SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Secretaría Académica, de Investigación e Innovación
Dirección de Posgrado, Investigación e Innovación

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Centro Nacional de Investigación y Desarrollo Tecnológico

Subdirección Académica

Departamento de Ciencias Computacionales

TESIS DE MAESTRÍA EN CIENCIAS

Sistema semiautomático para extraer atributos de documentos semiestructurados y su inserción en un repositorio

presentada por
Ing. Karen Jannete Jaime Díaz

como requisito para la obtención del grado de
Maestra en Ciencias en Ciencias de la Computación

Director de tesis
Dr. Juan Gabriel González Serna

Cuernavaca, Morelos, México. Enero de 2018.

"Año del Centenario de la Promulgación de la Constitución Política de los Estados Unidos Mexicanos"

Asunto: Aceptación de documento de tesis

DR. GERARDO V. GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial de la **Ing. Karen Jannete Jaime Díaz**, con número de control M15CE084, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis profesional titulado "**Sistema semiautomático para extraer atributos de documentos semiestructurados y su inserción en un repositorio"** y hemos encontrado que se han realizado todas las correcciones y observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

DIRECTOR DE TESIS

Dr. Juan Gabriel González Serna
Doctor en Ciencias de la Computación
7820329

REVISOR 1

Dr. Noé Alejandro Castro Sánchez
Doctor en Ciencias de la Computación
08701806

REVISOR 2

Dra. Azucena Montes Rendón
Doctora en Ciencias
4001014

C.p. M.T.I. María Elena Gómez Torres - Jefa del Departamento de Servicios Escolares.
Estudiante
Expediente

NACS/lmz

Cuernavaca, Mor., 15 de enero de 2018
OFICIO No. SAC/053/2018

Asunto: Autorización de impresión de tesis

ING. KAREN JANNET JAIME DÍAZ
CANDIDATA AL GRADO DE MAESTRA EN CIENCIAS
DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Sistema semiautomático para extraer atributos de documentos semiestructurados y su inserción en un repositorio"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

ATENTAMENTE
"CONOCIMIENTO Y TECNOLOGÍA AL SERVICIO DE MÉXICO"

SEP TecNM
CENTRO NACIONAL
DE INVESTIGACIÓN
Y DESARROLLO
TECNOLÓGICO
SUBDIRECCIÓN
ACADÉMICA

DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO

C.p. M.T.I. María Elena Gómez Torres .- Jefa del Departamento de Servicios Escolares.
Expediente

GVGR/mcr

Dedicatoria.

A Dios, que siempre me dio la fuerza necesaria para superar de las mejor forma las adversidades que se me presentaron a lo largo de estos años de maestría y así como permitirme lograr culminar una meta más en mi vida.

A mis padres Arturo Jaime, Maria Díaz, y hermana Evelyn, gracias por haberme impulsado a entrar a estudiar la maestría y creer que a pesar de lo difícil que pudiera ser conseguiría mi nueva meta, gracias a ustedes soy esta persona medio loca pero dedicada y enfocada en lo que quiere; gracias por ayudarme a cumplir otro de mis sueños en este periodo, como fue el realizar una estancia en Avignon, Francia, en verdad voy a estarles agradecida toda la vida por haberme apoyado en ese sueño que surgió de manera inesperada y culmino de la mejor forma. Los amo!

A mi tía Rosa y mis primos que estuvieron conmigo dándome apoyo y un más cuando lo necesite, se que no tenían porque pero en verdad se los agradezco, los quiero mucho.

A mis amigos, Alberto, Loreli, Antonio, Michel, que siempre estuvieron a mi lado para darme ánimos, apoyo y momentos increíbles durante este periodo, gracias por aguantar todas mis locuras, cambios de humor e hiperactividad que solía tener, bueno en general gracias por estar a mi lado.

Agradecimientos

Primeramente agradezco Centro Nacional de Ciencia y Tecnología (cenidet), por haberme aceptado para formar parte de los estudiantes que se integran estudiar un posgrado (maestría), así como darme la posibilidad de ampliar mis conocimientos por medio de los planes de estudio que tienen, así mismo agradezco al Concejo Nacional de Ciencia y Tecnología (CONACyT) por su programa de posgrados de calidad a través del cual recibí apoyo económico durante el tiempo que fui estudiante de maestría.

A mi director de tesis el Dr. Juan Gabriel González Serna, que bajo su dirección se logro culminar este trabajo de tesis, proporcionando el conocimiento y la orientación que se necesita durante el desarrollo del trabajo.

A mis revisores el Dr. Noé Alejandro Castro Sánchez y la Dra. Azucena Montes Rendón, quienes desde el comienzo de este trabajo de tesis aportaron tiempo y comentarios para mejorar el trabajo que se realizaba, con cada una de las revisiones realizadas.

Al Laboratoire Informatique d'Avignon de l'Université d'Avignon et des Pays de Vaucluse, Francia, y al Dr. Juan-Manuel Torres-Moreno, que me permitieron realizar una estancia de investigación, durante la cual se obtuvieron aportaciones importantes para el trabajo, así como la obtención de nuevo conocimiento y nuevas experiencias como estudiante y persona.

Resumen

En el presente documento de tesis se muestra el trabajo realizado durante los dos años de maestría; este trabajo tiene como objetivo ayudar a la conformación del repositorio institucional del Tecnológico Nacional de México (TecNM).

El tema que se abordó en este trabajo, es la extracción de información (metadatos) en documentos PDF, específicamente documentos creados en alguna de las instituciones que conforman el TecNM, dichos documentos pueden ser tesis de los niveles licenciatura, maestría y doctorado y artículos científicos con la arquitectura IEEE o ACM.

En este documento de tesis se describe la metodología seguida para realizar la extracción semiautomática de información específica (metadatos) en los documentos PDF. Los metadatos que se extraen dependen del tipo de documento (tesis, artículo); algunos de los metadatos que son extraídos de dichos documentos son; autor, título, resumen, fecha, palabras clave, entre otros.

También se presentan las pruebas realizadas al sistema desarrollado, así como los resultados obtenidos de dichas pruebas, tomando para evaluar los resultados las medidas de precisión y cobertura, las cuelas muestran cuales fueron los metadatos con mayor y menor calidad de precisión en la extracción.

Abstract

In this thesis document shows the work done during the two years of masters; this document aims to help the formation of the institutional repository of the Tecnológico Nacional de México (TecNM).

The subject that was addressed in this work is the extraction of information (metadata) in PDF documents, specifically documents created in institutions that make up the TecNM, these documents can be thesis of the undergraduate, masters and doctorate levels and scientific articles with the IEEE or ACM architecture.

This thesis document describes the methodology used to perform the semi-automatic extraction of specific information (metadata) in PDF documents. The metadata extracted depends on the type of document (thesis, paper); some of the metadata that are extracted from such documents are; author, title, abstract, date, keywords, among others.

Also present the tests performed to the developed system, as well as the results obtained from these tests, was taken to evaluate the results the precision and coverage measures, which show the metadata with higher and lower quality of extraction precision.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE DIAGRAMAS

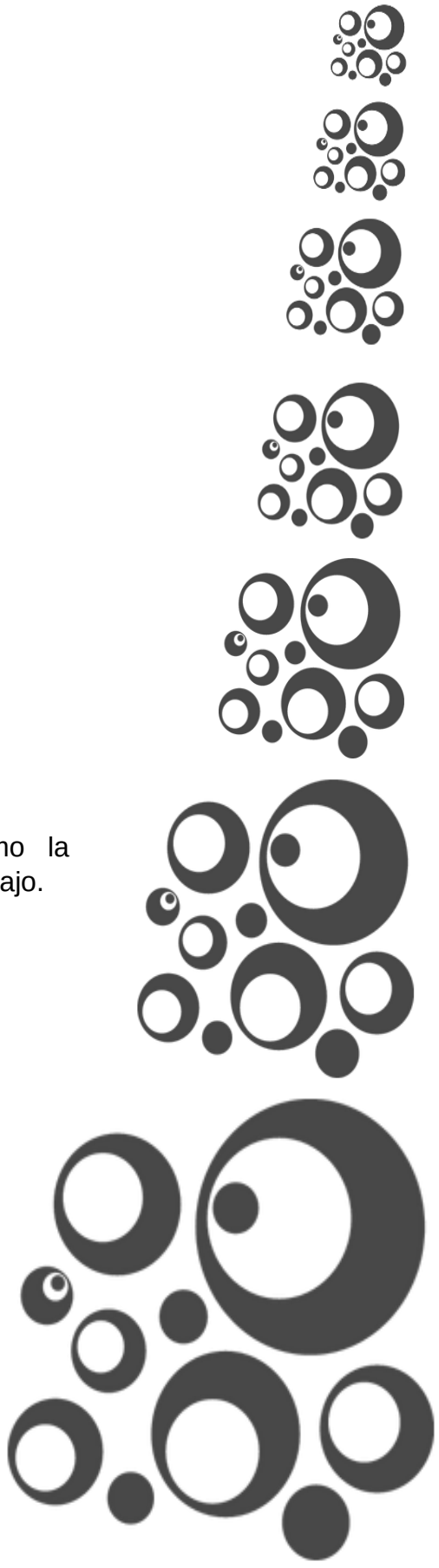
ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICAS DE DATOS

ÍNDICE DE GRÁFICA DE RESULTADOS

Capitulo I. Introducción

En este capítulo se da la introducción, así como la problemática, los objetivos y la justificación de este trabajo.



1. Introducción

En las últimas décadas, el número de documentos digitales creados ha sufrido un aumento considerable y más si nos enfocamos en los documentos de carácter científico o tecnológico, ya sean tesis de los diferentes niveles, licenciatura, maestría, doctorado; artículos publicados en revistas o congresos; patentes o incluso investigaciones médicas, entre muchos otros.

El inconveniente no es la generación de tal cantidad de información diaria en el país, el problema se encuentra específica a la hora de acceder a esta información, es decir, para poder acceder a dicha información es necesario realizar un proceso arduo, pero en muchos casos realizar esa labor es muy complicada o imposible de realizar. Este problema de acceso surge por varios factores entre ellos se encuentra, la mala organización a la hora de almacenar la información, la falta de estructura de los documentos que complican aún más la organización, entre otros más.

A que se refiere la mala organización, como ya se mencionó en el país diariamente se generan grandes cantidades de información, la cual no es almacenada ni clasificada de una manera adecuada por las instituciones que la generan, esté problema no es detectado por las instituciones hasta que la información generada es considerablemente grande y se les dificulta el manejo de esa información, por este motivo a la hora que se realizan búsquedas de información en ocasiones los resultados arrojado son nulos, mínimos o incorrectos.

Por el lado de la falta de estructura, el problema se enfoca en que los documentos científicos deben contener ciertas características o atributos para considerase documentos formales o estructurados y en ocasiones los autores no los toman

en cuenta, esto a su vez provoca que, si se desea realizar un trabajo de clasificación en los documentos el trabajo sea más complicado o imposible.

En este documento se hablará sobre la problemática y la solución que se dio para la clasificación por tipo de documento y para la extracción de información de los documentos PDF, tomando como base los metadatos sugeridos por el CONACYT y Dublin Core.

1.1. Antecedentes

A continuación, se muestran los trabajos realizados anteriormente en la institución, los cuales cuentan con cierto grado de similitud con el trabajo a realizar, los cuales se pueden tomar como base para el desarrollo de la tesis.

1.1.1. Clasificación automática de objetos de conocimiento con contenido no estructurado para el poblado semiautomático de ontologías multidimensionales

(Rendón Miranda, 2014)

En la última década ha ido aumentando la distribución de material científico (objetos de conocimiento) en formato PDF, el principal inconveniente de estos objetos de conocimiento es que no contienen una estructura estandarizada. Ha causa de esta problemática ha surgiendo el campo de minería de textos, para la extracción, recuperación y clasificación automática de los objetos de conocimiento.

Se requiere un sistema que permita en poco tiempo, la extracción de información, clasificación automática con base al dominio y el poblado semiautomático de una ontología de objetos de conocimiento. El fin de una ontología es modelar la información referente a un escenario organizacional con dimensiones contextuales, conocimiento de memorias organizacionales, memorias individuales y características de los usuarios.

El objetivo de trabajo es desarrollar un mecanismo que permita de manera unificada procesamiento de documentos en formato PDF, el cual permita el clasificado de manera automática con base al dominio de conocimiento al que se asocia con base a una taxonomía y realizar el poblado semiautomático de la ontología.

La ontología se refiere a un modelo abstracto de cierto fenómeno en el mundo, identificando el concepto relevante de este fenómeno. Un modelo de este tipo que este bien estructurado que es capaz de describir el mundo real.

Los algoritmos de clasificación utilizados para el desarrollo de este trabajo: Navie Bayes, máquina de soporte vectorial, arboles de decisión, k-vecinos más cercanos.

Similitud:
Debido a que los documentos con los cuales se trabajaran para realizar la clasificación so documentos que se encuentran en formato PDF, es necesario que se identifique la estructura que debe contener cada documento, esta estructura tiene que ser dependiendo del tipo que se clasificara (artículo, tesis). Como en este trabajo es necesario que el sistema sea lo suficientemente flexible para permitir la extracción y el clasificado de la información en diversas áreas de conocimiento.

1.1.2. Evaluación automática de ontologías de dominio restringido

(Tovar Vidal, 2015)

Las ontologías son recursos semánticos que almacenan el conocimiento de un determinado dominio, por medio de los elementos que la conforman. El objetivo de este trabajo es desarrollar una metodología para la evaluación automática de ontologías de dominio utilizando una métrica que involucre medidas léxico-sintácticas, estadísticas y de similitud.

Las etapas que se llevaron a cabo para la elaboración de este trabajo consisten en:

1. El procesamiento automático de la información
2. El descubrimiento automático de términos candidatos y/o relaciones ontológicas.
3. Evaluación de la ontología

La primera etapa consiste en la extracción de los conceptos y las relaciones ontológicas, es decir, la información que es extraída del corpus es asociada a los conceptos por medio de la recuperación de información.

La segunda etapa consiste en el uso de varios métodos como son los patrones léxico-sintáctico, análisis semántico latente, análisis de conceptos formales, análisis de dependencias; los cuales permiten la identificación de las relaciones semánticas con sus respectivos conceptos en el corpus de dominio definido.

La última etapa consiste en mostrar la calidad de la ontología, es decir propone una métrica para la evaluación de la ontología involucrando los resultados de las etapas anteriores.

Similitud:

Para poder realizar el análisis de los documentos es necesaria realizar una adecuada extracción de los metadatos de los documentos PDF con los que se trabajará, por otro lado, como en el trabajo descrito anteriormente es necesaria la extracción adecuada de términos o conceptos (metadatos) los cuales permitan realizar el análisis del contenido del documento para su posterior clasificación.

1.2. Planteamiento del problema

El proceso de extracción de atributos de un documento para su registro en un repositorio o base de datos es un proceso que requiere de la intervención de un experto humano. Actualmente el proceso para poblar un repositorio o base de datos institucional, se realiza de forma manual en tres etapas, la primera es identificar el tipo de documento, para el caso de instituciones de educación superior se podrían clasificar en tesis de diferentes grados académicos, artículo técnicos en congresos o revistas, reportes técnicos, entre otros; el segundo proceso es la extracción de atributos de cada uno de los documentos, por ejemplo: el título, el (los) autor(es), el resumen, las palabras clave, la institución, entre otros; el tercer proceso es la clasificación y almacenamiento de los documentos en la base de datos, para poder realizar dicha clasificación, es necesaria la intervención de un especialista, el cual realiza el análisis del contenido y de los atributos del documento para identificar el(las) área(s) de conocimiento y finalmente realizar el registro en la base de datos.

Un repositorio institucional es heterogéneo, es decir, está conformado por diferentes tipos de documentos, artículos de revista, artículos de congresos, tesis de los diferentes niveles licenciatura, maestría y doctorado; reportes técnicos, patentes, etc. Esta heterogeneidad de los repositorios hace que sea más complejo el análisis del contenido de los documentos y a su vez realizar la clasificación, esto se debe a que cada tipo de documento tiene diferentes patrones, estructura y atributos.

Al no contar los repositorios institucionales con mecanismos semiautomáticos los cuales permitan realizar la extracción de los atributos y el análisis del contenido de los documentos, surge la necesidad de desarrollar un método que cumpla con estas necesidades, utilizando las técnicas de procesamiento de lenguaje natural.

A continuación, se muestra un diagrama para hacer más clara la problemática que se desea atacar con este proyecto. (Diagrama1)

Documentos a analizar

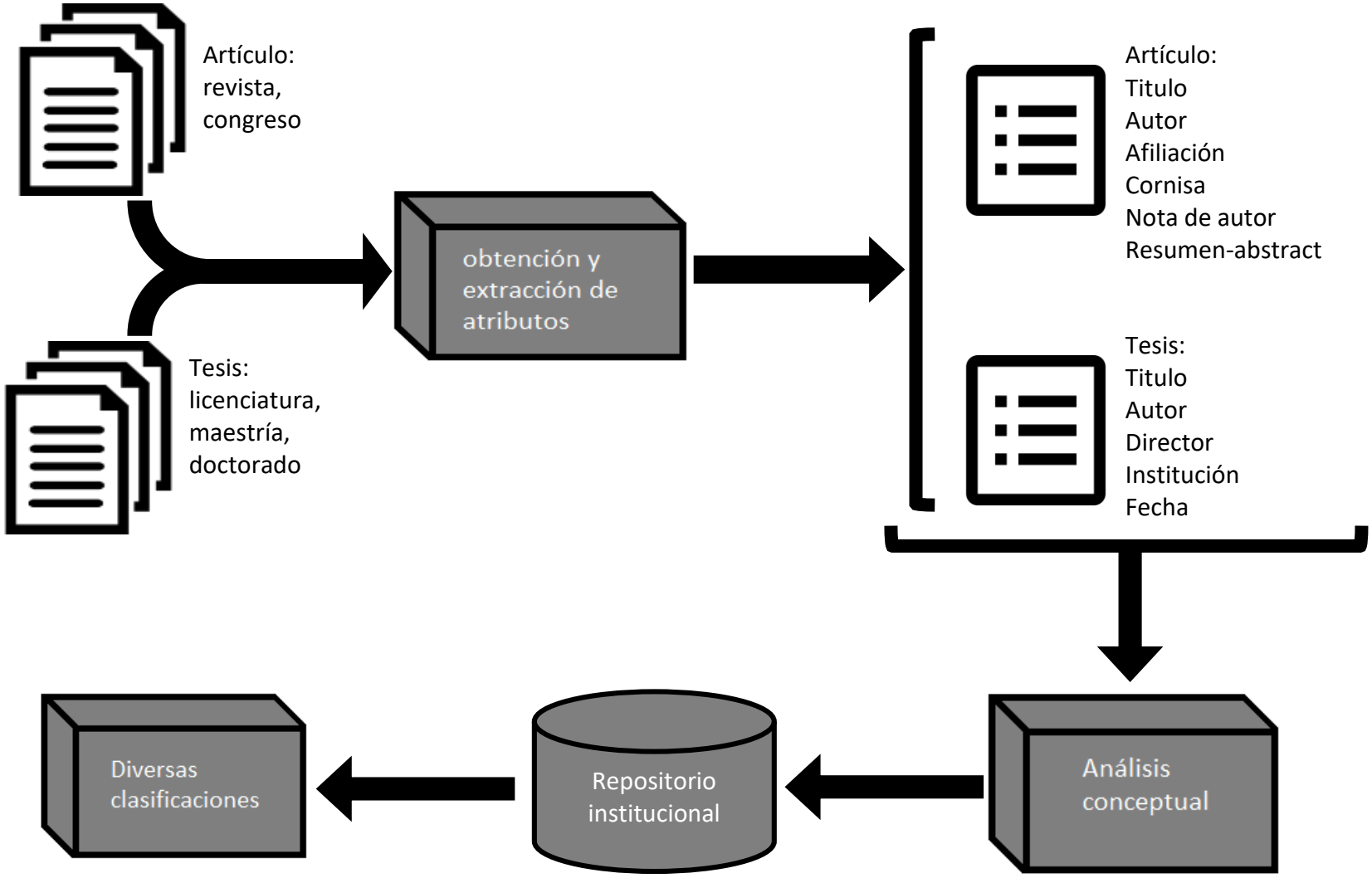


Diagrama 1 Problemática de proyecto

1.3. Objetivos

A continuación, se presenta el objetivo general y los objetivos específicos que se tienen para la realización del proyecto.

1.3.1. Objetivo general

Desarrollar un algoritmo para la extracción de atributos en documentos de tipo artículo científico y tesis de licenciatura, maestría y doctorado; el cual sea capaz de identificar el tipo de documento, así como información de los documentos, para semiautomatizar el proceso de poblado de un repositorio institucional.

1.3.2. Objetivos específicos

- Diseñar el patrón para documentos semiestructurados de tipo artículo en idioma español.
- Diseñar el patrón para documentos semiestructurados de tipo tesis en sus niveles licenciatura, maestría y doctorado en idioma español.
- Diseñar un algoritmo que permita la detección automática del tipo de documento ingresado.
- Diseñar un algoritmo que permita la extracción automática de los atributos específicos para cada uno de los tipos de documento artículo o tesis.
- Desarrollar un algoritmo para la inserción automática de los atributos de los documentos en un repositorio institucional.

1.4. Justificación

Actualmente las instituciones de educación superior no disponen de un repositorio institucional que almacene documentos con información científica, académica y tecnológica desarrollada en las instituciones, se debe realizar en una misma fuente de almacenamiento, a la cual se tendrá acceso por medio de tecnologías web, al no contar con este recurso en las instituciones, centros educativos y organismos gubernamentales, el consejo nacional de ciencia y tecnología CONACYT decidió fomentar la creación de los repositorios institucionales para tener un mejor y más rápido acceso a la información generada por éstos.

La función principal del repositorio institucional es el acopio, preservación, gestión y acceso electrónico a la información de calidad generada en las diversas instituciones, incluyendo aquella información que sea de interés social y cultural, esto con el fin de fortalecer las capacidades científicas del país.

Por lo mencionado anteriormente, el centro nacional de investigación y desarrollo tecnológico CENIDET optó por la creación del repositorio institucional del Tecnológico Nacional de México, con el objetivo de que los 266 institutos, centros y unidades que integran el Tecnológico Nacional de México conformen el repositorio institucional y de este modo se tenga un mejor acceso público a la información que se genera en estas instituciones.

1.5. Estructura del documento

La organización de este documento se encuentra dividida en siete capítulos, en los cuales se describe el proceso que se utilizó para el desarrollo de este trabajo, a continuación, se describe brevemente el contenido de cada capítulo.

Capítulo II

Marco teórico: En este capítulo se muestran los términos o conceptos con mayor relevancia que se utilizaran para para el entendimiento del presente documento.

Capítulo III

Estado del arte: En este capítulo se muestran algunos trabajos similares o relacionados a la extracción de metadatos en documentos PDFs.

Capítulo IV

Metodología: En este capítulo se muestra la metodología seguida para la detección, extracción e inserción de los metadatos de los PDFs en el repositorio institucional.

Capítulo V

Pruebas y resultados: En este capítulo se muestran las pruebas que se le realizaron al sistema con el objetivo de evaluar la precisión y cobertura con la cual se extrajeron los metadatos de los diferentes PDFs.

Capítulo VI

Conclusiones: En este capítulo se muestran las conclusiones del trabajo de igual forma si se cumplieron los objetivos de este trabajo.

Capitulo II. Marco teórico

En este capítulo se muestran los términos o conceptos con mayor relevancia que se utilizaran para para el entendimiento del presente documento.

2. Marco teórico

A continuación, se presentan los términos que tienen mayor relevancia para el entendimiento del desarrollo de esta investigación.

2.1. Lenguaje natural (LN)

El lenguaje natural está formado por un conjunto de símbolos finitos tomados de una colección, estas colecciones son las lenguas o idiomas y los símbolos son las letras del alfabeto y los diferentes símbolos utilizados (`´-_) en estas lenguas para la construcción de las palabras (Montero Martínez, 2001)

Se puede definir el lenguaje natural como la manera en la que se comunica el ser humano entre sí para expresar necesidades, ideas, emociones, etc., para el análisis de esta comunicación se debe tomar en cuenta la influencia que tiene el medioambiente en el que habita. (Peña Ayala, 2008)

2.2. Procesamiento de lenguaje natural (PLN)

El campo del procesamiento de lenguaje natural combina diferentes tecnologías que se encuentran dentro del área de ciencias computacionales como, el aprendizaje automático, inferencia estadística y la inteligencia artificial; y la lingüística aplicada, para poder realizar la comprensión y análisis del lenguaje natural por medio de un sistema computacional. (visual interaction & communication technologies, 2009)

El procesamiento de lenguaje natural se encarga de la creación de mecanismos computacionales, los cuales ayuden al entendimiento o comunicación entre personas y maquinas, esto quiere decir que las maquinas entiendan las oraciones escritas en lenguaje natural por las personas; algunas de las actividades que realiza el PLN son: la traducción, búsquedas semánticas, extracción de información, etc. (Mateos & Ruiz, 2013)

Entre las actividades en las que se puede aplicar del PLN son: (Mateos & Ruiz, 2013)

- Entendimiento del lenguaje
- Recopilación de la información
- Obtención de la información
- Búsqueda de datos
- Traducción
- Reconocimiento del habla
- Entre otros.

2.3. Extracción de información

La Extracción de información, es una disciplina que se encuentra dentro del área del Procesamiento de Lenguaje Natural. En esta disciplina se encuentran las tecnologías especializadas en la extracción de información, las cuales tienen como propósito recabar información útil para el usuario, a partir de datos no estructurados. (Sáez Guerrero, 2009)

Los datos no estructurados, es aquella información que es difícil de ser interpretada por los ordenadores, es decir, carecen de significado, algunos ejemplos de datos no estructurados son las imágenes, texto, video, audio, etc., con ayuda de las técnicas de extracción de información, estos datos no estructurados pasan a ser estructurados o semiestructurados, esto quiere decir que se le añade significado a los datos extraídos. (Sáez Guerrero, 2009)

Algunos de las técnicas de extracción de información son el reconocimiento de patrones (patrones léxicos, patrones sintácticos, patrones semánticos, patrones de discurso), aprendizaje supervisado (arboles de decisión, redes neuronales, modelos de markov, k-vecinos), aprendizaje no supervisado (clustering, auto-entrenamiento, auto-entrenamiento paralelo, aprendizaje competitivo). (Sáez Guerrero, 2009)

2.4. Similitud coseno

Es una medida de la similitud entre dos vectores en un espacio vectorial que posee un producto interior con el que se evalúa el valor del coseno con el ángulo comprendido entre ambos productos. Esta función proporciona un valor igual a 1 si ambos vectores apuntan a un mismo lugar. Cualquier ángulo existente entre los vectores, el coseno arrojaría un valor inferior a uno. (Computacion y Sistemas, Na)

Esta distancia se emplea frecuentemente en la búsqueda y recuperación de información los documento en un espacio vectorial. Se aplica la similitud coseno al análisis de textos con el objetivo de establecer una métrica de semejanza entre estos. (Computacion y Sistemas, Na)

2.5. K-vecinos cercanos

Es uno de los algoritmos más analizados en el aprendizaje automático. Este algoritmo permite aproximar funciones de salida para los calores discretos o continuos. Toma las instancias correspondientes a un punto en un espacio métrico n-dimensional. (Bedoya Puerta, 2011)

Usualmente se utiliza la distancia euclidiana; sin embargo, por ser un algoritmo basado en distancias, es posible utilizar cualquier otra distancia: la distancia de Manhattan, la distancia de Chebychev, entre otras. (Bedoya Puerta, 2011)

Este algoritmo se basa en que el nuevo elemento o elementos se pueden asignar a alguna clase tomando como base los elementos previamente asignados a esa clase. Este algoritmo se realiza un agrupamiento de elementos muy bueno debido a que se calcula la similitud de un elemento con otro elemento y no con las características de una clase en particular. (Cover & Hart, 1967)

2.6. Support vector machine

Las máquinas de vector soporte o máquinas de soporte vectorial, es una técnica de clasificación de patrones, e inicialmente se diseñó para realizar una clasificación bi-clase, con la finalidad de proyectar los datos en espacios con más dimensiones, de este modo se conseguiría una mayor distancia entre las posibles clases y así lograr un mejor resultado a la hora de la clasificación. El proceso que se lleva a cabo está formado por dos etapas: entrenamiento y decisión. (Pajares Martinsanz & de la Cruz García, 2008)

El entrenamiento tiene como objetivo encontrar una función de decisión capaz de separar las clases a clasificar, si estas clases no son separables por medio de esos vectores de entrenamiento, se cambia a una dimensión superior mediante funciones de transformación para lograr la separación.

Esta etapa se base en el análisis de un grupo X de n muestras que dará como resultado dos valores simbólicos $y \in \{+1, -1\}$, donde el conjunto o grupo de entrenamiento esta conformado por $(x_i, y_i), i = 1, \dots, n$, donde x_i es un vector de entrenamiento y los valores de $y \in \{+1, -1\}$ es la clase a la cual pertenece ese vector. (Pajares Martinsanz & de la Cruz García, 2008)

La decisión se realiza cuando se anexa un nuevo elemento x y se debe determinar a que clase pertenece de acuerdo a la prioridad de

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i H(x_i, x) - b \qquad \text{(Formula 1)}$$

2.7. Clustering

Para poder definir que es clustering es necesario definir primero que un cluster es la recopilación de objetos que cumplen cierta similitud entre ellos. Por tanto clustering es la división o clasificación de datos entre objetos similares o una técnica estadística para la generación estructural de categorías entre documentos, agrupándolos de acuerdo a esa clasificación dando como resultado un alto grado de asociación entre dichos documentos que se encuentren en el mismo grupo. (clustering jpmonge, 2012)

El análisis clustering radica en la segmentación de la información en grupos de objetos con cierto grado de similitud, para medir el grado de similitud que tiene la información con el grupo de objetos, es necesario utilizar algún algoritmo de distancia semántica, como son: distancia euclidiana, de Manhatan, vecinos cercanos, entre otros. (Garre, Cuadrado, & Sicilia, Na)

Clustering es una de técnicas de las máquinas de aprendizaje (Machine Learning), donde el aprendizaje que se realiza es no supervisado, esto hace referencia a que el modelado que se lleva a cabo con relación a conjunto de entradas del sistema, sin tener en cuenta su clasificación, y su finalidad es que el sistema pueda realizar el reconocimiento de patrones para las nuevas entradas. (Sancho Caparrini, 2015)

2.8. Repositorio

Es una plataforma digital centralizada, la cual almacena, mantiene y preserva la información científica, tecnológica y de innovación, el cual es conformado por investigaciones, productos académicos y desarrollos tecnológicos.
(CONACYT, 2016)

2.9. Repositorio nacional

Es una plataforma digital centralizada de acceso abierto cuya coordinación y modelos de operación serán emitidos por el CONACYT, el cual, siguiendo estándares internacionales, almacena, mantiene y preserva esa información. Esta información es generada a través de las investigaciones, productos académicos y educativos. (CONACYT, 2016)

2.10. Repositorio institucional

Es la plataforma digital de las instituciones de los diferentes sectores: social, privado, gubernamental; el cual contiene su información académica, científica, tecnológica y de innovación, la cual deberá ser vinculada con el repositorio nacional siguiendo los estándares marcados por éste. (CONACYT, 2016)

2.11. Grobid

(Lopez, 2016)
Es una librería de las máquinas de aprendizaje, que se utiliza para la extracción, restructuración, y análisis sintáctico de documentos que se encentren en formato PDF (no escaneados), se enfoca principalmente en publicaciones de carácter científico, específicamente en artículos.

Fue creada en el año 2008 y hasta el 2011 se volvió *open source*, algunas de las funcionalidades con las que cuenta esta herramienta para los documentos científicos son:

- Análisis y extracción de cabeceras en artículos científicos.
- Análisis y extracción de referencias en artículos científicos.
- Extracción de información en las publicaciones de patentes.
- Análisis de los nombres (títulos, apellidos, nombres, etc.), en específico de los que se encuentran dentro de las cabeceras y referencias.
- Análisis de las afiliaciones y direcciones.
- Análisis de la fecha, esto quiere decir que se encuentren en formato reglamentario conforme a la norma ISO.

- Cuenta con 55 etiquetas utilizadas para la conformación de la estructura del XML que genera.

2.12. PDFMiner

(Python Software Foundation, 2014)

Es una herramienta para la extracción de información en documentos que se encuentren en formato PDF (no documentos escaneados o imágenes). Esta herramienta a diferencia de otras que se enfocan en solo en el análisis y la extracción de datos de los documentos PDF, también permite la extracción del número de línea, tipo de fuente y la ubicación exacta de los datos, además contiene un convertidos de documentos, el cual permite cambiar el documento en formato PDF a un formato XML, HTML o TXT.

2.13. RAKE

Por sus siglas Rapid Automatic Keyword Extraction algorithm, es un algoritmo de extracción automática de palabras clave de documentos, dichas palabras clave son secuencias de una o n palabras, las cuales en conjunto representan el contenido del documento.

El módulo de Python está desarrollado tomando como base dicho algoritmo, por tal motivo al igual que el algoritmo, los resultados que arroja dependen mucho del idioma y la estructura del contenido del documento.

2.14. Imagick

Es una extensión nativa de php para crear y modificar imágenes utilizando la API ImageMagick. (PHP, NA)

2.15. ImageMagick

Es software para crear y editar imágenes de mapa de bits, también esta puede convertir y escribir imágenes en una variedad de formatos, como JPG, PNG, GIF, etc. (PHP, NA)

Capitulo III. Estado del arte

En este capítulo se muestran algunos trabajos similares o relacionados a la extracción de metadatos en documentos PDFs.

3. Estado del arte

A continuación, se presenta el estado del arte que tiene relación con el tema que se desarrolla en este trabajo.

3.1. Extracción de información con algoritmos de clasificación.

(Téllez Valero, 2007)
INAOE: Instituto Nacional de Astrofísica, Óptica y Electrónica.
La problemática es que existe una cantidad excesiva de documentos técnicos escritos en lenguaje natural, los cuales están disponibles en formato electrónico, por tanto, hace imposible el análisis completo de toda esta información. La construcción de sistemas capases de realizar el análisis y extracción de información específica de grandes cantidades de datos es una tarea muy costosa en tiempo y recursos.

El aprendizaje automático es la disciplina que estudia cómo se construyen los sistemas computacionales que logren mejorar los resultados arrojados a partir de conocimiento o experiencia recabada.
El autor propone la clasificación como la tarea de aproximar una función objeto por medio de una función de clasificación, esta clasificación es un conjunto de categorías definidas; para poder definir a que categoría pertenece es necesario un proceso previo de aprendizaje, por medio del cual se obtienen los atributos que debe contener la nueva instancia o documento, este aprendizaje se da por medio de la observación de los atributos del conjunto de instancias que contiene cada categoría.

La efectividad del clasificador se puede observar al momento de verificar que tan certera fue la clasificación de las instancias y de este modo obtener un grado de confiabilidad, la forma tradicional de ver la exactitud del clasificador es observar la exactitud y en número de decisiones correctas que éste tomó.
Los algoritmos de clasificación utilizados son:
Navie Bayes: este es considerado el padre de los clasificadores probabilísticos, se basa en suposiciones a partir de cantidades de datos, las decisiones óptimas pueden ser tomadas por medio de del razonamiento, las ecuaciones establecidas y la observación de los datos previamente clasificados.

C4.5: Es una extensión del algoritmo ID3, el cual forma parte de los algoritmos de clasificación por medio de árboles de decisión. En este tipo de clasificación los nodos internos son los que contienen las etiquetas o tributos, las ramas son las posibles pruebas para los valores posibles a asignar a los atributos y las hojas representan las categorías finales en las cuales entraría la instancia.
K-vecinos más cercanos: método de aprendizaje basado en instancias más básicas y con resultados más aceptables para el análisis de textos, no contiene fases de entrenamiento.

Similitud:
Debido a la cantidad excesiva de documentos técnicos que existen el autor propone un sistema de clasificación de los documentos, la cual se basa en categorías definidas. Al igual que en este artículo, se propone la creación de un sistema capaz de realizar la clasificación y detección automática de documentos de tipo tesis o artículo científico, con el fin de tener un mejor control y acceso a la información generada en las instituciones que conforman el TecNM.

3.2. A modular metadata extraction system for born-digital articles

(Tkaczyk, Bolikowski, Czeczko, & Rusek, 2012)
Se presenta un sistema para la extracción de metadatos de artículos escolares, este sistema está basado en un flujo de trabajo modular, por medio del cual se inspecciona el documento completo.

En ocasiones las bibliotecas digitales almacenan documentos que contienen información de poca confiabilidad, esto debido a que esa información puede estar incompleta, incorrecta o no contienen metadatos.
El objetivo principal es la extracción de la mayor cantidad de información posible, esta información debe incluir al menos el título, autor, asociaciones, resumen, bibliografía o referencias, entre otros.

El flujo de trabajo diseñado en este trabajo para la extracción de los metadatos es flexible y fácil de utilizar, se basa en algunas técnicas de aprendizaje automático, el cual toman como ayuda para la conformación del documento.

En la actualidad existen muchos documentos digitales lo cuales no requieren de una identificación individual de caracteres, a diferencia de los documentos escaneados. De igual modo existen muchos métodos para la extracción de información, a continuación, se mencionan algunos de estos métodos.

- Giuffrida: Extrae el contenido del documento por medio de PostScript files, utilizando la herramienta basada en postotext.
- Esposito: Sigue un proceso de extracción de metadatos de los documentos en formato natural PDF/PS (documentos no escaneados), utiliza la segmentación por medio de un método basado en kernel, para poder realizar la clasificación en una zona, es necesario además basarse en los resultados de las máquinas de aprendizaje.
- Marinai: Extrae las características de los documentos en formato PDF utilizando la paquetería JPedal, por medio de esta se realiza la segmentación, la cual se basa en reglas establecidas y clasificadores neuronales para la asignación de las zonas de clasificación.

El flujo de trabajo para la extracción de los metadatos de un documento que se encuentra en formato PDF de forma natural, obtiene como resultado el almacenamiento de todos los metadatos extraídos. Esta extracción consiste en la creación de un árbol de decisión, el cual incluye diferentes características de los documentos; posteriormente se realiza un análisis y el respectivo mejoramiento de algunos datos, todo en base a los datos del árbol previamente creado; el último proceso que se realiza es la clasificación y extracción en las zonas correspondientes.

El objetivo de la extracción de características es la obtener de las características individuales de los documentos PDF, así como la posición de cada uno de ellos (ubicación en el documento). Utiliza la librería open-source iText, con esta librería el

texto que es extraído se divide en caracteres individuales y se almacena la posición que este tenía en la página.

El objetivo de la segmentación de las páginas es la creación de una estructura de árbol. Una vez creada la segmentación, el documento es representado como una lista, esta contiene las zonas, cada zona contiene las líneas del texto, estas líneas contienen las palabras que la conforman y las palabras contienen los caracteres individuales de cada una de ellas. Esta segmentación en el árbol se utiliza para la ubicación en la zona de clasificación.

El objetivo de la zona de clasificación es identificar el rol que juega cada sección del documento, algunas de las etiquetas utilizadas para la clasificación son: resumen, afiliación, titulo autor bibliografía, palabras clave, número de páginas, entre otros.

Algunos de los algoritmos de clasificación que se mencionas son:

- Las máquinas de aprendizaje incluyen los modelos de Markov (HMM) y las máquinas de soporte vectorial (SVM).
- Wang usa las HMM y los arboles de decisión para la ubicación en las zonas de clasificación.
- Han usa para realizar la clasificación de los textos, los clasificadores de soporte vectorial (SVM).
- El algoritmo de Viterbi es utilizado para calcular la ubicación más probable de las etiquetas mencionadas anteriormente.
- Se utiliza naïve para la prevención de los errores.

Similitud:
El objetivo de este artículo es la extracción de la información de artículos, esta extracción se realiza con base a ciertas características que debe contener un artículo, algunas de ellas son título, autor, asociaciones, resumen, bibliografía o referencias. En el trabajo que se pretende realizar, existe un módulo en el cual es necesaria la extracción especifica de atributos, esta extracción dependerá del tipo de documento ingresado, esto debido a que se manejaran diferentes tipos de documentos (tesis,

artículos científicos) los cuales no cuentan con las mismas características, por tanto, es necesario que previo a la extracción de la información se realice la detección del tipo de documento.

3.3. Una ontología basada en un sistema de clasificación de documentos totalmente automático utilizando un sistema semiautomático existente

(Manjula, 2013)

La cantidad de información disponible tanto en formato impreso como en electrónico se ha incrementado radicalmente en los últimos años, esto conlleva también a la ambigüedad del vocabulario utilizado en esos documentos, por tanto, aumenta la inexactitud a la hora de realizar la clasificación.

La clasificación automática de documentos (ATC), está dedicada a la investigación acerca de la categorización automática de documentos digitales en clases predefinidas. Los algoritmos de clasificación más frecuentemente utilizados se basan principalmente en las máquinas de soporte vectorial, métodos probabilísticos, algoritmos genéticos, métodos de aprendizaje a distancia, modelos ocultos de Markov, arboles de decisión, métodos de regresión, redes neuronales, algunas reglas de decisión y funciones TF-IDF. Para este trabajo se decidió mejorar un sistema existente de clasificación basándose en las funciones TF-IDF.

Como ya se mencionó anteriormente, la existencia de la ambigüedad en el lenguaje natural es un gran conflicto a la hora de realizar la clasificación de los documentos, y este ha sido uno de los temas de discusión en el área del procesamiento de lenguaje natural.

La universidad de Princeton comenzó el proyecto WordNet, en el cual se desarrollaba un recurso léxico u ontología para el idioma inglés, gracias a este se evitaría la ambigüedad en el lenguaje a la hora de la clasificación. Las antiguas ontologías se basan en Library of Congress Classification y en los esquemas de clasificación Dewey

Decimal Classification (DDC), el mayor problema es que estas no eran adaptables a la creación de una clasificación más compleja.

La metodología utilizada se basa en 3 fases de clasificación.

1. Fase 1
    a. Detección y eliminación de palabras vacías. Se detectan y limitan los términos de indicación a los términos fundamentales del tema y se eliminan las palabras vacías.
    b. Colección de entrenamiento, se almacenan los documentos de entrenamiento, los cuales serán utilizados para realizar la comparación con el documento ingresado.
    c. Algoritmo de clasificación de textos, esto determinan que tanto se refiere un documento a una área o materia específica, para esta clasificación se utilizó el algoritmo clasificador con la función de frecuencia de pesos → TF-IDF.
2. Fase 2
    a. Ontología, ayuda a la eliminación de la ambigüedad en los textos; se construyó una ontología utilizando el esquema DDC y enriquecida utilizando listas Sears. Fue limitada al área de la filosofía y más materias.
3. Fase 3
    a. Filtrado final, se filtran todos los posibles resultados obtenidos en las fases anteriores. Se comprarán los términos obtenidos en la fase anterior con los del documento muestra original y el que obtenga mayor puntuación es el que mayor semejanza tiene.

Las herramientas utilizadas para la implementación de este sistema, en la primera etapa API Lucene, el cual se utiliza para elegir la materia o documento más adecuado para entrada; para la segunda etapa en la creación de la ontología se utilizó la ontología OWL pre-construida, por medio del editor ontologías Protégé y se utilizó la herramienta API Protégé-OWL para recuperar la información de dicha ontología; para

la última fase se utilizó nuevamente la API Lucene para la obtención de la información más relevante.

Los resultados que se obtuvieron fueron de dos maneras. En la primera, se comparó la precisión de la clasificación de los métodos manuales, semiautomático y completamente automático; y en el segundo, se midió la relación entre la imprecisión de los documentos y la inexactitud de la clasificación. Con base a estos resultados se evaluó que tan preciso son los resultados arrojados por el sistema.

Similitud:
Para poder realizar la clasificación de los documentos técnicos (artículos y tesis), es necesario la realización de diversos filtros, clasificaciones y detección de atributos, al igual que en este artículo; la detección de las características antes mencionadas y la obtención de las colecciones de entrenamiento para la realización de las comparaciones entre dos documentos es necesaria para la realización de las posteriores clasificaciones.

3.4. Categorización automática de documentos

(Yasotha & Charles, 2015)

Durante las últimas 2 décadas el número de documentos digitales ha aumentado considerablemente, por lo tanto, es necesaria la categorización de estos documentos dentro de áreas y sub-áreas para su mejor recuperación y acceso. Para realizar esta clasificación existen 2 enfoques, el primero es basado en reglas y el segundo es basado en máquinas de aprendizaje.

Existe una gran variedad de algoritmos de aprendizaje, los cuales se enfocan o aplican a la categorización de textos, algunos de estos algoritmos son: clasificador bayesiano, arboles de decisión, modelos de regresión, método rocchio, redes neuronales, máquinas de soporte vectorial, análisis semántico latente (LSA), “máximum entropy modelling”, entre otras.

La categorización estadística utiliza métodos de aprendizaje para aprender las reglas de clasificación de manera automática, tomando en cuenta las etiquetas con las que cuentan los documentos. El modelo LDA se ha convertido en uno de los modelos probabilísticos más populares, en este modelo el documento muestra varios temas, los cuales se observan en el TD topic-per-document, en esta cada palabra de cada documento es asignada a una TD. El objetivo es descubrir automáticamente el tema del documento por medio de un aprendizaje automático, esto con ayuda de las etiquetas con las que cuenta el documento a clasificar.

Clasificación de documentos: un artículo tiene en su estructura muchas características, en general estas características son: título, subtitulo, texto, imágenes, graficas, mapas, diagramas, encabezados y ecuaciones, estas características son incluidas por los autores dependiendo de tipo de artículo, con el objetivo de proveer información más clara del tema del artículo. Por otro lado, los autores también prestan mucha atención a los estilos que tiene cada parte del documento, es decir, el color, el tamaño y el estilo de la fuente que se usa en cada sección del documento; esto debido a que los estilos en las diferentes secciones pueden ayudar al lector a tener un mejor entendimiento y a detectar mejor la información del documento.

A continuación, se muestra la vista general del proceso de clasificado de los documentos. (Diagrama2)

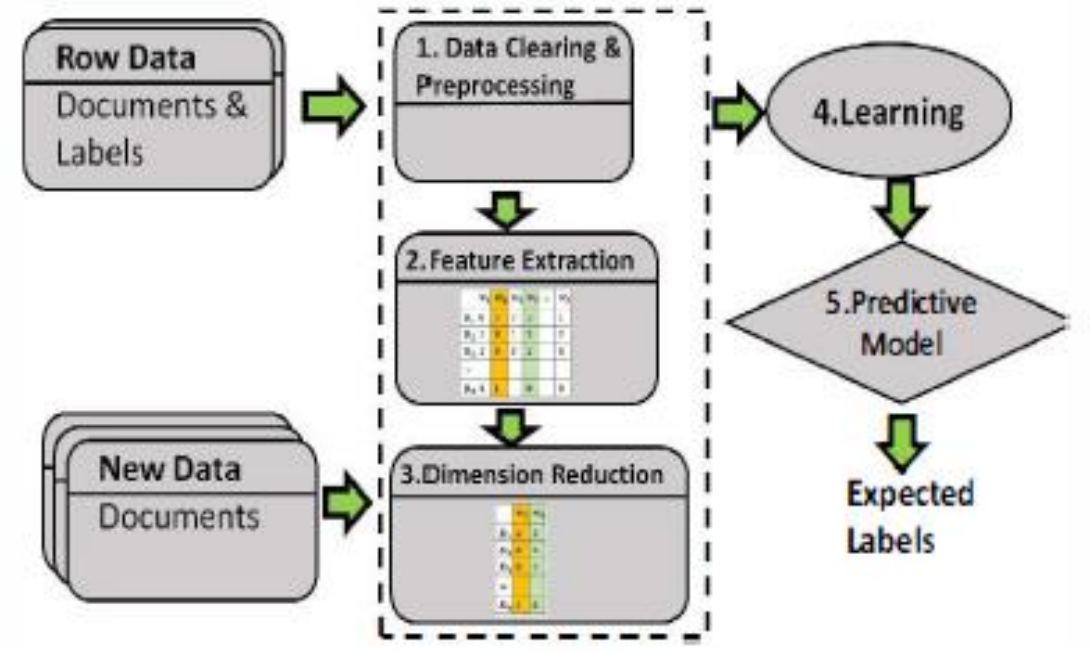


Diagrama 2 Categorización automática

A continuación, se explica un poco mejor parte del proceso mostrado en el diagrama anterior. Limpiado y procesado de la información: Las características de los documentos son extraídas de acuerdo a una lista de estructura de datos.

1. Codifica la información de a formato utf-8.
2. Unificar el texto a minúsculas.
3. Dividir las palabras, remplazar “-“,”_” por espacios.
4. Eliminar puntuación.
5. Realizar toquenización.
6. Eliminar palabras repetidas.

Extracción de características:

Para reducir el tamaño de las características (información extraída) es necesario realizar un filtro y una lematización. El filtro es necesario para la eliminación de ciertas palabras, por medio de la stoplist, esta es una lista de palabras que contiene conjunciones, y preposiciones, las cuales no contienen información importante o solo

es utilizada para unir oraciones. La lematización es la agrupación de las palabras que tienen significados similares (sinónimos), para ello se utiliza la WordNet el cual contiene una base de datos que contiene esa información.

1. Seleccionar un documento.
2. Eliminación de stopword o stoplist.
3. Stemword: cambiar a minúsculas.
4. Agrupación de palabras en orden alfabético.
5. Construcción del vector de características de acuerdo al total de palabras.

Similitud:

El trabajo que se pretende desarrollar es realizar la clasificación de documentos técnicos semiestructurados de tipo artículo de revista y congreso, tesis de licenciatura, maestría y doctorado; al igual que en este trabajo se deben detectar los atributos específicos para cada tipo de documentos, una vez obtenidos se debe realizar la eliminación de datos que podrían causar ruido a la hora de la realizar el clasificado.

3.5. PDF article metadata harvester

(Abdillah, 2013)

En la actualidad la difusión de los artículos y documentos de carácter científico que se encuentran en el internet se da por medio de revistas científicas, las cueles se encuentran en formato PDF. Los documentos que usualmente se pueden encontrar dentro de las revistas científicas son: artículos, los reportes técnicos, manuales de algún programa, notas de laboratorio.

Los artículos que se encuentran en las revistas científicas por lo general contienen caracteres, diagramas, ecuaciones, formulas, graficas, imágenes, tablas de contenido y muchos otros atributos, por tanto, las revistas las revistas definen estructuras claras para cada elemento, esto con el objetivo de que la información que se plasme sea entendible para los lectores.

El termino metadato es utilizado para hacer referencia a los datos o información contenida en los documentos, el registro de los metadatos consiste en la representación de los atributos en cierto número de elementos, dichos elementos pueden contener uno o más valores.

Dublin core es un estándar internacional e interdisciplinario para metadatos, este estándar ha sido tomado por varios institutos, para que la información y el entorno sea homogéneo en los documentos. Cuenta con 15 elementos estándar, los cueles deben estar dentro del documento, esto con el objetivo de facilitar la recuperación de la información de dichos documentos. Esos elementos son:

- Título
- Autor
- Área
- Descripción
- Revista
- Contribuidores
- Fecha
- Tipo
- Formato
- Identificador
- Materia
- Idioma
- Relación
- Cobertura
- Referencia

Dada esa información el autor determino que la estructura de los artículos científicos es la siguiente: título, autor, resumen, palabras clave (key words), metadatos (el contenido del documento) y las referencias. Con base a esa información el autor considera únicamente para la extracción tres elementos, el título, autor, y año, esto debido a que esos elementos son esenciales a la hora de realizar la búsqueda de documentos; para realizar la extracción utilizó la herramienta XMP de adobe ya que esta le permitía extraer esos elementos con mayor grado de precisión.

Similitud:

En el trabajo que se pretende desarrollar se realizar la extracción de metadatos en documentos técnicos semiestructurados de tipo artículo de revista y congreso, tesis de licenciatura, maestría y doctorado; al igual que en este trabajo se van a extraer ciertos elementos necesarios de los artículos científicos que se encuentren en formato PDF.

3.6. Tabla comparativa de los trabajos relacionados

A continuación, en la tabla 1 se muestra la comparación de los trabajos mencionados en el estado del arte, trabajos relacionados.

Para realizar la presente comparación se tomaron en cuanta los siguientes puntos:

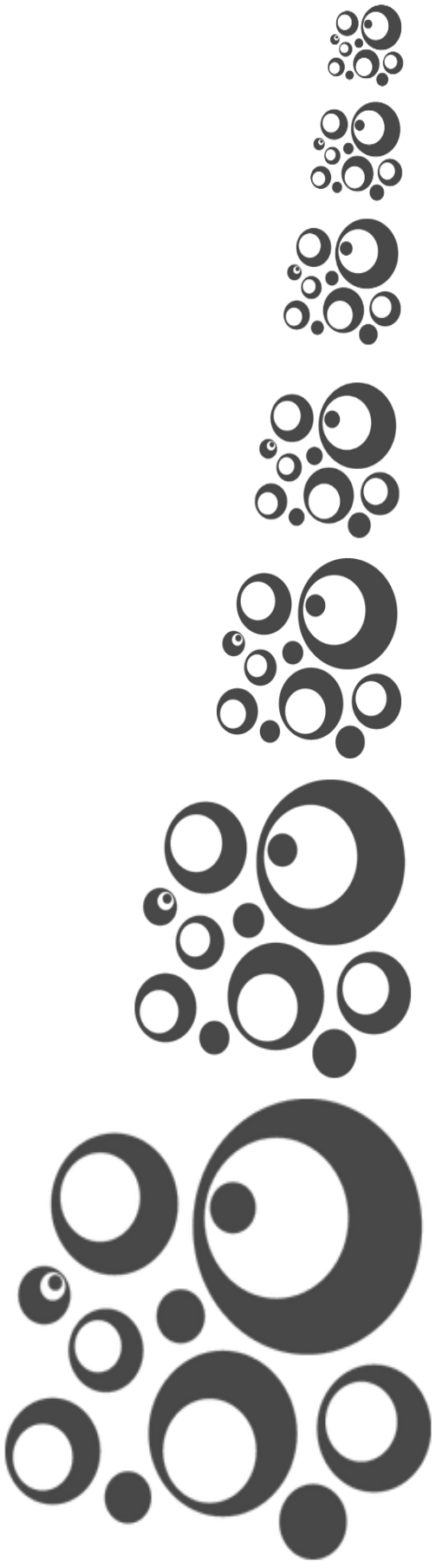
- Título del trabajo.
- Año.
- Tipo de documentos.
- Metodología de solución.
- Objetivo.

Tabla Comparativa				
Título	Año	Tipo Documento	Metodología	Objetivo
Extracción de información con algoritmos de clasificación	2007	Semi-Estructurados	Extracción de información de artículos científicos, análisis y procesamiento de la información, de acuerdo a temas	Extraer, procesar documentos de texto, de acuerdo a un tema.
A modular metadata extraction system for born-digital articles	2012	Estructurados	Extracción de información de artículos escolares, clasificación de por medio de Algoritmo Docstrum, SVM y simple rule-based.	Extraer la mayor cantidad de los atributos (título, autor, asociaciones, abstract, bibliografía o referencias) de los artículos escolares
Una ontología basada en un sistema de clasificación de documentos totalmente automático utilizando un sistema semiautomático existente	2013	Estructurados	*Detección y eliminación de palabras vacías, colección de entrenamiento, algoritmo de clasificación de textos *Uso de ontología eliminación de ambigüedad *Filtrado de información	Clasificar documentos mediante el uso de ontologías
Categorización automática de documentos	2015	Estructurados	Modelo LDA para la categorización estadística	Descubrir automáticamente el tema del documento para su clasificación
PDF article metadata harvester	2013	Estructurados	Extracción de los metadatos a partir de artículos científicos, por medio de otros algoritmos obtener el número de páginas y peso del archivo.	La extracción de los metadatos título, autor, año, peso del archivo y número de páginas.
Sistema semiautomático para clasificar, enlazar y almacenar documentos técnicos semiestructurados mediante técnicas de Procesamiento de Lenguaje Natural	2016	Semi-Estructurados	Detección de tipo de documento, extracción especifica de atributos, clasificación de acuerdo a un área de conocimiento y a un sector económico, inserción en repositorio institucional	Detectar el tipo de documento ingresado al sistema para la extracción específica de atributos para su posterior clasificación y almacenado en el repositorio institucional, para su posterior enlace con el repositorio nacional.

Tabla 1 Comparación trabajos relacionados.

Capitulo IV. Metodología

En este capítulo se muestra la metodología seguida para la detección, extracción e inserción de los metadatos de los PDFs en el repositorio institucional.



4. Metodología

A continuación, se presenta la metodología seguida para el desarrollo de este proyecto de tesis.

4.1. Diagrama general de la metodología

La metodología que se siguió para el desarrollo de este proyecto de tesis está dividida en 3 módulos principales, detección de tipo de tipo de documento, extracción de metadatos y la generación de imágenes (Diagrama3).

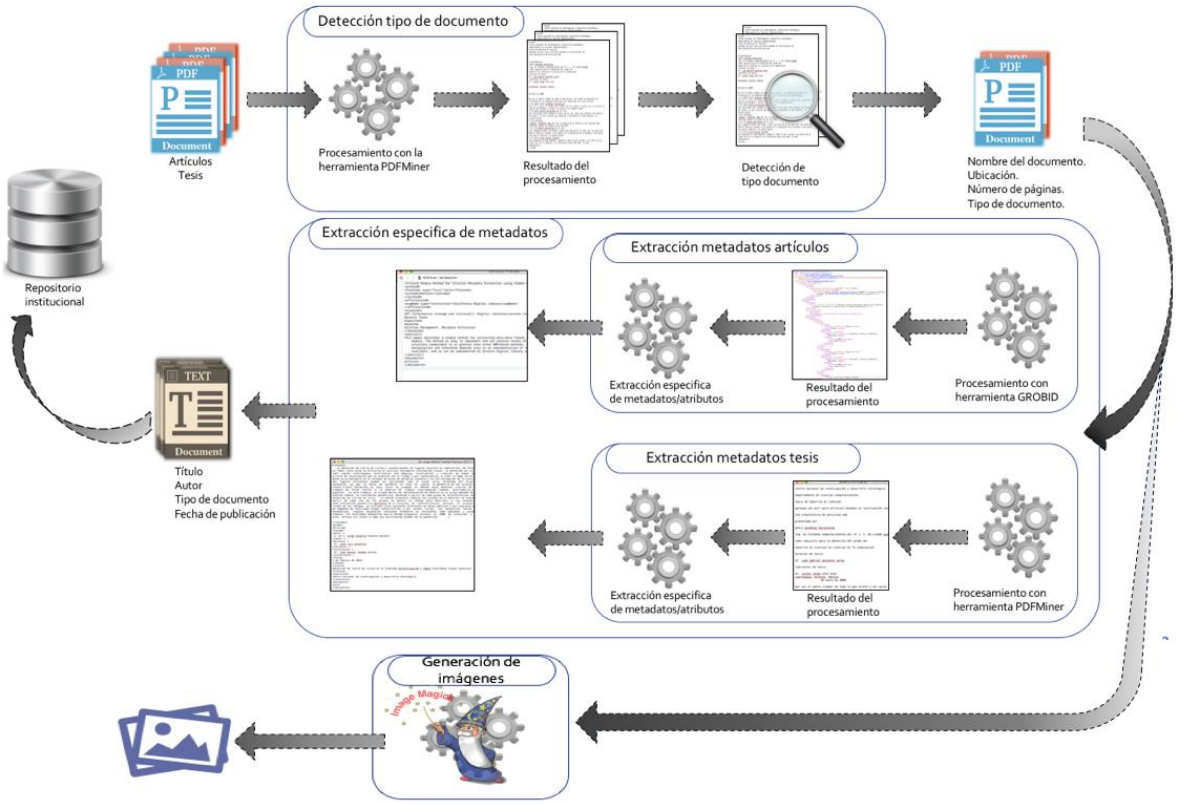


Diagrama 3 Metodología de solución.

La salida que se tendrá finalizado el proceso, son los documentos TXT con los atributos específicos de cada documento, así como las imágenes de las primeras páginas de cada uno de ellos, la información que se genera es almacenada posteriormente en el repositorio institucional.

A continuación se describen a detalle cada una de los módulos que conforman esta metodología.

4.2. Creación del corpus

Antes de poder iniciar con el proceso de la metodología, fue necesaria la creación de un corpus de documentos PDF.
La creación de dicho corpus se pretendía realizar mediante la descarga de tesis creadas en los diferentes tecnológicos que conforman el Tecnológico Nacional de México, sin embargo la creación de un corpus que contuviera tal variedad de tesis fue imposible debido a que desde hace muchos años los tecnológicos crearon la titulación por medio de reportes técnicos, por tal motivo los alumnos ya no tienen la obligación de escribir una tesis para obtener la titulación, lo que ha provocado que la creación de tesis bajara considerablemente.

Otro de las dificultades que surgieron durante la creación de dicho corpus, es la obtención de las tesis que han sido desarrolladas, debido a que los tecnológicos no cuentan con el documento en formato PDF o en caso de que tener el documento, este se encuentra escaneado tipo imagen y el sistema que se desarrolló necesita que los documentos cuenten con OCR para la extracción de información.

Después de una larga búsqueda, se obtuvieron un poco más de 600 documentos de tesis, a los cuales se limpiaron para finalmente dejar solo aquellos que hayan sido creados 10 años atrás.
Al final de la búsqueda y depuración de la información se conformó un corpus con 400 documentos.

4.3. Detección tipo de documento

El módulo “Detección tipo de documento” es la primera fase del sistema, en este se hace la detección del tipo de documento, es decir el sistema tiene por entrada un documento PDF el cual cuenta OCR, a tal documento se le procesaran las primeras dos páginas que lo conforman; solamente se toman las primeras dos páginas del documento en esta fase del sistema debido a que la información necesaria para identificar si es una tesis o un artículo se encuentra en esas páginas.

Los datos necesarios para la identificación del documento son, para el caso de las tesis, en la portada que usualmente es la primera si no la segunda página, se buscan las palabras: tesis, tesina, licenciatura, maestría y doctorado.
Para la identificación de los artículos se busca: resumen, introducción, palabras clave (key words) y resumen; adicionalmente se toma en cuenta el número de páginas que conforman los documentos ya que los artículos normalmente son de un número muy reducido de páginas a diferencia de las tesis, que son documentos de gran tamaño comparado con los artículos.

Los módulos se desarrollaron en lenguaje Python y por medio algunas herramientas como PDFMiner y Grobid se realizó el procesamiento de los documentos para realizar las diferentes fases del sistema.
En la Diagrama 4 se muestra el proceso para la detección del tipo de documento, el punto 1 es el documento PDF de entrada, cuando entra al módulo (punto 2) pasa por el sub-punto A que es el procesamiento con la herramienta PDFMiner.

En este sub-punto se obtiene el número total de páginas que conforman el documento, la ubicación en el sistema del archivo y se hace la extracción en formato TXT de las primeras dos páginas del documento, las cuales se almacenan para su análisis en la siguiente sección.

En sub-punto B se hace formalmente la detección del tipo de documento, tomando como archivo a analizar el generado por el PDFMiner, se buscan las palabras clave previamente mencionadas y se corrobora el tipo de documento verificando si el número de páginas total del documento concuerda con el tipo de documento detectado.

Finalmente, el punto 3 son los datos del documento que se tienen como salida del módulo, estos son:

- Tipo de documento.
- Nombre.
- Número de páginas.
- Ubicación

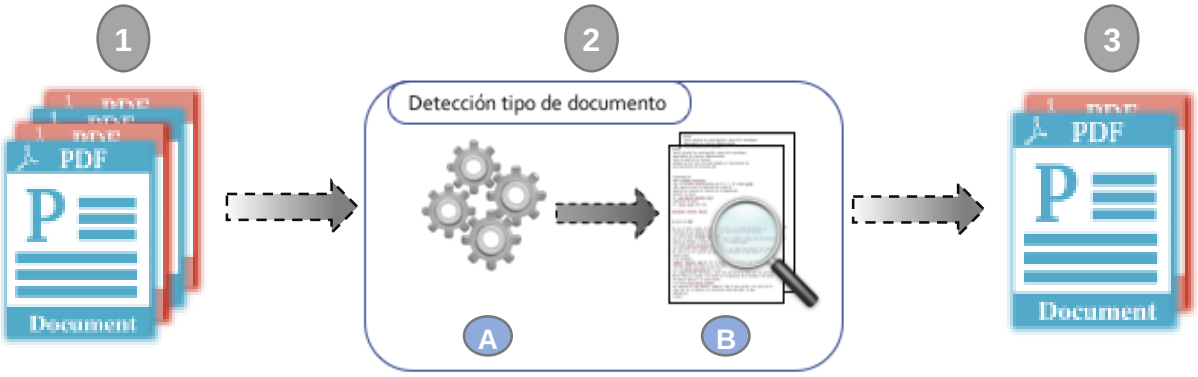


Diagrama 4 Fase detección tipo de documento.

4.4. Extracción específica de metadatos

El módulo “extracción específica de metadatos” es la segunda fase del sistema (Diagrama5), en esta fase se toma la información obtenida previamente (punto1) y dependiendo del tipo de documento que se haya detectado se realiza la extracción específica de los metadatos (punto 2), cuando termina la extracción de los metadatos específicos para cada tipo de documento ingresado al sistema, la información generada por cada documento debe ser almacenada en un archivo individual.

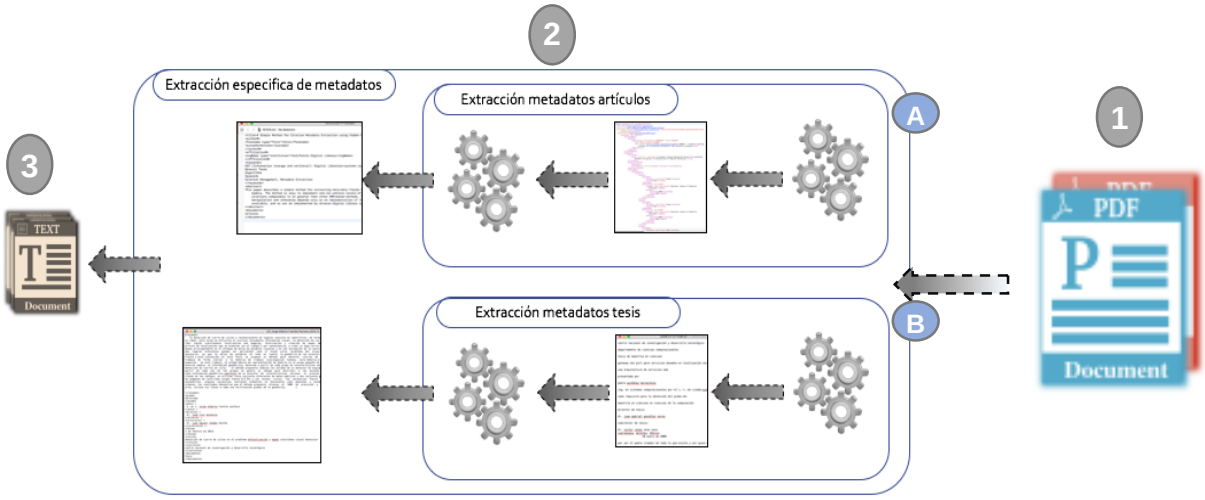


Diagrama 5 Fase Extracción específica de metadatos.

A continuación, se explica a más detalle el funcionamiento de este módulo dependiendo del tipo de documento detectado.

4.4.1. Extracción específica de metadatos en artículos científicos

Supongamos que el documento que se ingresó en el sistema se detectó que es un artículo, cuando pasa a la segunda fase, toma el camino del punto A (Diagrama5).

Una vez dentro (Diagrama6), el módulo recibe como entrada un documento en formato PDF (punto1), en el punto 2 se realiza todo proceso de extracción de metadatos, para ello es necesario procesar el documento mediante la herramienta GROBID (punto A). GROBID es una librería para el procesamiento de artículos científicos, la cual recibe por entrada un artículo en formato PDF, lo procesa y da como resultado en consola el etiquetado de las diferentes secciones del artículo. El resultado es almacenado en un archivo e formato XML (punto B).

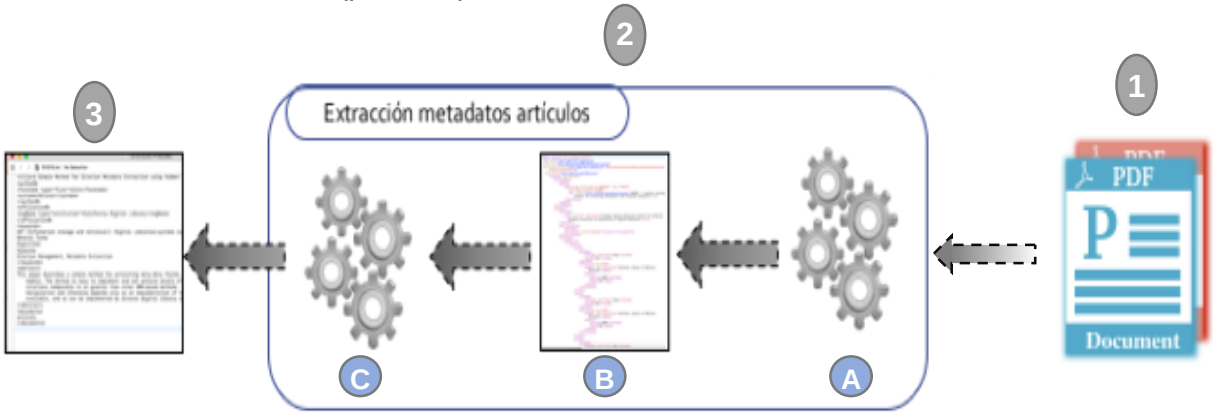


Diagrama 6 Extracción específica de metadatos en artículos.

En el punto C se realiza formalmente la detección y extracción de metadatos, mediante el uso de heurísticas se identifica la ubicación de los metadatos en el archivo XML previamente generado.

Una vez que se extrajeron los metadatos son almacenados en un archivo el cual contiene únicamente los metadatos seleccionados (punto 3); los metadatos que son extraídos en los artículos son:

- Título
- Autor
- Institución
- palabras clave
- Abstract / resumen
- Adicionalmente se almacena el tipo de documento.

4.4.2. Extracción específica de metadatos en tesis

Supongamos que el documento que se ingresó en el sistema se detectó que es una tesis, cuando pasa a la segunda fase, toma el camino del punto B (Diagrama5).

Una vez que el sistema ingresa a la fase de extracción específica de metadatos en

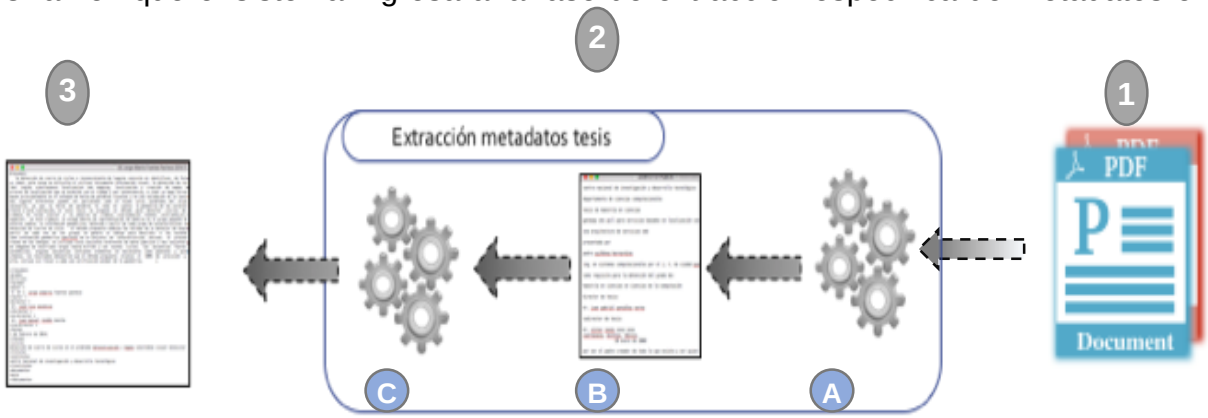


Diagrama 7 Extracción específica de metadatos en tesis.

tesis (Diagrama7), es necesario al igual que con los artículos tener el documento original en PDF y el número total de páginas que tiene (punto 1).

Para iniciar la extracción de los metadatos (punto 2), es necesario realizar el procesamiento del PDF (punto A) y convertirlo a un archivo TXT (punto B), esto se realiza con ayuda de la herramienta PDFMiner.

A diferencia de la primera etapa del sistema donde se procesan solo dos páginas del documento, en esta sección se procesa todo el documento, por este motivo es necesario saber el número total de páginas que lo conforman.

Cuando el documento PDF ha sido procesado, se procede a la extracción de los metadatos específicos de la tesis a partir del archivo TXT. Para poder realizar la extracción se desarrollaron heurísticas por medio de las cuales se identifican las diferentes secciones del documento y se descartan las secciones que no son relevantes (punto C), es decir, solo se almacenan los datos que se están buscando.

En el caso de las tesis no cuentan con palabras clave y debido a que uno de los requisitos del CONACYT es la extracción de las palabras clave, fue necesario utilizar el módulo RAKE de Python para realizar dicha obtención, para ello fue necesario identificar las secciones del texto que contuvieran la información más relevante del documento, en el caso de las tesis esas secciones son el resumen y las conclusiones.

Se toman únicamente esas secciones porque en el caso del resumen su objetivo es compactar en pocas líneas de que va a tratar el documento y la conclusión por su parte habla de lo que se realizó en el documento, si se cumplieron objetivo o completar un poco de información faltante en el resumen.
Una vez que se tiene identificadas dichas secciones, se envía la información al módulo RAKE y se generan las palabras clave.

Para finalizar el módulo, una vez que se extrajeron los metadatos y se generaron las palabras clave, se almacena la información en un archivo, el cual contiene únicamente los metadatos requeridos (punto 3).
Los metadatos que son extraídos en las tesis son:

- Autor.
- Director.
- Co-director.
- Área.
- Grado.
- Fecha.
- Título.
- Resumen.
- Palabras clave.
- Institución.
- Tipo de documento.

4.4.3. Generación Imágenes

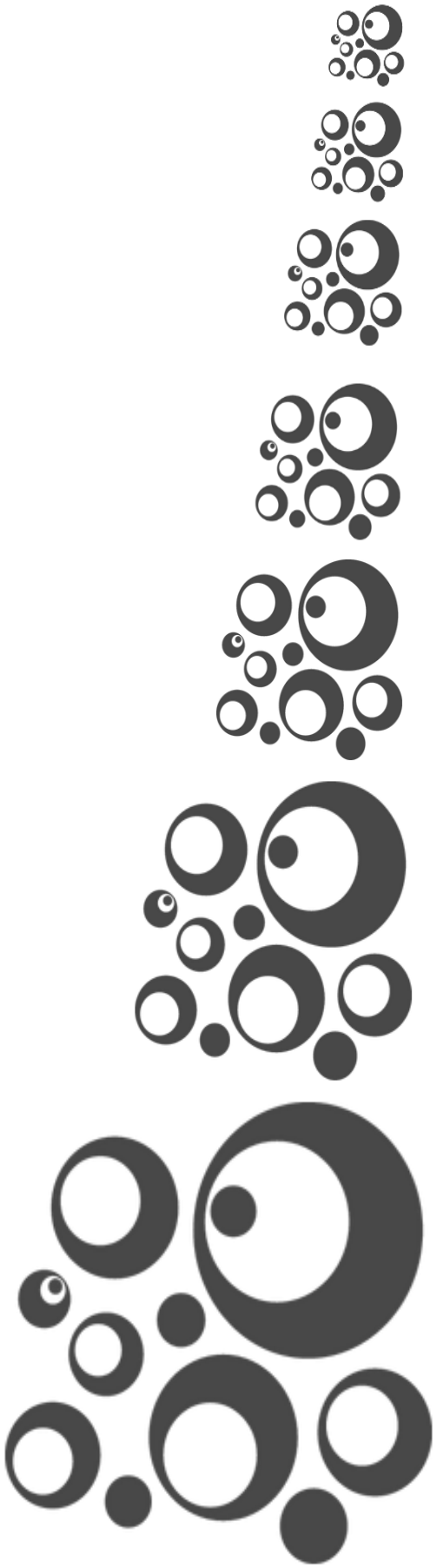
Debido a que este sistema es semiautomático, surge la necesidad de verificar que los datos obtenidos sean correctos, a causa de esto se creó el modulo para la generación de imágenes de las primeras páginas de los documentos.

Este módulo (Figura 5) tiene como estrada el documento PDF al cual se le generaran las imágenes, por medio de la herramienta Imagick se procesa el documento y se obtienen las imágenes en el formato solicitado ya sea PNG o JPG, para su posterior almacenado.

Diagrama 8 Generación de imágenes.

Capitulo V. Pruebas y resultados

En este capítulo se muestran las pruebas que se le realizaron al sistema con el objetivo de evaluar la precisión y cobertura con la cual se extrajeron los metadatos de los diferentes PDFs.



5. Pruebas y resultados

A continuación, se muestras las pruebas que se realizaron al sistema, así como los resultados obtenidos de ellas

5.1. Pruebas

Para la ejecución de la fase de pruebas se realizó una comparación de los resultados obtenidos por el sistema y los obtenidos de forma manual de los archivos PDF, Seanalizaron 300 archivos. En la tabla 2 se muestran ejemplos de los metadatos extraídos de forma manual de algunos de los documentos que se analizaron, en la tabla3 se muestran algunos ejemplos de los metadatos obtenido por el sistema desarrollado.

Autor	Director	Codirector	Area	Grado	Fecha	Título	Institución
Juan Carlos Rendón Miranda	Juan Gabriel González Serna	Azucena Montes Rendón	CIENCIAS COMPUTACIONALES	Maestría	01/07/2014	Clasificación Automática de Objetos de Conocimiento con Contenido No Estructurado Para el Poblado Semiautomático de Ontologías Multidimensionales	Centro Nacional de Investigación y Desarrollo Tecnológico
Julia Yazmín Arana Llanes	Juan Gabriel González Serna	-	CIENCIAS COMPUTACIONALES	Maestría	01/06/2014	Metodología para Evaluación de SRSC Centrada en el Usuario, Basada en Características de Efectividad, Confianza y Satisfacción Mediante Interfaces Multimodales sobre Dispositivos Móviles Multisensoriales	Centro Nacional de Investigación y Desarrollo Tecnológico
Félix Ortiz Carreón	Juan Gabriel González Serna	Azucena Montes Rendón	CIENCIAS COMPUTACIONALES	Maestría	01/06/2016	Metodología de Comunicación Aumentativa y Alternativa para Personas con Parálisis Cerebral Mediante Mecanismos Heterogéneos de Interacción Humano Computadora	Centro Nacional de Investigación y Desarrollo Tecnológico

Tabla 2 Ejemplo de metadatos extraídos manualmente de los documentos.

Autor	Director	Codirector	Area	Grado	Fecha	Título	Institución
juan carlos rendón miranda	juan gabriel gonzález serna	azucena montes rendón	ciencias computacionales	maestría	01/07/2014	clasificación automática de objetos de conocimiento con contenido no estructurado para el poblado semiautomático de ontologías multidimensionales	centro nacional de investigación y desarrollo tecnológico
julia yazmín arana llanes	juan gabriel gonzález serna	-	ciencias computacionales	maestría	01/06/2014	metodología para evaluación de srsc centrada en el usuario, basada en características de efectividad, confianza y satisfacción mediante interfaces multimodales sobre dispositivos móviles multisensoriales	centro nacional de investigación y desarrollo tecnológico
félix ortiz carreón	juan gabriel gonzález serna	azucena montes rendón	ciencias computacionales	maestría	01/06/2016	metodología de comunicación aumentativa y alternativa para personas con parálisis cerebral mediante mecanismos heterogéneos de interacción humano computadora	centro nacional de investigación y desarrollo tecnológico

Tabla 3 Ejemplo de metadatos de los documentos extraídos mediante el sistema.

Para realizar la evaluación de los resultados obtenidos por el sistema se tomaron los resultados obtenidos con la extracción manual y se compararon con los obtenidos, posteriormente se aplicaron las métricas de evaluación Precisión y Cobertura para verificar la exactitud de los resultados.

5.2. Resultados

En esta sección se muestran los resultados alcanzados de forma individual para cada uno metadatos extraídos los diferentes documentos.

Dichos resultados están divididos en los metadatos extraídos para las tesis y los extraídos para los artículos.

5.2.1. Resultados en tesis

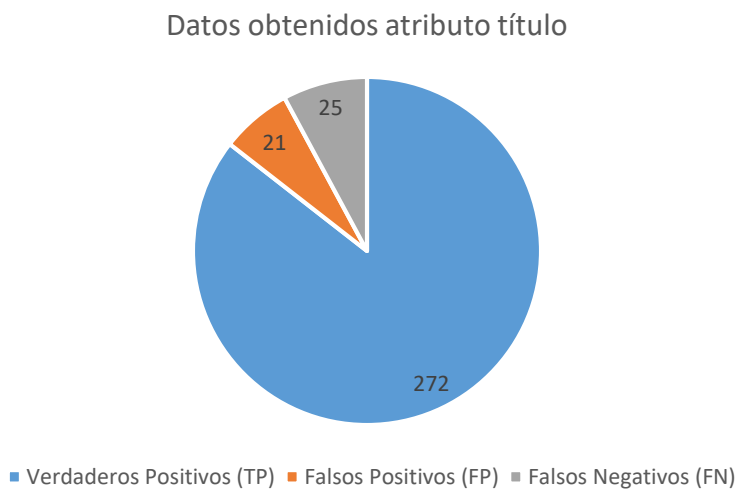
A continuación, se muestran los resultados obtenidos en la extracción de metadatos en los documentos de tipo tesis.

5.2.1.1. Resultados en autor

Se procesaron un total de 297 archivos de tipo tesis. Al revisar los diferentes archivos PDF se encontró que contenían un total de 272 datos en los autores.

Verdaderos Positivos (TP)	272
Falsos Positivos (FP)	21
Falsos Negativos (FN)	25

Tabla 4 Resultados obtenidos en la extracción de autores.



Gráfica de datos 1 Tesis autor.

De acuerdo a los resultados de autores mostrados en la tabla 4 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las siguientes formulas.

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)}$$ *(Formula 2)*

$$Precisión = \frac{TP}{(TP + FP)}$$ *(Formula 3)*

TP: Representa los datos verdaderos positivos, es decir, son aquellos datos generados por el sistema los cuales al ser comparados con los obtenidos de manera manual resultaron ser correctos del documento.

FP: Representa los datos falsos positivos, es decir, son aquellos datos generados por el sistema los cuales al ser comparados con los obtenidos de manera manual se identificaron como incorrectos.

FN: Representa los datos falsos negativos, es decir, son aquellos datos que el sistema no identifico de manera correcta como autor, es decir obtuvo otra cosa como dato o simplemente no identifico ese atributo.

Tomado las formulas anteriores y los resultados mostrados en la tabla 4 se obtuvieron los siguientes resultados (gráfica de resultados 1):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{272}{(272 + 25)} = \frac{272}{297} = 0.9158 * 100 = 91.58$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{272}{272 + 21} = \frac{272}{293} = 0.9283 * 100 = 92.83$$

Gráfica de resultados 1 Tesis Autor.

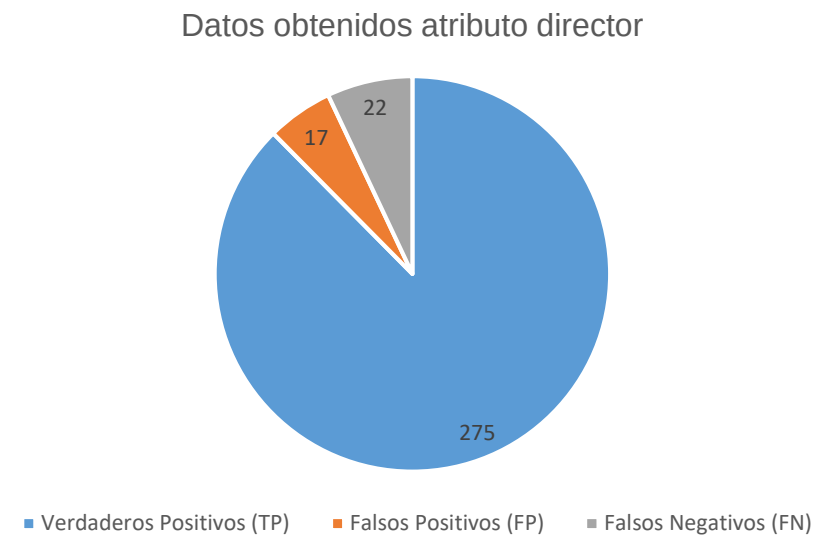
El motivo por el cual algunos de los autores no fueron extraídos de manera correcta se debe en algunos casos por la distribución del texto dentro del documento, o la tipografía utilizada no permitía que el sistema identificara la sección del documento que contenía el autor, adicionalmente algunos documentos no contaban con permisos de lectura y escritura lo cual dificultaba o impedía la extracción de información de los documentos.

5.2.1.2. Resultados en director

Se procesaron un total de 297 archivos de tipo tesis. Al revisar los diferentes archivos PDF se encontró que contenían un total de 275 datos en los directores.

Verdaderos Positivos (TP)	275
Falsos Positivos (FP)	17
Falsos Negativos (FN)	22

Tabla 5 Resultados obtenidos en la extracción de directores.

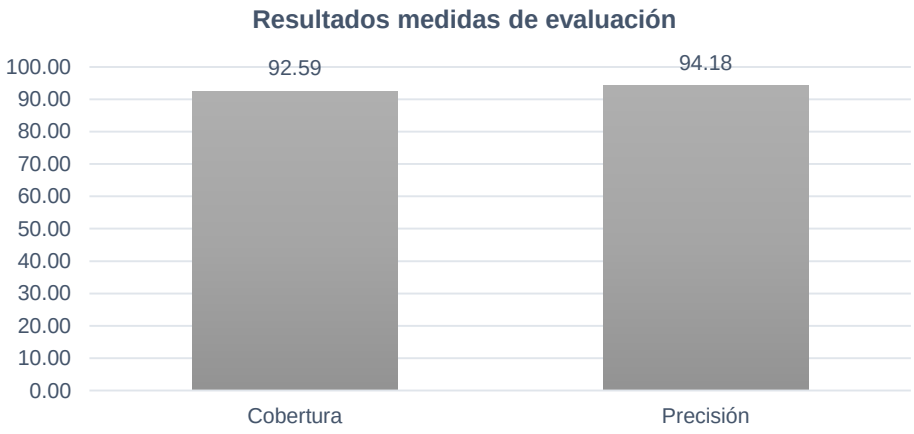


Gráfica de datos 2 Tesis director.

De acuerdo a los resultados obtenidos del atributo director de tesis mostrados en la tabla 5 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, dando como resultado lo siguiente (gráfica de resultados 2):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{275}{(275 + 22)} = \frac{275}{297} = 0.9259 * 100 = 92.59$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{275}{275 + 17} = \frac{275}{292} = 0.9417 * 100 = 94.18$$



Gráfica de resultados 2 Tesis Director.

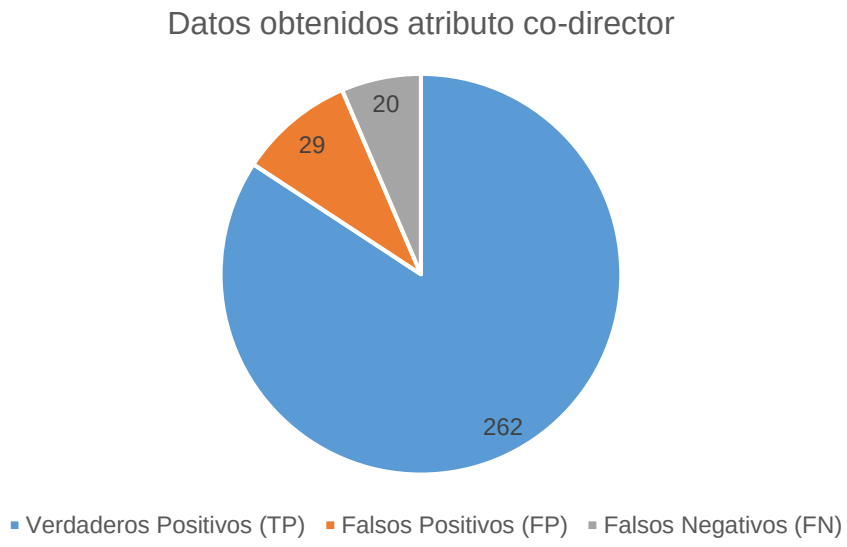
El motivo por el cual algunos de los directores no fueron extraídos de manera correcta o simplemente no se lograron extraer se debe en algunos casos la distribución del texto dentro del documento o los títulos de identificación no permitían identificar el director, adicionalmente algunos documentos no contaban con permisos de lectura y escritura lo cual dificultaba o impedía la extracción de información de los documentos.

5.2.1.3. Resultados en co-director

Se procesaron un total de 297 archivos de tipo. Al revisar los diferentes archivos PDF se encontró que contenían un total de 262 datos en los co-directores.

Verdaderos Positivos (TP)	262
Falsos Positivos (FP)	29
Falsos Negativos (FN)	20

Tabla 6 Resultados obtenidos en la extracción de Co-directores.

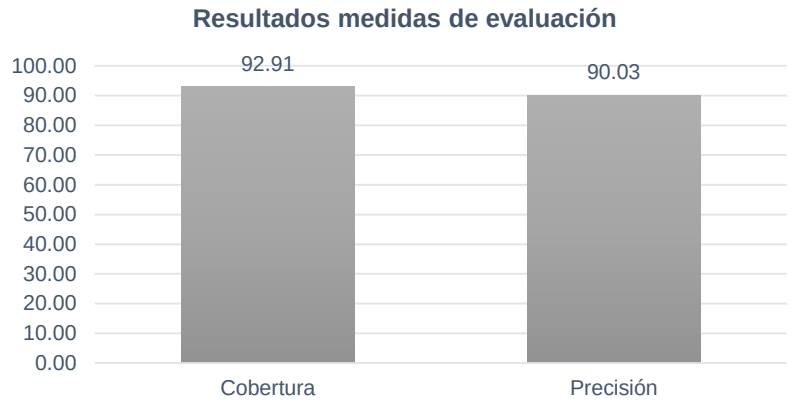


Gráfica de datos 3 Tesis co-director.

De acuerdo a los resultados obtenidos del atributo co-director de tesis mostrados en la tabla 6 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, dando como resultado lo siguiente (gráfica de resultados 3):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{262}{(262 + 20)} = \frac{262}{282} = 0.9290 * 100 = 92.9$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{262}{262 + 29} = \frac{262}{291} = 0.9003 * 100 = 90.03$$



Gráfica de resultados 3 Tesis Co-director.

El motivo por el cual algunos de los co-directores no fueron extraídos de manera correcta o simplemente no se lograron extraer se debe en algunos casos la distribución del texto dentro del documento o los títulos de identificación no permitían identificar el co-director, adicionalmente algunos documentos no contaban con permisos de lectura y escritura lo cual dificultaba o impedía la extracción de información de los documentos.

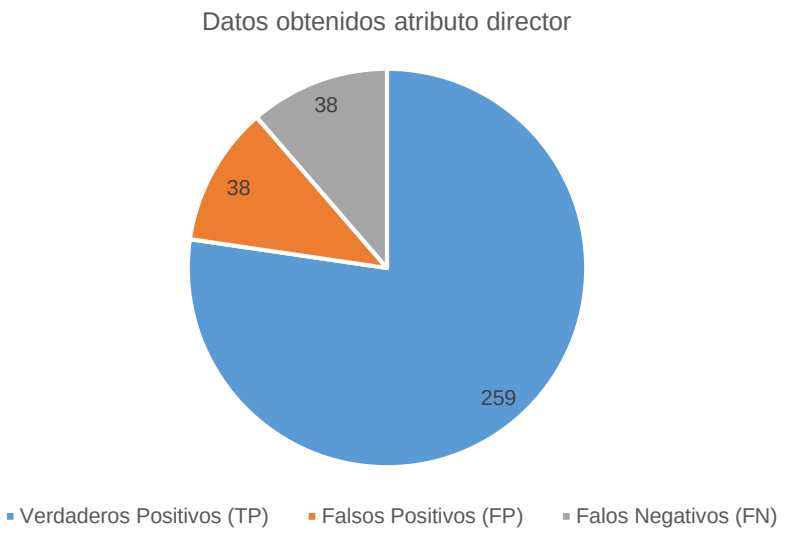
En otros casos el sistema identificaba igual al director como el co-director, debido a que algunos documentos contenían duplicada la portada lo cual provocaba algunos conflictos a la hora del análisis de los datos.

5.2.1.4. Resultados en área

Se procesaron un total de 297 archivos de tipo tesis. Al revisar los diferentes archivos PDF se encontró que contenían un total de 259 datos en el área.

Verdaderos Positivos (TP)	259
Falsos Positivos (FP)	38
Falos Negativos (FN)	38

Tabla 7 Resultados obtenidos en la extracción de áreas.

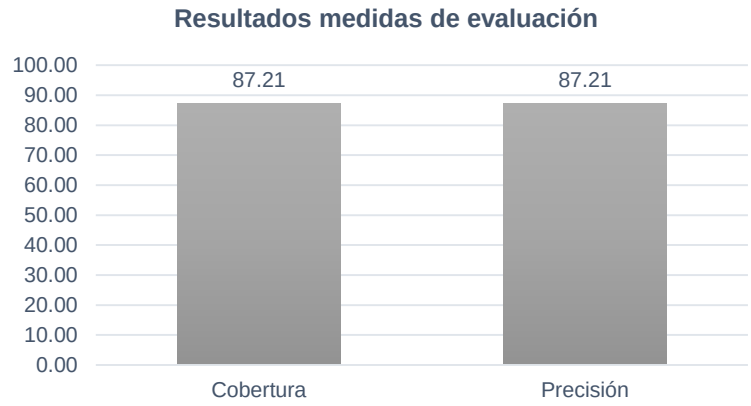


Gráfica de datos 4 Tesis área.

De acuerdo a los resultados del área mostrados en la tabla 7 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, dando como resultado lo siguiente (gráfica de resultados 4):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{259}{(259 + 38)} = \frac{259}{297} = 0.8720 * 100 = 87.21$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{259}{259 + 38} = \frac{259}{297} = 0.8720 * 100 = 87.21$$



Gráfica de resultados 4 Tesis área.

El motivo por el cual algunas de áreas no fueron extraídas de manera correcta se debe en algunos casos por la distribución del texto dentro del documento, o la tipografía utilizada no permitía que el sistema no identificara la sección del documento

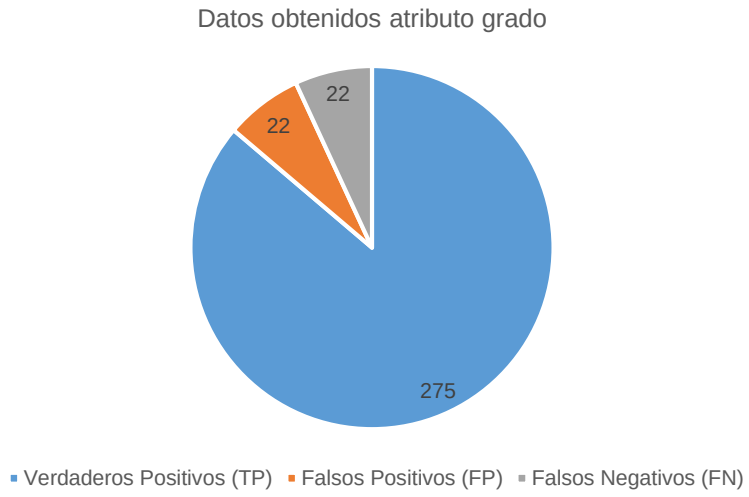
que contenía el área y en otros casos la sección que contenía el área o se concentraba en ninguna de las posibilidades, adicionalmente algunos documentos no contaban con permisos de lectura y escritura lo cual dificultaba o impedía la extracción de información de los documentos.

5.2.1.5. Resultados en grado

Se procesaron un total de 297 archivos de tipo tesis. Al revisar los diferentes archivos PDF se encontró que contenían un total de 275 datos en el grado.

Verdaderos Positivos (TP)	275
Falsos Positivos (FP)	22
Falsos Negativos (FN)	22

Tabla 8 Resultados obtenidos en la extracción de grados.

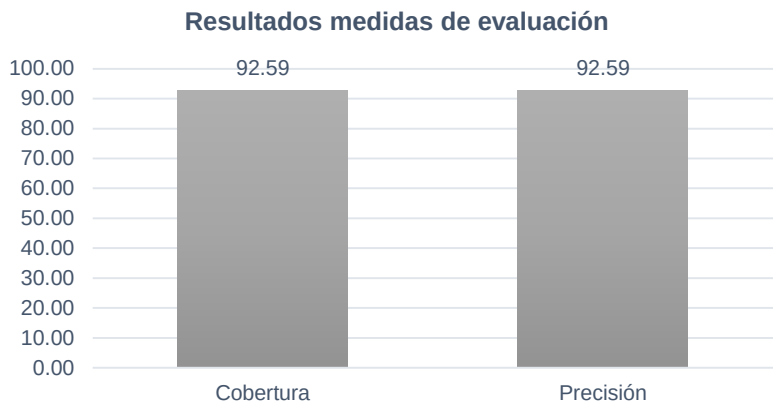


Gráfica de datos 5 Tesis grado académico.

De acuerdo a los resultados del grado académico mostrados en la tabla 8 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, dando como resultado lo siguiente (gráfica de resultados 5):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{275}{(275 + 22)} = \frac{275}{297} = 0.9259 * 100 = 92.59$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{275}{275 + 22} = \frac{275}{297} = 0.9259 * 100 = 92.59$$



Gráfica de resultados 5 Tesis grado académico.

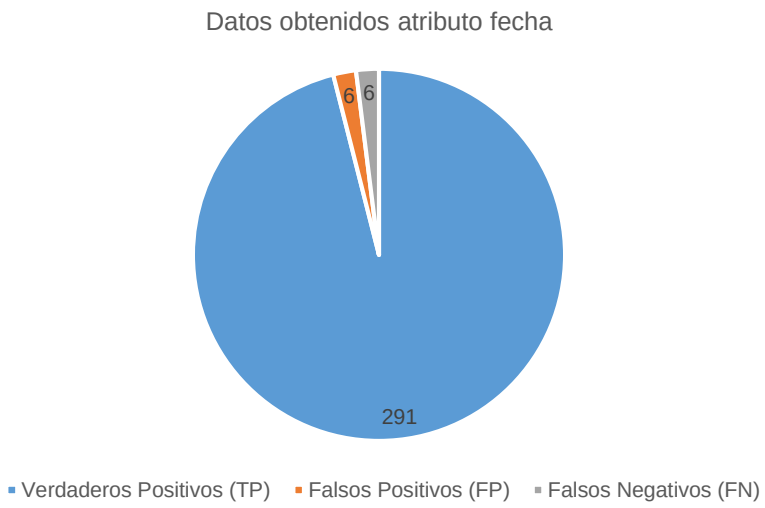
El motivo por el cual no se obtuvo de manera correcta el grado académico en algunos casos se debe a la distribución del texto dentro de algunos documento, o la tipografía utilizada no permitía que el sistema no identificara la sección del documento que contenía el área y en otros casos el nombre de los autores contenían su grado (ingeniería o licenciatura), lo cual causaba cierto conflicto al sistema, adicionalmente algunos documentos no contaban con permisos de lectura y escritura lo cual dificultaba o impedía la extracción de información de los documentos.

5.2.1.6. Resultados en fecha

Se procesaron un total de 297 archivos de tipo tesis. Al revisar los diferentes archivos PDF se encontró que contenían un total de 291 datos en la fecha.

Verdaderos Positivos (TP)	291
Falsos Positivos (FP)	6
Falsos Negativos (FN)	6

Tabla 9 Resultados obtenidos en la extracción de fechas.

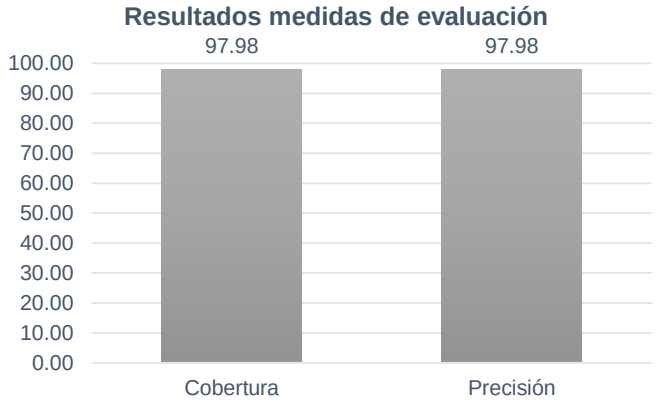


Gráfica de datos 6 Fecha de tesis.

De acuerdo a los resultados obtenidos de la fecha mostrados en la tabla 9 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, dando como resultado lo siguiente (gráfica de resultados 6):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{291}{(291 + 6)} = \frac{291}{297} = 0.9797 * 100 = 97.98$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{291}{291 + 6} = \frac{291}{297} = 0.9797 * 100 = 97.98$$



Gráfica de resultados 6 Fecha de tesis.

El motivo por el cual no se obtuvo de manera correcta la fecha en algunos casos se debe a la distribución del texto dentro del documento, o la tipografía utilizada no permitía que el sistema no identificara la sección del documento que contenía la fecha, en otros casos se debía a que la estructura de la fecha utilizada no permitía

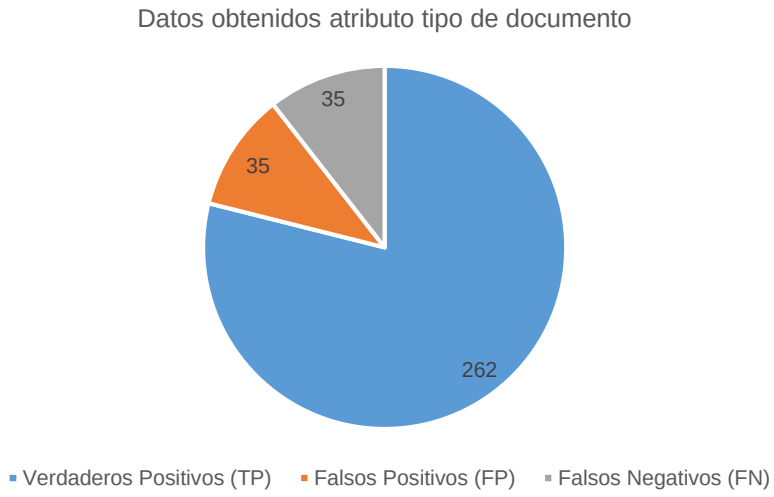
que el sistema delimitara la fecha concreta del documento, adicionalmente algunos documentos no contaban con permisos de lectura y escritura lo cual dificultaba o impedía la extracción de información de los documentos.

5.2.1.7. Resultados en título

Se procesaron un total de 297 archivos de tipo. Al revisar los diferentes archivos PDF se encontró que contenían un total de 262 datos en el título.

Verdaderos Positivos (TP)	262
Falsos Positivos (FP)	35
Falsos Negativos (FN)	35

Tabla 10 Resultados obtenidos en la extracción de título.

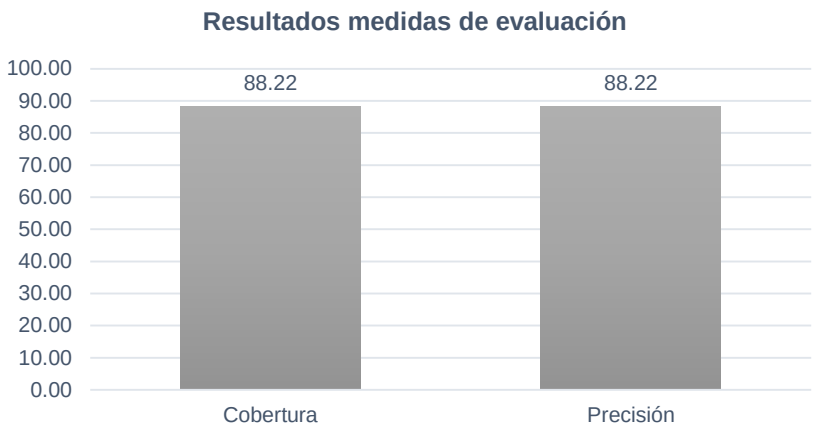


Gráfica de datos 7 Título de tesis.

De acuerdo a los resultados obtenidos del título mostrados en la tabla 10 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, dando como resultado lo siguiente (gráfica de resultados 7):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{262}{(262 + 35)} = \frac{262}{297} = 0.8821 * 100 = 88.22$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{262}{262 + 35} = \frac{262}{297} = 0.8821 * 100 = 88.22$$



Gráfica de resultados 7 Título de tesis.

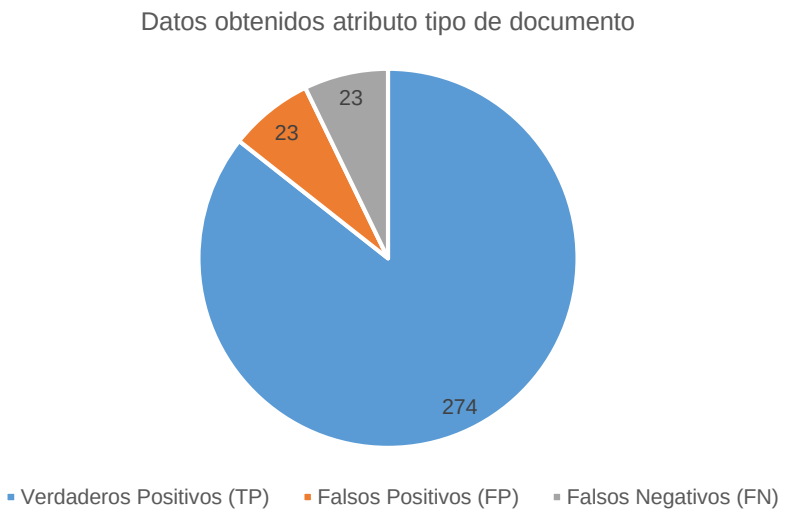
El motivo por el cual no se obtuvo de manera correcta el título es que en algunos casos la distribución del texto dentro del documento, o la tipografía utilizada no permitía que el sistema no identificara la sección del documento que contenía el título, adicionalmente algunos documentos no contaban con permisos de lectura y escritura lo cual dificultaba o impedía la extracción de información de los documentos.

5.2.1.8. Resultados en institución

Se procesaron un total de 297 archivos de tipo tesis. Al revisar los diferentes archivos PDF se encontró que contenían un total de 274 datos en el instituto.

Verdaderos Positivos (TP)	274
Falsos Positivos (FP)	23
Falsos Negativos (FN)	23

Tabla 11 Resultados obtenidos en la extracción de institutos.

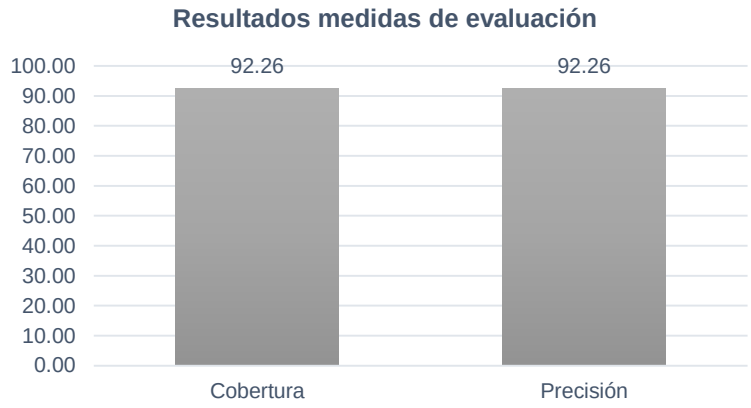


Gráfica de datos 8 Institución de tesis.

De acuerdo a los resultados obtenidos del instituto mostrados en la tabla 11 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, dando como resultado lo siguiente (gráfica de resultados 8):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{274}{(274 + 23)} = \frac{274}{297} = 0.9225 * 100 = 92.26$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{274}{274 + 23} = \frac{274}{297} = 0.9225 * 100 = 92.26$$



Gráfica de resultados 8 Institución de la tesis.

El motivo por el cual no se obtuvo de manera correcta el instituto es que en algunos casos la distribución del texto dentro del documento, o la tipografía utilizada no permitía que el sistema no identificara la sección del documento que contenía el instituto, en otros casos el instituto llegaba a venir su nomenclatura o venia el logo de

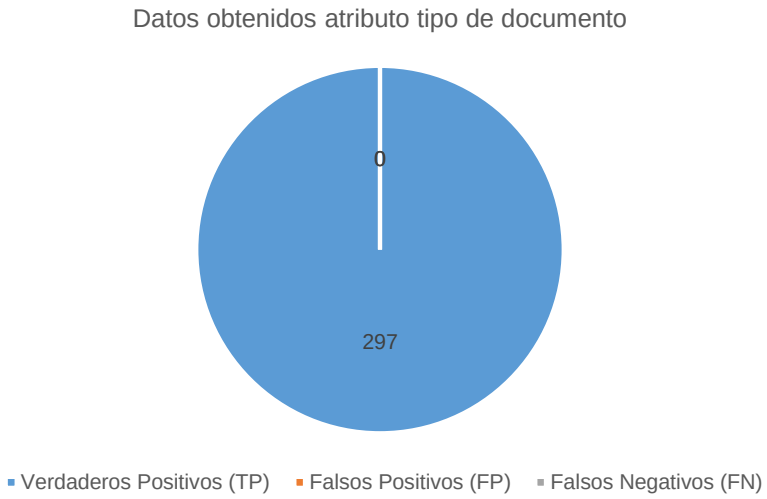
la institución lo cual no permitía al sistema que extrajera dichos datos, por otro lado algunos documentos no contaban con permisos de lectura y escritura lo cual dificultaba o impedía la extracción de información de los documentos.

5.2.1.9. Resultados en tipo de documento

Se procesaron un total de 297 archivos de tipo tesis. Al revisar los diferentes archivos PDF se encontró que contenían un total de 297 datos en el tipo de documento.

Verdaderos Positivos (TP)	297
Falsos Positivos (FP)	0
Falsos Negativos (FN)	0

Tabla 12 Resultados obtenidos en la extracción del tipo documento.

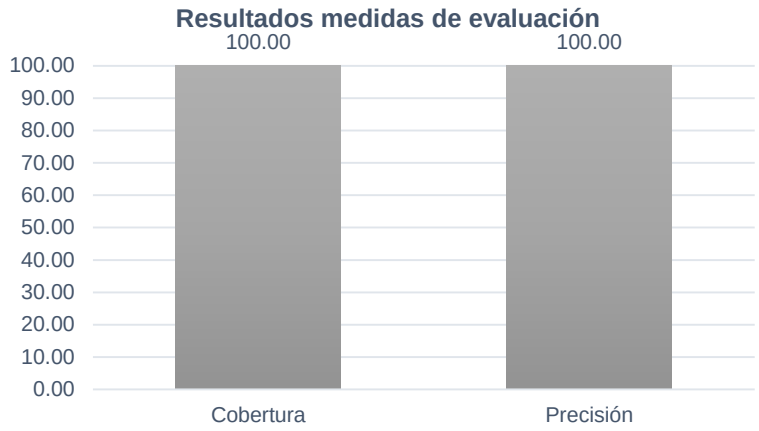


Gráfica de datos 9 Tipo de documento (Tesis).

De acuerdo a los resultados obtenidos del tipo de documento mostrados en la tabla 12 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, dando como resultado lo siguiente (gráfica de resultados 9):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{297}{(297 + 0)} = \frac{297}{297} = 1 * 100 = 100$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{297}{297 + 0} = \frac{297}{297} = 1 * 100 = 100$$



Gráfica de resultados 9 Tipo de documento (Tesis).

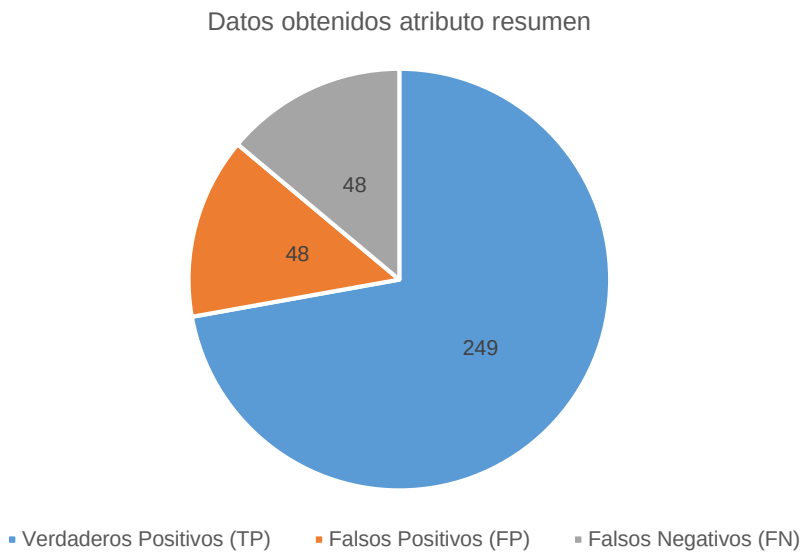
En este caso se obtuvo de manera correcta la identificación del tipo de documento debido a que se buscan en las primeras páginas del documento palabras clave que ayudan a identificar y diferenciar un artículo de una tesis, una vez que se obtiene el tipo de documento se podrece a la extracción especifica de lo metadatos.

5.2.1.10. Resultados en resumen

Se procesaron un total de 297 archivos de tipo tesis. Al revisar los diferentes archivos PDF se encontró que contenían un total de 249 datos en los resultados obtenidos.

Verdaderos Positivos (TP)	249
Falsos Positivos (FP)	48
Falsos Negativos (FN)	48

Tabla 13 Resultados obtenidos en la extracción del resumen.

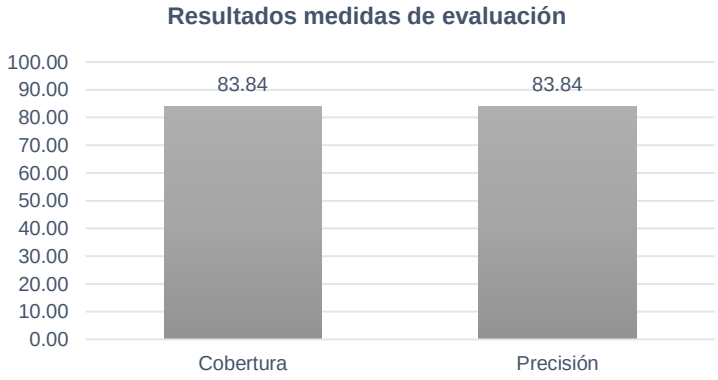


Gráfica de datos 10 Resumen de tesis.

De acuerdo a los resultados obtenidos del tipo de documento mostrados en la tabla 13 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, dando como resultado lo siguiente (gráfica de resultados 10):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{249}{(249 + 48)} = \frac{249}{297} = 0.8383 * 100 = 83.83$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{249}{249 + 48} = \frac{249}{297} = 0.8383 * 100 = 83.83$$



Gráfica de resultados 10 Resumen de tesis.

El motivo por el cual no se obtuvo de manera correcta el resumen, se debe a que en algunos casos la distribución del texto dentro del documento, o la tipografía utilizada no permitía que el sistema no identificara la sección del documento que contenía el

resumen, en otros casos los índices causaban interferencias a la hora de la extracción de la sección del resumen, adicionalmente algunos documentos no contaban con permisos de lectura y escritura lo cual dificultaba o impedía la extracción de información de los documentos.

5.2.1.11. Resultados en palabras clave

Para la extracción de las palabras clave en las tesis fue necesaria la extracción de la sección de “Resumen y Conclusión” debido a que en estas secciones el autor de la tesis trata de resumir lo que se va a ver o se vio en el trabajo que se presenta, por medio de la herramienta RAKE y de estas secciones se hace la extracción de las palabras clave.

En total se extrajeron a 272 archivos las palabras clave, los 25 archivos restantes no se les pudo extraer las palabras clave debido a la estructura del documento, ya que en algunos casos no permitía encontrar las secciones mencionadas ya se por la tipografía o alguna otra cuestión a la hora de la extracción de la información, e incluso en algunos casos por los permisos de lectura y escritura que contenían los documentos. El porcentaje de palabras clave obtenidas es de 91.52% de un total de 297 documentos de tesis.

5.2.2. Resultados en artículos

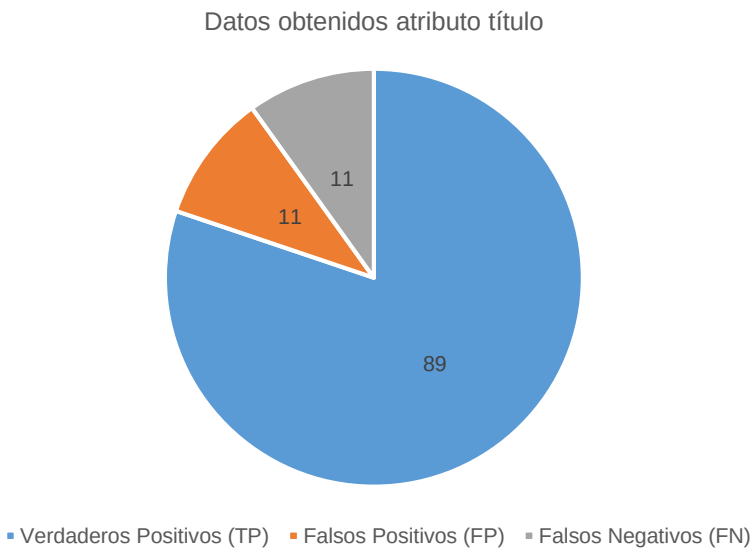
A continuación, se muestran los resultados obtenidos en la extracción de metadatos en los documentos de tipo artículo.

5.2.2.1. Resultados en título

Se procesaron un total de 100 archivos de tipo artículo científico. Al revisar los diferentes archivos PDF se encontró que contenían un total de 89 datos en el título.

Verdaderos Positivos (TP)	89
Falsos Positivos (FP)	11
Falsos Negativos (FN)	11

Tabla 14 Resultados obtenidos en la extracción de título en Artículos.

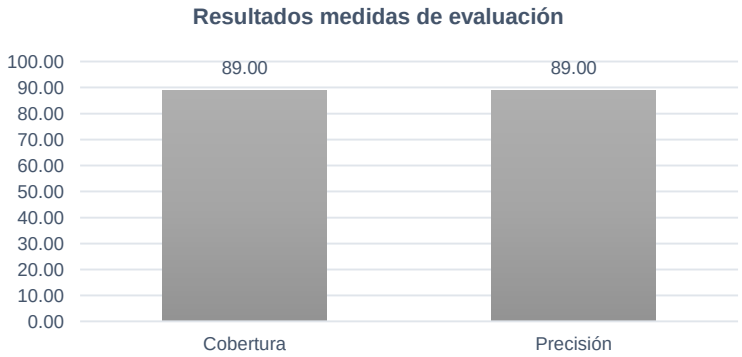


Gráfica de datos 11 Artículo título.

De acuerdo a los resultados obtenidos del título mostrados en la tabla 14 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, dando como resultado lo siguiente (gráfica de resultados 11):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{89}{(89 + 11)} = \frac{89}{100} = 0.89 * 100 = 89$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{89}{89 + 11} = \frac{89}{11} = 0.89 * 100 = 89$$



Gráfica de resultados 11 Artículo título.

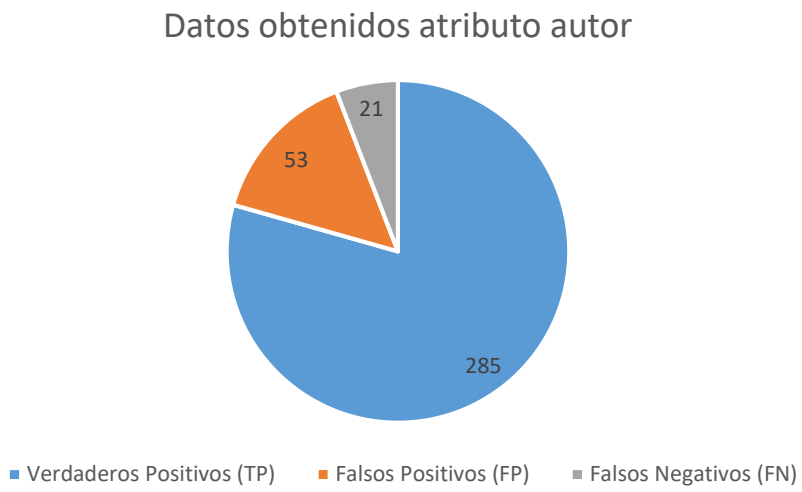
El motivo por el cual no se obtuvo de manera correcta el título es que en algunos casos la distribución del texto dentro del documento o el formato del artículo no permitían a la herramienta realizar correctamente el análisis del documento, esto cual dificultaba o impedía la extracción de la información.

5.2.2.2. Resultados en autor

Se procesaron un total de 100 archivos de tipo artículo científico. Al revisar los diferentes archivos PDF se encontró que contenían total de 311 datos de tipo autor.

Verdaderos Positivos (TP)	285
Falsos Positivos (FP)	53
Falsos Negativos (FN)	21

Tabla 15 Resultados obtenidos en la extracción de autores en Artículos.

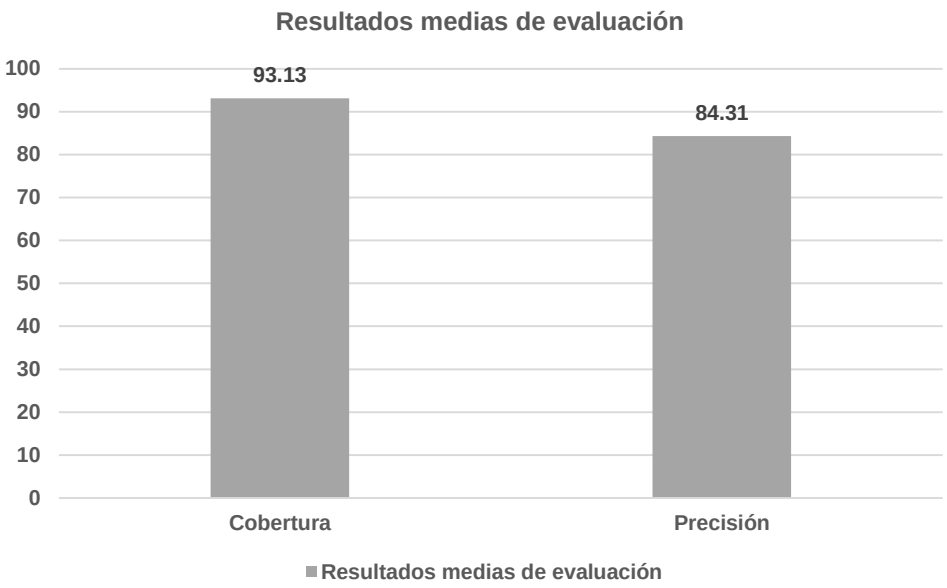


Gráfica de datos 12 Autor artículo.

De acuerdo a los resultados de autores mostrados en la tabla 15 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, resultado lo siguiente (gráfica de resultados 12):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{285}{(285 + 21)} = \frac{285}{306} = 0.9313 * 100 = 93.13$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{285}{285 + 53} = \frac{285}{338} = 0.8431 * 100 = 84.31$$



Gráfica de resultados 12 Autor artículo.

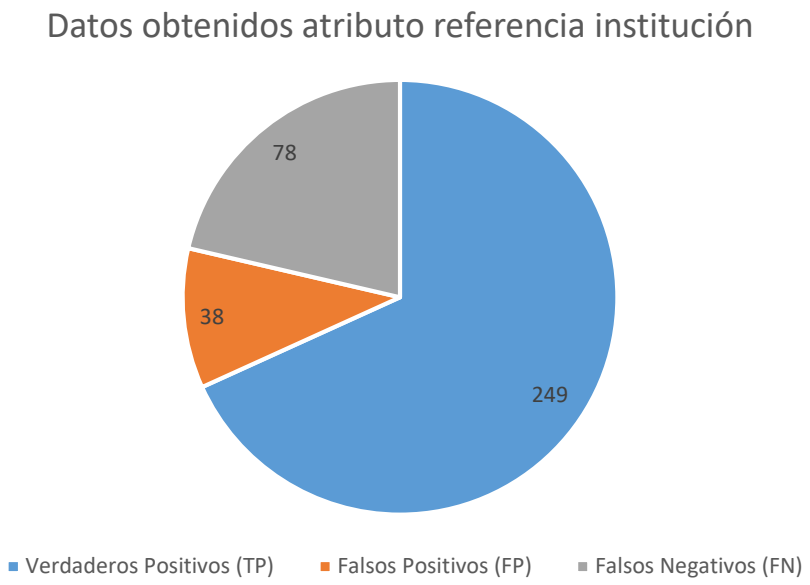
El motivo por el cual el resultado de cobertura fue de mayor calidad comparado con el resultado de precisión se debe a que el sistema en muchos casos además de detectar los autores del artículo, en algunos de ellos detectaba duplicadamente algunos autores, en otros casos los indicadores que se utilizan en los artículos para identificar las instituciones o correos con sus respectivos autores provocaban ruido a la herramienta, por lo cual arrojaba autores inexistentes en algunos artículos.

5.2.2.3. Resultados en referencia a institución

Se procesaron un total de 100 archivos de tipo artículo. Al revisar los diferentes archivos PDF se encontró que contenían un total de 293 referencias a institución, No se tienen el mismo número de referencias a institución que autores debido a que algunos archivos no contienen dichas referencias.

Verdaderos Positivos (TP)	249
Falsos Positivos (FP)	38
Falsos Negativos (FN)	78

Tabla 16 Resultados obtenidos en la extracción de instituciones en Artículos.

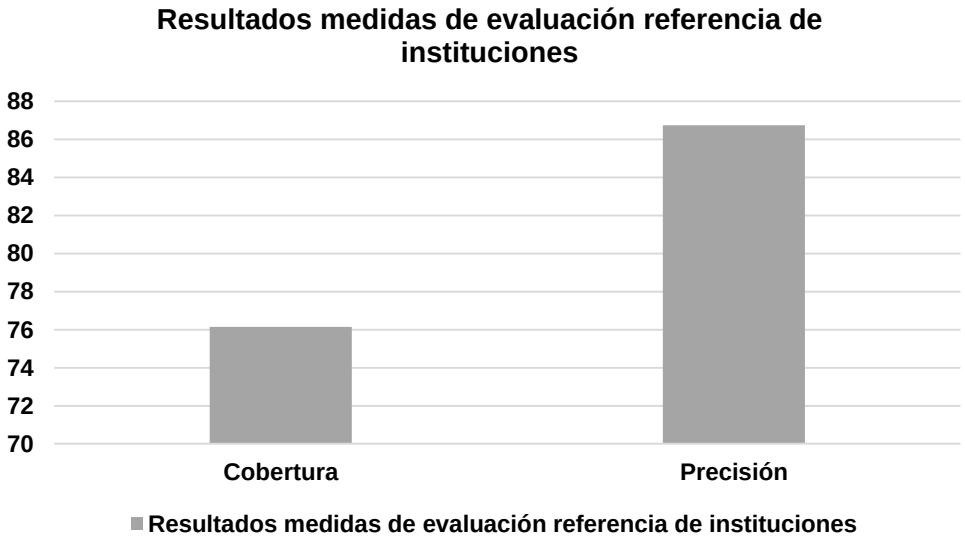


Gráfica de datos 13 Institución del artículo.

De acuerdo a los resultados obtenidos del instituto mostrados en la tabla 16 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, resultado lo siguiente (gráfica de resultados 13):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{249}{(249 + 78)} = \frac{249}{327} = 0.7478 * 100 = 76.14$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{249}{249 + 38} = \frac{249}{287} = 0.7978 * 100 = 79.78$$



Gráfica de resultados 13 Institución del artículo.

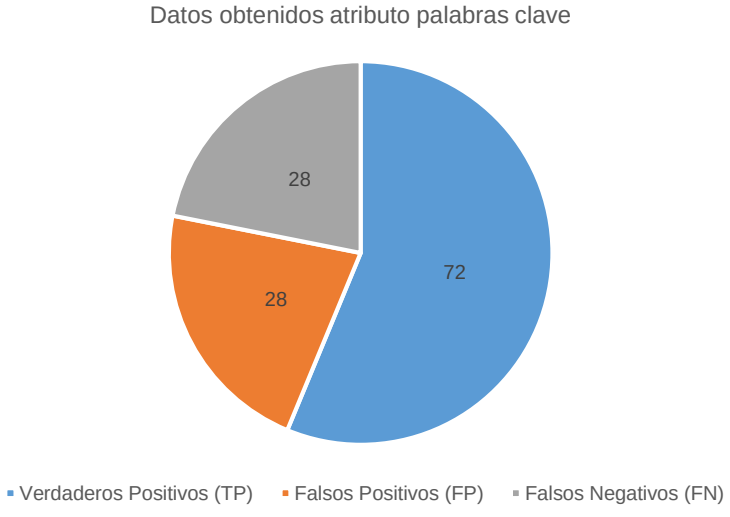
El motivo por el cual el resultado de Precisión fue de mayor calidad comparado con el resultado de cobertura se debe a que la herramienta en muchos casos al detectar una mayor cantidad de autores identifica instituciones para cada uno de los autores detectados, en otros casos debido a la estructura del archivo, no logra identificar instituciones para los autores.

5.2.2.4. Resultados en palabras clave

Se procesaron un total de 100 archivos de tipo artículo científico. Al revisar los diferentes archivos PDF se encontró que un total de 79 archivos si contenían palabras clave y el resto no, con base a esa información se verifico si la herramienta extraía en los archivos que si contenían palabras clave el dato y en los archivos que no las contenían en verdad no extrajeran el dato.

Verdaderos Positivos (TP)	72
Falsos Positivos (FP)	28
Falsos Negativos (FN)	28

Tabla 17 Resultados obtenidos en la extracción del resumen/abstract en Artículos.

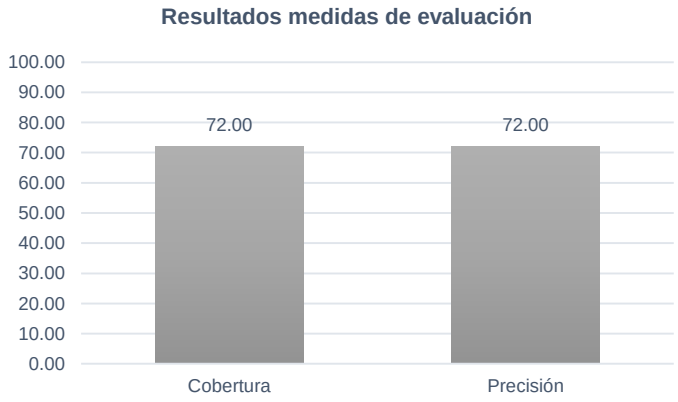


Gráfica de datos 14 Palabras clave de artículo.

De acuerdo a los resultados obtenidos de las palabras clave mostrados en la tabla 17 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, resultado lo siguiente (gráfica de resultados 14):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{72}{(72 + 28)} = \frac{72}{100} = 0.72 * 100 = 72$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{72}{72 + 28} = \frac{72}{100} = 0.72 * 100 = 72$$



Gráfica de resultados 14 Palabras clave artículo.

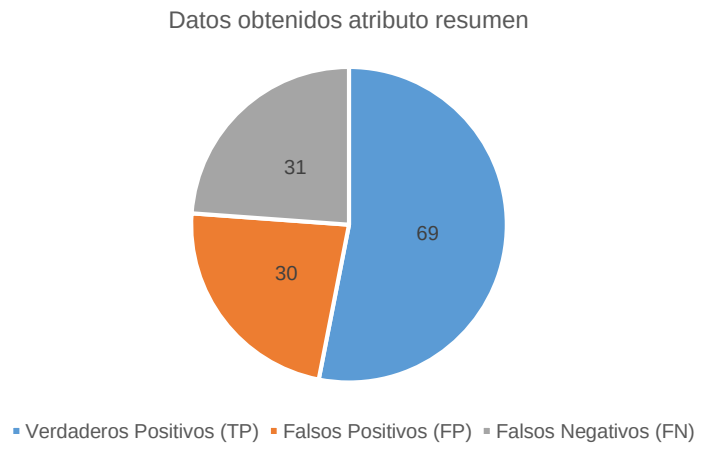
El motivo por el cual no se obtuvieron de manera correcta las palabras clave que pueden contener los artículos científicos, se debe a que en algunas ocasiones la herramienta no lograba identificar de manera correcta los límites del resumen/Abstract, lo cual provocaba que este atributo fuera obtenido como parte del resumen extraído.

5.2.2.5. Resultados en resumen

Se procesaron un total de 100 archivos de tipo artículo científico. Al revisar los diferentes archivos PDF se encontró que contenían un total de 69 datos en los resultados.

Verdaderos Positivos (TP)	69
Falsos Positivos (FP)	30
Falsos Negativos (FN)	31

Tabla 18 Resultados obtenidos en la extracción del resumen/abstract en Artículos.

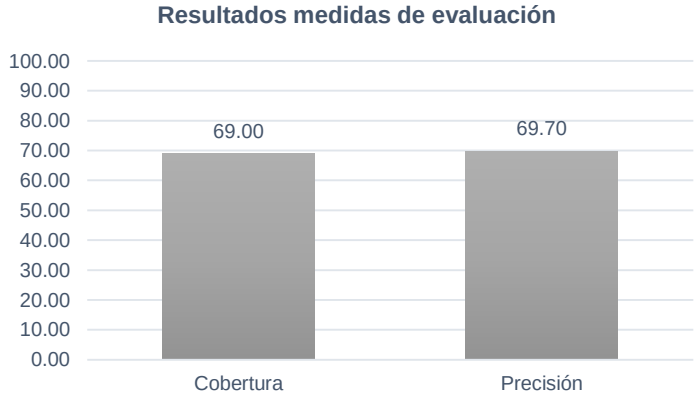


Gráfica de datos 15 Resumen artículo.

De acuerdo a los resultados obtenidos del resumen mostrados en la tabla 18 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, resultado lo siguiente (gráfica de resultados 15):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{69}{(69 + 31)} = \frac{69}{100} = 0.69 * 100 = 69$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{69}{69 + 30} = \frac{249}{99} = 0.6969 * 100 = 69.70$$



Gráfica de resultados 15 Resumen artículo.

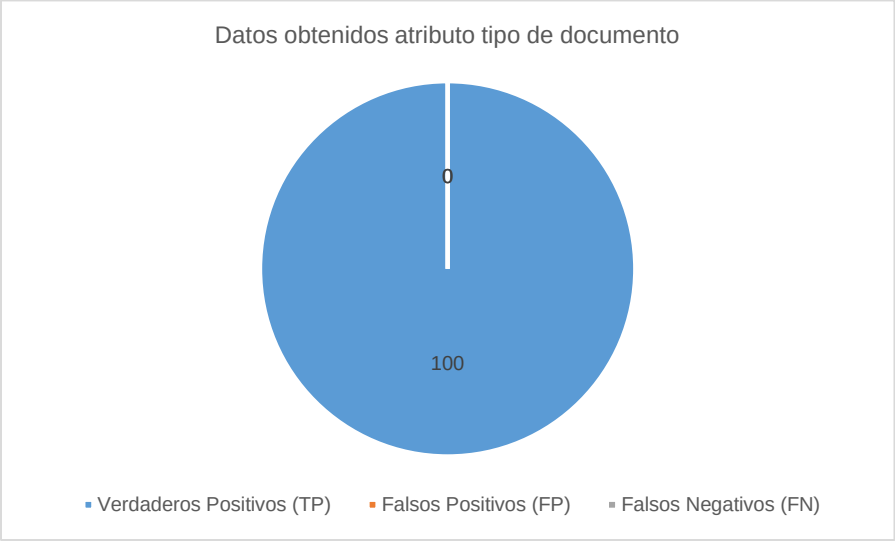
El motivo por el cual no se obtuvo de manera correcta el resumen, se debe a que en algunos casos la herramienta no identificaba los límites del resumen o abstract o definitivamente no identificaba la existencia de alguno de estos dos atributos en el documento.

5.2.2.6. Resultados en tipo de documento

Se procesaron un total de 100 archivos de tipo artículo científico. Al revisar los diferentes archivos PDF se encontró que contenían un total de 100 datos en el tipo de documento.

Verdaderos Positivos (TP)	100
Falsos Positivos (FP)	0
Falsos Negativos (FN)	0

Tabla 19 Resultados obtenidos en la extracción del tipo documento.



Gráfica de datos 16 Tipo de documento (artículo).

De acuerdo a los resultados obtenidos del tipo de documento mostrados en la tabla 19 se realizó el cálculo de las medidas de precisión y de cobertura tomando como base las ecuaciones1-2, resultado lo siguiente (gráfica de resultados 16):

$$Cobertura\ (Recall) = \frac{TP}{(TP + FN)} = \frac{100}{(100 + 0)} = \frac{100}{100} = 1 * 100 = 100$$

$$Precisión = \frac{TP}{(TP + FP)} = \frac{100}{100 + 0} = \frac{100}{100} = 1 * 100 = 100$$

Gráfica de resultados 16 Tipo de documento (artículo).

En este caso se obtuvo de manera correcta la identificación del tipo de documento debido a que se buscan en las primeras páginas del documento palabras clave que ayudan a identificar y diferenciar un artículo de una tesis, una vez que se obtiene el tipo de documento se podrece a la extracción especifica de lo metadatos.

A continuación, se muestra la forma en la que se realizan las correcciones manuales de los datos generados por el sistema.

Figura 1 Índex para seleccionar el documento a modificar.

Como se muestra en la Figura 1, lo primero que se hacer para realizar la modificación de los datos extraídos es seleccionar el documento que se modificará o revisará, para

ello se debe buscar el título del documento. Para realizar la modificación se abrirá una página como la que se muestra en la Figura 2, con los datos del documento a revisar.

Figura 2 Plantilla para modificar información del documento.

Se tomará como ejemplo el documento que se muestra en la Figura 3, en este se puede ver con diferentes colores cuales son los respectivos datos correctos (lado izquierdo, la portada del documento) y cueles son los datos obtenidos por el sistema que fueron erróneos (Lado derecho).

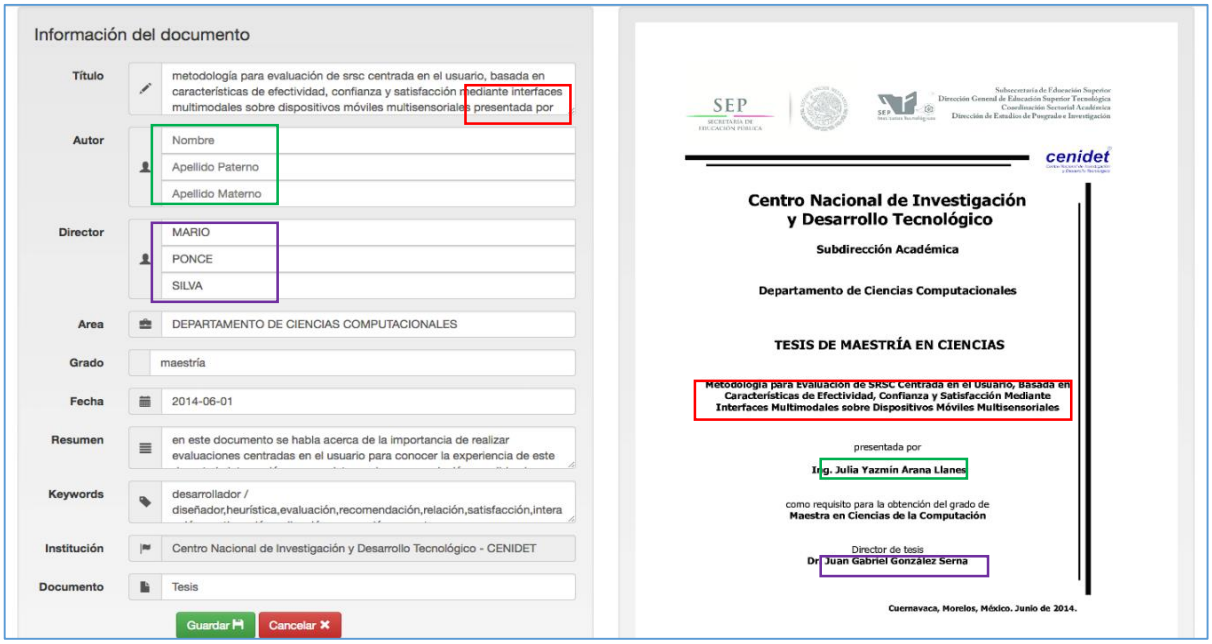


Figura 3 Ejemplo datos extraídos del sistema.

Como se puede observar algunos de los campos están vacíos como es el caso del autor (color verde), esto se puede deber a dos razones, en uno de los casos si el autor fue extraído de manera correcta, pero este no se encuentra registrado en la base de datos, este no se mostrará en la plantilla, el otro caso es que el autor no haya sido extraído de manera correcta, si este fuera el caso el autor puede ser o no localizado dentro de la base de datos.

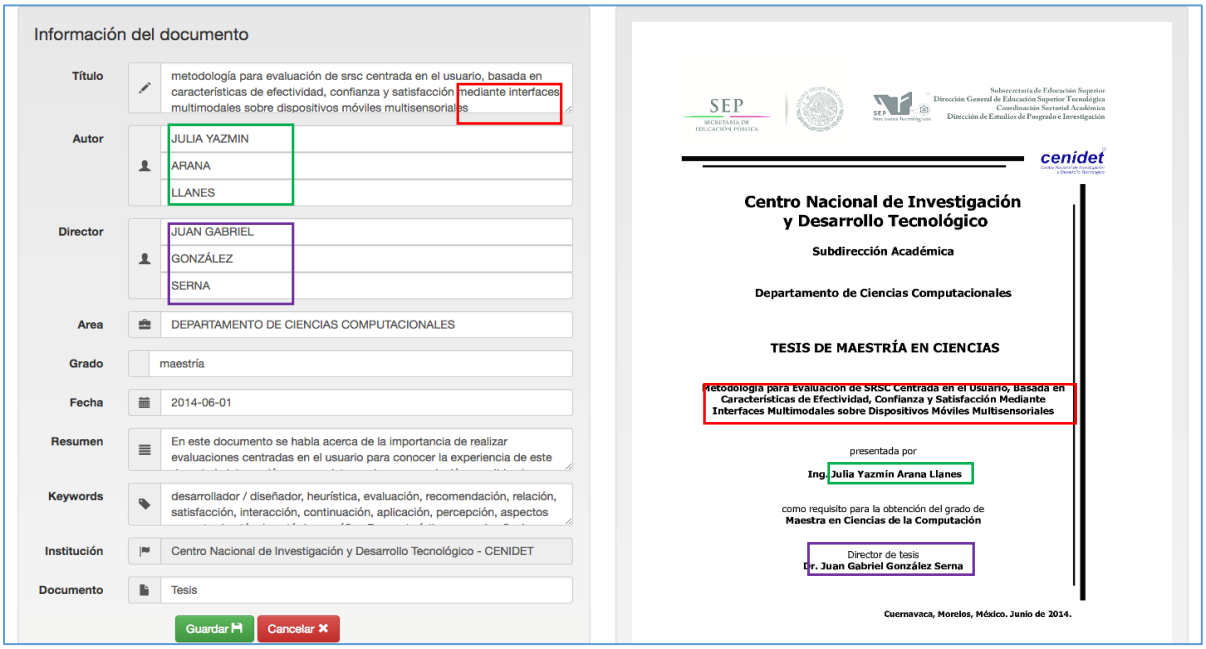


Figura 4 Ejemplo de datos modificados.

Estos casos pueden suceder de igual forma para el campo director (color morado), en este caso a pesar de que el director fue extraído de manera incorrecta si existía dentro de la base de datos, por tanto, este se muestra dentro de la plantilla.

En la Figura 4 se muestran los datos modificados por el usuario (cada uno con su respectivo color), para facilitar el trabajo del lado derecho de la página se muestra la portada del documento, en la cual el usuario puede observar si los datos que se obtuvieron por el sistema son correctos, en caso de lo contrario los puede modificar si es necesario, sin embargo si para el caso de los datos de autor y director como se mencionó arriba, si no existieran la nueva información en la base de datos, los cambios que se realicen dichos campos no se realizan.

Capitulo VI. Conclusiones

En este capítulo se muestran las conclusiones del trabajo de igual forma si se cumplieron los objetivos de este trabajo.

6. Conclusiones

Con este trabajo de investigación se logró desarrollar un sistema semiautomático con la capacidad para realizar la extracción semiautomática de metadatos en documentos de tipo tesis (licenciatura, maestría y doctorado) y artículos científicos.

Para realizar el trabajo de extracción de metadatos en primer lugar se realizó la búsqueda de documentos que conformarían el repositorio, la búsqueda de estos documentos fue complicada debido a que dichos documentos debían cumplir con ciertas características como:

- Ser escritas por instituciones que conforman el tecnológico nacional de México (TecNM).
- Estar escritos en idioma español.
- En caso de los artículos, deben tener la taxonomía IEEE o ACM.
- Que contuvieran OCR para el análisis de la información.

Dentro de los problemas que surgieron para la conformación del repositorio se puede mencionar, que desde hace varios años las instituciones educativas no están generando documentos de tesis, debido a que existen más métodos para la obtención del grado sin necesidad de escribir dicho documento, a pesar de este percance se lograron conseguir 400 archivos los cuales fueron limpiados para su análisis.

Para la extracción de metadatos se realizaron diversas heurísticas, las cuales permitirían determinar la distribución de cada metadato en los diferentes documentos. Los resultados obtenidos no fueron 100% exactos, una de las causas de esto fue la tipografía utilizada en la redacción del documento, la cual afectaba en algunos casos la identificación de los metadatos, por tanto, los resultados que se obtenían en esos casos era incorrectos, o contenían información de más.
Sin embargo, los resultados obtenidos al aplicar las medidas de precisión y cobertura a cada uno de los metadatos fueron en la mayoría de los casos mayores a un 85%, lo cual se considera adecuado ya que, al ser un sistema semiautomático, se le pueden

realizar las correcciones a los resultados de forma manual, mediante un sistema web, de este modo obtener los resultados al 100%.

6.1. Trabajos futuros

Este trabajo de investigación cuenta con algunos aspectos que se pueden optimizar para mejorar los resultados obtenidos en la precisión y cobertura de los metadatos extraídos en los diversos documentos.

Como trabajos futuros para los documentos de tipo tesis, se pueden optimizar algunas actividades o crear algunos módulos para mejorar los resultados.

- **Mejora para la definición de heurísticas**: Las heurísticas definidas para este trabajo pueden ser optimizadas en la generación de palabras clave, como se vio en el trabajo desarrollado durante la estancia, se comprobó que existe una forma más óptima para generar palabras clave de mayor calidad.
  Otra de las heurísticas que se podría mejorar es la de extracción del título, esta se podría optimizar verificando la tipografía utilizada en la portada para así verificar adicionalmente, la información del título, así como la de los demás metadatos que son extraídos de la portada.

- **Obtención de resumen**: Se pueden realizar algunas de las siguientes dos acciones, la primera es la creación de un módulo especializado en la generación automática de resumen y el segundo es optimizar la heurística para la extracción del resumen que contiene el documento.
- **Extracción y almacenamiento de múltiples autores en las tesis.**

Como trabajos futuros para los documentos de tipo artículo científico, se pueden optimizar algunas actividades o crear algunos módulos para mejorar los resultados.

- **Mejorar heurística en la extracción del metadato autor en artículos:**
  - Eliminar el truncamiento de nombres
  - Identificación de nombres completos de los autores a partir de las posibles siglas o abreviaturas.

- **Mejora de obtención de resumen.**
  - Creación de un módulo especializado en la generación automática de resumen.
  - Optimizar la heurística para la extracción del resumen que contiene el documento.

Los trabajos futuros que se pueden realizar en el sistema en general son:

- Extracción de metadatos en documentos que se encuentren en otros idiomas.
- Creación de heurísticas adicionales para extraer otro tipo de metadatos en estos documentos.
- Creación de módulos para la extracción de metadatos de otro tipo de documentos.

Bibliografía

clustering jpmonge. (2012). R.I : Clustering. Recuperado el Sep de 2016, de http://clustering.jpmonge.com/definicion

Computacion y Sistemas. (Na). soft similarity and aoft Cosine Measure: Similarity of features in Vector Space Model. Recuperado el Jun de 2016, de http://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2043

CONACYT. (2016). CONACYT. Recuperado el MAR de 2016, de http://www.conacyt.mx/siicyt/index.php/centros-de-investigacion-conacyt/2393-lineamientos-generales-para-el-repositorio-nacional-y-los-repositorios-institucionales?path=

CONACYT. (JULIO de 2016). CONACYT. Recuperado el 29 de AGOSTO de 2016, de http://conacyt.gob.mx/index.php/el-conacyt/normatividad

Lopez, P. (2016). GROBID - GeneRation of Bibliography Data. Recuperado el Sep de 2016, de GROBID: grobid.readthedocs.io/en/latest/introduction

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. En T. Cover, & P. Hart, Nearest neighbor pattern classification (Vol. 13, págs. 21-27).

Abdillah, L. A. (2013). PDF ARTICLE METADATA HARVESTER. Jurnal Komputer dan Informatika , 1, 6.

Bedoya Puerta, J. A. (Sep de 2011). Universidad Politecnica de Valencia. Recuperado el Jun de 2016, de http://users.dsic.upv.es/~flip/papers/TFM-JorgeBedoya.pdf

Garre, M., Cuadrado, J. J., & Sicilia, M. A. (Na). Comparación de diferentes algoritmos de clustering en la estimacion de coste en el desarrollo de software. Recuperado el Jun de 2016, de sc.ehu.es: http://www.sc.ehu.es/jiwdocoj/remis/docs/GarreAdis05.pdf

KDE. (s.f.). Okular - más que un lector. Recuperado el Mayo de 2017, de Okular: https://okular.kde.org/

Manjula, W. C. (2013). Una ontología basada en un sistema de clasificación de documentos totalmente automático utilizando un sistema semiautomático existente. IFLA WLIC , 14.

Mateos, M., & Ruiz, R. (2013). Universidad de Sevilla. Recuperado el 12 de MAR de 2016, de https://www.cs.us.es/cursos/ia2/temas/tema-06.pdf

Molina, A., da Cunha, I., Torres-Moreno, J.-M., & Velázquez-Morales, P. (Diciembre de 2010). La compresión de frases: un recurso para la optimización de resumen automático de documentos. LinguaMÁTICA .

Montero Martínez, J. M. (15 de FEB de 2001). Grupo de Tecnología del Habla . Recuperado el 12 de MAR de 2016, de http://lorien.die.upm.es/juancho/pfcs/DPF/capitulo2.pdf

ORCID. (NA). ORCID. Recuperado el OCT de 2016, de https://orcid.org/content/initiative?locale_v3=es

Pajares Martinsanz, G., & de la Cruz García, J. M. (2008). Ejercicios resueltos de visión por computador . En G. Pajares Martinsanz, & J. M. de la Cruz García, Ejercicios resueltos de visión por computador (págs. 181-183). México: Alfaomega.

Peña Ayala, A. (17 de Enero de 2008). Instituto Politecnico Nacional. Recuperado el Sep de 2016, de http://www.wolnm.org/apa/articulos/Lenguaje_Natural.pdf?target=

PHP. (NA). PHP. Recuperado el Junio de 2016, de http://php.net/manual/es/intro.imagick.php

Python Software Foundation. (28 de Mar de 2014). Python. Recuperado el Oct de 2016, de Python: https://pypi.python.org/pypi/pdfminer/

Rendón Miranda, J. C. (2014). Clasificación automática de objetos de conocimiento con contenido No Estructurado Para el poblado semiautomático de ontologías multidimensionales. Cuernavaca, Morelos, México.

Rose, S., Engel, D., Cramer, N., & Cowel, W. (2010). Automatic key extraction from individual documents. Text Mining: Aplications and Theory , 9.

Sáez Guerrero, M. (OCT de 2009). e-Archivo. Recuperado el MAR de 2016, de http://e-archivo.uc3m.es/bitstream/handle/10016/5874/PFC_Miguel_Saez_Guerrero.pdf?sequence=1

Sancho Caparrini, F. (Jul de 2015). Departamento de ciencias computacionales e inteligencia artificial. Recuperado el Jun de 2016, de Universidad de Sevilla: http://www.cs.us.es/~fsancho/?e=75

Téllez Valero, A. (2007). INAOE.edu.mx. Recuperado el FEB de 2016, de https://ccc.inaoep.mx/~mmontesg/tesis%20estudiantes/TesisMaestria-AlbertoTellez.pdf

Tkaczyk, D., Bolikowski, l., Czeczko, A., & Rusek, K. (2012). Una ontología basada en un sistema de clasificación de documentos totalmente automático utilizando un sistema semiautomático existente. IEEE conputer society , 6.

Tovar Vidal, M. (2015). Evaluación automática de ontologías de dominio restringido. Cuernavaca, Morelos, México.

visual interaction & communication technologies. (2009). vicomtech. Recuperado el Sep de 2016, de http://www.vicomtech.org/t4/e11/procesamiento-del-lenguaje-natural

Yasotha, R., & Charles, E. (2015). Automated Text Document Categorization. IEEE , 7.