UNIVERSIDAD NACIONAL DE MOQUEGUA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA



"RECOMENDACIÓN DE ARTÍCULOS CIENTÍFICOS UTILIZANDO TÉCNICAS DE MINERÍA DE TEXTOS PARA FORMULACIÓN DE PROYECTOS DE INVESTIGACIÓN"

TESIS:

PRESENTADO POR:

CARMEN MARIA CHINO CERVANTES

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS E INFORMÁTICA

MOQUEGUA, PERÚ

2022

UNAM
UNIVERSIDAD NACIONAL DE MOQUEGUA

ESCUELA PROFESIONAL DE
INGENIERÍA DE SISTEMAS E INFORMÁTICA

Folio Nº 009

ACTA DE SUSTENTACIÓN

I. Modalidad:

Tesis : [X]
Trabajo de Suficiencia Profesional : [ ]
Otro: ..... : [ ]

II. Documento de aprobación para la sustentación:

Resolución Nº 215-2022-FACIS-UNAM

Fecha: 26 setiembre 2022

III. El jurado ha dictaminado la (el) Tesis, TITULADO:

"Recomendación de Artículos Científicos utilizando Técnicas de Minería de Textos para formulación de Proyectos de Investigación"

Bachiller: Carmen María Chino Cervantes

Asesor Principal: Dr. Aníbal Fernando Flores García

Co-Asesor: Mg. Honorio Apaza Alanoca

IV. Se dictamina:

Como: Aprobado por Unanimidad la sustentación de manera:

Presencial [X] Virtual [ ]

Con nota: catorce (14) por lo que; Sí [X] No [ ]

Opta el Título Profesional de INGENIERO DE SISTEMAS E INFORMÁTICA.

Siendo las 17:31 horas, del 06 de octubre del año 2022 se procede con la firma en conformidad de lo dictaminado por el Jurado.

M.Sc. Hugo Eulo Tito Ch.
Presidente

M.Sc. Victor Yana Mamani
Primer Miembro

Mgt. Juan C. Flores Perca
Segundo Miembro

Carmen María Chino Cervantes
Bachiller

09

Dedicatoria

A Dios, por estar a mi lado en todo momento. A mis padres por su confianza, y acompañamiento en todo el proceso. A mis hermanas y familia por apoyo indirecto, ya estando cerca o lejos.

A mis docentes que fueron parte de mi formación universitaria, a mis compañeros de la universidad, quienes se convirtieron en amigos estos últimos años.

A mis amistades cercanas por sus palabras de motivación, y que por circunstancias coincidía en investigación.

Agradecimiento

Agradezco a Dios, por guiarme en todo este proceso de aprendizaje y hacerme mejorar como persona.

A mi madre, por sus sabias palabras, su amor incondicional y aliento día a día.

A mi padre, por su esfuerzo en enseñarme que el trabajo dignifica a las personas.

A mis hermanas y cuñados, por sus consejos de superación y apoyo en todo momento, ya estando cerca o lejos.

A mis coach de vida, que han contribuido en mi desarrollo personal y que todo cambio siempre tiene una razón de ser.

A mis asesores, Dr. Anibal Flores y Co asesor Ing. Honorio Apaza, por sus enseñanzas en transmitir sus conocimientos de investigación. A los miembros del jurado por su apoyo en el proceso del desarrollo del trabajo de investigación.

Finalmente, agradezco a la Universidad Nacional de Moquegua, mi alma mater y a todo el personal por su confianza.

Índice general

I. Planteamiento del problema **12**

1.1. Descripción y formulación del problema 12

1.2. Objetivos 16

1.2.1. Objetivo general 16

1.2.2. Objetivos específicos 16

1.3. Justificación e importancia 16

1.4. Variables 17

II. Marco Teórico **18**

2.1. Antecedentes 18

2.2. Bases teóricas 20

2.2.1. Inteligencia artificial 20

2.2.1.1. Minería de Textos 21

2.2.1.2. Procesamiento Lenguaje Natural 22

2.2.2. Metodología 24

2.2.3. Algoritmos propuestos 24

2.2.3.1. Jaro Winkler 24

2.2.3.2. Término Frecuencia - Frecuencia inversa del documento (TF-IDF) 25

2.2.3.3. Word2vec 26
2.2.4. Contexto de investigación científica en el Perú 29
2.2.4.1. Producción científica en la universidades 30
2.3. Definición de términos 31

III. Metodología **33**

3.1. Tipo y nivel de investigación 33
3.2. Ámbito temporal y espacial 33
3.3. Población y muestra 34
3.4. Técnicas e instrumentos de recolección de datos 34
3.5. Procedimientos 34
3.5.1. Recuperación de información 34
3.5.2. Procesamiento Lenguaje Natural 37
3.5.2.1. Jaro Winkler 37
3.5.2.2. Término Frecuencia - Frecuencia inversa del documento (TF-IDF) 39
3.5.2.3. Word2vec 40
3.5.3. Extracción de datos 41
3.5.4. Minería de Datos y resultados 43
3.6. Análisis de datos 47

IV. Resultados y discusión **54**

V. Conclusiones **58**

VI. Recomendaciones **59**

Referencias **60**

Índice de figuras

II.1. *Metodología de investigación para proyectos Minería de Textos.* . . . 24

II.2. *Arquitectura de modelo CBOW.* 26

II.3. *Modelo continua de bag-of-word.* 27

II.4. *Modelo Skip-Gram.* 28

III.1. *Artículos recolectadas.* 35

III.2. *Base de datos de Artículos recolectadas.* 36

III.3. *Base de datos de tesis de la EPISI, UNAM.* 37

III.4. *Estructura final de datos.* 43

III.5. *Relevancia de las Áreas de investigación en las tesis.* 44

III.6. *Relevancia de los temas según los algoritmos en los 10 tesis de EPISI.* 45

Índice de cuadros

III.1. *Promedio de valores de similitud de documentos por algoritmos.* . . 42

III.2. *Diez artículos científicos mas recomendados para tesis N° 0.* 50

III.3. *Diez artículos científicos mas recomendados para tesis N° 1.* 50

III.4. *Diez artículos científicos mas recomendados para tesis N° 2.* 50

III.5. *Diez artículos científicos mas recomendados para tesis N° 3.* 51

III.6. *Diez artículos científicos mas recomendados para tesis N° 4.* 51

III.7. *Diez artículos científicos mas recomendados para tesis N° 5.* 51

III.8. *Diez artículos científicos mas recomendados para tesis N° 6.* 52

III.9. *Diez artículos científicos mas recomendados para tesis N° 7.* 52

III.10 *Diez artículos científicos mas recomendados para tesis N° 8.* 52

III.11 *Diez artículos científicos mas recomendados para tesis N° 9.* 53

IV.1. *Leyenda de tesis.* 55

IV.2. *Precisión de los algoritmos.* 56

IV.3. *Áreas de artículos científicos mas recomendados.* 57

Resumen

El desarrollo de proyectos de investigación es una actividad que busca descubrir nuevos conocimientos, aplicar conocimientos actuales para resolver problemas de la sociedad. La presente investigación tiene como objetivo recomendar artículos científicos para la formulación de proyectos de investigación con técnicas de minería de textos. En la primera etapa de este proceso, es la recuperación de información, para ello se ha usado el método de recolección manual de los datos, se ha seleccionado de manera aleatoria y clasificado por áreas de investigación en computación los artículos científicos y tesis de investigación. En la segunda etapa se detalla el proceso de experimentación con los tres algoritmos propuestos Jaro Winkler, TF-IDF y Word2vec. Como parte de la tercera etapa, es la extracción de datos con los algoritmos desarrollados nos retorna como resultados valores cuantitativos de relevancias y/o similitud de textos. En la cuarta etapa se aplica minería de datos con lo extraído en la anterior etapa para realizar el análisis de los documentos procesados. Y finalmente se concluye que el algoritmo con mayor número de aciertos es Jaro Winkler.

Abstract

The development of research projects is an activity that seeks to discover new knowledge, apply current knowledge to solve problems in society. The objective of this research is to recommend scientific articles for the formulation of research projects with text mining techniques. In the first stage of this process, it is the retrieval of information, for this the method of manual data collection has been used, the scientific articles and research theses have been selected randomly and classified by areas of research in computing. The second stage details the experimentation process with the three algorithms proposed by Jaro Winkler, TF-IDF and Word2vec. As part of the third stage, it is the data extraction with the developed algorithms that return quantitative values of relevance and/or similarity of texts as results. In the fourth stage, data mining is applied with what was extracted in the previous stage to carry out the analysis of the processed documents. And finally it is concluded that the algorithm with the highest number of hits is Jaro Winkler.

Introducción

Para aprender a formular un proyecto de investigación, comprende de diversas etapas formativas, de las cuales deben comenzar a desarrollarse desde los primeros años de vida, cosechando un espíritu de observación, curiosidad, indagación y crítica de la realidad (Ortiz Rivera, s.f.). Una sólida formación y un creciente conocimiento sobre un área especifica de la realidad, es la base para poder aportar nuevos conocimientos.

Para un estudiante universitario empezar a investigar sin tener una base de formación sólida o el conocimiento sobre un área especifica, se hace largo el proceso de aprendizaje, y el desarrollo de un trabajo de investigación. Actualmente es una política de las universidades Peruanas desarrollar una investigación en todas las áreas del conocimiento aplicada al desarrollo, con la finalidad de cumplir su rol en la sociedad, como los centros de investigación que están vinculadas a las instituciones educativas, al ámbito político y económico.

Hoy en día la información científica se encuentra en aumento, es importante seguir de cerca y estar al día en diversos métodos de aprendizaje, o el dominio de una metodología de investigación científica, existen diferentes plataformas tecnológicas que contribuyen a este aprendizaje, como la base de datos bibliográficos te resumes y artículos de revistas científicas.

El presente trabajo ha sido desarrollado en mi alma mater, la Universidad Nacional de Moquegua (UNAM), y se encuentra estructurada de la siguiente manera:

En el primer capítulo se documenta el planteamiento y definición del problema, donde se explica la descripción de la realidad problemática, los objetivos alcanzar, justificación e Importancia y variables.

En el segundo capítulo se explica el marco teórico, el cual comprende los antecedentes de la investigación, las bases teóricas y definición de términos necesario para comprender el estudio de esta investigación.

En el tercer capítulo se da a conocer el tipo y nivel de la metodología de investigación, el ámbito temporal, espacial, población, muestra, técnicas e instrumentos de recolección de datos, procedimientos y análisis de datos.

Finalmente, en el capítulo cuarto se documenta todos los resultados, conclusiones y recomendaciones de la misma experimentación realizada para cumplir con los objetivos plateados en el proyecto.

Capítulo I

Planteamiento del problema

1.1. Descripción y formulación del problema

El punto de partida de un proyecto de investigación es la identificación del problema, pero este problema puede haber sido abordado por otros investigadores en el contexto nacional o internacional, por lo que resulta de suma importancia identificar los antecedentes del mismo para determinar en que situación se encuentra la solución del problema. A partir de ello, se puede plantear un nuevo proyecto de investigación cuyos resultados serán innovadores y contribuirán en el avance de la solución del problema. En ese sentido, en este proyecto se propone la implementación de la RECOMENDACIÓN DE ARTÍCULOS CIENTÍFICOS UTILIZANDO TÉCNICAS DE MINERÍA DE TEXTOS PARA FORMULACIÓN DE PROYECTOS DE INVESTIGACIÓN, para que un estudiante pueda identificar con mayor facilidad los antecedentes de un problema de investigación.

En el contexto Peruano, a través de Superintendencia Nacional de Educación Superior Universitaria (SUNEDU), recientemente se implementó políticas de control en la calidad de investigación en las universidades del estado, una de las características de esta política es el desarrollo obligatorio de una investigación

para la obtención de grado y título de una profesión. Por lo tanto, es muy importante realizar estudios en este aspecto, porque antes de esta política existió otras modalidades de obtención del grado, sin exigencia de una investigación, esta realidad es contrastante para los nuevos estudiantes universitarios. Es decir el sistema educativo Peruano en la actualidad esta en proceso de formación a la nueva generación de estudiantes con habilidades de investigación y para asegurar aquello se debe hacer estudios de la realidad, también desde punto de vista de tecnología debe hacerse estudios para como ayudar a aprender a desarrollar en ejecutar proyectos de investigación sin mayor problema en diferentes aspectos de la vida de un estudiante universitario.

El "I Censo Nacional de Investigación y Desarrollo a Centros de Investigación" realizado el año 2016 por el Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (Concytec) y el Instituto de Estadística e Informática (INEI), se determinó que en el Perú, solo existían 3,374 investigadores en el 2015, de los cuales el 71 % desarrollaba actividades de investigación y desarrollo (I+D) en universidades públicas (35 %) y privadas (36 %), confirmaron que, a pesar de los esfuerzos de los últimos años, el Perú se ubica en el último puesto respecto a sus pares internacionales en Investigación + Desarrollo (Luis Cervantes Liñán, 2019).

Si se analiza el número de investigación que se dedican a I+D en países de la región miembros de la Organización para Cooperación y el Desarrollo Económico (OCDE), se observa que en Argentina, en el 2015, existían 82,407 investigadores (MCTeIP, 2017); en Brasil, 329 706 en el 2016 (MCTIC, 2018); en Chile, 6000 en el 2013 (Conicyt, 2015); en Colombia, 13 001 en el 2017 (OCyT, 2018); en Costa Rica, 3753 en el 2016 (MICITT, 2017); y en México, en el 2016 se encontraban registrados 6,2 millones de personas en el Acervo de Recursos Humanos en Ciencia y Tecnología que trabajan en Ciencia y Tecnología, de manera que existen 64,4 investigadores por cada 1000 integrantes de la población económicamente activa (PEA) (Conacyt, 2017) (Luis Cervantes Liñán, 2019).

Estas cifras ubican al Perú por debajo del promedio de los países de la región y por debajo del promedio de América Latina y El Caribe en cuanto al número de investigadores, ya que solo hay 0,2 investigadores por cada mil integrantes de la PEA (Concytec, 2017). Por otra parte, el número de investigadores en el Perú disminuye aún más si se tiene en cuenta que a la fecha solo hay 2289 investigadores calificados como tal por Concytec que forman parte del Registro de Investigadores en Ciencia y Tecnología del Sistema Nacional de Ciencia, Tecnología e Innovación Tecnológica (Regina, 2018) (UNIGV, 2019).

Visto las cifras y estadísticas dadas en el estudio (Luis Cervantes Liñán, 2019) con datos de Concytec, organismo regulador de investigación en el Perú, se evidencia el problema de poca cantidad de desarrollo de la investigación en Perú, por tanto la baja cantidad de profesores investigadores en las universidades, por con-secuente la falta preparación de estudiantes para investigación es un problema serio.

Los estudiantes universitarios Peruanos en su gran mayoría carecen de una sólida formación en la formulación y ejecución de proyectos de investigación para optar el título profesional. Estas dificultades se presentan debido a la poca información y conocimiento de áreas de investigación en el círculo universitario.

Una de las causas es la ausencia de asesoramiento al estudiante, esto sucede debido a la poca experiencia o formación en la investigación científica. Es así que en efecto los estudiantes en el proceso de elaboración de un Plan de tesis tienen dificultades para identificar y redactar un problema de investigación.

En el otro contexto "la lista de nuevas tecnologías crece cada día. Robots, realidad aumentada, algoritmos y las comunicaciones de máquina a máquina ayudan a las personas en una gran variedad de tareas. Estas tecnologías son de amplio espectro y trascendentales por su potencial para transformar las empresas y las vidas humanas. Tienen la capacidad de facilitar la vida de los individuos y

mejorar sus relaciones personales y laborales. La tecnología se vuelve cada vez más avanzada y ello tiene una repercusión considerable en la población activa"(West, s.f.).

Por otra parte, el mundo ha tenido grandes avances científicos en diferentes áreas, es decir los científicos del mundo desarrollan investigaciones, cuyo resultados producto final es un artículo científico. Para realizar el planteamiento de un nuevo proyecto de investigación se debe estudiar el estado de arte de una linea de investigación, por el estado de arte se comprende estudiar muchos artículos científicos de la linea de investigación de interés, y los artículos científicos se encuentran en plataformas como IEEE library, ACM library, Scopus, Scielo y otros en el ámbito de computación.

El presente proyecto pretende proponer un método de ayuda a los estudiantes, la recomendación de artículos científicos basada en su idea inicial de la propuesta de proyecto de tesis. Es decir, el estudiante inicia con la definición del problema, descripción del posible escenario de investigación, métodos y técnicas para resolver un problema. La recomendación que se plantea está basada en procesar con algoritmos de Procesamiento de Lenguaje Natural (NLP), los términos escritos en la idea inicial del estudiante para identificar los artículos científicos con contenido de términos muy similares a la idea inicial del estudiante. La finalidad es reforzar su conocimiento del estado de arte con los artículos científicos, el cual contiene información acerca de metodologías, técnicas, y diferentes temas relacionados a un área de investigación, de tal forma que el estudiante obtienen mayor conoci-miento sobre el estado de arte del área de investigación en la que desea formular una tesis de alta calidad e impacto.

1.2. Objetivos

1.2.1. Objetivo general

Recomendar artículos científicos para la formulación de proyectos de investigación con técnicas de minería de textos.

1.2.2. Objetivos específicos

- Implementar algoritmos de minería de textos para recomendación de artículos científicos para la formulación de proyectos de investigación.
- Comparar las técnicas de minería de textos para recomendación de artículos científicos a través de diversas métricas de modelos de recomendación.

1.3. Justificación e importancia

La presente tesis se justifica que es importante, porque busca experimentar métodos y técnicas de procesamiento lengua natural para ayudar a los estudiantes y egresados en el planteamiento de un proyecto de investigación de calidad en la Universidad Nacional de Moquegua y en lo posterior a estudiantes de otras universidades del país.

El presente caso esta basada en el uso de inteligencia artificial para resolver problemas de planteamiento de proyectos de investigación, usando experiencias pasadas similares, los algoritmos buscan descubrir la similaridad de propuestas en la base de artículos científicos. En este sentido sería una gran herramienta de apoyo desde un punto de vista científico, metodológico y técnico. Además, los resultados de la tesis serán un gran aporte a la comunidad científica, sobre todo a la comunidad peruana, universidades, autoridades, quienes buscan indicadores y

herramientas para mejorar la calidad de investigación en las universidades del país, así mismo la innovación tecnológica, de tener buenos resultados, el presente trabajo se convertiría en una herramienta de mucha utilidad.

1.4. Variables

- **Variable independiente**: Recomendación de artículos científicos.
- **Variable dependiente**: Formulación de proyectos de investigación.

Capítulo II

Marco Teórico

2.1. Antecedentes

Los antecedentes de estudio relacionados al tema de investigaciones más similares se describen a continuación:
En el trabajo (Sun y cols., 2013), los autores proponen la integración de tres tipos de medidas de similaridad para solucionar el problema de resultados de búsqueda inadecuados para los investigadores. Estas medidas de similaridad son, la similaridad de la palabra clave, similaridad de el artículo y la similaridad del autor. La medida de similaridad integrada se utiliza para hacer un ranking de artículos y seleccionar los más adecuados para el autor.

Para abordar algunos de los problemas que surgen de la falta de coincidencia entre las diferentes formas de las palabras utilizadas en las consultas y los documentos relevantes, los investigadores han propuesto durante mucho tiempo el uso de varios algoritmos de derivación para reducir los términos a una forma base común, la mejora casi lineal en el rendimiento con el aumento de la cobertura sugiere que pueden ser posibles mejoras adicionales en la clasificación TFIDF para IR al incluir léxicos especializados y específicos de dominio en las clases de raíz (Kantrowitz, Mohit, y Mittal, 2000).

Se usa TFIDF para volver a ponderar las características de conteo para reducir la influencia de términos más frecuentes pero menos valiosos y mejorar la influencia de términos más raros pero más valiosos. A través del modelado de temas con asignación de Dirichlet latente, se extraen diez temas y las tres palabras características correspondientes, y cada tema se caracteriza por las tres palabras especiales (Zhao, Liu, Zhang, y Wang, 2018).

Una forma de reducir el plagio es detectar la similitud de los textos de las oraciones en un documento. Hay varios programas de detección de similitudes, incluidos Turnitin, Eve2, CopyCatchGold, WodCheck, Glatt. Un algoritmo utilizado para detectar similitudes de texto es Jaro Winkler Distance. Este algoritmo tiene una buena precisión al hacer coincidir una cadena relativamente corta y puede verificar con precisión las copias de los documentos después del proceso de tokenización (Manaf y cols., 2019). El algoritmo de Jaro-Winkler (J-W) es un comparador de cadenas bien conocido y ampliamente utilizado que se emplea a menudo en estos problemas de vinculación de registros. A pesar de la popularidad de J-W, no existe un consenso académico sobre el umbral de puntuación en el que las cadenas deben considerarse coincidencias probables. En la práctica, los umbrales J-W se seleccionan de forma algo arbitraria o con poca justificación. Este documento utiliza un enfoque de simulación para identificar los umbrales J-W óptimos en función de las longitudes de cadena de un par de entidades, mejorando así el proceso de descubrimiento de clientes potenciales para los profesionales de CTFI (Kloo, Dabkowski, y Huddleston, 2019).

Los algoritmos tradicionales de detección de texto largo duplicado son difíciles de aplicar en las situaciones actuales, por lo que se necesita un algoritmo de detección de duplicado más efectivo para texto corto. Basado en el modelo de bolsa de palabras Word2vec, este artículo propone una especie de algoritmo de detección de duplicados con semántica embebida para texto breve. Las palabras se

incrustan en vectores que son elementos de entrada en el algoritmo Simhash para adquirir una secuencia de 64 bits, luego comparar dos secuencias con distancias de Hamming y devolver el resultado filtrado por valor de umbral preestablecido (Gao, He, Zhang, y Xia, 2017). Con el fin de mejorar el grado de coincidencia de texto y la precisión de cálculo del método de clasificación de texto breve, este artículo estudia la optimización del método de clasificación de texto breve del algoritmo Word2Vec (Lei, 2020).

2.2. Bases teóricas

2.2.1. Inteligencia artificial

La Inteligencia Artificial (IA) es un campo de la ciencia y la ingeniería que se ocupa de la compresión, desde el punto de vista informático, de lo que denomina comúnmente comportamiento inteligente. También se ocupa de la creación de artefactos que exhiben este comportamiento.
En cualquier caso, desde distintas perspectivas, en el seno de la IA como ciencia y tecnología se han ido acumulando conocimientos sobre como emular las diversas capacidades del ser humano para exhibir comportamientos inteligentes y se han desarrollado sistemas cada vez más perfeccionados que reproducen parcialmente dichas capacidades. Así, el estudio de sensores y mecanismos de intercambio de información con el exterior constituyen, por sí mismos, pujantes al área de investigación y aplicaciones prácticas. También la robótica, como especialidad relacionada con máquinas móviles capaces de interactuar inteligentemente con su entorno y manipular objetos, suele considerarse parcialmente incluida en el ámbito de la IA. Y naturalmente el conjunto da habilidades cognitivas e intelectivas realizadas por el cerebro humano para comprender el mundo que rodea a una persona, interpretar sus señales e informaciones, resolver los constantes

problemas de diverso tipo que plantea el quehacer diario, representar y acumular experiencia y conocimiento, aprender y generar nuevo conocimiento a partir del bagaje disponible mediante diversos modos de razonamientos (Díez, Gómez, y de Abajo Martínez, 2001).

La IA se representa por diversas disciplinas, con mayor o menor medida. A continuación hablaremos brevemente de algunas de sus ramas:

2.2.1.1. Minería de Textos

La minería de textos busca extraer información útil e importante de formatos de documentos heterogéneos, tales como páginas web, correos electrónicos, medios sociales, artículos de artículos, etc. Es una forma específica de minería de datos que se relaciona con el texto, es decir que el resultado de minería de texto es minaría de datos.

La minería de texto se encarga de aprovechar y desarrollar métodos automatizados para procesar documentos digitales no estructurados de forma rápida y eficiente, con la finalidad de organizarlos y analizarlos. Por lo que en este estudio se emplea la minería de texto en el procesamiento de documentos digitales para ser usada en la clasificación de documentos no estructurados partiendo de la siguiente hipótesis: La minería de texto facilita el procesamiento de documentos digitales por medio de los métodos de extracción de información y clasificación, permitiendo la organización de documentos. (Di Giuda y cols., 2020).

La extracción de textos puede dividirse en cinco pasos:

- Recolección: Recopilación de datos de diferentes recursos, tales como sitio web, correos electrónicos, comentarios de clientes, archivo de documentos. Dependiendo de la aplicación, este proceso puede ser completamente automatizado o guiado por una persona encargada de realizar este proceso.

- Preprocesamiento: La identificación del contenido y la extracción de características representativas
- Limpieza de textos: Eliminación de cualquier información innecesaria o no deseada, como los anuncios de las páginas.
- Tokenización: Un ordenador solo «ve» una cadena de caracteres, sin poder identificar, por ejemplo, párrafos, frases o palabras. La Tokenización divide el texto en entidades significativas (palabras, oraciones, etc.) dados los espacios en blanco presentes y las puntuaciones.
- Extracción de características (también llamada selección de atributos): es el proceso de caracterización.

2.2.1.2. Procesamiento Lenguaje Natural

El procesamiento del lenguaje natural (NLP) es un proceso de transformación de la información del texto en datos numéricos (Di Giuda y cols., 2020).

Un aspecto fundamental a considerar es que la mayor parte del conocimiento humano se codifica en lenguaje natural escrito o hablado, por lo que la compresión de como manejar el lenguaje y cómo construir programas de ordenador que lo utilicen con la misma habilidad y soltura que los seres humanos es crucial para el avance de la inteligencia artificial (IA).

Este es uno de los campos en que los primeros investigadores de la IA creyeron que se conseguirían avances mucho más rápidamente de lo que se han producido en la realidad. Hace años que tanto los especialistas como los potenciales usuarios esperan tener sistemas de traducción automática de carácter general que agilicen los procesos administrativos y comerciales internacionales. Otra meta es ese tipo de sistemas es permitir la interacción de las personas con los ordenadores en el lenguaje cotidiano y así poder hacer consultas a las bases de datos mediante el

tipo de preguntas o peticiones que se usarían con otras personas y no teniendo que escribir complejas instrucciones.

En NLP las computadoras analizan el lenguaje humano, lo interpretan y dan significado para que pueda ser utilizado de manera práctica. Usando NLP podemos hacer tareas como resumen automático de textos, traducción de idiomas, extracción de relaciones, Análisis de sentimiento, reconocimiento del habla y clasificación de artículos por temáticas.

Algunos de los usos más frecuentes:

- Resumen de textos, el algoritmo deberá encontrar la idea central de un artículo e ignorar lo que no sea relevante.
- ChatBots, deberán ser capaces de mantener una charla fluida con el usuario y responder a sus preguntas de manera automática.
- Generación automática de keywords y generación de textos siguiendo un estilo particular.
- Reconocimiento de Entidades, como encontrar Personas, Entidades comerciales o gubernamentales ó Países, Ciudades o marcas.
- Análisis de Sentimientos, deberá comprender si un tweet, una review o comentario es positivo ó negativo y en qué magnitud (ó neutro). Muy utilizado en Redes Sociales, en política, opiniones de productos y en motores de recomendación.
- Traducción automática de Idiomas
- Clasificación automática de textos en categorías pre-existentes ó a partir de textos completos, detectar los temas recurrentes y crear las categorías.

Resumiendo, podemos procesar de diversas maneras al lenguaje, sus componentes como gramática, sintaxis e intentar crear estructuras de apoyo que nos servirán

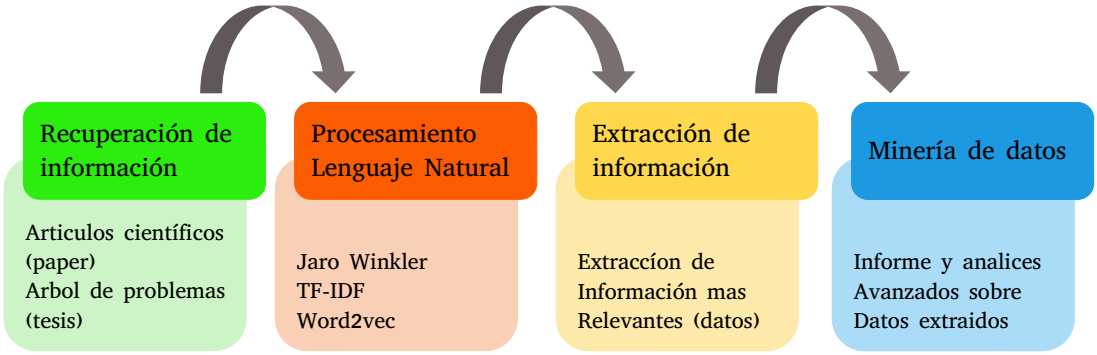
como entradas para aplicar modelos de inteligencia artificial. El procesamiento del lenguaje natural (NLP) es un proceso de transformación de la información del texto en datos numéricos (Di Giuda y cols., 2020).

2.2.2. Metodología

Es necesario e importante seguir una metodología que nos permita planear estratégicamente los hitos del trabajo hasta cumplir con el objetivo plantada. Sin embargo, en términos generales, el objetivo general de la minería de textos es convertir el texto en datos para que sea adecuado para el análisis (Zanini y Dhawan, 2015). El presente trabajo esta basado en la siguiente metodología:

Figura II.1:

Metodología de investigación para proyectos Minería de Textos.



2.2.3. Algoritmos propuestos

2.2.3.1. Jaro Winkler

La similitud de Jaro-Winkler, a menudo llamada erróneamente distancia de Jaro-Winkler, es una medida de similitud para dos cadenas propuesta en (Winkler, 1990). Se basa en la similitud de Jaro. donde $|s1|, |s2|$ son las longitudes de ambas

cadenas, m es el número de caracteres coincidentes y t es el número de transposiciones. Los caracteres coincidentes son caracteres comunes en ambas cadenas con una distancia máxima (Keil, 2019).

$$Sim_j(s_1, s_2) = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}(\frac{m}{s_1} + \frac{m}{s_2} + \frac{m-t}{m}) & \end{cases} \quad \text{(II.1)}$$

Un carácter coincide como máximo con un carácter en la otra cadena, seleccionado seleccionando el primer candidato durante una iteración anidada sobre ambas cadenas. El número de transposiciones t es la mitad del número de posiciones desiguales en las cadenas concatenadas de todos los caracteres coincidentes en el orden de aparición original (Keil, 2019).

2.2.3.2. Término Frecuencia - Frecuencia inversa del documento (TF-IDF)

Frecuencia del término-Frecuencia inversa del documento (TF-IDF) representa una relevancia numérica de un término para cada documento y conjunto de documentos (corpus) (Apaza, Rubin de Celis Vidal, y Chire Saire, 2021). El valor TF-IDF es proporcional a la frecuencia con el que se repite el término en un documento, pero es equilibrada por el número de documentos en que se repite el mismo término, esto hace que algunos términos sean más populares que otros en un determinado corpus de datos (Apaza Alanoca, 2019).

$$tf_idf_{t,d} = tf_{t,d} \cdot \log \frac{N}{df_t} \quad \text{(II.2)}$$

Donde:

tf_idf = Término Frecuencia - Frecuencia inversa del documento

$tf_{t,d}$ = Frecuencia ocurrencias de término t en documento d.

N = Total número de documentos.

df_t = Número de documentos con el termino t.

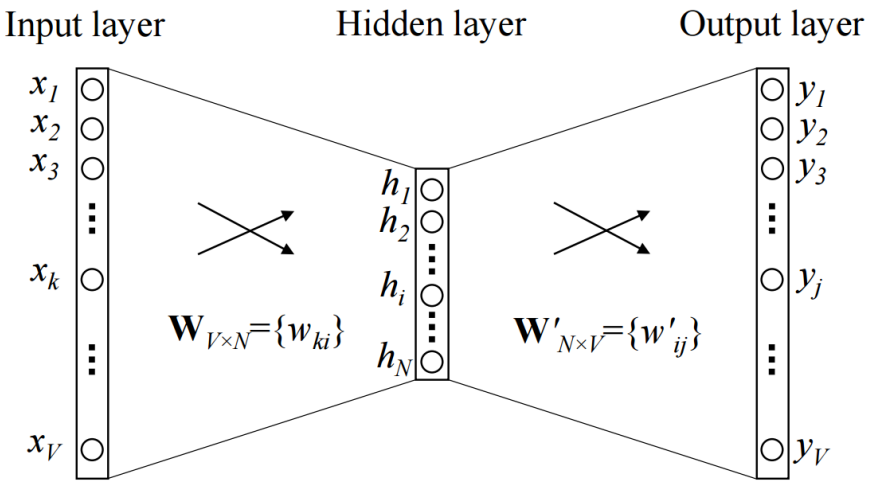
2.2.3.3. Word2vec

Word2vec toma un corpus de texto como entrada y produce los vectores de palabras como salida. Primero construye un vocabulario a partir de los datos del texto de entrenamiento y luego aprende la representación vectorial de las palabras. El archivo vectorial de palabras resultante se puede utilizar como características en muchas aplicaciones de procesamiento de lenguaje natural y aprendizaje automático (Mikolov, Chen, Corrado, y Dean, 2013) (Mikolov, Sutskever, Chen, Corrado, y Dean, 2013).

Los siguiente párrafos de conceptos están basados en (Karani, 2018) y (Rong, 2016), Word2Vec es un método para construir tal incrustación. Se puede obtener usando dos métodos (ambos involucrando redes neuronales): Skip Gram y Common Bag Of Words (CBOW)

Figura II.2:

Arquitectura de modelo CBOW.



Nota: (Rong, 2016)

La entrada o la palabra de contexto es un vector codificado en caliente de tamaño V. La capa oculta contiene N neuronas y la salida es nuevamente un vector

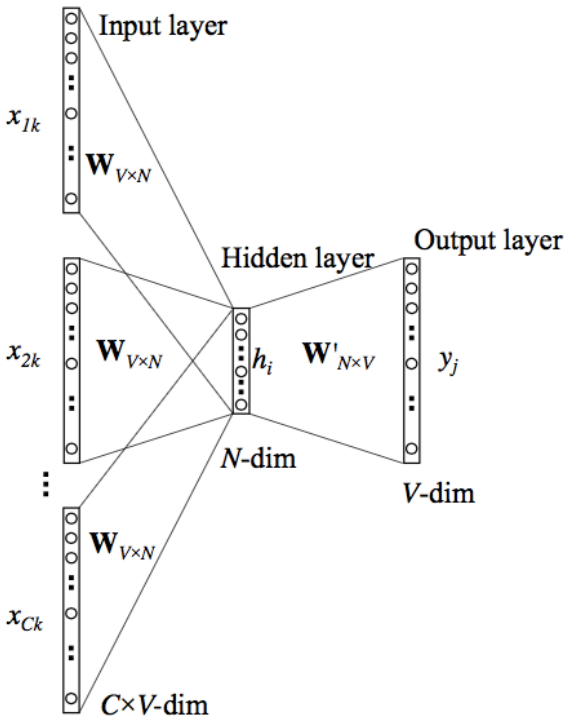
de longitud V con los elementos siendo los valores softmax. Veamos los términos de la imagen correctamente:

- $W\,vn$ es la matriz de ponderación que asigna la entrada x a la capa oculta (matriz dimensional $V * N$)
- $W\text{‘}nv$ es la matriz de ponderación que asigna las salidas de la capa oculta a la capa de salida final (matriz dimensional $N * V$)

Las neuronas de la capa oculta simplemente copian la suma ponderada de entradas a la siguiente capa. No hay activación como sigmoid, tanh o ReLU. La única no linealidad son los cálculos de softmax en la capa de salida. Pero, el modelo anterior usó una sola palabra de contexto para predecir el objetivo. Podemos usar varias palabras de contexto para hacer lo mismo.

Figura II.3:

Modelo continua de bag-of-word.



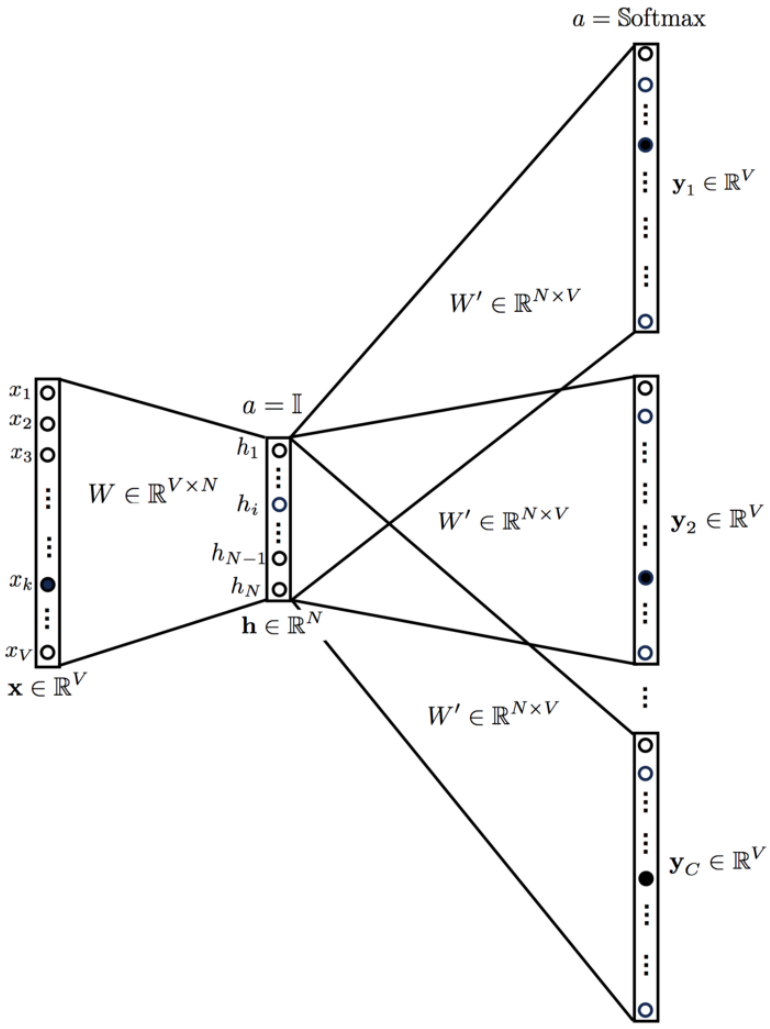
Nota: (Rong, 2016)

El modelo anterior toma palabras de contexto C. Cuando se usa $W\,vn$ para

calcular entradas de capas ocultas, tomamos un promedio sobre todas estas entradas de palabras de contexto C. Entonces, hemos visto cómo se generan representaciones de palabras usando las palabras de contexto. Pero hay una manera más en la que podemos hacer lo mismo. Podemos usar la palabra objetivo (cuya representación queremos generar) para predecir el contexto y, en el proceso, producimos las representaciones. Otra variante, llamada modelo Skip Gram, hace esto. Parece que el modelo CBOW de contexto múltiple acaba de cambiarse. Hasta

Figura II.4:

Modelo Skip-Gram.



Nota: (Rong, 2016)

cierto punto eso es cierto. Introducimos la palabra objetivo en la red. El modelo genera distribuciones de probabilidad C. ¿Qué significa esto?

Para cada posición de contexto, obtenemos C distribuciones de probabilidad de V probabilidades, una para cada palabra.

En ambos casos, la red utiliza la retropropagación para aprender. Para mayor detalle de la matemática del modelo Wor2vec se puede ver en (Rong, 2016).

2.2.4. Contexto de investigación científica en el Perú

El 9 de julio del 2014 se crea la ley universitaria N° 30220, que tiene como objetivo de crear, funcionamiento, supervisión y cierre de universidades, promueve el mejoramiento de la calidad de educación en las universidades del Perú, principalmente el desarrollo de la investigación y de la cultura.
En el 2016, el Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (Concytec), institución rectora del Sistema Nacional de Ciencia y Tecnología e Innovación Tecnológica del Perú, realizó el Ï Censo Nacional de Investigación y Desarrollo a Centros de Investigación en el Perú"(Concytec, 2017). Los resultados evidenciaron el poco desarrollo de la investigación y desarrollo tecnológico a nivel nacional respecto a los países de América Latina miembros de la Organización para la Cooperación y el Desarrollo Económico (OCDE), de la cual Perú desea formar parte.

En este sentido, el objetivo de este trabajo es analizar la situación de la investigación y el desarrollo en el Perú, en comparación con otros países de la región que son miembros de la OCDE, de acuerdo con el rol de las universidades en la creación de conocimiento. Para esto se analizan indicadores tales como: recursos humanos dedicados a la investigación y el desarrollo, inversión en investigación y desarrollo y producción científica, entre otros (Luis Cervantes Liñán, 2019).
Al analizar la producción científica de diferentes países, se observa que existe una relación entre la inversión en I+D y la generación de conocimiento. La producción científica de Brasil (es el país que más invierte en I+D en la región) en el 2017

representó el 52,82 % de toda la producción científica de América Latina, de modo que logran la publicación de 68 741 artículos científicos en artículos indexadas en Scopus (MCTIC, 2018). Asimismo, México logró en el 2016 la publicación de 12 000 artículos científicos en artículos indexadas (Conacyt, 2017), y Argentina logró la publicación de 8638 artículos científicos en artículos indexados en el 2015 (MCTeIP, 2017)(Luis Cervantes Liñán, 2019).

2.2.4.1. Producción científica en la universidades

La producción científica de una institución depende de varios aspectos, siendo el financiamiento un factor importante, ya que posibilita realizar investigaciones más complejas o publicar en artículos de mayor impacto que en algunos casos requieren pago por publicación. En Perú, las universidades públicas reciben de los gobiernos regionales hasta un 20 % del total de dinero percibido por canon y este debe ser destinado exclusivamente a la inversión en investigación científica y tecnológica que potencien el desarrollo regional (Purizaca-Rosillo, Cardoza-Jiminez, y Herrera-Añazco, 2016).

La mitad de docentes universitarios de investigación de escuelas de Medicina del Perú nunca han publicado un artículo en una revista, tres de cada cuatro nunca han publicado un original en una revista indexada en Medline y solo uno de cada seis lo hicieron en los dos últimos años. Los factores asociados con haber publicado un artículo original en Medline fueron el tener menos de 40 años, ser profesor de una universidad en la que la tesis es obligatoria para titularse y ser un profesor de una universidad productiva científica (Pereyra-Elias, Huaccho-Rojas, Taype-Rondan, Mejia, y Mayta-Triston, 2014).

Una de las principales características de las tesis desarrolladas entre los años 2014 y 2018 en las tres universidades en estudio es la orientación hacia el enfoque cuantitativo, de diseño no experimental con alcance descriptivo simple, lo cual

implica que la investigación solo se centra en el aspecto teórico y no plantean soluciones concretas de los problemas que la sociedad requiere (Incio, Morales, Cruzata-MartÃnez, y Ãguila, 2020).

La investigación abre un nuevo e interesante mundo en el cual se puede experimentar un continuo mejoramiento personal y profesional, y es una garantía de calidad para las instituciones de educación superior. Se hace necesario fomentar una cultura investigativa desde el pregrado en las universidades de la región latinoamericana en general, y particularmente en Ecuador, que forme y desarrolle estas habilidades a fin de potenciar la producción científica en vínculo con la solu-ción de las problemáticas sociales (Daher Nader, Panunzio, y Hernandez Navarro, 2018).

Por diversas razones no se alcanza buen nivel de investigación, pero una tarea fundamental para tener alcanzar el buen nivel de investigación es tomando en conocimiento el estado arte, es decir adquirir conocimiento de otras investigaciones de una misma línea de investigación.

2.3. Definición de términos

- Regresión Lineal, es una técnica estadística utilizada para estudiar la relación entre variables.
- Regresión Logística, la regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable.
- Naive Bayes, es un algoritmo clasificador probabilístico simple con fuerte suposición de independencia.
- Lineal support machire learning, estos algoritmos de machine learning son el alma que mueven los procesos de aprendizaje, el cual podemos

obtener la información que necesitamos para tomar decisiones o predecir el comportamiento de los datos.

- Árbol de problema, es una técnica que se emplea para identificar una situación problemática (un problema central).

- Redes Neuronales, son un modelo de creación cuyo sistema se basa en el funcionamiento del cerebro humano, formadas por diferentes nodos que funcionan como neuronas, y que transmiten señales e información entre sí.

- Minería de Datos, es el proceso de hallar anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados.

- Algoritmos, Un algoritmo es una serie de pasos organizados, que describe el proceso que se debe seguir, para dar solución a un problema específico.

- Similitud, Semejanza que presenta una producción científico-intelectual con otra u otras, orienta el desarrollo de la tesis o trabajo de investigación.

Capítulo III

Metodología

3.1. Tipo y nivel de investigación

El presente investigación es de tipo aplicativo, porque se prueban métodos (algoritmos) existentes en el estado de arte de inteligencia artificial, específicamente en la disciplina procesamiento lenguaje natural.

El nivel de investigación del presente se define como descriptivo, porque se estiman frecuencias, probabilidades y se estiman parámetros de confianza.

3.2. Ámbito temporal y espacial

La investigación es realizada en el ámbito temporal entre 2017 a 2020 años de calendario, en el espacio del continente de América. Porque se obtuvo un total de 500 artículos científicos desarrollados en continente en mención y 97 casos de árbol de problemas correspondientes a las tesis de titulados realizados en el Perú. Estos documentos fueron desarrollados entre los años 2017 a 2020.

3.3. Población y muestra

La población de la presente investigación se considera 597 documentos de investigación, 500 corresponden a artículos científicos publicados en diferentes revistas y 97 pertenecen a tesis de titulados con el contenido de árbol de problemas. La experimentación se realizó con 597 documentos, por lo tanto, se considera el total de la población como muestra de la presente investigación..

3.4. Técnicas e instrumentos de recolección de datos

La recolección de datos se realiza de forma manual, bajo el criterio del autor. Los instrumentos y técnicas usadas en el presente investigación son los algoritmos de procesamiento lenguaje natural; Jaro Winkler, Término Frecuencia - Frecuencia inversa del documento (TF-IDF) y Word2vec, los cuales no permiten medir la similitud de documentos. La fiabilidad de estas técnicas son validadas por autores propios de las técnicas y por otros autores de la comunidad científica que usaron y evidenciaron su eficacia.

3.5. Procedimientos

En esta sección se documenta todo el procedimiento de las etapas planificados según la metodología presentada en el anterior capítulo del presente informe. Las fases se dividen en cuatro principales etapas de minería de textos.

3.5.1. Recuperación de información

La primera etapa de este proceso, es la recuperación, estructuración de la información, para alcanzar este objetivo se ha usado el método de recolección

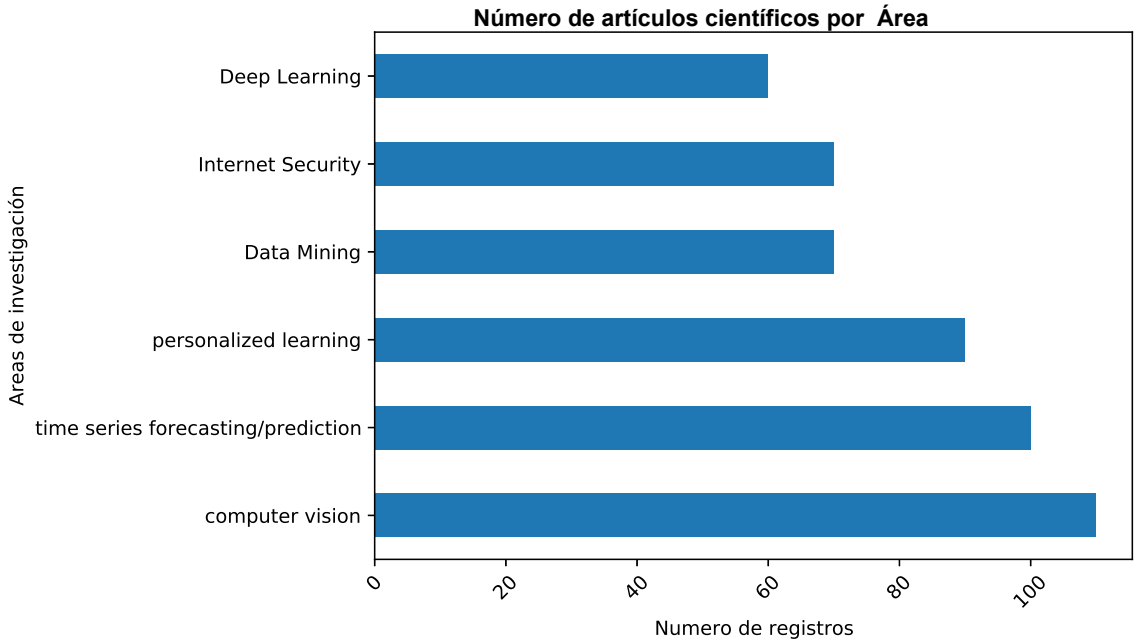
manual de los datos. Se ha realizado el trabajo de ingresar manualmente a los artículos científicos de IEEE, ACM, SCielo y Scopus, en donde sea realizado la búsqueda de manera manual y se ha seleccionado artículos de manera aleatoria y clasificando por áreas de investigación en computación. En el presente trabajo se experimentó con dos tipos información, las cuales son:

- Artículos Científicos:

  La primera set de datos corresponde a recolección de artículos científicos de 6 áreas de computación: Computer vision, Personalized Learning, Time series forecasting / predicction, Internet Security, Data Mining y Deep Learnin, como se puede apreciar en la Figura III.1.

Figura III.1:

Artículos recolectadas.



Los documentos de artículos constan de varias partes importantes, considerando el contexto de la propuesta del presente trabajo se ha seleccionado los siguientes campos: Área, Titulo, Keywords y Abstrac, como se puede apreciar en la Figura III.2. Estos campos poseen textos muy descriptivos del contexto de cada artículo científico, es decir describe el

problema que abor-da, los métodos o algoritmos usados en el artículo científico, estos términos son muy importantes porque nos permite identificar los algoritmos usados, áreas de investigación y otros de interés del buscador.

Figura III.2:

Base de datos de Artículos recolectadas.

Area	Titulo	Keywords	ABSTRACT
computer vision	Vision-Face Recognition Attendance Monitoring ...	Face Attendance, Face detection, face recognit...	Nowadays, Artificial Neural networks can be tr...
computer vision	Cost effective and reliable mobile solution fo...	Wireless body area networks, physical layer se...	Attendance marking is a critical and time-cons...
computer vision	Development of model for automatic tutor in e-...	Face detection, face recognition, viola Jones,...	In e-learning environment, the main function o...
computer vision	An Advancement towards Efficient Face Recognit...	Face recognition, Live video feed, face detect...	Image based or live video feed based face reco...
computer vision	A Survey on Computer Vision for Assistive Medi...	Computer vision, face analysis, facial symptom...	Automatic medical diagnosis is an emerging cen...
...	...	...	...
Deep Learning	Applying Deep Learning to Balancing Network In...	network intrusion detection , deep learning , ...	In this investigation, we apply deep learning ...
Deep Learning	Deep Learning for Augmented Reality	deep learning , augmented reality , tracking ,...	Augmented reality aims to mix real-world visua...
Deep Learning	Traffic signs recognition with deep learning	Artificial Intelligence , artificial Neural Ne...	In this paper, a deep learning based road traf...
Deep Learning	Deep Learning with Multi Modal Ensemble Fusion...	Deep Learning , Multi modal Ensemble Fusion , ...	Epilepsy is a neurological condition affecting...
Deep Learning	Iterative Deep Learning Based Unbiased Stereol...	Unbiased Stereology, Active learning, Deep lea...	Lack of enough labeled data is a major problem...

- Tesis de egresados de EPISI:

  Por otra parte se han recolectado tesis aprobados, sustentadas y publicadas pertenecientes a los egresados de la Escuela Profesional de Ingeniería de Sistemas e Informática de la Universidad Nacional de Moquegua, sede Ilo, porque el presente investigación busca recomendación de Artículos para formulación de proyectos de investigación.

  Los datos de tesis recolectados fueron validados y aprobados por diversas etapas del procedimiento por parte del jurado, consideramos que son datos idóneos y reales que se pueden poner en evaluación para recomendación de

nuevos artículos parecidos con algoritmos que en la presente investigación propone.

Los campos recolectados con titulo, área y abstrac de la tesis, como se puede apreciar en la Figura III.3.

Figura III.3:

Base de datos de tesis de la EPISI, UNAM.

Autor	Area	Titulo	abstract
Mamani Padilla, Diego Ismael	Data mining	Data mining model based on associated factors ...	The following research project aims to develop...
Valero GÃ³mez, Juan Carlos	Computer vision	Analysis of predictive models based on compute...	The following final thesis report has the purp...
Condori Bellido, Sadan Eusebio	Data mining	Data mining model for predicting cases of anem...	At present, anemia is a disease that affects 2...
Apaza Alanoca, Honorio	Text mining, Natural Language Processing	Text mining to build a data mining model for j...	In the XXI century digital technology and the ...
Castro Alvarez, Gregorio Tomas	Computer vision	Application of intelligent algorithms for auto...	We developed a research work: "IMPLEMENTATION ...
Mamain Laqui, Marcos Rufino	Data mining	Optimization of distribution unit routes apply...	The work presented in this document is aimed a...
Huerta Laura, Luis Enrique	Data mining	Prediction Model Of Customers Who Will unscrib...	The research arises from observing the behavio...
Machaca Huanca, Anamelba Lissette	Computer vision	Diagnosis of diabetic retinopathy using classi...	Diabetic Retinopathy is a complication of Diab...
Cruz Diaz, Jafet Andres	System information	Information System for the Control and Documen...	The present investigation entitled: â¯Informa...
Moron Cuaquira, Alonso	Data mining	Predition of irregular consumiption of drinkki...	This research work focuses on the experimentat...

3.5.2. Procesamiento Lenguaje Natural

En esta sección se detalla el proceso de experimentación con los tres algoritmos propuestos en el presente informe de investigación.

3.5.2.1. Jaro Winkler

Jaro Winkler mide en porcentajes la similitud de dos cadenas, más similares son las cuerdas. La puntuación se normaliza de manera que 0 significa una

coincidencia exacta y 1 significa que no hay similitud. En el presente caso aplicamos sobre las siguientes dos cadenas de texto: la primera cadena textos son los textos correspondientes al título y abstract de tesis, la segunda cadena corresponde a área, título, keywords y abstract de los papers, nuestro objetivo es medir en porcentajes la similitud de las cadenas de tesis vs. la cadena de textos de artículos científicos.

Listing III.1:

Codigo de programación de Jaro Winkler en python

```
def jaro_distance(s1, s2):
    if (s1 == s2):
        return 1.0
    len1 = len(s1)
    len2 = len(s2)
    max_dist = floor(max(len1, len2) / 2) - 1
    match = 0
    hash_s1 = [0] * len(s1)
    hash_s2 = [0] * len(s2)
    for i in range(len1):
        for j in range(max(0, i - max_dist),
                       min(len2, i + max_dist + 1)):
            if (s1[i] == s2[j] and hash_s2[j] == 0):
                hash_s1[i] = 1
                hash_s2[j] = 1
                match += 1
                break
    if (match == 0):
        return 0.0
    t = 0
    point = 0
    for i in range(len1):
        if (hash_s1[i]):
            while (hash_s2[point] == 0):
                point += 1
```

```
            if (s1[i] != s2[point]):
                point += 1
                t += 1
    t = t//2
    return (match/ len1 + match / len2 +
            (match - t + 1) / match)/ 3.0
```

3.5.2.2. Término Frecuencia - Frecuencia inversa del documento (TF-IDF)

TF-IDF (término frecuencia-frecuencia de documento inversa) es valor estadístico que evalúa la relevancia de términos de un documento en una colección de documentos, esta técnica tiene dos métricas importantes, primero el cálculo de la frecuencia de términos en el mismo documento, el segundo es la frecuencia inversa del documento de la palabra en la colección de documentos.

En la presente investigación se mide la relevancia de cada uno de los términos escritos en las tesis con relación a textos escritos en los artículos científicos, el objetivo es calcular promedio de relevancia de cada unos de los documentos de tesis en relación a documentos de artículos científicos.

Listing III.2:

Código de programación de TFIDF en python

```
#Calculando frecuencia de terminos
    def computeTF(wordDict, bow):
      tfDict = {}
      bowCount = len(bow)
      for word, count in wordDict.items():
          tfDict[word] = count/float(bowCount)
      return tfDict
    def computeIDF(docList):
      import math
      idfDict = {}
      N = len(docList)
```

```
    idfDict = dict.fromkeys(docList[0].keys(), 0)
    for doc in docList:
        for word, val in doc.items():
            if val > 0:
                idfDict[word] += 1
    for word, val in idfDict.items():
        idfDict[word] = math.log10(N / float(val))
    return idfDict
 idfs = computeIDF([wordDictR, wordDictA, wordDictB,
wordDictC, wordDictD])
 def computeTFIDF(tfBow, idfs):
    tfidf = {}
    relvants={}
    zumm=0
    for word, val in tfBow.items():
        tfidf[word] = val*idfs[word]
        if (tfidf[word]!=0):
          relvants[word]=val*idfs[word]
          zumm=zumm+tfidf[word]
    avrg=round(zumm/len(relvants),6)
    return (tfidf, relvants,avrg)
```

3.5.2.3. Word2vec

Word2Vec es una de las técnicas más populares para aprender incrustaciones de palabras utilizando redes neuronales superficiales (Mikolov, Sutskever, y cols., 2013). word2vec no es un algoritmo singular, sino una familia de arquitecturas modelo y optimizaciones que se pueden usar para aprender incrustaciones de palabras a partir de grandes conjuntos de datos. Las incrustaciones aprendidas a través de word2vec han demostrado ser exitosas en una variedad de tareas de procesamiento de lenguaje natural posteriores (www.tensorflow.org, 2022).

En el presente investigación usamos el word2vec de vectores de palabras

preentrenados desde gensimdata ($word_vectors = api.load("glove-wiki-gigaword-300")$), vectores pre-entrenados es un conjunto de datos de Google News (alrededor de 100 mil millones de palabras). El modelo contiene vectores de 300 dimensiones para 3 millones de palabras y frases (Mikolov, Suts-kever, y cols., 2013). Nuestro objetivo es consultar la similitud de términos es-critos entre los documentos de tesis y artículos científicos, se usó la función: $model.similarity(w1, w2)$, en donde w1 corresponde al término de tesis y w2 corresponde a términos de artículos científicos, la funcion model.semilarity mide la similitud de los términos candidatos.

Listing III.3:

Codigo de programación de Wor2vec en python.

```
from gensim.models import Word2Vec
from gensim.models import KeyedVectors
model = gensim.models.KeyedVectors.load_word2vec_format('
    GoogleNews-vectors-negative300.bin.gz', binary=True)
model = gensim.models.Word2Vec(documents, size=150, window=10,
    min_count=2, workers=10)
model.train(documents,total_examples=len(documents),epochs=10)
```

3.5.3. Extracción de datos

El procesamiento de textos con los algoritmos desarrollados en el presente investigación nos retorna como resultados los valores cuantitativos de relevancias y/o similitud de textos, tales representa al conjunto de textos escritos en las tesis y artículos científicos procesados en el presente. Por lo tanto es importante extraer estos datos cuantitativos representativos y dar una estructura comprensible para análisis y conclusiones.
En la Tabla III.1 se puede apreciar una estructura de datos compuestos por los valores cuantitativos calculados por los tres algoritmos desarrollados en la presente investigación, estos valores permiten diferenciar la relevancia y/o

Cuadro III.1:

Promedio de valores de similitud de documentos por algoritmos.

Id_Tesis	Id_Paper	Jaro_W	TF-IDF	Word2vec
0	0	0.580725	0.006807	0.997061
0	1	0.599007	0.006864	0.997410
0	2	0.596804	0.006490	0.996877
0	3	0.584585	0.006541	0.997589
0	4	0.594991	0.006766	0.997089
0	5	0.592197	0.006151	0.997583
0	6	0.591997	0.006419	0.991127
0	7	0.589695	0.006692	0.997012
0	8	0.589153	0.006835	0.997376
0	9	0.594836	0.006404	0.997199

similitud de entre los diferentes documentos, por tanto es un método para recomendar al in-vestigador los artículos científicos muy similares o más relevantes a su búsqueda. En el presente investigación y en los siguientes párrafos se analiza los datos de los 10 primeros Artículos científicos y tesis más relevantes según los tres algoritmos desarrollados en el presente.

Leyenda de la Tabla III.1:

- Id_Tesis = Clave primaria de tesis.
- Id_Paper = Clave primaria de las Artículos científicas.
- Jaro_W = Valores cuantitativo de similitud entre los documentos procesados calculado por algoritmos Jaro Winckler.
- TF-IDF = Valores cuantitativo de relevancia entre los documentos procesados calculado por algoritmos TFIDF.
- Word2vec = Valores cuantitativo de similitud entre los documentos procesados calculado por algoritmos Word2vec.

Por tanto, cumplimos con extraer los datos más importantes de un conjunto de

datos que se obtiene en el procesamiento de textos, estos campos se completan con datos originales, no referimos a los campos de textos de los documentos, como: título, are, keywords y abstract de los artículos científicos y tesis, como se puede apreciar en la Figura III.4 (La dataset original se adjunta en la sección de anexos).

Figura III.4:

Estructura final de datos.

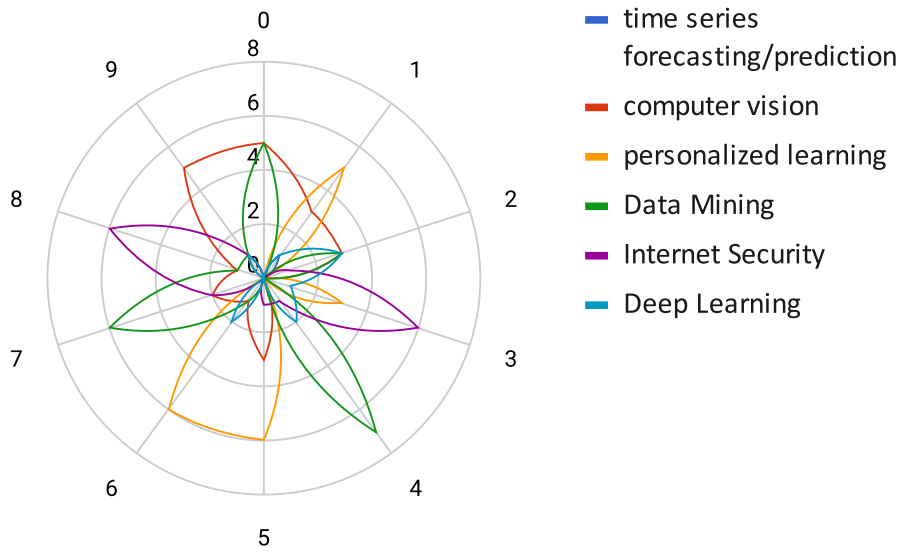
Id_Tesis	Id_Paper	Jaro_W	TF-IDF	Word2vec	Area_x	Titulo_x	Keywords	ABSTRACT	Area_y	Titulo_y	abstract
9	380	0.550366	0.011211	0.996022	Internet Security	SDN-Based Data Transfer Security for Internet ...	Dataflow management , Internet of Things (IoT)...	The exponential growth of devices connected to...	Data mining	Predition of irregular consumiption of drinkki...	This research work focuses on the experimentat...
9	423	0.536769	0.011087	0.995957	Internet Security	Practical Evaluation of Internet Systems' Secu...	web-based services , Internet security policie...	A proposed Internet systems security layer wit...	Data mining	Predition of irregular consumiption of drinkki...	This research work focuses on the experimentat...
9	428	0.553421	0.010988	0.995857	Internet Security	A review on biologically inspired approaches t...	Biologically Inspired , Human Immune System , ...	Internet of Things (IoT) paradigm involves new...	Data mining	Predition of irregular consumiption of drinkki...	This research work focuses on the experimentat...
9	300	0.567080	0.010802	0.995599	time series forecasting/prediction	ScaleNet-multiscale neural-network architectur...	History , Wavelet coefficients , Time series a...	The effectiveness of a multiscale neural net a...	Data mining	Predition of irregular consumiption of drinkki...	This research work focuses on the experimentat...
9	453	0.544494	0.010848	0.995583	Deep Learning	Deep Learning of Facial Depth Maps for Obstruc...	Obstructive Sleep Apnea , Transfer learning , ...	Obstructive Sleep Apnea (OSA) occurs when obst...	Data mining	Predition of irregular consumiption of drinkki...	This research work focuses on the experimentat...
9	384	0.551394	0.010926	0.995572	Internet Security	A note on security management of the Internet ...	IOT , security management system , cybersecuri...	Cyber-attackers are threatening the security o...	Data mining	Predition of irregular consumiption of drinkki...	This research work focuses on the experimentat...
9	66	0.560501	0.010962	0.995518	computer vision	Computer vision based detection and localizati...	Computer vision, pothole detection, histogram ...	Asphalt pavement distresses have significant i...	Data mining	Predition of irregular consumiption of drinkki...	This research work focuses on the experimentat...
9	167	0.551311	0.011087	0.995516	computer vision	Olive batches automatic classification in mill...	On-line olives classification , computer visio...	Olive batches classification is crucial to obt...	Data mining	Predition of irregular consumiption of drinkki...	This research work focuses on the experimentat...
9	404	0.541397	0.011149	0.995484	Internet Security	Security protocols in the Internet new framework	Protocols , Information security , Authenticat...	The Internet is one of the most important adva...	Data mining	Predition of irregular consumiption of drinkki...	This research work focuses on the experimentat...
9	108	0.568045	0.011025	0.995458	computer vision	Automation of a wheelchair mounted robotic arm...	Assistive Technology , Automation , Computer V...	Assistive robotic devices have great potential...	Data mining	Predition of irregular consumiption of drinkki...	This research work focuses on the experimentat...

3.5.4. Minería de Datos y resultados

Con los datos extraídos en el anterior etapa se puede realizar algunos análisis de los documentos procesados, en la Figura III.5 se puede apreciar la presencia de los temas en los 10 tesis procesados, es decir podemos apreciar la presencia del contenidos de los 6 temas o áreas de investigación, las cuales pertenecen a las Artículos científicos procesadas en el presente.

Figura III.5:

Relevancia de las Áreas de investigación en las tesis.

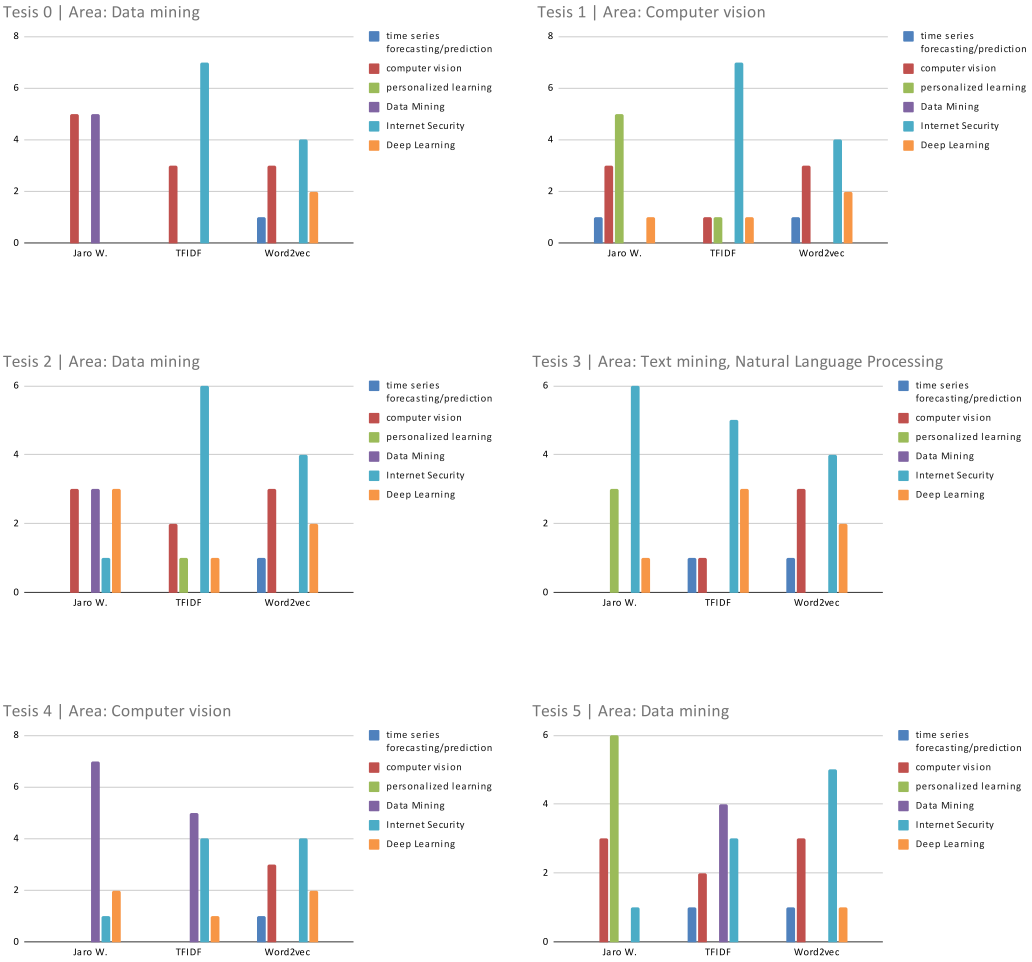


En lo que respecta a la tesis número uno, se puede apreciar que cuyo contenido pertenece al tema personalized learning, sin embargo el mismo tema se hace presente en las tesis cinco y seis. Con relación a la tesis número dos se puede apreciar que no hay una clara alineación sobre algún tema en especifico, sin embargo hay leve evidencia de contenido de tema relacionado a deep learning, data mining y computer vision. Con respecto a la tesis número tres, se aprecia una clara evidencia que cuyo contenido del internet security, ese mismo tema está presente en la tesis número ocho. El análisis se puede realizar con cada una de las variables presente en la investigación y datos, en este caso solo se ha considerado los 10 primeros más relevantes por cada algoritmo

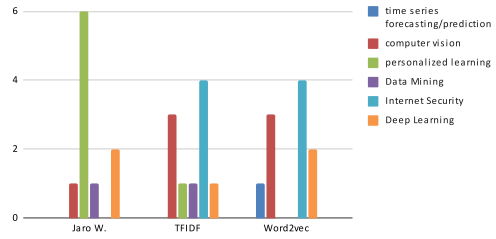
A continuación, realizamos un análisis mas avanzado y explicito para cada uno de los 10 tesis procesados, en la Figura III.6 se puede apreciar figuras estadísticos de las relevancia de contenidos de los documentos, en la cabecera de cada uno de los sub gráfico esta la clave principal de las tesis y el área respectiva, al costado se encuentra todas las áreas tomadas en cuenta en la presente investigación.

Figura III.6:

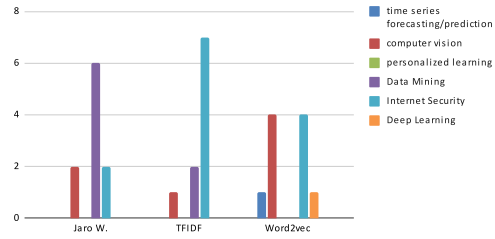
Relevancia de los temas según los algoritmos en los 10 tesis de EPISI.



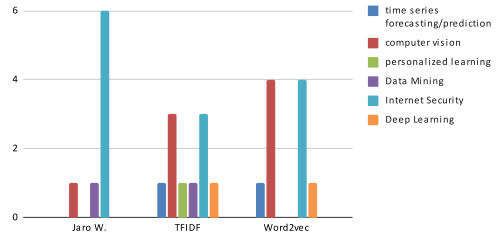
Tesis 6 | Area: Data mining



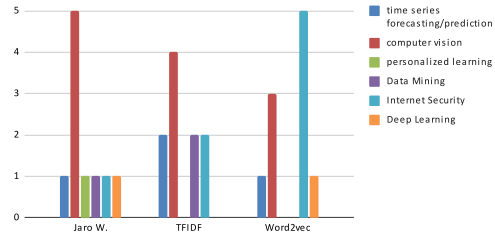
Tesis 7 | Area: Computer vision



Tesis 8 | Area: System information



Tesis 9 | Area: Data mining



3.6. Análisis de datos

A continuación, se describe y se analiza los resultados obtenidos por cada tesis procesado, así mismo las recomendaciones respectivas, las tablas citadas se encuentran al final de las descripciones.

- Tesis 0: el contenido de la tesis corresponde a la área Data mining, por tanto se espera que la recomendación basada en relevancia y/o similitud sean Artículos científicos del área data mining, sin embargo en la sub figura de III.6 y en la Tabla III.2 se aprecia que: según el algoritmo Jaro Winckler las Artículos mas similares a la tesis con Artículos científicas del área computer vision y Data Mining. Según el algoritmo TFIDF los artículos mas recomendadas son del área Internet Security y Computer Vision. Segun el algoritmo Word2vec los artículos mas recomendados son del área Iternet Security y Computer Vision.

- Tesis 1: el contenido de la tesis corresponde al área Computer Vision, segun el algoritmo Jaro winckler los artículos científicos mas recomendados son de área personalized leargning, por parte de algoritmos TFIDF y Word2vec el área mas recomendada son del área Internet Security, sin embargo se puede apreciar que la segunda área mas recomendada es Computer Vision, para mayor detalle ver la Figura III.6 y la Tabla III.3.

- Tesis 2: el contenido de la tesis es del contexto predicción de caso de anemia haciendo uso de herramientas, métodos de minería de datos. Por tanto se esperan recomendaciones temas relacionados a minería de datos, Sin embargo en la Figura III.6 y en la Tabla III.4, se aprecia que el algoritmo Jaro Winkler recomienda 3 artículos científicos del área computer vision, data mining y deep learning, por otro lado el algoritmo TFIDF recomienda 6 artículos científicos de internet security, 2 de computer vision y otros, el algoritmo Word2vec recomienda 4 de internet security, seguido por computer

vision con 3 artículos científicos.

- Tesis 3: el contenido de la tesis corresponde al área procesamiento lenguaje natural y minería de textos, según el algoritmo Jaro winkler los artículos científicos mas recomendados son de área internet sicurity, personalized learning, por parte de algoritmo TFIDF recomienda artículos científicos del área internet sicurity, seguido por deep learning y finalmente el algoritmo Word2vec recomienda artículos del área internet security, seguido por computer vision, para mayor detalle ver la Figura III.6 y la Tabla III.5.

- Tesis 4: el contenido de la tesis corresponde al área computer vision, según el algoritmo Jaro winkler los artículos científicos mas recomendados son de área internet data mining seguido por deep learning, por parte de algoritmo TFIDF recomienda artículos científicos del área data mining, seguido por internet security y finalmente el algoritmo Word2vec recomienda artículos del área internet security, seguido por computer vision, para mayor detalle ver la Figura III.6 y la Tabla III.6.

- Tesis 5: el contenido de la tesis corresponde al área data mining, según el algoritmo Jaro winkler los artículos científicos mas recomendados son de área personalized learning seguido por computer vision, por parte de algoritmo TFIDF recomienda artículos científicos del área data mining, seguido por internet security y finalmente el algoritmo Word2vec recomienda artículos del área internet security, seguido por computer vision, para mayor detalle ver la Figura III.6 y la Tabla III.7.

- Tesis 6: el contenido de la tesis corresponde al área data mining, según el algoritmo Jaro winkler los artículos científicos más recomendados son de área personalized learning seguido por deep learning, por parte de algoritmo TFIDF recomienda artículos científicos del área internet security, seguido por computer vision y finalmente el algoritmo Word2vec recomienda ar-tículos del área internet security, seguido por

computer vision, para mayor detalle ver la Figura III.6 y la Tabla III.8.

- Tesis 7: el contenido de la tesis corresponde al área computer vision, según el algoritmo Jaro winkler los artículos científicos mas recomendados son de área data mining, por parte de algoritmo TFIDF recomienda artículos científicos del área internet security, seguido por data mining y finalmente el algoritmo Word2vec recomienda artículos del área internet security, seguido por deep learning, para mayor detalle ver la Figura III.6 y la Tabla III.9.

- Tesis 8: el contenido de la tesis corresponde al área System information, según el algoritmo Jaro winkler los artículos científicos mas recomendados son de área internet security, por parte de algoritmo TFIDF recomienda artículos científicos del área internet security, seguido por computer vision y finalmente el algoritmo Word2vec recomienda artículos del área internet security, seguido por computer vision, para mayor detalle ver la Figura III.6 y la Tabla III.10.

- Tesis 9: el contenido de la tesis corresponde al área data mining, según el algoritmo Jaro winkler los artículos científicos mas recomendados son de área computer vision, por parte de algoritmo TFIDF recomienda artículos científicos del área computer vision y finalmente el algoritmo Word2vec recomienda artículos del área internet security, seguido por computer vision, para mayor detalle ver la Figura III.6 y la Tabla III.11.

Cuadro III.2:

Diez artículos científicos mas recomendados para tesis N° 0.

Título : Data mining model based on associated factors for predicting college dropout.			
Áreas / Algoritmos	Jaro W.	TFIDF	Word2vec
time series forecasting/prediction	0	0	1
computer vision	5	3	3
personalized learning	0	0	0
Data Mining	5	0	0
Internet Security	0	7	4
Deep Learning	0	0	2
Total	10	10	10

Cuadro III.3:

Diez artículos científicos mas recomendados para tesis N° 1.

Título : Analysis of predictive models based on computer vision applied to parallelism.			
Áreas / Algoritmos	Jaro W.	TFIDF	Word2vec
time series forecasting/prediction	1	0	1
computer vision	3	1	3
personalized learning	5	1	0
Data Mining	0	0	0
Internet Security	0	7	4
Deep Learning	1	1	2
Total	10	10	10

Cuadro III.4:

Diez artículos científicos mas recomendados para tesis N° 2.

Título : Data mining model for predicting cases of anemia in pregnant women in the province of Ilo.			
Áreas / Algoritmos	Jaro W.	TFIDF	Word2vec
time series forecasting/prediction	0	0	1
computer vision	3	2	3
personalized learning	0	1	0
Data Mining	3	0	0
Internet Security	1	6	4
Deep Learning	3	1	2
Total	10	10	10

Cuadro III.5:

Diez artículos científicos mas recomendados para tesis N° 3.

Título : Text mining to build a data mining model for job offers recommendation.			
Áreas / Algoritmos	Jaro W.	TFIDF	Word2vec
time series forecasting/prediction	0	1	1
computer vision	0	1	3
personalized learning	3	0	0
Data Mining	0	0	0
Internet Security	6	5	4
Deep Learning	1	3	2
Total	10	10	10

Cuadro III.6:

Diez artículos científicos mas recomendados para tesis N° 4.

Título : Application of intelligent algorithms for automatic recognition of foliar diseases of avocado crops.			
Áreas / Algoritmos	Jaro W.	TFIDF	Word2vec
time series forecasting/prediction	0	0	1
computer vision	0	0	3
personalized learning	0	0	0
Data Mining	7	5	0
Internet Security	2	4	4
Deep Learning	1	1	2
Total	10	10	10

Cuadro III.7:

Diez artículos científicos mas recomendados para tesis N° 5.

Título : Optimization of distribution unit routes applying reinforced learning: Los Pieros SRL case.			
Áreas / Algoritmos	Jaro W.	TFIDF	Word2vec
time series forecasting/prediction	0	1	1
computer vision	3	2	3
personalized learning	6	0	0
Data Mining	0	4	0
Internet Security	1	3	5
Deep Learning	0	0	1
Total	10	10	10

Cuadro III.8:

Diez artículos científicos mas recomendados para tesis N° 6.

Título : Prediction Model Of Customers Who Will unscribe the HFC Service Based on neuronal networks in the city of Moquegua.			
Áreas / Algoritmos	Jaro W.	TFIDF	Word2vec
time series forecasting/prediction	0	0	1
computer vision	1	3	3
personalized learning	6	1	0
Data Mining	1	1	0
Internet Security	0	4	4
Deep Learning	2	1	2
Total	10	10	10

Cuadro III.9:

Diez artículos científicos mas recomendados para tesis N° 7.

Título : Diagnosis of diabetic retinopathy using classification algorithms.			
Áreas / Algoritmos	Jaro W.	TFIDF	Word2vec
time series forecasting/prediction	0	0	1
computer vision	2	1	4
personalized learning	0	0	0
Data Mining	6	2	0
Internet Security	2	7	4
Deep Learning	0	0	1
Total	10	10	10

Cuadro III.10:

Diez artículos científicos mas recomendados para tesis N° 8.

Título : Information System for the Control and Documentary Monitoring of the Provincial Subprefecture of Ilo, 2019.			
Áreas / Algoritmos	Jaro W.	TFIDF	Word2vec
time series forecasting/prediction	0	1	1
computer vision	1	3	4
personalized learning	0	1	0
Data Mining	1	1	0
Internet Security	6	3	4
Deep Learning	0	1	1
Total	10	10	10

Cuadro III.11:

Diez artículos científicos mas recomendados para tesis N° 9.

Título : Predition of irregular consumiption of drinkking water service based on machine learning .			
Áreas / Algoritmos	Jaro W.	TFIDF	Word2vec
time series forecasting/prediction	1	2	1
computer vision	5	4	3
personalized learning	1	0	0
Data Mining	1	2	0
Internet Security	1	2	5
Deep Learning	1	0	1
Total	10	10	10

Capítulo IV

Resultados y discusión

Según los datos de resultados descritos en el anterior sección, en la tabla IV.1 se muestra las tesis procesadas para el presente investigación, a las cuales se le ha asignado una variable que comienza con la letra T mayúscula, como se puede apreciar en el campo variable. Cada una de las tesis también tienen una área de estudio, el cual también se aprecia en el campo área.
De ahora en adelante nos referimos a las tesis por la variable asignada, si bien se busca la recomendación de artículos científicos similares a las tesis, en la tabla IV.2 se aprecia el número de aciertos de recomendación por área de tesis, el cual significa que el algoritmo Jaro Winkler ha acertado recomendar a 19 de 300 artículos científicos, esta cantidad es la suma de número acertado en los 10 tesis, en la misma tabla se puede apreciar de manera más detallado. Sin embargo los algoritmos TFIDF y Word2vec solo han acertado 6 de 300 tesis respectivamente. Los 300 artículos científicos comprende ser solo los 10 artículos científicos con mayor valor de relevancia/similitud por cada tesis, en total se tiene 10 tesis, es decir que para cada tesis se ha filtrado los 10 artículos científicos más recomenda-dos por cada algoritmo, Jaro Winkler, TFIDF y Word2vec, en total 30 artículos por cada tesis.

Cuadro IV.1:

Leyenda de tesis.

N°	Variable	Área	Titulo de tesis
0	T0	Data Mining	Data mining model based on associated factors for predicting college dropout.
1	T1	Computer vision	Analysis of predictive models based on computer vision applied to parallelism.
2	T2	Data mining	Data mining model for predicting cases of anemia in pregnant women in the province of Ilo.
3	T3	NLP, Text mining	Text mining to build a data mining model for job offers recommendation.
4	T4	Computer vision	Application of intelligent algorithms for automatic recognition of foliar diseases of avocado crops
5	T5	Reinforcement Learning,	Optimization of distribution unit routes applying reinforced learning: Los Pieros SRL case.
6	T6	Data mining	Prediction Model Of Customers Who Will unscribe the HFC Service Based on neuronal net-works in the city of Moquegua.
7	T7	Data mining	Diagnosis of diabetic retinopathy using classification algorithms.
8	T8	System Information	Information System for the Control and Documentary Monitoring of the Provincial Subpre-fecture of Ilo, 2019.
9	T9	Data mining	Predition of irregular consumiption of drinking water service based on machine learning.

Entonces se puede concluir que en este caso el algoritmo con mayor número

de aciertos es el Jaro Winkler, en seguida empatados por TFIDF y Word2vec, el algoritmo Jaro Winkler estaba basada en la búsqueda de similitudes de oraciones, es decir un conjunto de palabras que describen la acción, sin embargo TFIDF y Word2vec se basan en buscar relevancia y similitud (respectivamente) de palabras de manera individual, más no oraciones completas. En este caso particular funciona mejor el método basado en buscar similitudes de oraciones compuestas por más de una palabra.

Cuadro IV.2:

Precisión de los algoritmos.

Algoritmos / Tesis	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Total
Yaro Winkler	5	3	3	-	0	-	1	6	-	1	19
TFIDF	0	1	0	-	0	-	1	2	-	2	6
Word2vec	0	3	0	-	3	-	0	0	-	0	6
Total	5	7	3	-	3	-	2	7	-	3	

Adicionalmente analizamos el área mas recomendado, porque 7 de 10 tesis corresponden a áreas de artículos científicos que han sido procesados en la presente investigación, en la tabla IV.3 se aprecia que el área mas recomendado es internet security, con un total de 118 artículos científicos, sin embargo, no hay tesis de esta área en los tesis procesados. Los siguientes áreas mas recomendados son computer vision y data mining, esto incluye los artículos acertadas y no acertadas, en este tabla nos centramos las interacción de los áreas de artículos científicos.

Cuadro IV.3:

Áreas de artículos científicos mas recomendados.

Áreas / Tesis	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	Total
Time series forecasting / prediction	1	2	1	2	1	2	1	1	2	4	17
computer vision	11	7	8	4	3	8	7	7	8	12	75
personalized learning	0	6	1	3	0	6	1	0	1	1	19
Data Mining	5	0	3	0	12	4	2	7	2	3	38
Internet Security	11	11	11	15	10	9	8	13	7	8	118
Deep Learning	2	4	6	6	4	1	5	1	2	2	33
Total	30	30	30	30	30	30	30	30	30	30	300

Capítulo V

Conclusiones

- Se recomienda como parte de la implementación, el algoritmo Jaro Winkler con la mayor cantidad de aciertos de artículos científicos de la disciplina de PLN, basado en la búsqueda de similitud de oraciones o conjunto de palabras que describe la acción.
- Se han implementado tres algoritmos de la disciplina de PLN, Jaro Winkler, TFIDF y Word2vec, la selección de los algoritmos están basados en el estado de arte de la linea de investigación, y con el objetivo de detectar y determinar el valor de similitud entre los documentos de tesis y artículos científicos.
- No se obtuvo resultados esperados de recomendación de artículos científicos basado en minería de textos, en análisis de los resultados se evidencia que los 10 mejores recomendaciones no siempre concuerdan con el área planteada en las tesis, los campos tomado en cuenta fueron titulo y el ab-tract de las tesis, los campos procesado de los artículos científicos fueron: titulo, keywords y abstract.

Capítulo VI

Recomendaciones

- Se recomienda construir una base de datos de mas de 5 mil tesis y artículos científicos para entrenar el algoritmo Word2vec, porque en la presente investigación se usó los vectores pre entrenados de noticias de google.
- Probar otro métodos avanzados basados en aprendizaje profundo, tal vez mejore en cuanto a agrupación de documentos por patrones comunes o similares, este modelo se podría usar para realizar recomendaciones mas acertadas.

Referencias

Apaza, H., Rubin de Celis Vidal, A. A., y Chire Saire, J. E. (2021). Job recommendation based on curriculum vitae using text mining. , 1051–1059.

Apaza Alanoca, H. (2019). *Minería de textos para construir un modelo de minería de datos para recomendación de ofertas laborales.* Universidad Nacional de Moquegua.

Daher Nader, J. E., Panunzio, A. P., y Hernandez Navarro, M. I. (2018, 12). La investigacion cientifica: una funcion universitaria a considerar en el contexto ecuatoriano. *EDUMECENTRO*, *10*, 166 - 179. Descargado de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2077-28742018000400011&nrm=iso

Díez, R., Gómez, A., y de Abajo Martínez, N. (2001). *Introducción a la inteligencia artificial: sistemas expertos, redes neuronales artificiales y computación evolutiva.* Servicio de Publicaciones, Universidad de Oviedo. Descargado de https://books.google.com.pe/books?id=RKqLMCw3IUkC

Di Giuda, G. M., Locatelli, M., Schievano, M., Pellegrini, L., Pattini, G., Giana, P. E., y Seghezzi, E. (2020). Natural language processing for information and project management. En B. Daniotti, M. Gianinetto, y S. Della Torre (Eds.), *Digital transformation of the design, construction and management processes of the built environment* (pp. 95–102). Cham: Springer International Publishing. Descargado de https://doi.org/10.1007/978-3-030-33570-0_9 doi: 10.1007/978-3-030-33570-0_9

Gao, J., He, Y., Zhang, X., y Xia, Y. (2017). Duplicate short text detection based on word2vec. En *2017 8th ieee international conference on software engineering and service science (icsess)* (p. 33-37). doi: 10.1109/ICSESS.2017.8342858

Incio, F. A. R., Morales, W. R. P., Cruzata-MartÃnez, A., y Ãguila, S. C. A. d. (2020, 05). AnÃ¡lisis de la produccion cientÃfica de universidades en Comunicacion Social (2014 - 2018). *Propósitos y Representaciones*, *8*. Descargado de http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S2307-79992020000300030&nrm=iso

Kantrowitz, M., Mohit, B., y Mittal, V. (2000). Stemming and its effects on tfidf ranking (poster session). En *Proceedings of the 23rd annual international acm sigir conference on research and development in information retrieval* (p. 357–359). New York, NY, USA: Association for Computing Machinery. Descargado de https://doi.org/10.1145/345508.345650 doi: 10.1145/345508.345650

Karani, D. (2018). *Introduction to word embedding and word2vec.* Descargado de https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

Keil, J. M. (2019). Efficient bounded jaro-winkler similarity based search. En T. Grust y cols. (Eds.), *Btw 2019* (p. 205-214). Gesellschaft für Informatik, Bonn. doi: 10.18420/btw2019-13

Kloo, I., Dabkowski, M. F., y Huddleston, S. H. (2019). Improving record linkage for counter-threat finance intelligence with dynamic jaro-winkler thresholds. En *Proceedings of the winter simulation conference* (p. 2467–2478). IEEE Press.

Lei, S. (2020). Research on the improved word2vec optimization strategy based on statistical language model. En *2020 international conference on information science, parallel and distributed systems (ispds)* (p. 356-359). doi: 10.1109/ISPDS51347.2020.00082

Luis Cervantes Liñán, V. P. C., Ludisleydis Bermúdez Díaz. (2019, 06).

Situación de la investigación y su desarrollo en el perú: reflejo del es-tado actual de la universidad peruana. *Pensamiento y Gestión*, 311-322. Descargado de http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1657-62762019000100311&nrm=iso

Manaf, K., Pitara, S., Subaeki, B., Gunawan, R., Rodiah, y Bakhtiar. (2019). Comparison of carp rabin algorithm and jaro-winkler distance to determine the equality of sunda languages. En *2019 ieee 13th international conference on telecommunication systems, services, and applications (tssa)* (p. 77-81). doi: 10.1109/TSSA48701.2019.8985470

Mikolov, T., Chen, K., Corrado, G., y Dean, J. (2013). *Efficient estimation of word representations in vector space.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., y Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.*

Ortiz Rivera, M. L., Graciela Cervantes. (s.f.). La formación científica en los primeros años de escolaridad. *Universidad del Atlántico.*

Pereyra-Elias, R., Huaccho-Rojas, J. J., Taype-Rondan, A., Mejia, C. R., y Mayta-Triston, P. (2014, 07). Publicación y factores asociados en docentes universitarios de investigación científica de escuelas de medicina del Perú. *Revista Peruana de Medicina Experimental y Salud Publica*, *31*, 424-430. Descargado de http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1726-46342014000300003&nrm=iso

Purizaca-Rosillo, N., Cardoza-Jiminez, K., y Herrera-Añazco, P. (2016, 01). Producción científica en una universidad publica peruana beneficiaria del canon. *Anales de la Facultad de Medicina*, *77*, 73 - 74. Descargado de http://www.scielo.org.pe/scielo.php?script=sci _arttext&pid=S1025-55832016000100016&nrm=iso

Rong, X. (2016). *word2vec parameter learning explained.*

Sun, J., Ma, J., Liu, X., Liu, Z., Wang, G., Jiang, H., y Silva, T. (2013). A novel approach for personalized article recommendation in online scientific communities.

En *2013 46th hawaii international conference on system sciences* (p. 1543-1552). doi: 10.1109/HICSS.2013.48

West, D. M. (s.f.). Avance tecnológico: riesgos y desafíos. Descargado de https://www.bbvaopenmind.com/articulos/avance-tecnologico-riesgos-y-desafios

Winkler, W. (1990, 01). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*.

www.tensorflow.org. (2022). *word2vec*. Descargado de https://www.tensorflow.org/tutorials/text/word2vec

Zanini, N., y Dhawan, V. (2015, 01). Text mining: An introduction to theory and some applications. *Research Matters*, 38-44.

Zhao, G., Liu, Y., Zhang, W., y Wang, Y. (2018). Tfidf based feature words extraction and topic modeling for short text. En *Proceedings of the 2018 2nd international conference on management engineering, software engineering and service sciences* (p. 188–191). New York, NY, USA: Association for Computing Machi-nery. Descargado de https://doi.org/10.1145/3180374.3181354 doi: 10.1145/3180374.3181354

Anexos



(a) Bases de datos.



(b) Código de programación de Jaro Winkler.



(c) Código de programación de TFIDF.



(d) Código de programación de Wor2Vec

FORMATO DE AUTORIZACIÓN DE PUBLICACIÓN EN EL REPOSITORIO INSTITUCIONAL UNAM

Formulario de autorización para la publicación electrónica en la página web del Repositorio Institucional de la Universidad Nacional de Moquegua, de conformidad con la Ley N° 30035 Ley que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto y Resolución de Comisión Organizadora N° 041-2022-UNAM que aprueba el Reglamento que regula el Repositorio Institucional Digital de la UNAM.

1. DATOS DEL AUTOR

Nombres y Apellidos : CARMEN MARIA CHINO CERVANTES
DNI/Carnet de Extranjería/Pasaporte N° : 46133531
Correo electrónico : cmcervants@gmail.com
Teléfono/Celular : +51 953 524844

2. GRADO, TITULO O ESPECIALIZACIÓN (Marque con una X)

Bachiller	Título Profesional	Magister	Doctor	Segunda Especialidad
	X			

3. TIPO DE PUBLICACIÓN (Marque con una X)

a) Articulo ( )
b) Tesis de pregrado (X)
c) Tesis de Maestría ( )
d) Tesis de Doctorado ( )
e) Monografía ( )
f) Reporte de caso científico o hallazgos ( )
g) Libro ( )
h) Parte de un libro ( )
i) Revisión ( )
j) Conferencia ( )
k) Presentación ( )
l) Artículo preliminar ( )
m) Articulo antes de ser publicado ( )
n) Artículo de prensa ( )
o) Documentos internos ( )
p) Imagen ( )
q) Documentos técnicos ( )
r) Ponencias presentadas en eventos científicos (Jornada, Congreso, Simposio Conferencias, etc.) ( )
s) Informes de suficiencia profesional ( )
t) Patentes producidas por académicos/as ( )

04

4. TIPO DE TRABAJO DE INVESTIGACIÓN (Marque con una X)

a) Tesis (X)
b) Trabajo de Investigación ( )
c) Trabajo de suficiencia profesional ( )
d) Trabajo académico ( )

5. DATOS DE LA PUBLICACION

a) Título : RECOMENDACIÓN DE ARTÍCULOS CIENTÍFICOS UTILIZANDO TÉCNICAS DE MINERÍA DE TEXTOS PARA FORMULACIÓN DE PROYECTOS DE INVESTIGACIÓN.
b) Campo del conocimiento OCDE (Indicar el código de clasificación OCDE de la obra : https://catalogos.concytec.gob.pe/vocabulario/ocde_ford.html) : 1.02.01
c) Fecha de publicación : 2022

6. DATOS DEL ASESOR:

a) Nombres y Apellidos : ANIBAL FERNANDO FLORES GARCIA
b) DNI/Pasaporte/Cédula : 04743476
c) ORCID : 0000-0003-4638-8242

FIRMA : .....

DOCUMENTO DE IDENTIDAD : 46133531
NOMBRES Y APELLIDOS : CARMEN MARIA CHINO CERVANTES