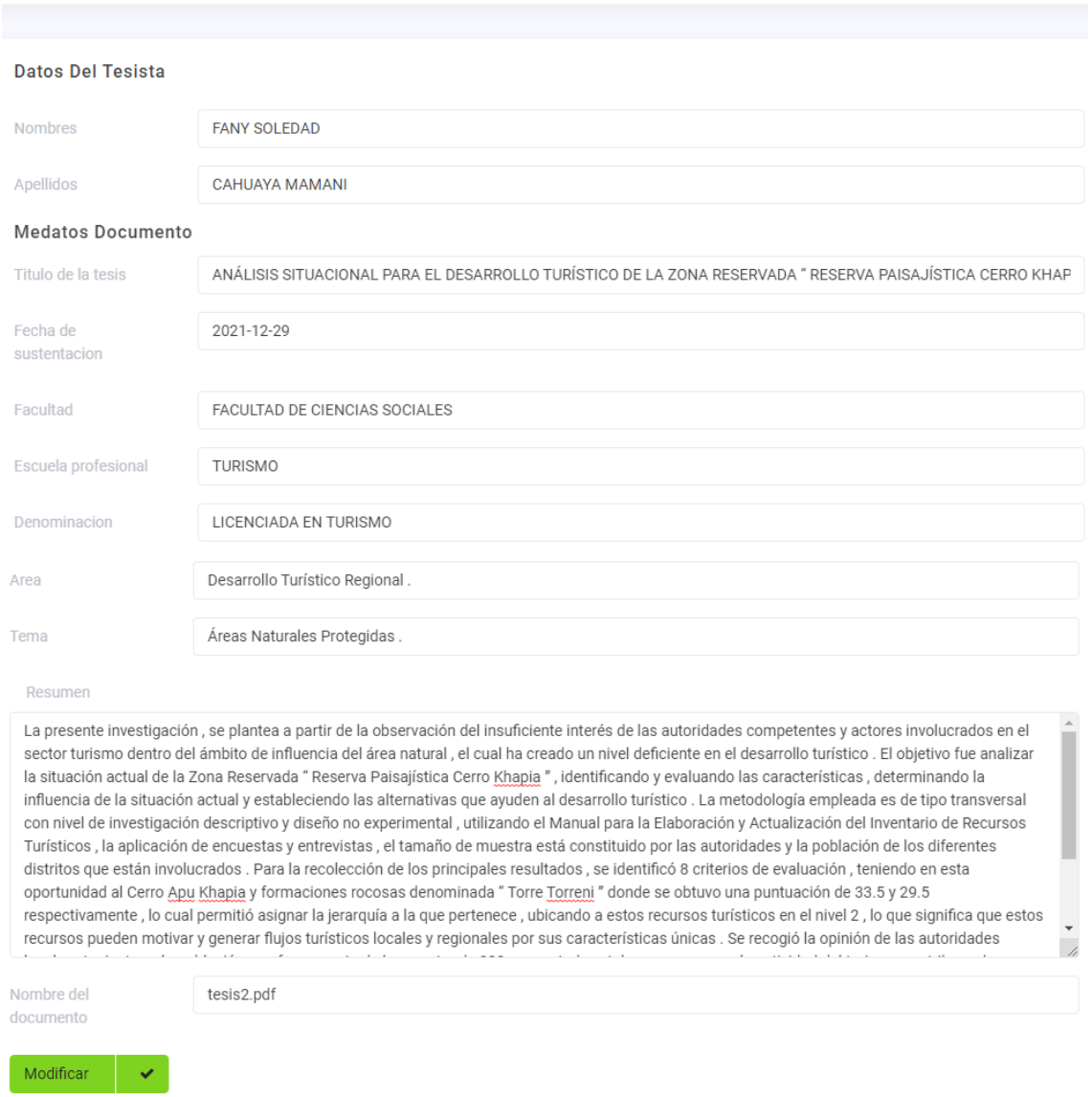


Figura 35. Metadatos extraídos automáticamente del documento de investigación

La opción publicación en el menú documentos subidos (ver figura 36), muestra todos los documentos de investigación a los que se les aplico la extracción automática de metadatos. En este módulo se realiza la publicación de los documentos dando clic en el botón publicar, una vez efectuada la acción, el estado cambia a publicado, a la misma vez el documento figura en la plataforma del DSpace.

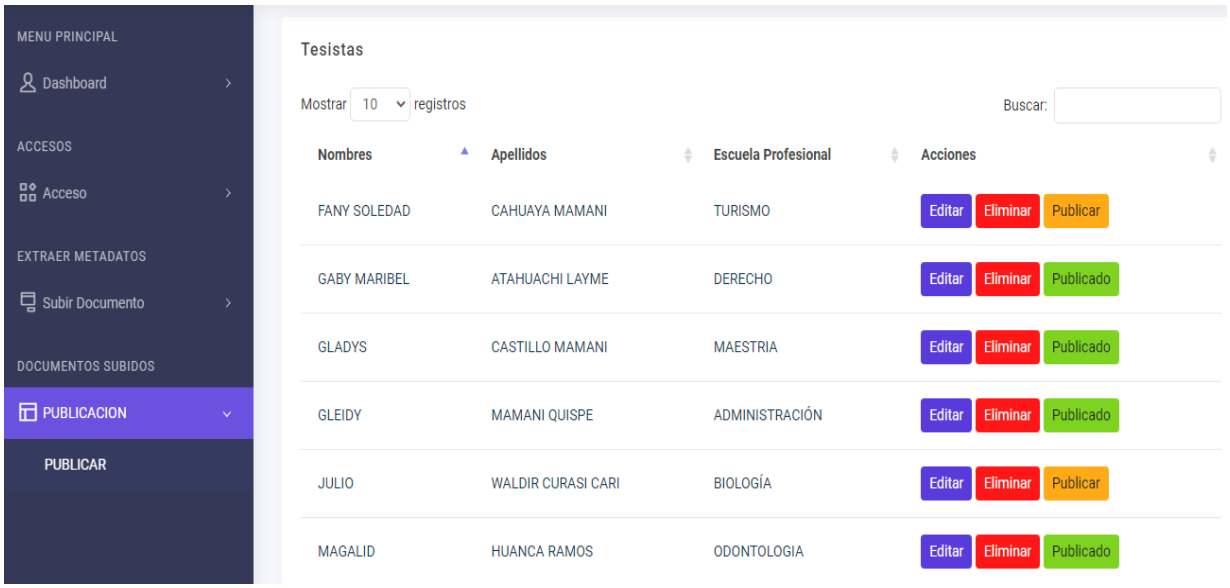


Figura 36. Vista del módulo para la publicación de los documentos

- **Nivel de precisión del Algoritmo**

  Para determinar el rendimiento del software se utilizó la métrica de precisión en la extracción de los metadatos, de los 380 documentos de investigación en formato PDF procesados, el resultado se clasifico en cuatro categorías. TP (true positive), representa los valores denominados verdaderos positivos, es decir, son aquellos valores que fueron identificados por el software como los metadatos correspondientes y de hecho lo son. TN (true negative) son los metadatos que no se han podido extraer y realmente tampoco aparecen en el documento. FP (false positive) representan los valores denominados falsos positivos, es decir, son los valores identificados por el software como metadatos, pero no son los correctos. Finalmente, FN (false negative) representa los valores denominados falsos negativos, es decir, son aquellos valores que el software no identifico como los metadatos correspondientes, pero deben de considerarse correctos.

Tabla 8

Resultados de la precisión y cobertura para cada uno de los metadatos extraídos

METADATO	TP	TN	FP	FN	Precisión	Cobertura
Titulo	368	0	9	3	0.98	0.99
Autor	359	0	10	11	0.97	0.97
Fecha de publicación	349	10	12	9	0.96	0.97
Tipo de publicación	366	0	8	6	0.98	0.98
Institución que otorga el grado académico(Facultad)	368	0	9	3	0.98	0.99
Escuela profesional, nombre de maestría o doctorado	372	0	6	2	0.98	0.99
Denominación	366	0	7	7	0.98	0.98
Nivel de educación	368	0	6	6	0.98	0.98
Área	342	18	10	10	0.97	0.97
Tema	338	16	13	13	0.96	0.96
Resumen	370	0	5	5	0.99	0.99

Fuente: Equipo de trabajo

En la tabla 8, se muestran los resultados de la precisión y cobertura de cada uno de los metadatos extraídos sobre los 380 documentos procesados, se puede observar el desempeño del software “E-MeRI”: 0.99 de precisión para el resumen, lo que significa que obtuvo un 99% de resultados correctos en la extracción del metadato. Una precisión del 0.98 para títulos; tipo de documento a publicar; institución que otorga el grado académico (facultades); escuela profesional, nombre de maestría o doctorado; denominación y nivel de educación; lo cual indica que los metadatos extraídos obtuvieron un nivel del 98% de coincidencias correctas. Respecto a la extracción del metadato autor y área alcanzaron una precisión del 0.97 por lo que existe un 97% de coincidencias correctas. Finalmente, con una precisión del 0.96 para el metadato fecha de publicación y tema, lo que implica que existe un 96% de resultados correctos en la extracción de ambos metadatos.

4.3 Resultados conforme al objetivo específico 3

Para evaluar la diferencia de tiempo en la tabla 9, se muestra los resultados del diseño experimental propuesto que tiene como finalidad, analizar la mejora del tiempo de extracción de los metadatos previos y posteriores a la implementación del software. Una vez realizada la medición del tiempo que demora el personal encargado en extraer los metadatos de 50 documentos de investigación, se obtiene un tiempo medio de 05:28.54 minutos por documento. En cuanto al proceso de extracción de metadatos utilizando el estímulo, el software de extracción automática de metadatos para el Repositorio Institucional "E-MeRI", se obtiene un tiempo promedio de 00:15.80 minutos por documento de investigación. A la misma vez se muestra los resultados obtenidos para determinar cuánto tiempo demora en extraer los metadatos al personal encargado y al software de extracción automática en el procesado de varias cantidades de documentos, en nuestro caso desde 50 hasta 380 documentos de investigación.

Tabla 9

Resultados del diseño experimental propuesto

Grupo	Tiempo en horas antes del estimulo	Estimulo	Tiempo en horas después del estimulo
Procesar 50 documentos	04:33:46	Software de extracción automática de metadatos (E-MeRI)	00:13:09
Procesar 150 documentos	14:01:41		00:39:14
Procesar 300 documentos	28:13:45		01:19:48
Procesar 380 documentos	35:35:40		01:40:14

Fuente: Equipo de trabajo

Para 380 documentos de investigación el software realiza el proceso de extracción de metadatos de los documentos en 01h 40min 14 seg, lo que significa que alrededor de un minuto extrae los metadatos de 4 documentos de investigación.

4.4 Prueba de calidad de software para determinar el grado de satisfacción del usuario según norma ISO/IEC 25000

Para la aplicación del estándar en el software, se utilizó el modelo de calidad ISO 25010, donde se detalla las particularidades de la calidad en uso, utilizando las características: Efectividad, Eficiencia y Satisfacción. A la misma vez, las subcaracteristicas que se

midieron son: Efectividad, Eficiencia y Utilizada respectivamente. Por otro lado, se utilizó las métricas ISO 25022 para realizar la medición de la calidad en uso del producto. Luego se recabo los resultados sobre las opiniones del personal encargado de la publicación de documento de investigación en el Repositorio Institucional UNA Puno.

En la Tabla 10, se presenta la ponderación determinada estableciendo el orden fundamental a la Satisfacción, siendo la utilidad primordial al momento de utilizar las funciones del software sin complicaciones con un 50%, debido a que es necesario evaluar si el software al momento de utilizarlo satisface las necesidades del usuario. Con respecto a la efectividad se estableció un 30%, debido a que se requiere que el software cumpla con los objetivos para el que fue creado, es decir completando los objetivos de las tareas sin fallas de funcionalidad. Finalmente, la eficiencia se pondera con un 20%, para evaluar que tan eficientes son los usuarios, basado en el tiempo de concluir una tarea.

Tabla 10

Métricas y ponderación de las características de calidad en uso

Características	**Sub característica**	**Métrica**	**Nivel de importancia**	**Ponderación**
Efectividad	Efectividad	Efectividad de la tarea Frecuencia de error	A	30%
Eficiencia	Eficiencia	Tiempo de la tarea Eficiencia de la tarea Tiempo relativo de la tarea	A	20%
Satisfacción	Utilidad	Nivel de satisfacción Uso discrecional de las tareas Porcentaje de quejas de los usuarios	A	50%

Las practicas tomadas como referencia del modelo de evaluación de la ISO/IEC 25040 permitieron establecer rangos de medición, los cuales fueron adaptados al criterio de las necesidades de solución del software. Los valores están comprendidos entre 0 y 10, en donde se asignó cuatro niveles de puntuación y tres grados de satisfacción establecidos como niveles de puntuación final.

Tabla 11

Niveles de puntuación final

Valor de medición	Nivel de puntuación	Grado de satisfacción
8.75 – 10	Cumple con los requisitos	Muy satisfactorio
5 – 8.74	Aceptable	Satisfactorio
2.75 – 4.9	Mínimamente aceptable	Insatisfactorio
0 – 2.74	Inaceptable	

Fuente: (ISO/IEC, 2022)

Para la evaluación de las métricas, se utilizó el registro de valoración en el cual el valor deseado por todas las métricas, corresponde a un valor mayor o igual a cero y menor o igual a 1. Se obtuvieron los datos mediante una encuesta a los usuarios del software, los cuales vienen a ser, el personal encargado de la publicación de documentos de investigación en el Repositorio Institucional de la UNA Puno. En la figura 36, se presentan un resumen de la evaluación de calidad de uso.

Característica	Subcaracterística	Métrica	Fórmula A/B	Valor deseado	Datos obtenidos	Ponderación	Valor parcial total (/10)	Nivel de importancia	Porcentaje de Importancia	Valor Final
Efectividad	**Efectividad**	Efectividad de la tarea	X = A/B A = Número de tareas completadas. B = Número total de tareas intentadas. Dónde B > 0	1	A = 5 B = 5 X = 1	10	10	A	30%	3
		Frecuencia del error	X = A/B A = Cantidad de errores cometidos. B = Numero de tareas. Dónde B > 0	1	A = 1 B = 5 X = 0.2	10				
Eficiencia	**Eficiencia**	Tiempo de publicacion de documentos de investigacion	X = A/B A = Timepo planteado (Min). B = Tiempo actual (Min). Dónde B > 0	<=1	A = 1 B = 7 X = 0.14	9	6.33	A	20%	1.26666667
		Eficiencia de la tarea	X = A/B A = Tareas efectivas. B= Tiempo de la tarea. Dónde B > 0	1	A = 4 B = 5 X = 0.5	5				
		Tiempo relativo de la tarea	X = A/B A = Tiempo que completa una tarea un usuario experto (seg). B= Tiempo que completa una tarea un usuario normal(seg). Dónde B > 0	1	A = 30 B = 60 X = 0.5	5				
Satisfacción	**Utilidad**	Nivel de satisfacción	X = A/B A = Número de respuestas satisfactorias. B = Núméor total de pregntas realizadas en el Cuestionario. Dónde B>0	1	A = 12 B = 13 X = 0.9	9	9.33333333	A	50%	4.66666667
		Uso discrecional de las funciones	X = A/B A = Nro de Funciones especificas. B = Nro de funciones implementadas. Dónde B > 0	1	A = 2 B = 2 X = 1	10				
		porcentaje de quejas de los clientes	X = A/B A = Número de usuario que se quejaron. B = Número total de usuarios. Dónde B > 0	0	A = 0 B = 3 X = 0	9				

Figura 37. Matriz de calidad de uso

Como se muestra en la tabla 12, los valores obtenidos sobre las características de calidad de uso evaluadas, se logró un resultado “muy satisfactorio”, de manera que en nivel de uso del software de extracción automática de metadatos “E-MeRI”, el personal encargado de la publicación de documentos de investigación se encuentra muy satisfecho con la utilización.

Tabla 12

Resumen y valor total obtenido de calidad en uso

Característica	Valor parcial total (/10)	Nivel de importancia	Ponderación	Valor final	Calidad en uso del sistema
Efectividad	10	A	30%	3	
Eficiencia	6.33	A	20%	1.27	**8.93**
Satisfacción	9.33	A	50%	4.67	

Fuente: Equipo de trabajo

4.5 Hipótesis prueba t para muestras relacionadas

El estudio de esta hipótesis tiene como finalidad determinar la mejora de la administración del Repositorio Institucional mediante la publicación de documentos de investigación antes y después de la implementación del software. Para dicho fin se consideraron las medias muéstrales.

Tabla 13

Prueba de la normalidad

Medidas	Publicación de documentos de investigación en el Repositorio Institucional UNAP	
	Antes	**Después**
Media	10:10.38	00:54.04
Desviación estándar	01:41.32	00:05.38
Kolgomorov-Smirnov	**0.350**	**0.460**

Fuente: Equipo de trabajo

Como P-valor publicación de documentos antes (0.350) > 0.05 y P-valor publicación de documentos después de la implementación del software (0.460) > 0.05, provienen de una distribución normal, por lo tanto, se utilizará la prueba t de muestras relacionadas.

i. Planteamiento de hipótesis

$H_0 : \mu_1 = \mu_2$, con la implementación del software de extracción automática de metadatos, no se reduce el tiempo de publicación de documentos de investigación en la administración del Repositorio Institucional

$H_a : \mu_1 \neq \mu_2$, con la implementación del software de extracción automática de metadatos, si se reduce el tiempo de publicación de documentos de investigación en la administración del Repositorio Institucional

ii. Nivel de significancia

Se usó un nivel de significancia del 5%, es decir $\alpha = 0.05$

iii. Prueba estadística

Prueba t para muestras relacionadas

Tabla 14

Prueba t de muestras relacionadas sobre la publicación de documentos antes y después

		95% de intervalo de confianza de la diferencia		t	gl	Sig. (bilateral)
		Inferior	Superior			
Par 1	Tiempo de publicación antes- Tiempo de publicación después	09:06.09	09:26.59	106.718	379	0.000

Fuente: Equipo de trabajo

iv. Regla de decisión

p-valor$(0.000) < 0.05$, entonces se rechaza la H_0 y se acepta la H_a

v. Decisión

Según la tabla 14, la prueba t para muestras relacionadas obtuvo una diferencia significativa (p=0.000 < 0.05); de modo que, se rechaza la hipótesis nula (H_0) y se acepta la hipótesis alterna (H_a), es decir, que las medias de los tiempos del antes y después de la implementación del software de extracción automática "E-MeRI" son significativamente diferentes. Por lo tanto, se concluye que la implementación reduce significativamente el tiempo de publicación de documentos de investigación por consiguiente logra mejorar la administración del Repositorio Institucional.

4.6 Discusión de Resultados

El procesamiento de lenguaje natural usada como técnica en esta investigación permite la extracción automática de metadatos, estos resultados coinciden con lo obtenido por Pan (2020), quien encontró que el procesamiento de lenguaje natural, así como la clasificación de entidades combinado con el uso de expresiones regulares permitieron la extracción de información de sentencias judiciales. Con respecto al algoritmo desarrollado consigue una complejidad de O(l*n*m) en el cual "l" es la longitud de la lista de entidades, "n" es la longitud de texto y siendo "m" el tamaño de la expresión regular, este resultado es diferente a la notación asintótica presentada en la investigación, la cual consigue una complejidad algorítmica lineal O(n), esta variación en la notación asintótica se debe a que el procesamiento de lenguaje natural en el antecedente lo realiza a través de Textserver que es una plataforma para el NLP en la nube.

Según Iturbe Herrera et al. (2019) en su investigación "Extracción semiautomática de metadatos en documentos no estructurados utilizando procesamiento de lenguaje natural y propiedades tipográficas", señala que la extracción de metadatos es una tarea compleja de realizar debido a que carece de un orden en la distribución, por lo que su investigación se centró solo en la extracción de metadatos: título, autores, editorial y fecha de publicación. Metadatos que fueron extraídos utilizando técnicas de procesamiento de lenguaje natural sobre 300 documentos nos estructurados en formato PDF, de los cuales se obtuvieron 300 títulos, 549 autores, 265 editoriales y 282 fechas de publicación con una precisión de 74.66%, 83.36%, 79.66% y 82.33% respectivamente. En comparación a lo obtenido en la presente investigación, la extracción de los mismos metadatos fue del 99% para títulos, 97% para autores y 97% para las fechas de publicación. Esta diferencia en la precisión de extracción de metadatos se debe a que en el antecedente no fue posible determinar un patrón en la organización de metadatos en ese tipo de documentos, además que al ser un proceso semiautomático se presentan errores en la extracción y clasificación de la información por lo que debe ser verificado y corregido por el usuario final.

Respecto a los resultados obtenidos de la diferencia de tiempo de extracción del personal especialista en extraer los metadatos de documentos de investigación y el software, dentro de la metodología propuesta se realiza el proceso de extracción en cuatro tareas: procesar 50, 150, 300 y 380 documentos de investigación, donde se observa que el tiempo medio en el proceso de extracción sobre 50 documentos es 15 segundos y 80 centésimos minutos por documento. En Flores *et al.* (2017) presenta un "Componente para la extracción

automática de metadatos bibliográficos desde corpus textuales en formato PDF", donde se plantea que el tiempo medio que demora el componente en el proceso de extracción de metadatos sobre la misma cantidad de documentos es de 03 minutos y 12 segundos por artículo científico. Esta variación en tiempo que se presenta en el componente y el software de extracción automática de metadatos desarrollado en la presente investigación se debe al tipo de documento, además de lo metadatos extraídos del documento de investigación son muchos más en comparación de un artículo científico. Por otra parte Pinilla (2017) realiza un caso de estudio sobre la extracción de metadatos mediante su algoritmo sobre tres tipos de documentos: artículo de revista, libro y tesis de posgrado, donde el tiempo de respuesta del procesamiento de extracción es de 8 segundos para un artículo científico, 48 segundos sobre un libro y 41 segundos en respuesta para tesis de posgrado. Estos resultados evidencian que el uso de una herramienta de extracción de metadatos reduce el tiempo de este proceso, además que facilita su conservación posterior mediante el depósito de los mismos en diferentes plataformas de conservación y difusión de documentos.

Según Lipinski *et al.* (2013), en su investigación "Evaluation of header metadata extraction approaches and tolos for scientific", en el cual realiza una comparación de herramientas en el proceso de extracción de metadatos bibliográficos, GROBID, Mendeley Desktop y Parscit, Lipinski para este proceso selecciono aleatoriamente una colección de 1153 artículos en PDF. De las herramientas analizadas Grobid obtuvo el mejor desempeño con: 0.92 para títulos, 0.83 para los autores, 0.90 para el apellido de los autores, 0.74 para el resumen y 0.69 para el año de publicación. Por otra parte Casali et al. (2015) en su investigación compara las herramientas extractoras KEA, Alchemy y Mr. DLib con el objetivo de diseñar una herramienta extractora para la extracción de metadatos, de las cuales KEA y Alchemy tienen una precisión similar en la extracción del título, autores y palabras claves, además de realizar la combinación de las herramientas ParsCit y Alchemy, con la cual se logró incrementar el nivel de precisión de extracción de metadatos a un 70%. Algo similar se ve Torres (2022) en cuanto a la extracción de palabras claves sobre los proyectos de investigación obteniendo un 72% de precisión. En relación a estos resultados, en esta investigación se realizó la comparación de herramientas extractoras KEA, Mr.Dlib y Parscit, permitiendo la extracción de metadatos: autor, titulo, resumen de los documentos de investigación, los cuales fueron parcialmente a diferencia del software desarrollado que logra extraer todos los metadatos

del documento con una precisión superior al 96%, con respecto al tiempo de extracción a las herramientas extractoras les toma más tiempo en comparación del software propuesto.

CONCLUSIONES

Se logró optimizar el proceso de extracción de metadatos y publicación de documentos de investigación, los cuales permiten mejorar la administración del Repositorio Institucional de la Universidad Nacional del Altiplano mediante la implementación del software “E-MeRI”, reduciendo el tiempo de publicación antes y después de forma significativa con un valor $p\ (0.000) < \alpha = 0.05$. Además, en la evaluación del software basado en la norma ISO 25000 se obtuvo una valoración de 8.93 de 10 puntos como calidad total, logrando un nivel de “Cumple con los requisitos” y un grado “Muy satisfactorio”.

El procesamiento de lenguaje natural (NLP) permitió el desarrollo del algoritmo de extracción automática de metadatos mediante la librería NLTK del lenguaje de programación Python normalizando el texto y haciendo uso de expresiones regulares para buscar patrones e identificar partes del texto que ayudaron a obtener los metadatos a extraer, además se evaluó la eficiencia del algoritmo, obteniendo una complejidad algorítmica lineal O(n), el cual probó extraer más metadatos de documentos de investigación y ser mucho más rápido en comparación de otras herramientas extractoras.

El algoritmo basado en procesamiento de lenguaje natural (NLP) para la extracción automática de metadatos obtuvo un nivel de precisión entre 96% a 99% de resultados correctos, lo que indica que el algoritmo es eficiente para la extracción de metadatos en los documentos de investigación del Repositorio Institucional de la Universidad Nacional del Altiplano.

El software de extracción automática de metadatos "E-MeRI" realiza el proceso de extracción de metadatos sobre 380 documentos de investigación en una hora y cuarenta minutos, lo que significa que alrededor de un minuto extrae los metadatos de 4 documentos. Además, que reduce el tiempo medio que demora el personal encargado en llevar a cabo el proceso de extracción de metadatos en 5 minutos y 21 segundos por documento, logrando un alto desempeño.

RECOMENDACIONES

Se recomienda el uso del software de extracción automática de metadatos “E-MeRI” propuesto en esta investigación, para la administración de Repositorios Institucionales que usan las directrices de ALICIA CONCYTEC.

Ampliar los métodos de extracción, en el software propuesto en esta investigación se utilizó el procesamiento de lenguaje natural (NLP) para la extracción de metadatos, el cual proporciono resultados satisfactorios. Podría resultar interesante realizar la extracción de metadatos utilizando otras técnicas para los mismos documentos, con el fin de realizar una comparativa entre las diferentes técnicas y determinar cuáles son las más adecuadas para el proceso de extracción de metadatos.

Desarrollar algoritmos utilizando técnicas de Optical Character Recognition (OCR) para la extracción de metadatos en documentos de investigación.

Se recomienda utilizar el procesamiento de lenguaje natural (NLP) para la extracción de información en textos estructurados que contienen secciones y subsecciones, capítulos y párrafos.

Ampliar la extracción de metadatos haciendo uso de diferentes métodos de extracción de información como minería de texto sobre documentos de investigación en tipo de contenido monografía, tesinas, trabajo académico, trabajo de suficiencia profesional y el examen de suficiencia de competencia profesional.

Al personal encargado de la publicación de documentos de investigación en Repositorios Institucionales establecer documentos en formato digital mediante etiquetas a los cuales se les pueda extraer la información de otros metadatos como los nombres de los jurados revisores: presidente de jurado; primer miembro; segundo miembro; director o asesor; ORCID y DNI.

BIBLIOGRAFIA

Adam, U. A., & Kiran, K. (2021). Driving forces behind the management of Institutional Repositories: Qualitative evidences. *Malaysian Journal of Library and Information Science*, *26*(3), 33–56. https://doi.org/10.22452/mjlis.vol26no3.2

Aggarwal, C., & Zhai, C. (2012). *Mining Text Data*. Springer.

Alfano, M., Lenzitti, B., & Visalli, N. (2007). SAXEF: A System for Automatic eXtraction of E-learning object Features. *Journal of E-Learning and Knowledge Society*, *3*(june 2007), 83–92. Recuperado de http://www.math.unipa.it/~lenzitti/papers/Jelks_Saxef.pdf

ALICIA CONCYTEC. (2017). Directrices para el procesamiento de información en los repositorios institucionales | Repositorio CONCYTEC. Recuperado Marzo 2, 2022, de http://repositorio.concytec.gob.pe/handle/20.500.12390/2165

ALICIA CONCYTEC. (2021). *DIRECTRICES PARA REPOSITORIOS INSTITUCIONALES DE LA RED NACIONAL DE REPOSITORIOS DIGITALES DE CIENCIA, TECNOLOGÍA E INNOVACIÓN DE ACCESO ABIERTO (RENARE)*. Recuperado de http://repositorio.concytec.gob.pe/handle/20.500.12390/2231

Arias, F. G. (2012). *El proyecto de investigación Introducción a la metodologia de investigación* (Sexta). Caracas: EPISTEME, C.A.

Barrueco, J. M., & Garcia, C. (2009). Repositorios institucionales universitarios: evolución y perspectivas. Recuperado Agosto 19, 2019, de http://www.fesabid.org/zaragoza2009/Libro_Actas_Fesabid_2009.pdf

Beel, J., Gipp, B., Langer, S., Genzmehr, M., Wilde, E., Nürnberger, A., & Pitman, J. (2011). Introducing Mr. DLib, a Machine-readable Digital Library. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 463–464. https://doi.org/10.1145/1998076.1998187

Bermudez, C. (2010). Metadata: Definicion hoy.

Billhardt, H. (2007). *Teoria de Automatas y Lenguajes Formales* (U. R. J. Carlos, Ed.). Recuperado de https://docplayer.es/28658988-Capitulo-7-expresiones-regulares.html

Burnim, J., Jukevar, S., & Sen, K. (2009). WISE: Automated test generation for worst-case complexity. *International Conference on Sofware Engineering*, 463–473.

Bustos-Gonzalez, A., & Porcel, A. (2007). Directrices para la creación de repositorios institucionales en universidades y organizaciones de educación superior. *Alfa Network Babel Library*, (January 2007). Recuperado de http://eprints.rclis.org/13512/1/%0ADirectrices_RI_Espa_ol.pdf

Cartic, Ramakrishnan; Abhishek, Patnia; Eduard, Hovy;Gully, A. B. (2012). Layout-aware text extraction from full-text pdf of scientific articles. *Source Code for Biology and Medicine*, 7.

Casali, A., Deco, C., Bender, C., & Fontanarrosa, S. (2015). Extracción Automática de Metadatos de Objetos Digitales Educativos. *Conferencias LACLO*, (October), 23–30. Recuperado de http://www.laclo.org/papers/index.php/laclo/article/viewFile/236/218

Chazarra, J., Requena, V. M., & Valverde, S. (2010). Desarrollo de un repositorio de objetos de aprendizaje usando DSpace. Recuperado Agosto 19, 2019, de

http://eprints.ucm.es/11078/1/MemoriaSI.pdf

Chomsky, N. (1956). Three models for description of language. *IRE Transactions on Information Theory*, *2*, 113–124.

Chomsky, N. (2011). Language and other cognitive systems. What is special about language? *Language Learning and Development*, *7*, 263–278.

Choudhury, M. H., Jayanetti, H. R., Wu, J., Ingram, W. A., & Fox, E. A. (2021). *Automatic Metadata Extraction Incorporating Visual Features from Scanned Electronic Theses and Dissertations*. 1–7. Recuperado de http://arxiv.org/abs/2107.00516

Congreso de la República del Perú. (2013). *Ley 30035: La Ley que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de acceso abierto*. Diario El Peruano.

Congreso De La República Del Perú. (2013). *Ley número 30035: Ley que regula el repositorio nacional digital de ciencia, tecnología e innovación de acceso abierto*. 496508–496509. Recuperado de https://portal.concytec.gob.pe/images/stories/images2013/portal/areas-institucion/dsic/ley-30035.pdf

Copara Zea, J. L. (2017). *Reconocimiento de entidades nombradas para el idioma español utilizando Conditional Random Fields con características no supervisadas*. 51. Recuperado de https://repositorio.ucsp.edu.pe/bitstream/UCSP/15404/1/COPARA_ZEA_JEN_REC.pdf

Cortez, E., Da Silva, A. S., Gonçalves, M. A., Mesquita, F., & De Moura, E. S. (2009).

A flexible approach for extracting metadata from bibliographic citations. *Journal of the American Society for Information Science and Technology*, *60*(6), 1144–1158. https://doi.org/10.1002/asi.21049

Cui, B. G., & Chen, X. (2010). An improved hidden markov model for literature metadata extraction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6215 LNCS*, 205–212. https://doi.org/10.1007/978-3-642-14922-1_26

DCMI. (1995). The Dublin Core Metadata Initiative. Retrieved September 29, 2021, from http://dublincore.org/specifications/dublin-core/

DCMI. (2018). Dublin Core Metadata Element Set, Version 1.1. Recuperado de https://www.dublincore.org/specifications/dublin-core/dces/

De Giusti, M. R., Lira, A. J., Oviedo, N. F., Villarreal, G. L., & Texier, J. (2012). Las actividades y el planeamiento de la preservación en un repositorio institucional. Recuperado de Conferencia Internacional Acceso Abierto, Comunicación Científica y Preservación Digital, Colombia website: http://hdl.handle.net/10915/26045

Deng, L., & Yu, D. (2014). *Deep learning : methods and applications*.

Flores-Ruiz, E., Miranda-Novales, M. G., & Villasís-Keever, M. Á. (2017). El protocolo de investigación VI: cómo elegir la prueba estadística adecuada. Estadística inferencial. *Revista Alergia México*, *64*(3), 364–370. https://doi.org/10.29262/ram.v64i3.304

Flores Reira, L., Mariño Molerio, A., Mojena Roman, L., & Hidalgo Delgado, Y. (2017). Componente para la extracción automática de metadatos bibliográficos

desde corpus textuales en formato PDF. *Revista Cubana de Ciencias Informáticas*, *11*(4), 85–98. Recuperado de https://www.researchgate.net/publication/320765091_Componente_para_la_extraccion_automatica_de_metadatos_bibliograficos_desde_corpus_textuales_en_formato_PDF

Flynn, P., Zhou, L., Maly, K., Zeil, S., & Zubair, M. (2007). Automated Template-Based Metadata Extraction Architecture. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers* (pp. 327–336). https://doi.org/10.1007/978-3-540-77094-7_42

France, P. A. H., & Allen, J. F. (1997). *INCORPORATING POS TAGGING INTO LANGUAGE MODELING*.

Granitzer, M., Hristakeva, M., Jack, K., & Knight, R. (2011). *A Comparison of Metadata Extraction Techniques for Crowdsourced Bibliographic Metadata Management*.

Greene, J. (2010). Project management and institutional repositories: A case study at University College Dublin Library. *New Review of Academic Librarianship*, *16*, 98–115.

Halascy, P. (2006). *Benefits of deep NLP-based lemmatization for information retrieval*. Working Notes for the CLEF-2006 Workshop.

Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., & Fox, E. A. (2003). Automatic document metadata extraction using support vector machines. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, *2003-Janua*, 37–48. https://doi.org/10.1109/JCDL.2003.1204842

Han, Y., & Thorup, M. (2002). Integer Sorting in O(n log n) Time and Linear Space. *Proceedings of the 43erd Symposium on Foundations of Computer Science*, 135–144.

Hasan, M., Wu, C. J., Ingram, W. A., & Fox, E. A. (2020). A heuristic baseline method for metadata extraction from scanned electronic theses and dissertations. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 515–516. https://doi.org/10.1145/3383583.3398590

Heinrichs, B., Preuß, N., Politze, M., Müller, M., & Pelz, P. (2021). *Automatic General Metadata Extraction and Mapping in an HDF5 Use-case*. *1*(Ic3k), 172–179. https://doi.org/10.5220/0010654100003064

Hernández Sampieri, R., Fernández Collado, C., Baptista Lucio, P., Mendoza Torres, C. P., & Méndez Valencia, S. (2014). *Metodología de la investigación sexta edición* (McGrawHill).

ISO/IEC. (2022). ISO 25000. Recuperado de https://iso25000.com/index.php/normas-iso-25000/iso-25040

ISO 25000. (2022). NORMAS ISO/IEC 25000. Retrieved February 17, 2022, from https://iso25000.com/index.php/en/iso-25000-standards/iso-25010

Iturbe Herrera, A., Montes Rendón, A., Torres-Moreno, J.-M., Sierra Martínez, G., Castro Sánchez, N. A., & González Serna, J. G. (2019). Extracción semiautomática de metadatos en documentos no estructurados utilizando procesamiento de lenguaje natural y propiedades tipográficas. *Research in Computing Science*, *148*(7), 331–345. https://doi.org/10.13053/rcs-148-7-25

Júnior Grossi, J. A. (2016). *Análise Comparativa de Ferramentas de Extração de*

Metadados em Artigos Científico. Belo Horizonte.

Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J. & Nithya, M. (2015). Preprocessing Techniques for Text Mining Preprocessing Techniques for Text Mining. *International Journal of Computer Science & Communication Networks*, *5*(October 2014), 7–16.

Kern, R., Jack, K., & Hristakeva, M. (2012). TeamBeam - Meta-Data Extraction from Scientific Literature. *D-Lib Magazine*, *18*(7/8). https://doi.org/10.1045/july2012-kern

Kovačević, A., Ivanović, D., Milosavljević, B., Konjović, Z., & Surla, D. (2011). Automatic extraction of metadata from scientific publications for CRIS systems. *Program*, *45*(4), 376–396. https://doi.org/10.1108/00330331111182094

Kuznetsov, S. O., & Obiedkov, S. A. (2010). Comparing performance of algorithms for generating concept lattices. *Journal of Experimental & Theoretical Artificial Intelligence*, *14*, 189–216. https://doi.org/https://doi.org/10.5325/gestaltreview.14.2.0189

Lin, X. (2011). *Fine-grained Named Entity Classification in Machine Reading.* University of Oxford.

Lipinski, M., Yao, K., Breitinger, C., Beel, J., & Gipp, B. (2013). Evaluation of header metadata extraction approaches and tools for scientific PDF documents. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, (July), 385–386. https://doi.org/10.1145/2467696.2467753

Luhn, H. P. (2014). Association for Information Science and Technology. Recuperado Agosto 19, 2019, de https://www.asist.org/pioneers/hans-peter-luhn/

Lynch, C. (2005). Institutional repositories: Essential infrastructure for scholarship in the digital age. *Portal: Libraries and Academy*, *3*(2), 327–336.

Marinai, S. (2009). Metadata extraction from PDF papers for digital library ingest. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 251–255. https://doi.org/10.1109/ICDAR.2009.232

Martínez Arellano, F. F. (2017). Metadatos y repositorios institucionales. *Bibliotecas y Archivos (México, D.F.)*, *2*(4), 44–52. Recuperado de https://biblat.unam.mx/es/revista/bibliotecas-y-archivos-mexico-d-f/articulo/metadatos-y-repositorios-institucionales

Miranda, S., & Ritrovato, P. (2015). Supporting Learning Object Repository by automatic extraction of metadata. *Journal of E-Learning and Knowledge Society*, *11*(1). https://doi.org/10.20368/1971-8829/988

Nabe, J. A. (2012). Starting, strengthening and managing institutional repositories. *Journal of Librarianship and Scholarly Communication*, *1*. Recuperado de http://dx.doi.org/10.7710/2162-3309.1024

NISO. (2019). Metadata and the Institutional Repository. Recuperado de https://www.niso.org/events/2017/02/metadata-and-institutional-repository

Nkiko, C., Bolu, C., & Michael-Onuoha, H. (2014). Managing a sustainable institutional repository. *The Covenant University Experience. Samaru Journal of Information Studies*, *14*, 1-2:1-6.

Ochando, M. B. (2013). *Técnicas avanzadas de recuperación de información Procesos, técnicas y métodos mblazquez.es*.

OpenDoar. (2021). Directory of Open Access Repositories. Recuperado Agosto 18, 2019, de http://v2.sherpa.ac.uk/view/repository_visualisations/1.html

Pan, Y. (2020). *Extracción de información de sentencias judiciales*.

Paradelo Luque, A. M. (2009). *Preservación documental en repositorios institucionales* (Vol. 23). Recuperado de http://www.scielo.org.mx/pdf/ib/v23n49/v23n49a9.pdf

Pavani, A. (2009). Interoperabilidad de bibliotecas digitales y protocolos internacionales = Interoperabilidade de bibliotecas digitais. Recuperado de http://www.unesco.org.uy/informatica/publicaciones/bibliotecasdigitales2005/01-intro-agenda.pdf %0D

Peinado Rodriguez, J. (2003). Lematización para palabras médicas complejas: Implementación de un algoritmo en LISP . *Revista Medica Herediana*, *14*(4), 223–228. Recuperado de http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1018-130X2003000400013&nrm=iso

Pérez Velandia, M., & Felipe Silva, L. (2007). *Como funciona el protocolo OAI-PMH en la recuperación de información*. Recuperado de http://eprints.rclis.org/10677/1/COMO_FUNCIONA_EL_PROTOCOLO_OAI_-_PMH_EN_LA_RECUPERACION_DE_INFORMACION.pdf

Pinilla, A., Gutiérrez, M., & Ballejos, L. (2014). Extracción Automática de Metadatos a partir de Objetos de Aprendizaje en un Repositorio Institucional : Estado del Arte. *Simposio Argentino de Tecnología y Sociedad*, 67–82. Recuperado de http://hdl.handle.net/10915/41759

Pinilla Gómez, A. C. (2017). *AMELOIR: algoritmo para la extracción automática de*

metadatos a partir de objetos de aprendizaje en un repositorio institucional. Recuperado de http://ria.utn.edu.ar/handle/123456789/2513

Pire, T., Deco, C., Casali, A., & Espinase, B. (2011). *Extracción Automática de Metadatos de Objetos de Aprendizaje: un estudio comparativo*. Recuperado de http://www.ieee.org

Polanco-Cortés, J. (2016). *Repositorios digitales. Definición y pautas para su creación*. Recuperado de https://ucrindex.ucr.ac.cr/docs/repositorios-digitales-definicion-y-pautas-para-su-creacion.pdf

Pressman, R. S. (2022). *Ingenieria de Software* (Septima). Recuperado de http://cotana.informatica.edu.bo/downloads/ld-Ingenieria.de.software.enfoque.practico.7ed.Pressman.PDF

Rivera, A. C. (2009). *Creación de un repositorio digital con la producción intelectual de la Dra. María Eugenia Bozzoli Vargas, en el Laboratorio de Etnología de la Universidad de Costa Rica Title* (Universidad de Costa Rica, San José). Recuperado de http://repositorio.sibdi.ucr.ac.cr:8080/jspui/bitstream/123456789/261/1/30231.pdf

Rodríguez Gairín, J. M., & Sulé, A. (2008). DSpace: un manual específico para gestores de la información y la documentación. *BiD Textos Universitaris de Biblioteconomia i Documentaci*, (20), 1–15. Recuperado de http://www2.ub.edu/bid/consulta_articulos.php?fichero=20rodri2.htm

Samsudin, N. (2020). Calculation on Euler Arithmetic Complexity using Big O Notation. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(1.4), 460–464. https://doi.org/10.30534/ijatcse/2020/6591.42020

Sanz de Diego, A. (2021). LA NOTACIÓN O GRANDE CON EJEMPLOS EN JAVASCRIPT. Recuperado de https://www.asanzdiego.com/2018/12/la-notacion-o-grande-con-ejemplos-en-javascript.html

Scikit-Learn. (2022). Machine Learning in Python. Recuperado Enero 26, 2022, de https://scikit-learn.org/stable/modules/model_evaluation.html

Senso, J. A., & Rosa Piñero, A. de la. (2003). El concepto de metadato: algo más que descripción de recursos electrónicos. *Ciência Da Informação*, *32*(2), 95–106. https://doi.org/10.1590/s0100-19652003000200011

Singh, N., & Tiwari, R. G. (2015). Basics of Algorithm Selection: A Review. *International Journal of Computer Science Trends and Technology*, *3*, 139–142.

Testa, P., & Degiorgi, H. (2013). *Esquemas de metadatos para los repositorios institucionales de las universidades nacionales argentinas*. Recuperado de http://bdigital.uncu.edu.ar/5881

Texier, J., De Giusti, M. R., Oviedo, N., Villarreal, G. L., & Lira, A. J. (2012). *El uso de repositorios y su importancia para la educación en Ingeniería*. 1–10. Recuperado de http://hdl.handle.net/10915/22943%5Cnhttp://sedici.unlp.edu.ar/handle/10915/22943

The OAI Executive. (2015). The Open Archives Initiative Protocol for Metadata Harvesting. Recuperado de https://www.openarchives.org/OAI/openarchivesprotocol.html

Tkaczyk, D. (2017). *New Methods for Metadata Extraction from Scientific Literature*. Recuperado de http://arxiv.org/abs/1710.10201

Tkaczyk, D., Szostek, P., Dendek, P. J., Fedoryszak, M., & Bolikowski, L. (2014). CERMINE - Automatic extraction of metadata and references from scientific literature. *Proceedings - 11th IAPR International Workshop on Document Analysis Systems, DAS 2014*, 217–221. https://doi.org/10.1109/DAS.2014.63

Torres Cruz, F. (2022). *ALGORITMOS DE APRENDIZAJE AUTOMÁTICO NO SUPERVISADO PARA LA EXTRACCIÓN DE PALABRAS CLAVE EN TRABAJOS DE INVESTIGACIÓN DE PREGRADO*. Recuperado de http://repositorio.unap.edu.pe/handle/UNAP/18372

UNAP. (2021). Universidad Nacional del Altiplano. Retrieved November 11, 2021, from https://portal.unap.edu.pe/historia-de-la-universidad

Vásquez, A. C., Huerta, H. V., Quispe, J. P., & Huayna, A. M. (2009). Procesamiento de lenguaje natural robusto. *Reviste de Ingeniería de Sistemas e Informática*, *6*(2), 45–54. Recuperado de https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/5923

VRI UNAP. (2022). Repositorio Institucional Digital de la Universidad Nacional del Altiplano. Recuperado Febrero 17, 2022, de http://repositorio.unap.edu.pe/

W3schools. (2021). Python RegEx. Retrieved December 10, 2021, from https://www.w3schools.com/python/python_regex.asp

Webster, J. J., & Kit, C. (1992). *TOKENIZATION AS THE INITIAL PHASE IN NLP* (Vol. 4). Recuperado de https://bit.ly/2SktsCP

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). *KEA: Practical Automatic Keyphrase Extraction*. Recuperado de http://www.nzdl.org/

ANEXOS

Anexo N° 1. Documentos procesados mediante el software

Nro	Fecha publicación	Tipo de obra	Escuela Profesional	Facultad
1	2021-07-16	Tesis	Biología	Facultad de Ciencias Biológicas
2	2021-07-23	Tesis	Biología	Facultad de Ciencias Biológicas
3	2021-08-31	Tesis	Biología	Facultad de Ciencias Biológicas
4	2021-11-11	Tesis	Biología	Facultad de Ciencias Biológicas
5	2021-11-22	Tesis	Biología	Facultad de Ciencias Biológicas
6	2021-09-03	Tesis	Biología	Facultad de Ciencias Biológicas
7	2021-09-03	Tesis	Biología	Facultad de Ciencias Biológicas
8	2021-09-02	Tesis	Enfermería	Facultad de Enfermería
9	2021-08-27	Tesis	Enfermería	Facultad de Enfermería
10	2021-09-02	Tesis	Enfermería	Facultad de Enfermería
11	2021-09-01	Tesis	Enfermería	Facultad de Enfermería
12	2021-08-11	Tesis	Enfermería	Facultad de Enfermería
13	2021-08-04	Tesis	Enfermería	Facultad de Enfermería
14	2021-09-03	Tesis	Enfermería	Facultad de Enfermería
15	2021-09-02	Tesis	Enfermería	Facultad de Enfermería
16	2021-12-07	Tesis	Enfermería	Facultad de Enfermería
17	2021-07-02	Tesis	Enfermería	Facultad de Enfermería
18	2021-11-12	Tesis	Enfermería	Facultad de Enfermería
19	2021-11-17	Tesis	Enfermería	Facultad de Enfermería
20	2021-11-12	Tesis	Medicina Humana	Facultad de Medicina Humana
21	2021-07-09	Tesis	Medicina Humana	Facultad de Medicina Humana
22	2021-11-17	Tesis	Medicina Humana	Facultad de Medicina Humana
23	2021-07-01	Tesis	Medicina Humana	Facultad de Medicina Humana
24	2021-07-16	Tesis	Medicina Humana	Facultad de Medicina Humana
25	2021-07-14	Tesis	Medicina Humana	Facultad de Medicina Humana
26	2021-10-18	Tesis	Medicina Humana	Facultad de Medicina Humana
27	2021-07-19	Tesis	Medicina Humana	Facultad de Medicina Humana
28	2021-11-29	Tesis	Medicina Humana	Facultad de Medicina Humana
29	2021-08-25	Tesis	Medicina Humana	Facultad de Medicina Humana
30	2021-08-19	Tesis	Medicina Humana	Facultad de Medicina Humana
31	2021-11-26	Tesis	Medicina Humana	Facultad de Medicina Humana
32	2021-11-09	Tesis	Medicina Humana	Facultad de Medicina Humana
33	2021-09-03	Tesis	Medicina Veterinaria y Zootecnia	Facultad de Medicina Veterinaria y Zootecnia
34	2021-08-24	Tesis	Medicina Veterinaria y Zootecnia	Facultad de Medicina Veterinaria y Zootecnia
35	2021-07-30	Tesis	Medicina Veterinaria y Zootecnia	Facultad de Medicina Veterinaria y Zootecnia
36	2021-07-02	Tesis	Medicina Veterinaria y Zootecnia	Facultad de Medicina Veterinaria y Zootecnia
37	2021-08-10	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
38	2021-07-19	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud

39	2021-08-19	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
40	2021-09-03	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
41	2021-08-11	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
42	2021-07-23	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
43	2021-08-11	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
44	2021-11-17	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
45	2021-07-16	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
46	2021-08-31	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
47	2021-09-03	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
48	2021-12-06	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
49	2021-08-09	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
50	2021-08-18	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
51	2021-11-10	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
52	2021-08-09	Tesis	Nutrición Humana	Facultad de Ciencias de la Salud
53	2021-08-23	Tesis	Odontología	Facultad de Ciencias de la Salud
54	2021-11-30	Tesis	Odontología	Facultad de Ciencias de la Salud
55	2021-09-03	Tesis	Odontología	Facultad de Ciencias de la Salud
56	2021-08-06	Tesis	Odontología	Facultad de Ciencias de la Salud
57	2021-07-09	Tesis	Odontología	Facultad de Ciencias de la Salud
58	2021-08-12	Tesis	Odontología	Facultad de Ciencias de la Salud
59	2021-08-24	Tesis	Odontología	Facultad de Ciencias de la Salud
60	2021-07-23	Tesis	Odontología	Facultad de Ciencias de la Salud
61	2021-08-13	Tesis	Odontología	Facultad de Ciencias de la Salud
62	2021-09-02	Tesis	Odontología	Facultad de Ciencias de la Salud
63	2021-10-22	Tesis	Odontología	Facultad de Ciencias de la Salud
64	2021-08-20	Tesis	Odontología	Facultad de Ciencias de la Salud
65	2021-08-13	Tesis	Odontología	Facultad de Ciencias de la Salud
66	2021-08-26	Tesis	Arquitectura y Urbanismo	Facultad de Ingeniería Civil y Arquitectura
67	2021-09-17	Tesis	Arquitectura y Urbanismo	Facultad de Ingeniería Civil y Arquitectura
68	2021-11-15	Tesis	Arquitectura y Urbanismo	Facultad de Ingeniería Civil y Arquitectura
69	2021-07-15	Tesis	Arquitectura y Urbanismo	Facultad de Ingeniería Civil y Arquitectura
70	2021-08-18	Tesis	Arquitectura y Urbanismo	Facultad de Ingeniería Civil y Arquitectura
71	2021-11-03	Tesis	Arquitectura y Urbanismo	Facultad de Ingeniería Civil y Arquitectura
72	2021-11-16	Tesis	Ciencias Físico Matemáticas	Facultad de Ingeniería Civil y Arquitectura
73	2021-11-19	Tesis	Ingeniería Agrícola	Facultad de Ingeniería Agrícola
74	2021-07-30	Tesis	Ingeniería Agrícola	Facultad de Ingeniería Agrícola
75	2021-08-27	Tesis	Ingeniería Agrícola	Facultad de Ingeniería Agrícola
76	2021-11-26	Tesis	Ingeniería Agrícola	Facultad de Ingeniería Agrícola
77	2021-08-26	Tesis	Ingeniería Agrícola	Facultad de Ingeniería Agrícola
78	2021-08-25	Tesis	Ingeniería Agrícola	Facultad de Ingeniería Agrícola
79	2021-08-19	Tesis	Ingeniería Agrícola	Facultad de Ingeniería Agrícola

80	2021-07-19	Tesis	Ingeniería Agrícola	Facultad de Ingeniería Agrícola
81	2021-09-03	Tesis	Ingeniería Agroindustrial	Facultad de Ciencias Agrarias
82	2021-11-05	Tesis	Ingeniería Agroindustrial	Facultad de Ciencias Agrarias
83	2021-12-09	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
84	2021-11-26	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
85	2021-07-09	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
86	2021-10-28	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
87	2021-07-02	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
88	2021-11-29	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
89	2021-08-27	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
90	2021-08-26	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
91	2021-11-30	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
92	2021-09-03	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
93	2021-09-03	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
94	2021-07-15	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
95	2021-08-18	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
96	2021-11-11	Tesis	Ingeniería Agronómica	Facultad de Ciencias Agrarias
97	2021-09-03	Tesis	Ingeniería Civil	Facultad de Ingeniería Civil y Arquitectura
98	2021-08-27	Tesis	Ingeniería Civil	Facultad de Ingeniería Civil y Arquitectura
99	2021-07-16	Tesis	Ingeniería Civil	Facultad de Ingeniería Civil y Arquitectura
100	2021-08-06	Tesis	Ingeniería Civil	Facultad de Ingeniería Civil y Arquitectura
101	2021-09-02	Tesis	Ingeniería de Minas	Facultad de Ingeniería de Minas
102	2021-07-16	Tesis	Ingeniería de Minas	Facultad de Ingeniería de Minas
103	2021-09-01	Tesis	Ingeniería de Minas	Facultad de Ingeniería de Minas
104	2021-11-05	Tesis	Ingeniería de Minas	Facultad de Ingeniería de Minas
105	2021-08-26	Tesis	Ingeniería de Minas	Facultad de Ingeniería de Minas
106	2021-08-17	Tesis	Ingeniería de Minas	Facultad de Ingeniería de Minas
107	2021-08-27	Tesis	Ingeniería de Minas	Facultad de Ingeniería de Minas
108	2021-09-03	Tesis	Ingeniería de Minas	Facultad de Ingeniería de Minas
109	2021-08-19	Tesis	Ingeniería de Minas	Facultad de Ingeniería de Minas
110	2021-07-19	Tesis	Ingeniería de Minas	Facultad de Ingeniería de Minas
111	2021-08-20	Tesis	Ingeniería de Minas	Facultad de Ingeniería de Minas
112	2021-08-27	Tesis	Ingeniería de Minas	Facultad de Ingeniería de Minas
113	2021-09-01	Tesis	Ingeniería de Sistemas	Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas
114	2021-08-27	Tesis	Ingeniería de Sistemas	Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas

115	2021-10-21	Tesis	Ingeniería de Sistemas	Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas
116	2021-08-06	Tesis	Ingeniería de Sistemas	Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas
117	2021-07-30	Tesis	Ingeniería de Sistemas	Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas
118	2021-07-02	Tesis	Ingeniería de Sistemas	Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas
119	2021-10-20	Tesis	Ingeniería de Sistemas	Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas
120	2021-08-06	Tesis	Ingeniería Electrónica	Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas
121	2021-08-25	Tesis	Ingeniería Estadística e Informática	Facultad de Ingeniería Estadística e Informática
122	2021-08-04	Tesis	Ingeniería Estadística e Informática	Facultad de Ingeniería Estadística e Informática
123	2021-08-27	Tesis	Ingeniería Estadística e Informática	Facultad de Ingeniería Estadística e Informática
124	2021-07-14	Tesis	Ingeniería Geológica	Facultad de Ingeniería Geológica y Metalúrgica
125	2021-09-03	Tesis	Ingeniería Geológica	Facultad de Ingeniería Geológica y Metalúrgica
126	2021-09-03	Tesis	Ingeniería Geológica	Facultad de Ingeniería Geológica y Metalúrgica
127	2021-07-16	Tesis	Ingeniería Geológica	Facultad de Ingeniería Geológica y Metalúrgica
128	2021-09-02	Tesis	Ingeniería Geológica	Facultad de Ingeniería Geológica y Metalúrgica
129	2021-11-12	Tesis	Ingeniería Geológica	Facultad de Ingeniería Geológica y Metalúrgica
130	2021-07-22	Tesis	Ingeniería Geológica	Facultad de Ingeniería Geológica y Metalúrgica
131	2021-11-18	Tesis	Ingeniería Mecánica Eléctrica	Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas
132	2021-09-06	Tesis	Ingeniería Mecánica Eléctrica	Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas
133	2021-09-03	Tesis	Ingeniería Mecánica Eléctrica	Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas
134	2021-09-06	Tesis	Ingeniería Mecánica Eléctrica	Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas
135	2021-10-26	Tesis	Ingeniería Metalúrgica	Facultad de Ingeniería Geológica y Metalúrgica
136	2021-12-06	Tesis	Ingeniería Metalúrgica	Facultad de Ingeniería Geológica y Metalúrgica
137	2021-11-12	Tesis	Ingeniería Metalúrgica	Facultad de Ingeniería Geológica y Metalúrgica
138	2021-08-24	Tesis	Ingeniería Metalúrgica	Facultad de Ingeniería Geológica y Metalúrgica
139	2021-11-15	Tesis	Ingeniería Metalúrgica	Facultad de Ingeniería Geológica y Metalúrgica
140	2021-11-22	Tesis	Ingeniería Metalúrgica	Facultad de Ingeniería Geológica y Metalúrgica
141	2021-08-20	Tesis	Ingeniería Metalúrgica	Facultad de Ingeniería Geológica y Metalúrgica
142	2021-08-18	Tesis	Ingeniería Química	Facultad de Ingeniería Química
143	2021-08-02	Tesis	Ingeniería Topográfica y Agrimensura	Facultad de Ciencias Agrarias
144	2021-10-29	Tesis	Ingeniería Topográfica y Agrimensura	Facultad de Ciencias Agrarias
145	2021-10-25	Tesis	Ingeniería Topográfica y Agrimensura	Facultad de Ciencias Agrarias

146	2021-07-02	Tesis	Administración	Facultad de Ciencias Contables y Administrativas
147	2021-08-10	Tesis	Administración	Facultad de Ciencias Contables y Administrativas
148	2021-11-30	Tesis	Administración	Facultad de Ciencias Contables y Administrativas
149	2021-08-20	Tesis	Administración	Facultad de Ciencias Contables y Administrativas
150	2021-09-08	Tesis	Administración	Facultad de Ciencias Contables y Administrativas
151	2021-07-30	Tesis	Administración	Facultad de Ciencias Contables y Administrativas
152	2021-10-15	Tesis	Administración	Facultad de Ciencias Contables y Administrativas
153	2021-09-01	Tesis	Administración	Facultad de Ciencias Contables y Administrativas
154	2021-08-13	Tesis	Administración	Facultad de Ciencias Contables y Administrativas
155	2021-08-04	Tesis	Administración	Facultad de Ciencias Contables y Administrativas
156	2021-11-05	Tesis	Administración	Facultad de Ciencias Contables y Administrativas
157	2021-09-03	Tesis	Ciencias Contables	Facultad de Ciencias Contables y Administrativas
158	2021-11-19	Tesis	Ciencias Contables	Facultad de Ciencias Contables y Administrativas
159	2021-11-11	Tesis	Ciencias Contables	Facultad de Ciencias Contables y Administrativas
160	2021-11-24	Tesis	Ciencias Contables	Facultad de Ciencias Contables y Administrativas
161	2021-07-01	Tesis	Ciencias Contables	Facultad de Ciencias Contables y Administrativas
162	2021-08-19	Tesis	Ciencias Contables	Facultad de Ciencias Contables y Administrativas
163	2021-11-25	Tesis	Ciencias Contables	Facultad de Ciencias Contables y Administrativas
164	2021-07-27	Tesis	Ciencias Contables	Facultad de Ciencias Contables y Administrativas
165	2021-07-22	Tesis	Ciencias Contables	Facultad de Ciencias Contables y Administrativas
166	2021-11-24	Tesis	Ciencias Contables	Facultad de Ciencias Contables y Administrativas
167	2021-08-19	Tesis	Ciencias Contables	Facultad de Ciencias Contables y Administrativas
168	2021-08-20	Tesis	Ciencias Contables	Facultad de Ciencias Contables y Administrativas
169	2021-11-11	Tesis	Ingeniería Económica	Facultad de Ingeniería Económica
170	2021-07-23	Tesis	Ingeniería Económica	Facultad de Ingeniería Económica
171	2021-10-28	Tesis	Ingeniería Económica	Facultad de Ingeniería Económica
172	2021-08-18	Tesis	Ingeniería Económica	Facultad de Ingeniería Económica
173	2021-09-02	Tesis	Ingeniería Económica	Facultad de Ingeniería Económica
174	2021-12-07	Tesis	Ingeniería Económica	Facultad de Ingeniería Económica
175	2021-08-06	Tesis	Ingeniería Económica	Facultad de Ingeniería Económica
176	2021-11-30	Tesis	Ingeniería Económica	Facultad de Ingeniería Económica
177	2021-08-19	Tesis	Ingeniería Económica	Facultad de Ingeniería Económica
178	2021-10-28	Tesis	Ingeniería Económica	Facultad de Ingeniería Económica
179	2021-11-24	Tesis	Ingeniería Económica	Facultad de Ingeniería Económica

180	2021-07-01	Tesis	Turismo	Facultad de Ciencias Sociales
181	2021-09-01	Tesis	Turismo	Facultad de Ciencias Sociales
182	2021-11-16	Tesis	Turismo	Facultad de Ciencias Sociales
183	2021-08-27	Tesis	Turismo	Facultad de Ciencias Sociales
184	2021-09-03	Tesis	Turismo	Facultad de Ciencias Sociales
185	2021-11-25	Tesis	Turismo	Facultad de Ciencias Sociales
186	2021-09-03	Tesis	Turismo	Facultad de Ciencias Sociales
187	2021-11-17	Tesis	Turismo	Facultad de Ciencias Sociales
188	2021-08-06	Tesis	Turismo	Facultad de Ciencias Sociales
189	2021-07-02	Tesis	Turismo	Facultad de Ciencias Sociales
190	2021-11-29	Tesis	Turismo	Facultad de Ciencias Sociales
191	2021-09-02	Tesis	Antropología	Facultad de Ciencias Sociales
192	2021-12-07	Tesis	Antropología	Facultad de Ciencias Sociales
193	2021-08-17	Tesis	Antropología	Facultad de Ciencias Sociales
194	2021-07-27	Tesis	Antropología	Facultad de Ciencias Sociales
195	2021-07-22	Tesis	Antropología	Facultad de Ciencias Sociales
196	2021-09-03	Tesis	Arte	Facultad de Ciencias Sociales
197	2021-08-06	Tesis	Arte	Facultad de Ciencias Sociales
198	2021-09-03	Tesis	Arte	Facultad de Ciencias Sociales
199	2021-11-08	Tesis	Ciencias de la Comunicación Social	Facultad de Ciencias Sociales
200	2021-08-25	Tesis	Ciencias de la Comunicación Social	Facultad de Ciencias Sociales
201	2021-08-19	Tesis	Ciencias de la Comunicación Social	Facultad de Ciencias Sociales
202	2021-08-20	Tesis	Ciencias de la Comunicación Social	Facultad de Ciencias Sociales
203	2021-10-22	Tesis	Ciencias de la Comunicación Social	Facultad de Ciencias Sociales
204	2021-08-24	Tesis	Ciencias de la Comunicación Social	Facultad de Ciencias Sociales
205	2021-11-26	Tesis	Derecho	Facultad de Ciencias Jurídicas y Políticas
206	2021-07-23	Tesis	Derecho	Facultad de Ciencias Jurídicas y Políticas
207	2021-11-05	Tesis	Derecho	Facultad de Ciencias Jurídicas y Políticas
208	2021-08-11	Tesis	Derecho	Facultad de Ciencias Jurídicas y Políticas
209	2021-11-18	Tesis	Derecho	Facultad de Ciencias Jurídicas y Políticas
210	2021-08-12	Tesis	Educación Física	Facultad de Ciencias de la Educación
211	2021-07-22	Tesis	Educación Física	Facultad de Ciencias de la Educación
212	2021-10-28	Tesis	Educación Física	Facultad de Ciencias de la Educación
213	2021-07-02	Tesis	Educación Física	Facultad de Ciencias de la Educación
214	2021-11-18	Tesis	Educación Física	Facultad de Ciencias de la Educación
215	2021-08-03	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
216	2021-11-30	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
217	2021-11-12	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
218	2021-07-01	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
219	2021-09-01	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
220	2021-07-22	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
221	2021-11-12	Tesis	Educación Inicial	Facultad de Ciencias de la Educación

222	2021-11-04	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
223	2021-11-12	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
224	2021-11-18	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
225	2021-08-06	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
226	2021-10-20	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
227	2021-07-01	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
228	2021-11-12	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
229	2021-11-30	Tesis	Educación Inicial	Facultad de Ciencias de la Educación
230	2021-11-19	Tesis	Educación Primaria	Facultad de Ciencias de la Educación
231	2021-11-25	Tesis	Educación Primaria	Facultad de Ciencias de la Educación
232	2021-11-25	Tesis	Educación Primaria	Facultad de Ciencias de la Educación
233	2021-12-01	Tesis	Educación Primaria	Facultad de Ciencias de la Educación
234	2021-12-09	Tesis	Educación Primaria	Facultad de Ciencias de la Educación
235	2021-08-23	Tesis	Educación Secundaria	Facultad de Ciencias de la Educación
236	2021-11-19	Tesis	Educación Secundaria	Facultad de Ciencias de la Educación
237	2021-11-26	Tesis	Educación Secundaria	Facultad de Ciencias de la Educación
238	2021-10-20	Tesis	Educación Secundaria	Facultad de Ciencias de la Educación
239	2021-11-16	Tesis	Educación Secundaria	Facultad de Ciencias de la Educación
240	2021-07-02	Tesis	Educación Secundaria	Facultad de Ciencias de la Educación
241	2021-08-06	Tesis	Educación Secundaria	Facultad de Ciencias de la Educación
242	2021-09-01	Tesis	Educación Secundaria	Facultad de Ciencias de la Educación
243	2021-11-25	Tesis	Educación Secundaria	Facultad de Ciencias de la Educación
244	2021-09-03	Tesis	Educación Secundaria	Facultad de Ciencias de la Educación
245	2021-07-16	Tesis	Educación Secundaria	Facultad de Ciencias de la Educación
246	2021-10-20	Tesis	Educación Secundaria	Facultad de Ciencias de la Educación
247	2021-08-03	Tesis	Educación Secundaria	Facultad de Ciencias de la Educación
248	2021-07-14	Tesis	Sociología	Facultad de Ciencias Sociales
249	2021-11-29	Tesis	Sociología	Facultad de Ciencias Sociales
250	2021-09-03	Tesis	Sociología	Facultad de Ciencias Sociales
251	2021-10-28	Tesis	Sociología	Facultad de Ciencias Sociales
252	2021-08-18	Tesis	Sociología	Facultad de Ciencias Sociales
253	2021-11-11	Tesis	Sociología	Facultad de Ciencias Sociales
254	2021-07-02	Tesis	Sociología	Facultad de Ciencias Sociales
255	2021-11-19	Tesis	Sociología	Facultad de Ciencias Sociales
256	2021-08-19	Tesis	Trabajo Social	Facultad de Trabajo Social
257	2021-12-09	Tesis	Trabajo Social	Facultad de Trabajo Social
258	2021-08-09	Tesis	Trabajo Social	Facultad de Trabajo Social
259	2021-07-09	Tesis	Trabajo Social	Facultad de Trabajo Social
260	2021-09-01	Tesis	Trabajo Social	Facultad de Trabajo Social
261	2021-07-09	Tesis	Trabajo Social	Facultad de Trabajo Social
262	2021-07-02	Tesis	Trabajo Social	Facultad de Trabajo Social
263	2021-08-20	Tesis	Trabajo Social	Facultad de Trabajo Social
264	2021-08-24	Tesis	Trabajo Social	Facultad de Trabajo Social
265	2021-11-05	Tesis	Trabajo Social	Facultad de Trabajo Social
266	2021-11-08	Tesis	Trabajo Social	Facultad de Trabajo Social

267	2021-10-14	Tesis	Administración	Doctorado
268	2021-08-27	Tesis	Administración	Doctorado
269	2021-06-18	Tesis	Administración	Doctorado
270	2021-10-25	Tesis	Ciencias de la Salud	Doctorado
271	2021-11-12	Tesis	Ciencias de la Salud	Doctorado
272	2021-09-02	Tesis	Ciencias de la Salud	Doctorado
273	2021-08-31	Tesis	Ciencias de la Salud	Doctorado
274	2021-09-07	Tesis	Ciencias de la Salud	Doctorado
275	2021-08-27	Tesis	Ciencias Políticas y Gobernanza	Doctorado
276	2021-12-03	Tesis	Ciencias Sociales, Gestión Pública y Desarrollo Territorial	Doctorado
277	2021-06-18	Tesis	Ciencias Sociales, Gestión Pública y Desarrollo Territorial	Doctorado
278	2021-10-22	Tesis	Ciencia, Tecnología y Medio Ambiente	Doctorado
279	2021-10-18	Tesis	Ciencia, Tecnología y Medio Ambiente	Doctorado
280	2021-12-03	Tesis	Ciencia, Tecnología y Medio Ambiente	Doctorado
281	2021-10-22	Tesis	Ciencia, Tecnología y Medio Ambiente	Doctorado
282	2021-09-09	Tesis	Ciencia, Tecnología y Medio Ambiente	Doctorado
283	2021-07-09	Tesis	Ciencia, Tecnología y Medio Ambiente	Doctorado
284	2021-06-10	Tesis	Ciencia, Tecnología y Medio Ambiente	Doctorado
285	2021-07-21	Tesis	Ciencia, Tecnología y Medio Ambiente	Doctorado
286	2021-06-17	Tesis	Contabilidad y Administración	Doctorado
287	2021-12-29	Tesis	Derecho	Doctorado
288	2021-06-04	Tesis	Economía y Políticas Públicas	Doctorado
289	2021-09-10	Tesis	Educación	Doctorado
290	2021-12-29	Tesis	Estadística Aplicada	Doctorado
291	2021-12-06	Tesis	Estadística e Informática	Doctorado
292	2021-08-31	Tesis	Ciencia Animal	Maestría
293	2021-10-29	Tesis	Ciencia Animal	Maestría
294	2021-10-29	Tesis	Ciencias de la Nutrición	Maestría
295	2021-07-22	Tesis	Ciencias de la Nutrición	Maestría
296	2021-07-05	Tesis	Ciencias de la Nutrición	Maestría
297	2021-07-27	Tesis	Ciencias de la Nutrición	Maestría
298	2021-12-01	Tesis	Ciencias - Ingeniería Química	Maestría
299	2021-11-26	Tesis	Ciencias Sociales	Maestría
300	2021-11-18	Tesis	Ciencias Sociales	Maestría
301	2021-07-07	Tesis	Ciencias Sociales	Maestría

302	2021-07-22	Tesis	Ciencias Sociales	Maestría
303	2021-06-25	Tesis	Ciencias Sociales	Maestría
304	2021-08-12	Tesis	Contabilidad y Administración	Maestría
305	2021-09-09	Tesis	Contabilidad y Administración	Maestría
306	2021-06-25	Tesis	Contabilidad y Administración	Maestría
307	2021-06-10	Tesis	Contabilidad y Administración	Maestría
308	2021-11-17	Tesis	Derecho	Maestría
309	2021-11-19	Tesis	Desarrollo Rural	Maestría
310	2021-07-30	Tesis	Ecología	Maestría
311	2021-07-27	Tesis	Ecología	Maestría
312	2021-10-14	Tesis	Economía	Maestría
313	2021-07-01	Tesis	Economía	Maestría
314	2021-06-17	Tesis	Economía	Maestría
315	2021-08-13	Tesis	Educación	Maestría
316	2021-08-26	Tesis	Educación	Maestría
317	2021-08-05	Tesis	Educación	Maestría
318	2021-07-09	Tesis	Educación	Maestría
319	2021-09-09	Tesis	Educación	Maestría
320	2021-06-11	Tesis	Educación	Maestría
321	2021-08-27	Tesis	Educación	Maestría
322	2021-07-30	Tesis	Educación	Maestría
323	2021-10-06	Tesis	Educación	Maestría
324	2021-11-08	Tesis	Informática	Maestría
325	2021-10-27	Tesis	Informática	Maestría
326	2021-07-21	Tesis	Informática	Maestría
327	2021-08-04	Tesis	Informática	Maestría
328	2021-10-12	Tesis	Investigación y Docencia Universitaria	Maestría
329	2021-07-26	Tesis	Tecnologías de Protección Ambiental	Maestría
330	2021-07-12	Tesis	Trabajo Social	Maestría
331	2021-06-11	Tesis	Trabajo Social	Maestría
332	2021-12-01	Tesis de segunda especialidad	Didáctica Universitaria	Título de Segunda Especialidad
333	2021-09-02	Tesis de segunda especialidad	Didáctica Universitaria	Título de Segunda Especialidad
334	2021-12-28	Tesis de segunda especialidad	Docencia en Idioma Extranjero Inglés	Título de Segunda Especialidad
335	2021-12-20	Tesis de segunda especialidad	Docencia en Idioma Extranjero Inglés	Título de Segunda Especialidad
336	2021-10-26	Tesis de segunda especialidad	Docencia en Idioma Extranjero Inglés	Título de Segunda Especialidad

337	2021-10-19	Tesis de segunda especialidad	Docencia en Idioma Extranjero Inglés	Título de Segunda Especialidad
338	2021-10-14	Tesis de segunda especialidad	Docencia en Idioma Extranjero Inglés	Título de Segunda Especialidad
339	2021-10-13	Tesis de segunda especialidad	Docencia en Idioma Extranjero Inglés	Título de Segunda Especialidad
340	2021-08-12	Tesis de segunda especialidad	Docencia en Idioma Extranjero Inglés	Título de Segunda Especialidad
341	2021-07-19	Tesis de segunda especialidad	Docencia en Idioma Extranjero Inglés	Título de Segunda Especialidad
342	2021-07-09	Tesis de segunda especialidad	Docencia en Idioma Extranjero Inglés	Título de Segunda Especialidad
343	2021-12-17	Tesis de segunda especialidad	Psicomotricidad	Título de Segunda Especialidad
344	2021-11-04	Tesis de segunda especialidad	Ciencias Sociales	Título de Segunda Especialidad
345	2021-10-11	Tesis de segunda especialidad	Ciencias Sociales	Título de Segunda Especialidad
346	2021-09-03	Tesis de segunda especialidad	Ciencias Sociales	Título de Segunda Especialidad
347	2021-09-03	Tesis de segunda especialidad	Educación Intercultural Bilingüe, Aymara y Quechua	Título de Segunda Especialidad
348	2021-08-17	Tesis de segunda especialidad	Investigación Educativa	Título de Segunda Especialidad
349	2021-06-16	Tesis de segunda especialidad	Tecnología Computacional e Informática Educativa	Título de Segunda Especialidad
350	2021-10-17	Trabajo académico	Pediatría	Título de Segunda Especialidad
351	2021-10-17	Trabajo académico	Pediatría	Título de Segunda Especialidad
352	2021-09-04	Trabajo académico	Pediatría	Título de Segunda Especialidad
353	2021-09-04	Trabajo académico	Anestesiología	Título de Segunda Especialidad
354	2021-09-04	Trabajo académico	Anestesiología	Título de Segunda Especialidad
355	2021-08-08	Trabajo académico	Ginecología y Obstetricia	Título de Segunda Especialidad
356	2021-08-08	Trabajo académico	Ortopedia y Traumatología	Título de Segunda Especialidad
357	2021-08-08	Trabajo académico	Radiologia	Título de Segunda Especialidad
358	2021-08-08	Trabajo académico	Medicina Interna	Título de Segunda Especialidad
359	2021-08-08	Trabajo académico	Medicina Interna	Título de Segunda Especialidad
360	2021-08-08	Trabajo académico	Radiologia	Título de Segunda Especialidad

361	2021-06-07	Trabajo académico	Pediatría	Título de Segunda Especialidad
362	2021-06-25	Trabajo académico	Anestesiología	Título de Segunda Especialidad
363	2021-12-22	Tesis de segunda especialidad	Educación Básica Alternativa	Título de Segunda Especialidad
364	2021-12-22	Tesis de segunda especialidad	Educación Básica Alternativa	Título de Segunda Especialidad
365	2021-12-21	Tesis de segunda especialidad	Educación Básica Alternativa	Título de Segunda Especialidad
366	2021-12-23	Tesis de segunda especialidad	Educación Básica Alternativa	Título de Segunda Especialidad
367	2021-12-10	Tesis de segunda especialidad	Educación Básica Alternativa	Título de Segunda Especialidad
368	2021-12-07	Tesis de segunda especialidad	Educación Básica Alternativa	Título de Segunda Especialidad
369	2021-12-07	Tesis de segunda especialidad	Educación Básica Alternativa	Título de Segunda Especialidad
370	2021-12-01	Tesis de segunda especialidad	Educación Básica Alternativa	Título de Segunda Especialidad
371	2021-08-11	Tesis de segunda especialidad	Educación Básica Alternativa	Título de Segunda Especialidad
372	2021-06-07	Tesis de segunda especialidad	Educación Básica Alternativa	Título de Segunda Especialidad
373	2021-12-20	Tesis de segunda especialidad	Educación Inicial	Título de Segunda Especialidad
374	2021-12-14	Tesis de segunda especialidad	Educación Inicial	Título de Segunda Especialidad
375	2021-08-04	Tesis de segunda especialidad	Educación Inicial	Título de Segunda Especialidad
376	2021-07-30	Tesis de segunda especialidad	Educación Inicial	Título de Segunda Especialidad
377	2021-12-15	Tesis de segunda especialidad	Gestión y Administración Educativa	Título de Segunda Especialidad
378	2021-07-08	Tesis de segunda especialidad	Gestión y Administración Educativa	Título de Segunda Especialidad
379	2021-08-20	Tesis de segunda especialidad	Psicología Educativa	Título de Segunda Especialidad
380	2021-08-06	Tesis de segunda especialidad	Psicología Educativa	Título de Segunda Especialidad

Anexo N° 2. Encuesta de satisfacción

ENCUESTA DE SATISFACCION

La encuesta tiene el propósito de evaluar el software para la administración de RI, se desea conocer su opinión sobre dicho software. Se le solicita a Ud. Tener en cuenta la siguiente escala de satisfacción establecida para su valoración:

1. Muy en desacuerdo
2. En desacuerdo
3. No estoy seguro
4. De acuerdo
5. Muy de acuerdo

Para responder maque con X según la escala de satisfacción con respecto a cada ítem

N°	Ítem	Escala de satisfacción				
		1	2	3	4	5
1	Los metadatos extraídos de un documento de investigación son confiables					
2	El software me resulto complejo de usar					
3	El sistema es bastante fácil de usar					
4	Pienso que necesitaría el soporte técnico para utilizar el software					
5	El software contiene funciones que abarcan las necesidades para la publicación de documentos de investigación					
6	La funcionalidad ofrecida por el software contribuye de manera completa a los procesos de administración del Repositorio Institucional					
7	Pienso que hubo demasiada inconsistencia en el software					
8	Opino que la mayoría de personas podrían utilizar el software rápidamente					
9	Pienso que para operar el software se requiere hacer una capacitación					
10	El software presenta errores continuamente mientras se opera con él, en los procesos que corresponden a la administración de RI					
11	La extracción automática de metadatos y la publicación de metadatos facilita la administración del RI					
12	El software es de suma importancia para mi trabajo diario					
13	Recomendaría el uso del software a otras Instituciones como herramienta para la administración de Repositorios Institucionales.					

¿Cuánto es el tiempo promedio en minutos que demora en publicar un documento de investigación en el Repositorio Institucional?

……(minutos)

GRACIAS!.. POR SU COLABORACION

Anexo N° 3. Ficha de validación

UNIVERSIDAD NACIONAL DEL ALTIPLANO
ESCUELA DE POSGRADO
MAESTRÍA EN INFORMÁTICA

FICHA DE VALIDACIÓN
INFORME OPINION DEL JUICIO DE EXPERTO

DATOS GENERALES

1.1. Nombre de los instrumentos motivo de evaluación: **Encuesta de Satisfacción sobre Extracción automática de metadatos para la administracion del Repositorio Institucional de la UNA PUNO**

1.2. Autor del instrumento: **Alain Paul Herrera Urtiaga**

ASPECTOS DE VALIDACION

INDICADORES	CRITERIOS	Deficiente				Regular				Bueno				Muy bueno				Excelente			
		0	6	11	16	21	26	31	36	41	46	51	56	61	66	71	76	81	86	91	96
		5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
1. CLARIDAD	Está formado con lenguaje apropiado														✓						
2. OBJETIVIDAD	Esta expresado en conductas observables																✓				
3. ACTUALIDAD	Adecuado al avance de ciencia y la tecnología																		✓		
4. ORGANIZACION	Existe una organización lógica															✓					
5. SUFICIENCIA	Comprende los aspectos en cantidad y calidad															✓					
6. INTENCIONALIDAD	Adecuado para valorar los instrumentos de investigación.																✓				
7. CONSISTENCIA	Basado en aspectos teóricos científicos.															✓					
8. COHERENCIA	Entre los índices e indicadores														✓						
9. METODOLOGIA	La estrategia responde al propósito del diagnostico															✓					
10. PERTINENCIA	Es útil y adecuado para la investigación.																✓				

PROMEDIO DE VALORACION: 77

OPINION DE APLICABILIDAD: a) Deficiente b) Regular c) Bueno d) Muy bueno e) Excelente

Nombres y apellidos:	Reynaldo Sucari León
Dirección domiciliaria:	Jr. Razuhuillca N° 388 Cercado de Huanta
Grados Académicos:	M.Sc. en Informática y Dr. en Administración de la Educación
Nro DNI:	01341544
Teléfono celular:	975126540

Lugar y fecha: Huanta, 02 de abril de 2022

Universidad Nacional Autónoma de Huanta
Firma Digital
Firmado digitalmente por Reynaldo Sucari León
Fecha: 2022.04.02 09:45:35 -05'00'

Dr. Reynaldo Sucari León
DOCENTE

UNIVERSIDAD NACIONAL DEL ALTIPLANO
ESCUELA DE POSGRADO
MAESTRÍA EN INFORMÁTICA

FICHA DE VALIDACIÓN
INFORME OPINION DEL JUICIO DE EXPERTO

DATOS GENERALES

1.1. Nombre de los instrumentos motivo de evaluación: **Encuesta de Satisfacción sobre Extracción automática de metadatos para la administracion del Repositorio Institucional de la UNA PUNO**

1.2. Autor del instrumento: **Alain Paul Herrera Urtiaga**

ASPECTOS DE VALIDACION

INDICADORES	CRITERIOS	Deficiente				Regular				Bueno				Muy bueno				Excelente			
		0	6	11	16	21	26	31	36	41	46	51	56	61	66	71	76	81	86	91	96
		5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
1. CLARIDAD	Está formado con lenguaje apropiado																X				
2. OBJETIVIDAD	Esta expresado en conductas observables															X					
3. ACTUALIDAD	Adecuado al avance de ciencia y la tecnología																		X		
4. ORGANIZACIÓN	Existe una organización lógica																	X			
5. SUFICIENCIA	Comprende los aspectos en cantidad y calidad																X				
6. INTENCIONALIDAD	Adecuado para valorar los instrumentos de investigación.																X				
7. CONSISTENCIA	Basado en aspectos teóricos científicos.															X					
8. COHERENCIA	Entre los índices e indicadores																X				
9. METODOLOGIA	La estrategia responde al propósito del diagnostico																X				
10. PERTINENCIA	Es útil y adecuado para la investigación.																X				

PROMEDIO DE VALORACION: 79.7

OPINION DE APLICABILIDAD: a) Deficiente b) Regular c) Bueno d) **Muy bueno** e) Excelente

Nombres y apellidos:	**JUAN CARLOS JUAREZ VARGAS**
Dirección domiciliaria:	**JIRON 7 DE JUNIO 329 PUNO**
Grados Académicos:	**Magister Scientiae en Informática y Doctor Scientiae en Administración**
Nro DNI:	**40419555**
Teléfono celular:	**951007384**

Lugar y fecha: Puno, 02 de abril 2022

Firmado digitalmente por JUAREZ VARGAS Juan Carlos FAU 20145496170 soft
Motivo: Soy el autor del documento
Fecha: 05.04.2022 12:18:47 -05:00

DR. JUAN CARLOS JUAREZ VARGAS

UNIVERSIDAD NACIONAL DEL ALTIPLANO
ESCUELA DE POSGRADO
MAESTRÍA EN INFORMÁTICA

FICHA DE VALIDACIÓN
INFORME OPINION DEL JUICIO DE EXPERTO

DATOS GENERALES

1.1. Nombre de los instrumentos motivo de evaluación: **Encuesta de Satisfacción sobre Extracción automática de metadatos para la administracion del Repositorio Institucional de la UNA PUNO**

1.2. Autor del instrumento: **Alain Paul Herrera Urtiaga**

ASPECTOS DE VALIDACION

INDICADORES	CRITERIOS	Deficiente				Regular				Bueno				Muy bueno				Excelente			
		0 5	6 10	11 15	16 20	21 25	26 30	31 35	36 40	41 45	46 50	51 55	56 60	61 65	66 70	71 75	76 80	81 85	86 90	91 95	96 100
1. CLARIDAD	Está formado con lenguaje apropiado																		X		
2. OBJETIVIDAD	Esta expresado en conductas observables																X				
3. ACTUALIDAD	Adecuado al avance de ciencia y la tecnología																		X		
4. ORGANIZACIÓN	Existe una organización lógica																		X		
5. SUFICIENCIA	Comprende los aspectos en cantidad y calidad																		X		
6. INTENCIONALIDAD	Adecuado para valorar los instrumentos de investigación.																X				
7. CONSISTENCIA	Basado en aspectos teóricos científicos.																	X			
8. COHERENCIA	Entre los índices e indicadores																		X		
9. METODOLOGIA	La estrategia responde al propósito del diagnostico																		X		
10. PERTINENCIA	Es útil y adecuado para la investigación.																				X

PROMEDIO DE VALORACION: 88.5

OPINION DE APLICABILIDAD: a) Deficiente b) Regular c) Bueno d) Muy bueno e) Excelente

Nombres y apellidos:	RAMIRO PEDRO LAURA MURILLO
Dirección domiciliaria:	URB. AZIRUNI III ETAPA MZ X1 L13
Grados Académicos:	DOCTOR EN CIENCIAS DE LA COMPUTACIÓN
Nro DNI:	41939172
Teléfono celular:	930654095

Lugar y fecha: Puno 31 de marzo de 2022

UNIVERSIDAD NACIONAL DEL ALTIPLANO

Anexo N° 4. Matriz de Consistencia

PROBLEMA	OBJETIVOS	HIPOTESIS	VARIABLES	INDICADOR
Problema General ¿De qué manera la extracción automática de metadatos mejora la administración del Repositorio Institucional de la Universidad Nacional del Altiplano	**Objetivo General** Optimizar la extracción de metadatos y publicación de documentos de investigación para la administración del Repositorio Institucional de la Universidad Nacional del Altiplano.	**Hipótesis General** La extracción automática de metadatos mejora la administración del Repositorio Institucional de la Universidad Nacional del Altiplano	**Variable independiente** Extracción automática de metadatos	**Calidad de uso ISO/IEC 25010** Dimensiones: • Efectividad • Eficiencia • Satisfacción **Precisión algoritmo** Dimensiones: • Precisión • Recall
Problemas específicos • ¿En qué medida el procesamiento de lenguaje natural contribuye al desarrollo del algoritmo para la extracción automática de metadatos en los documentos de investigación del Repositorio Institucional Universidad Nacional del Altiplano? • ¿En qué medida la precisión del algoritmo permite la extracción automática de metadatos de los documentos de investigación? • ¿Cuál es la diferencia del tiempo de ejecución en el proceso de extracción de metadatos antes y después de la implementación del software?	**Objetivos específicos** • Desarrollar un algoritmo para la extracción automática de metadatos basada en el procesamiento de lenguaje natural (NLP). • Implementar y determinar el nivel de precisión del algoritmo de extracción automática de metadatos basada en procesamiento de lenguaje natural (NLP). • Evaluar la diferencia del tiempo de extracción de metadatos antes y después de la implementación del software.	**Hipótesis especificas** • El procesamiento de lenguaje natural contribuye al desarrollo del algoritmo para la extracción automática de metadatos en los documentos de investigación del Repositorio Institucional UNA- Puno • El algoritmo basado en procesamiento de lenguaje natural logra un nivel de precisión adecuado para la extracción automática de metadatos de los documentos de investigación. • La implementación del software mejora el tiempo en el proceso de extracción de metadatos de los documentos de investigación	**Variable dependiente** Administración del Repositorio Institucional	Tiempo de extracción de metadatos y publicación de documentos de investigación en el RI UNA-Puno Dimensiones: • Tiempo

Anexo N° 5. Archivo app.py

```
from flask import Flask
from flask import render_template,request,jsonify,json,redirect
from flaskext.mysql import MySQL
from werkzeug.utils import secure_filename
import pymysql
import requests
import os

#Connection Database
app.config['UPLOAD_FOLDER']="./Documents"
mysql=MySQL()
app.config['MYSQL_DATABASE_HOST']='localhost'
app.config['MYSQL_DATABASE_USER']='root'
app.config['MYSQL_DATABASE_PASSWORD']=''
app.config['MYSQL_DATABASE_DB']='dbextract'
mysql.init_app(app)

#Module app
app=Flask(__name__)

#Routes application
app.add_url_rule(routes["index_route"],view_func=routes["index_cont
roller"])
app.add_url_rule(routes["documents_route"],view_func=routes["docume
nts_controller"])
app.add_url_rule(routes["users_route"],view_func=routes["users_cont
roller"])
app.add_url_rule(routes["thesis_route"],view_func=routes["documents
_controller"])
app.add_url_rule(routes["dictonaries_route"],view_func=routes["dict
ionaries_controller"])
app.add_url_rule(routes["advisers_route"],view_func=routes["adviser
_controller"])
app.add_url_rule(routes["uploads_route"],view_func=routes["uploads_
controller"])
app.add_url_rule(routes["update_route"],view_func=routes["update_co
ntroller"])
app.add_url_rule(routes["delete_route"],view_func=routes["delete_co
ntroller"])
app.register_error_handle(routes["not_found_route"],view_func=route
s["not_found_controller"])

#------MAIN--------------
if __name__=='__main__':

    app.run(debug=True)
```