Universidad Autónoma de Querétaro

Facultad de Informática

Algoritmo para la extracción del conocimiento (KDD) a través del análisis de textos aplicado en la investigación social

Tesis

Que como parte de los requisitos
para obtener el Grado de
Doctor en Ciencias de la Computación

Presenta

M.A. Esp. Jhonathan Quillo Espino

Dirigido por:

Dra. Rosa María Romero González

Querétaro, Qro. a 26 de Mayo de 2022

Universidad Autónoma de Querétaro

Facultad de Informática

Doctorado en Ciencias de la Computación

Algoritmo para la extracción del conocimiento (KDD) a través del análisis de textos aplicado en la investigación social

Tesis

Que como parte de los requisitos para obtener el Grado

Doctor en Ciencias de la Computación

Presenta

M.A. Esp. Jhonathan Quillo Espino

Dirigido por:

Dra. Rosa María Romero González

Dra. Rosa María Romero González
Presidente

Dra. Ana Marcela Herrera Navarro
Secretario

Dra. Diana Margarita Córdova Esparza
Vocal

Dra. Sandra Luz Canchola Magdaleno
Suplente

Dra. Rocío Edith López Martínez
Suplente

Centro Universitario, Querétaro, Qro.
Mayo 2022
México

Dedicatorias

Por todo el amor y cariño que desde niño me diste

por todo tu esfuerzo y dedicación para educarme

por todos tus consejos sabios

por todo lo que haces por mi

De todo corazón, amor y cariño para ti mamá…Te amo.

Para ti papá que estas en el cielo te quiero y te extraño…

Agradecimientos

Quiero agradecer a mi mamá porque siempre me ha apoyado ha sido mi guía y mi pilar motivándome todos los días a seguir, ser mejor y darme su ejemplo para salir adelante ante cualquier reto que la vida te ponga por delante. A mi papá un ángel que está en el cielo cuidándome todos los días. A mis hermanos: Michael, Ivonne y Edgar que siempre me han estado ahí apoyándome y cuidándome en todo momento.

A mi directora de tesis la Dra. Rosa María Romero González por su tiempo, dedicación, enseñanza y apoyo para que este proyecto se concluya. A mis docentes doctores y doctoras que me han asesorado durante todo el tiempo de mis estudios.

A Conacyt por su apoyo incondicional para el pago de mis estudios.

A la Universidad Autónoma de Querétaro por el apoyo con el lugar de trabajo e instalaciones siempre listas, limpias, seguras, con las herramientas necesarias para tomar clases.

Índice

Índice de Figuras

Índice de tablas

Abreviaturas

Abreviatura	Descripción	Pág.
MD	Minería de Datos	2
ML	Machine Learning	2
IA	Inteligencia artificial	2
MT	Minería de Textos	3
LN	Lenguaje Natural	9
RI	Recuperación de la Información	5
KDD	Knowledge Discovering in Databases	10
PLN	Procesamiento del Lenguaje Natural	15
EI	Extracción de la información	20
REN	Reconocimiento de entidad nombrada	20
RC	Resolución de conferencia	20
CEL	Construcción de elemento plantilla	20
PPE	Producción plantilla escenario	20
ACD	Análisis crítico del discurso	25
ACC	Análisis cualitativo de contenido	26
ACCD	Análisis cualitativo del contenido directo	26
SACC	Sumativa de Análisis Cualitativo de Contenido	26
ACCCC	Análisis Cualitativo de Contenido Convencional	27
RTA	Resumen en Textos Automático	28
TSUM	Text Sumarization	28
RTAE	Resumen de Textos Extractivos	28
RTAA	Resumen Automático de Textos Abstractivo	29
TF	Termino de Frecuencia-	30
FID o Idf	Frecuencia Inversa de Documento	30
Tf-Idf	Termino Frecuencia-Frecuencia Inversa de Documento	30
Isf	Frecuencia Inversa del enunciado	31
Sf	Sentence frequency o Frecuencia del enunciado	31
MEV	Modelo de Espacio Vectorial	32
CT	Clasificación de Texto	33

Resumen

El análisis de información a través de la investigación social es arduo, debido a que es un proceso que requiere tiempo y esfuerzo para la interpretación de grandes lecturas de texto importante, ¿De qué manera la minería de textos puede apoyar la investigación social para evitar la interpretación subjetiva de la información? Es la pregunta que guía esta investigación que es abordada desde la teoría de la minería de textos. A través del estudio y análisis de las técnicas de la minería de texto, se propone un algoritmo denominado *Quillo Espino* basado en el modelo de resúmenes de texto automático con el método extractivo, el cual consta de 3 fases en la ejecución, transformación de los datos, aplicación del método *TfIsf* y finalmente obtención del conocimiento. El algoritmo se encuentra compuesto por diferentes algoritmos: *tokenizador, adaptación del algoritmo snowball, stop words, conversión a vectores, creación de matrices, adaptación método TfIdf* etc. Mediante la aplicación del algoritmo *Quillo Espino* se logra una interpretación objetiva de la información analizada, sin perder la homogeneidad de los datos analizados. El estudio del caso es el resultado de análisis de entrevistas, entregadas a través del software Atlas.ti, realizadas por los investigadores del área cualitativa en la Universidad Autónoma de Querétaro. El algoritmo analiza, transforma, estructura la información, realiza operaciones de tipo estadístico asignando ponderación a cada uno de los términos analizados, por medio de discriminación y comparación de valores presenta resultados de acuerdo a solicitud del usuario. Los resultados mostraron evidencia de que, a través de la aplicación del algoritmo, los investigadores redujeron el tiempo de análisis en su investigación, teniendo todos los resultados en un solo lugar sin necesidad de cambiar de plataforma además contribuyo en agilización del proceso de obtención de conocimiento generada por la investigación social.

.

Palabras clave: Minería de textos, pre procesos, método extractivo

Summary/ Abstract

The analysis of information through social research is challenging, because it is a process that demands time and effort for the interpretation of large readings of important text. How can text mining support social research to avoid subjective interpretation of information? This is the question that leads this study, which is approached from the theory of text mining. Through the study and analysis of text mining techniques, an algorithm called J Quillo Espino based on the model of automatic text summaries with the extractive method is proposed, which consists of 3 phases in the execution, data transformation, application of the TfIsf method and finally obtaining knowledge. The algorithm is integrated by different algorithms: tokenizer, snowball algorithm adaptation, stop words, conversion to vectors, matrix creation, TfIdf method adaptation, etc. By employing the Quillo Espino algorithm, an objective interpretation of the analyzed information is achieved, without losing the uniformity of the analyzed data. The case study is the result of the analysis of interviews, delivered through the Atlas.ti software, carried out by the researchers of the qualitative area at the Autonomous University of Queretaro. The algorithm analyzes, transforms, structures the information, performs statistical operations assigning weight to each of the analyzed terms. By means of discrimination and comparison of values presents results according to the user's request. The results showed evidence that, through the application of the algorithm, the researchers reduced the analysis time in their study having all the results in one place without the need to change platforms and also contributed to speed up the process of obtaining knowledge generated by social research.

Keywords: text mining, preprocessing, extractive method

1. Introducción

El crecimiento desmesurado de la sociedad, y los avances tecnológicos fomentan el uso de grandes cantidades de datos y el acumulamiento de información en cantidades dramáticamente grandes, hay una necesidad propia del desarrollo de la sociedad en la búsqueda del uso de la tecnología como medio de asistencia para llevar a cabo tareas que faciliten y ayuden al desarrollo de una manera eficiente, sencilla y fácil ¿Qué pasa con la información que se genera todos los días? como los datos, textos, imágenes, etc.

No es suficiente con solo almacenarla, por tal motivo los científicos se han dado a la tarea de investigar y obtener algún resultado de la misma. El análisis de la información es una tendencia actual para las organizaciones privadas, públicas, y educativas, ya que permite conocer la orientación de las personas hacia productos o servicios o intereses particulares.

Tratar de comprender y entender es una tarea que requiere un análisis profundo debido a la existencia de diferentes factores que se encuentran implícitos en los textos, como los signos de diversa naturaleza y además, el uso de la lingüística como una herramienta básica para que las computadoras puedan entender y traducir el lenguaje humano al lenguaje de la computadora.

En la actualidad, la manipulación de la información es motivo para el desarrollo, los investigadores tienen necesidad por descubrir el pensamiento de la sociedad, descubrir ideas nuevas que ayuden a mejorar el desarrollo social estudiantil, la importancia para declarar un modelo que se útil, permitiendo colaborar con la investigación permite generar sugerencias como el uso de métodos provenientes de la minería de textos permitiendo la reducción de elementos como el tiempo, ayudando al entendimiento de lecturas de grandes cantidades textuales, de una forma, rápida, sencilla, fácil.

2. Aspectos Teóricos

2.1. Minería de datos (MD)

Diferentes autores definen la Minería de datos (MD) como: Leskovec, Rajaraman & Ulman. (2010), "el descubrimiento de modelos para los datos. Un modelo, sin embargo, puede ser uno de varios modelos. Se considera un modelo estático aquél por medio de cual se extraen los datos visibles*"*. Se refiere a la aplicación de técnicas de aprendizaje automático, entre otros métodos para encontrar importantes patrones en los datos. Fayyad et al. (1996), concluyen que "la búsqueda de patrones interesantes y de regularidades importantes en grandes bases de datos*"* (p. 3). Es considerada como sinónimo del concepto de aprendizaje por computadora o *Machine Learning* (ML) porqué se basan en modelos estructurados para la deducción de datos. Mena (mencionado en Aluja, 2001) define ML como "una rama de la inteligencia artificial (IA) que aborda el diseño y aplicación de algoritmos de aprendizaje" (p. 480).

Troche (2014, p.58), "consiste en extraer la información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior ". Los datos deben ser sólidos, permitiendo un manejo eficaz de la información. Han, Kamber & Pei (2012) descubrir patrones y conocimiento de una gran cantidad de datos. El origen puede ser de distintos lugares como bases de datos u otros y están mezclados dinámicamente. Shapiro, mencionado en Maracano & Talavera (2007), "como un proceso completo de extracción de información, encargado de la preparación de datos e interpretando los resultados, ayudando a encontrar las relaciones o patrones entre datos procesados" (p. 5).

Hand (2007) el procedimiento para descubrir estructuras ocultas en grandes conjuntos de datos. Pueden ser estructuras globales con objetivo principal modelar las formas para la distribución, o bien locales cuyo objetivo es detectar anomalías, eliminarlas y encontrar patrones. Fayyad et al. (1996a) "la extracción de las características más importantes de los datos e ignorar el resto*"* (p. 3). Un problema común en la minería de datos es descubrir en grandes cantidades de datos eventos inusuales ocultos en los datos.

Fayyad et al. (1996), declaran que consiste en aplicar algoritmos especiales para extraer patrones en bases de datos Hand mencionado en Aluja, (2001) *"*como el proceso de análisis

secundario de grandes bases de datos; para encontrar relaciones insospechadas que son de interés o valor en las bases de datos" (p. 480). Hand et al (2016, p. 6), concluyen que es "proceso de descubrir patrones en datos".

Usama et al. (1996), es aquella aplicación de algoritmos encargada de descubrir información relevante útil en los datos, y la se encuentra centrada en un único proceso particular.

La MD aborda cualquier problema o tema en el que existan datos históricos almacenados susceptibles de ser tratados mediante técnicas de MD, por ejemplo: puede buscar asociaciones, definición de tipologías, detección de ciclos temporales, predicciones, entre otras.

2.2. Minería de textos (MT)

La globalización ha provocado la necesidad de encontrar métodos y sistemas que tengan la capacidad de crear sistemas automatizados para la administración, organización y consulta de los textos donde nace la MT.

Definiciones para la MT: Contreras (2014) "el proceso encargado de descubrimiento de conocimientos que no existían explícitamente en ningún texto de la colección pero que surgen de relacionar el contenido de varios de ellos" (p. 131). Kao & Poteet (2007, p. 12) definen a la minería de textos "como el descubrimiento y extracción de conocimiento interesante no trivial de texto libre o no estructurado*"*. Inicialmente, se conocía como término para la extracción de información, pero no soportaba datos; después se consideró como máquina de aprendizaje, pero conforme fue avanzando el desarrollo, la ciencia la tomó como un problema algorítmico. En esta parte se consideraron los aspectos la de similitud y frecuencia de los datos, pero se dieron cuenta que la MT detecta el descubrimiento de eventos ocultos dentro de grandes cantidades de datos. Para Pushpa & Balamurugan (2015), "es el método de extracción de información significativa, conocimiento o patrones de los documentos de texto disponibles en varias fuentes*"* (p. 1). Sukanya & Biruntha (2012) la minería de datos cuando se descubre información nueva de diferentes fuentes escritas la MT es similar a la MD excepto las herramientas utilizadas ya que este puede funcionar con textos no estructurados, como correos, documentos, archivos HTML entre otros. En la Tabla 2.1 se muestran las etapas del proceso de MT.

Tabla 2.1.

Etapas de la MT

Número	Etapa
1	Determinar el propósito de estudio de la MT. Recolectar, identificar, recoger y validar información
2	En esta fase se realiza la recuperación de información en la cual se buscan identificar las fuentes más relevantes para el objetivo de estudio de la MT.
3	Procesamiento de texto: Se eliminan la información no relevante, realizando algunas acciones como: análisis léxico, tratamiento y separación de palabras vacías (artículos, preposiciones, conjunciones), normalización de palabras. Tratamiento de términos flexionados: términos relacionados morfológicamente, variaciones de género, número o tiempo verbal.
4	Extracción y análisis de clases, relaciones, asociaciones o secuencias con el fin de encontrar evidencias de conceptos y estructuras existentes. Los documentos se pueden presentar en un modelo de espacio vectorial, donde cada documento es modelado como un vector n y es presentado como: D=(di1, di2,di3...din) donde cada di representa el número de repeticiones de cada palabra en el documento.
5	Presentar resultados en resúmenes, en esta etapa se puede almacenar la información en bases de datos para su recuperación posterior.

Fuente: Contreras (2014, p. 132).

La MT es la forma de descubrir conocimiento de datos de texto ubicuo, basado en marco para aplicaciones, como tema de descubrimiento, extracción de información, resumen, recuperación de información. Hotho mencionado en Aghila (2010, p. 613), considera que "la MD y la MT son similares en la técnica de minería por que analizan y buscan grandes cantidades de información*"*. Hearts mencionado en Eíto & Senso (2004, p. 12) definen la MT como el descubrimiento semi-automatizado de patrones y tendencias en grandes conjuntos de datos. Eíto & Senso (2004) mencionan que "pretende obtener información a partir de patrones y tendencias que pueden observarse en grandes volúmenes de información estructurada" (p. 13). Vijayarani et al (2016, p. 7) aclaran que "es usada para encontrar nueva información previamente no identificada de diferentes recursos escritos".

A continuación, se muestra la Tabla 2.2 en la cual se detalla la diferencia entre MT y MD desde un enfoque computacional.

Tabla 2.2.

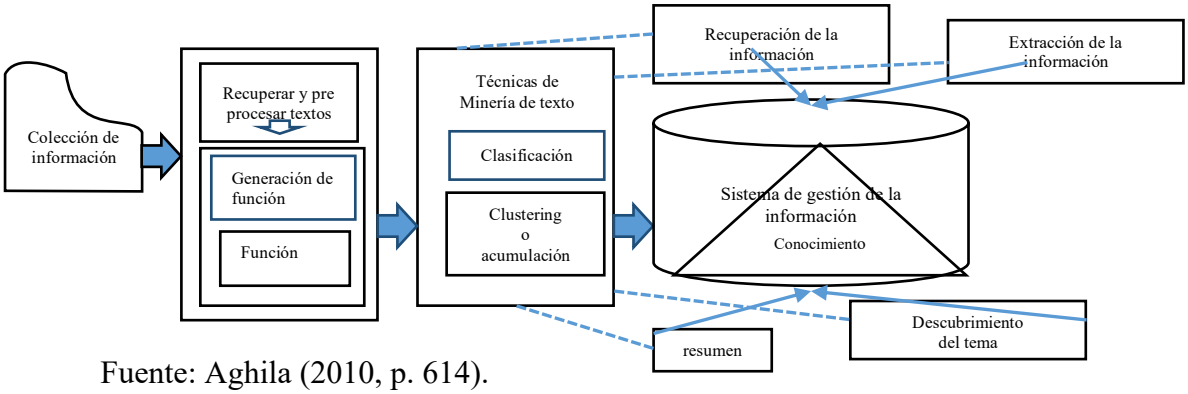
Diferencias entre MT y MD

Búsqueda de patrones		Búsqueda de Items y datos	
Datos No Textuales	Minería de datos	Nuevos	Ya conocidos
Datos Textuales	Lingüística computacional	Minería de textos	Búsqueda de BB.DD/ Recuperación Textual

Fuente: Eíto & Senso (2004, p. 12)

Gupta & Lehal mencionados en Aghila (2010, p. 613) determinan que "algunos de los significados en minería son tipos como datos, texto, web, procesos de negocios y servicios de minería y que la MT busca patrones en datos no estructurados como notas, pdf, archivos de texto, etc. En consecuencia, utiliza lenguaje semántico e inteligencia artificial*"*. Aghila (2010) determina que *"*el objetivo principal de la minería de textos es identificar patrones sin que exista duplicación de los mismos*"* (p. 613). En el proceso de MT juega un papel importante la Recuperación de Información (RI) por sus siglas en inglés. Vijayani et al (2016) la definen como el procedimiento de asociar y recuperar información de grandes volúmenes de textos. Por otra parte, Dursun & Crossland (2008) definen la minería de textos como "un proceso semi-automatizado de la extracción de conocimiento de una gran cantidad de datos no estructurados" (p. 1707). Vijay et al (2014, p. 42) mencionan que la MT además de descubrir información oculta también encuentra automáticamente diferentes recursos escritos por la computadora para extraer información nueva previamente desconocida. La Figura 2.1 explora a detalle los métodos de procesamiento en la minería de textos.

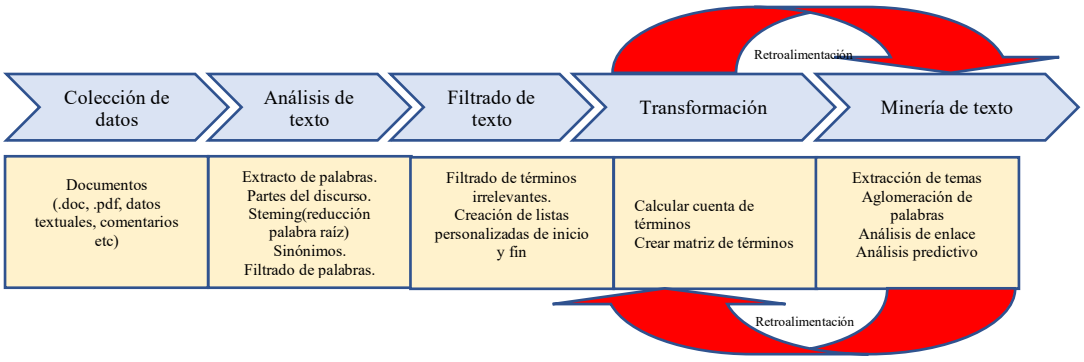
Figura 2.1. Proceso de MT



Fuente: Aghila (2010, p. 614).

A continuación, se muestra la Figura 2.2 en la cual se detalla el proceso de MT con su retroalimentación. Comenzando por la recolección de los datos que será sometidos a evaluación, a continuación, se realiza un análisis al texto (tokenización y *stemming*) después e realizan listas de personalizadas de términos relevantes, pasando a creación de matrices de términos y terminando por extracción de conocimiento.

Figura 2.2. Proceso de MT



Fuente: Chakraborty, Pagolu & Garla (2013, p. 13).

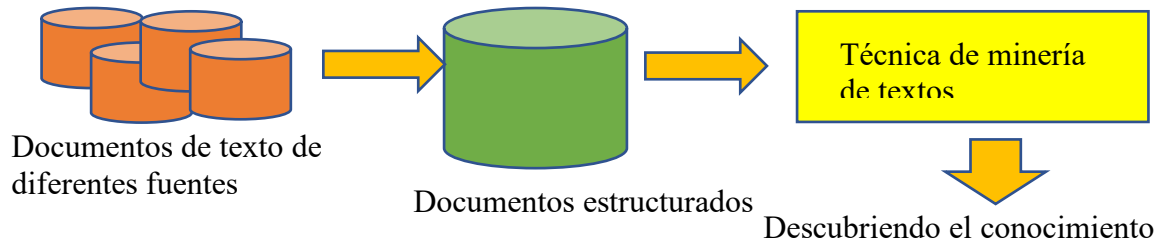
Chakraborty, Pagolu & Garla (2013, p. 13), determinan que el proceso de minería de textos típico envuelve las siguientes tareas:

1. Colección de datos**:** es el primer paso donde se recopilan los datos que son necesarios

para poder realizar el análisis.

2. Texto de análisis y transformación*:* se extraen, analizan, limpian, para crear un diccionario de palabras. Aquí se identifican las frases, y se retiran palabras que no se utilicen además se pueden crear variables asociadas. Lo más importante es que la transformación de texto sea acorde al tema que se está investigando.
3. Filtrado de textos*:* se filtra el texto y se elimina términos que no sean asociados e irrelevantes y se resumen los textos de una manera manual. Es un proceso relativamente lento, ya que se requiere gran conocimiento de la materia o experiencia. Transformación: los datos obtenidos deben de ser pertinentes a las necesidades de información del usuario.
4. Aplicación de método de MT*:* se aplican el método seleccionado de MT, como clasificación, análisis de asociación y análisis de enlace. En este proceso se obtiene el conocimiento. Pushpa & Balamurugan (2015) concluyen que es "es un proceso que emplea un grupo de algoritmos para convertir textos no estructurados, así como también a textos estructurados también métodos cuantitativos usados para analizar esos datos" (p. 1). En la Figura 2.3 se muestra un proceso simplificado propuesto por Pushpa & Balamurugan (2015, p. 841) para adquirir conocimiento a través de MT pasando desde colección de documentos textuales y terminando en la aplicación de la MT.

Figura 2.3. Extracción de la información.

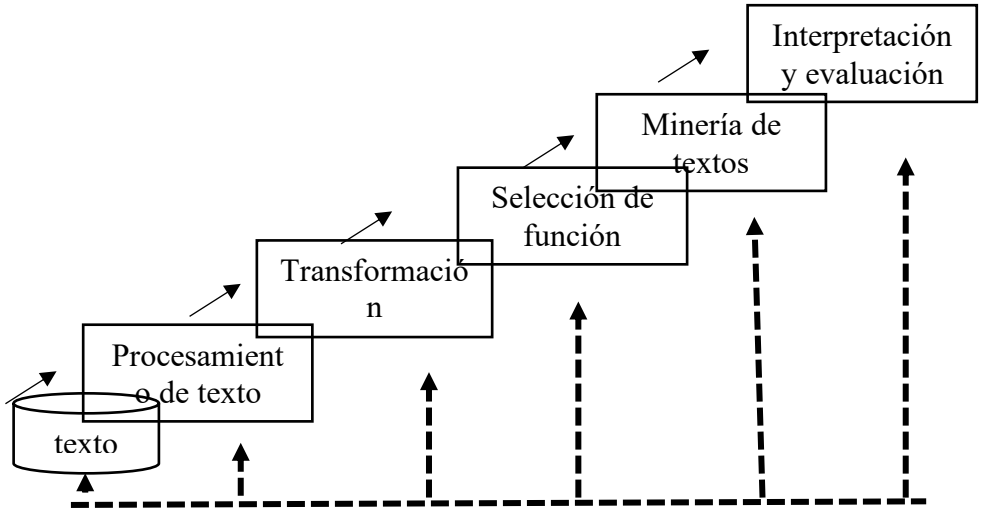


Fuente: Balamurugan & Pushpa (2015, p. 841).

En la Figura 2.4 se muestra el proceso de MD combinado con MT. Las bases de datos

estructuradas usan el proceso clásico de técnica de MD.

Figura 2.4. Proceso de MT



Fuente: Balamurugan y Pushpa (2015, p. 839).

Proceso de MT propuesto por Balamurugan & Pushpa (2015, pp. 839-841) consiste en: Recogida de documentos (*document gathering*): es la primera parte donde se recogen los datos que están presentados en diferentes formatos. Pre-procesamiento de documentos (*document pre-procesing*):se eliminan las redundancias, inconsistencias, palabras sueltas, y están preparados para el siguiente punto que está compuesto de otros puntos: Tokenización: se considera cada documento identificado con un token. Vijayarani & Janani (2016, p. 38) lo definen como "dividir una secuencia de texto y reducir en palabras, símbolos, frases en unidades significativas, pero debe de realizarse antes de cualquier otro procesamiento del texto".

El gran reto deriva de que los enunciados no se encuentran ortográficamente bien escritos, puntuados y acentuados. Cada día los investigadores buscan que los resultados sean más exactos, pero la tokenización se afecta por los patrones de las palabras y también en la forma como se almacena considerando la forma en como se hace la indexación de dichas palabras. El objetivo principal es realizar la identificación del token y contarlo a través de la

frecuencia. La tokenización se acompaña directamente con la eliminación de caracteres no deseados como lo son los puntos, acentos, signos y caracteres especiales como las comillas, símbolos y números entre otros. Esto puede reducir el tiempo del proceso. Remover palabras inusuales: se remueven palabras que no son usuales.

Stemming: se juntas palabras que tengan el mismo significado, utilizando el algoritmo Porter. Discriminación de palabras diferentes y dejar únicamente las que sean iguales.

Transformación (*tranformation*): la colección de palabras filtradas. MT o selección de patrón (*Pattern Selection*): se aplica el algoritmo a los resultados de los previos pasos.

Evaluación (*Evaluate*):se revisan los resultados y si es necesario se pueden volver a introducir al proceso. La Tabla 2.3 muestra las posibles diferentes aplicaciones para la minería de textos en la industria.

Tabla 2.3.

Diferentes aplicaciones de la minería de textos

Aplicaciones para la MT
Publicidad y media.
Telecomunicaciones, energía y otras industrias de servicios.
Tecnologías de la información e internet.
Seguros, bancos y servicios financieros.
Instituciones políticas, analistas políticos, administración pública y documentaciones legales.
Empresas farmacéuticas y de investigación y cuidado de la salud.

Fuente: Gupta y Lehal (2009, p. 69).

2.3. Extracción de la información

Karanokas, Tjortjis & Theodoulidis (2000), declaran que consiste en mapear textos del Lenguaje Natural (LN) como (informes de noticias, artículos periodísticos revistas, cualquier contenido de datos textual, en representaciones estructuradas predefinidas que representan un extracto del texto principal. Está compuesto por relaciones entre dichas entidades y eventos que forman parte de ellos mismos. Se almacenan en bases de datos y posterior mente se les aplica preprocesos de MT para una futura utilización.

2.4. Recuperación de la información (RI)

Nayak, Prasad & Senepati (2015), destacan que el proceso de RI consiste en reconocer los documentos que existen en una colección que coincidan con la consulta solicitada por el usuario. Dang & Ahmad (2013), proponen que la RI se encarga de encontrar coincidencias de búsqueda dentro de un conjunto de textos para su interpretación. Ahmad & Dang (2014), deducen que la RI es la encargada de obtener información importante de una gran colección de varios recursos como páginas web, pdfs, diapositivas, artículos etc.

2.5 Algoritmos KDD

El procesamiento de las bases de datos denominado (KDD) derivado por siglas en inglés (Knowledge Discovering in Databases) es el proceso que involucra el descubrimiento del conocimiento, es un proceso de identificación. Usama et al. (1996, p. 83) definen algoritmos KDD como el proceso de usar bases de datos aplicado a cualquier selección requerida, pre-procesamiento, sub-muestreo y transformaciones de éste.

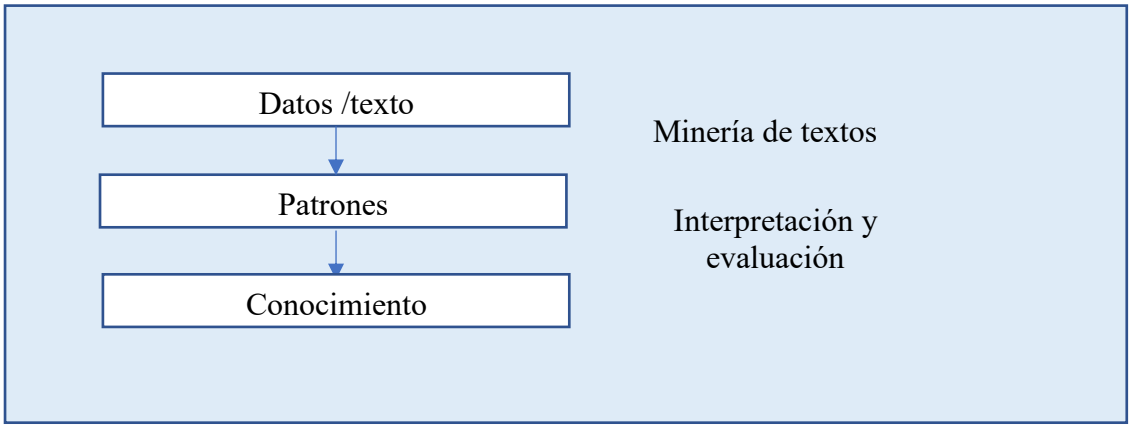
Simoudis mencionado en Fayad define los KDD como "el descubrimiento en las bases de datos, es un campo de la inteligencia artificial de rápido crecimiento, que combina técnicas de aprendizaje de máquina, reconocimiento de patrones, estadística, bases de datos y virtualización para automáticamente extraer conocimiento y/o información de un nivel bajo de datos" (*bases de datos*) (p. 1). Es de vital importancia la preparación o pre procesamiento de los datos, es el paso que más demora debido a que la información puede estar con inconsistencias**.** Fayyad et al (1996, p. 82) mencionan que los algoritmos KDD se centran en encontrar patrones comprensibles que puedan ser interpretados como conocimiento útil y además ponen gran énfasis con grandes cantidades de datos del mundo real, siendo de fundamental interés.

Molina mencionado en Valcárcel (2004, p. 83), define los algoritmos KDD como *la extracción no trivial de información potencialmente útil a partir de grandes volúmenes de datos en el cual la información está implícita, donde se trata de interpretar grandes cantidades de datos y encontrar relaciones y patrones*. Para conseguirlo harán falta técnicas de aprendizaje, estadística y bases de datos". Troche (2014, p. 59), declaran que *es proceso para identificar comportamientos o patrones, novedosos potencialmente útiles y con la*

característica que son comprensibles a partir de los datos, encontrar conocimiento útil, valido, y relevantes.

La Figura 2.5 se describe el proceso de patrones que se encuentran en un conjunto de datos para la interpretación, evaluación y obtención del conocimiento.

Figura 2.5. Proceso de descubrimiento de conocimiento (KDD)



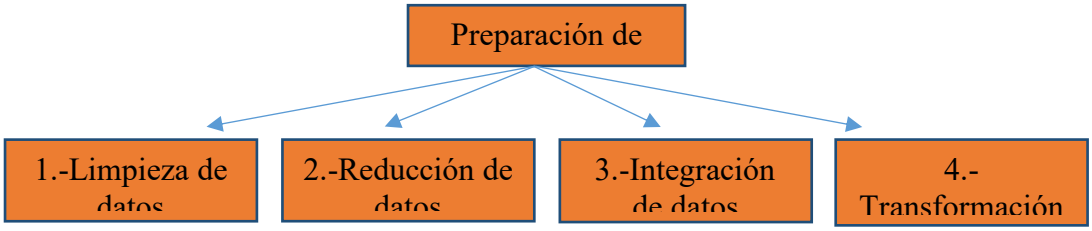
Fuente: Valcárcel (2004, p. 84).

El método tradicional de convertir datos en conocimientos se basa en la interpretación y el análisis manual pero cuando el volumen de información es sobresaturado, por lo tanto, los algoritmos KDD son un intento de herramienta digital para solventar el problema (Fayyad, Shapiro & Smyth; 1996). Witten et al mencionados en Quiroz & Valencia (2012, p. 98), sostiene que KDD son la manera para obtener nuevo conocimiento a partir de grandes volúmenes de datos almacenados en diferentes formatos.

Algunas de las aplicaciones de KDD se incluyen en áreas como: mercadotecnia para detectar, analizar su previo comportamiento de grupos de clientes y en finanzas e inversión para el desarrollo de estrategias económicas, también detección de fraudes para monitorear las cuentas, en manufactura para detectar problemas en los procesos de producción, en telecomunicaciones para detectar herramientas de recuperación apoyo e interactividad y establecer reglas de costos y realizar cálculos de costos en limpieza de datos *data cleaning* para detectar conocimiento a partir de su análisis. El proceso KDD se puede considerar como actividad multidisciplinaria ya que se incluye aprendizaje computadora ML y MT e

inteligencia artificial. La Figura 2.6 muestra el proceso o estrategias para mejorar la calidad de los datos.

Figura 2.6. Estrategias para el pre-proceso de datos



Fuente: Herrera y Cano (2006, p. 2).

Cano & Herrera (2006) definen el proceso de reducción de datos en KDD en cuatro pasos: Limpieza de datos: Se aumenta la calidad de datos requerida mediante técnicas de análisis selectivo. *Reducción de datos:* se decide qué datos deben de ser utilizados para el análisis, a criterio del investigador. *Integración de datos*: combinar múltiples tablas o registros para crear nuevos registros o valores. La combinación de datos también incluye la agregación que consiste en operaciones, donde se obtiene nuevos valores mediante la unión de información de varios registros o tablas. *Transformación*: son las modificaciones sintácticas llevadas a cabo sobre los datos, sin que supongan cambio en el significado del mismo.

El proceso KDD es interactivo ya que envuelve numerosos pasos con decisiones hechas por el usuario. Los algoritmos KDD ayudan en la toma de las decisiones importantes, ya que proporciona conocimiento de un proceso automatizado y puede ser utilizado en diferentes áreas como astronomía, aspectos climatológicos, medicina, mercadotecnia, análisis de mercados etcétera.

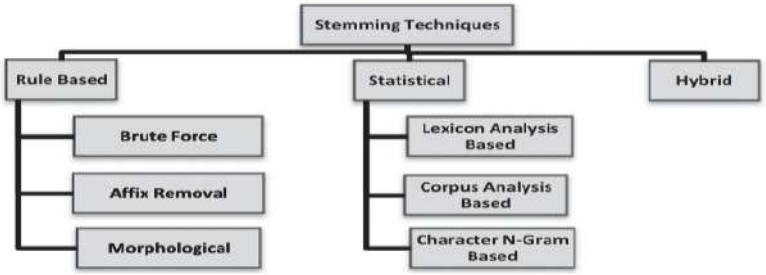
El proceso comienza determinado los objetivos de KDD, después se seleccionan cuáles datos deben de ser añadidos incluyendo que datos deberían ser añadidos e integrados en la selección. A continuación, se procesan y limpian a través de un filtrado de acuerdo a las especificaciones anteriores, siguiendo una transformación de los resultados obtenidos, según el algoritmo implementado se reducen o se expanden. Ésta es la parte crucial de todo el

proceso, pues dependiendo de los resultados se escoge la MT para realizarle el trabajo. Por último, se corre el algoritmo las veces que sean necesarias para encontrar el resultado deseado como mínimo deben ejecutarse 3 veces.

2.5.1. Stemming.

Para Singh & Gupta (2016, p. 45) lo definen como *el proceso en el cual las palabras son asignadas o llevadas a su forma de raíz*. Es una de las formas más utilizadas por *Procesamiento del Lenguaje Natural (PLN)* y *Recuperación de Información (RI)* para reducir el tamaño de las bases de datos ya que al existir palabras duplicadas o triplicadas en su forma de origen se pueden eliminar de manera sencilla reduciendo drásticamente los tamaños de almacenamiento, además permite a que el proceso de RI se realice de una manera más rápida eliminando tiempos de procesamiento. La Figura 2.7 describe la clasificación de tipos de *stemming* por técnica.

Figura 2.7. Clasificación de Stemming por técnica



Fuente: Sigh & Gupta (2016, p. 45.)

La Tabla 2.4 muestra la clasificación de técnicas de *Stemming* de acuerdo a Sigh & Gupta (2016, p. 45) y su posible aplicación basadas por reglas, por staticas o hibridas.

Tabla 2.4.

Técnicas de Stemming

Basada en reglas: es aquella que transforma la variante de la palabra en su raíz basada en una regla predefinida. Su mayor ventaja es su facilidad de uso por las

reglas específicas del idioma, una vez creada se pueden utilizar sin ningún procesamiento adicional.	
Algoritmo de fuerza bruta	
Algoritmo de eliminación de Affix	Utilizan tabla de búsqueda para devolver la raíz de la palabra.
Algoritmos morfológicos	Los que se enfocan al sufijo y prefijo de una palabra y utilizan técnicas para eliminar los mismos de las formas variantes.
Basada en raíz estática: son aquellos que se entrenan de manera supervisada o semi supervisada para aprender las reglas de derivación de un idioma determinado.	Aplican un análisis morfológico, requieren léxicos complejos para realizar el análisis.
Léxico basado en raíces de análisis	
Análisis basados en raíz	Analizan un conjunto de palabras obtenidas para agruparlas por conjuntos léxico gráficamente.
Carácter N- Gram	Aquellos que se agrupan morfológicamente relacionados por el contexto.
Algoritmos Híbridos: aquellos que combinan los anteriores para mejorar su funcionamiento, aquellos que se adaptan al idioma en el que se desarrolla.	Son los que aprenden basados en la derivación a través de la frecuencia de N-Gramas obtenidos a partir de las palabras.

Fuente: Sigh & Gupta (2016, p. 45).

2.5.2. Algoritmo Porter.

El algoritmo Porter se deriva de una técnica que permite extraer sufijos y prefijos. La notación fue escrita en 1979 y publicada en 1980 por Roberson y MF Porter, en el laboratorio de computación de la universidad de Cambridge Inglaterra. Bordingon (2017), describe que su funcionamiento *consiste en leer un archivo después tomar una serie de caracteres, y de ahí una palabra para verificar que esa serie contenga letras de ser así aplica la regla de extracción de sufijos.* Para que este proceso se pueda llevar a cabo, es necesario realizar un conjunto de reglas, cada una de ellas está formada por n criterios, y cada criterio debe

contener: los elementos descritos en la Tabla 2.5. Su objetivo principal es ayudar al investigador ajustando las reglas del idioma para su uso facilitando su aplicación casi a cualquier lengua.

Tabla 2.5.

Elementos de regla de un proceso de Stemming

Descripción del elemento
Un identificador de regla
El sufijo a identificar
El tamaño de sufijo
El tamaño de texto de remplazo
Tamaño mínimo que debe tener la raíz resultante luego de aplicar la regla.
Una función de validación (verifica si se debe aplicar la regla una vez que se ha encontrado el sufijo).

Fuente: Bordignon (2017, p. 5).

2.6. Procesamiento de lenguaje natural PLN

El procesamiento del lenguaje natural (PLN) comenzó a principios de 1950 como parte del proceso de recuperación de información RI, Manning et al mencionado en (Nadkarni, Ohno-Machado & Chapman, 2011) concluyen que el PLN toma prestados de varios campos de las ciencias computacionales sus técnicas por lo cual los desarrollados de hoy en día deben ampliar su base de conocimiento significativamente para poder implementarlo correctamente. Un desafío para las reglas escritas a mano en el PLN es la gramática, el PLN debe de extraer el significado de las palabras y así mismo encontrar la relación entre ellas, además al generarse el lenguaje de expresiones cotidianas del día, debe de aprenderlas para poder interpretarlas por lo cual genera que la primera vez que se realiza el proceso sea retardado además de que debe corregirlas para poder llegar a su raíz, se basa en reglas gramaticales previamente definidas.

Muñoz & Rodríguez (1996, p. 206) mencionan que el (PLN) consiste en el estudio y análisis de aspectos lingüísticos de un texto a través de programas informáticos. La evolución constante de la tecnología promueve que el análisis crítico del discurso gracias al PLN se pueda realizar de manera automática. También definen que el *LN se distingue de los*

lenguajes artificiales por su riqueza en vocabulario, en su flexibilidad por las múltiples excepciones, ambigüedad por diversos significados en las palabras y por la indeterminación debido a sus diversas interpretaciones. Por tal motivo se considera que el proceso de PLN, es un proceso complejo que necesita llevar a cabo un análisis de conocimiento y procesamiento de razonamiento.

Hirschberg & Manning (2015, p. 261), señalan que el PLN "emplea diferentes técnicas computacionales con el propósito de aprender, entender y producir contenido de lenguaje humano". Ordoñez & Gelbukh (2010, p. 7) definen el PLN como "una disciplina de apoyo para la lingüística computacional encargada de proveer soluciones para la interpretación y la gestión del lenguaje natural con soporte de herramientas y técnicas propias de la estadística, matemáticas, lingüística e inteligencia artificial". Cortez et al (2009, p. 48) lo definen como "la utilización de un lenguaje natural para comunicar con la computadora, debiendo ésta entender las oraciones que le sean proporcionadas, facilitando el desarrollo de modelos que ayuden a comprender los mecanismos humanos relacionados con el lenguaje".

Existen factores claves que han permitido los avances de desarrollo del PLN como el aumento de potencia en el procesamiento de las computadoras desarrollo de métodos de aprendizaje en la computadora ML, y, sobre todo, una comprensión del lenguaje humano por parte de la computadora. Durante los últimos años los científicos han intentado escribir reglas y vocabulario para las computadoras, comenzando formalmente en los ochentas, pero transformada por los científicos en los noventas con la finalidad de construir grandes cantidades de lenguajes lingüísticos empíricos. Finalmente, gracias a todos los cambios que se han realizado, se logra que el objetivo del PLN tenga una rica comprensión lingüística como herramientas de alto rendimiento que facilitan la información sintáctica y semántica para poder transformar, comprender el contexto para su uso. Hirschberg y Manning (2015, p. 261) consideran que *"*la mayor limitación del (PNL) es el hecho que los recursos y sistemas solamente están disponibles en pocos lenguajes como inglés, francés, español, alemán y se considera como un reto que en poco tiempo estén disponibles en todos los idiomas".

El PLN representa importantes oportunidades para la minería de textos de forma libre sobre todo en anotación automatizada de indexación. Hacia la comprensión de idiomas puede ayudar sustancialmente. El PLN puede contribuir significativamente como interfaz efectiva

para afirmar mejoras para algoritmos de minería de textos y explicar el conocimiento derivado de un sistema de KDD.

Las computadoras deben tener la habilidad para poder entender el lenguaje humano traduciéndolo en lenguaje de computadora (*machine translation*). En la lingüística se consideran diferentes sub-campos que se encargan de estudiar la estructura del lenguaje y están implícitas en el PLN, así mismos que se muestran en la Tabla 2.6.

Tabla 2.6.

División de sub-campos de lingüística

Sub-campo	Descripción
Fonética	Estudio de los sonidos del lenguaje humano.
Fonología	Se encarga de la organización sistemática de los sonidos en las lenguas.
Morfología	Estudia la estructura interna de las palabras para delimitar clasificar y definir sus unidades
Sintaxis	Estudia la formación y estructura interna de los enunciados.
Semántica	Estudia el significado, sentido o interpretación de signos lingüísticos.
Pragmática	Estudia la forma en que los enunciados con sus significados semánticos son usados para comunicación en particular.

Fuente: Bender (2013, p. 21).

Kao & Poteet (2007), concluyen que el PLN es el intento de extraer una mayor parte de significado de texto libre (p. 12). Se encuentra ligado directamente con la estructura de textos, ya que utiliza y analiza la estructura gramatical de las palabras y frases. Cambria (2014) establece que los algoritmos del procesamiento del lenguaje natural están limitados por el hecho de que solo pueden procesar información que solo pueden ver. Los modelos computacionales intentan emular de manera artificial como el cerebro humano procesa, por tal motivo es importante la comunicación y traducción entre lenguaje humano y lenguaje computadora.

Existe otro enfoque de PLN considerado como análisis de opiniones ya que estudia la opinión expresada en particular y puede ser subjetiva. Saggion (2010, p. 247) señala que el análisis de la opinión se enfrenta a los siguientes problemas:

- ✓ Identificar si un fragmento de texto expresa opinión o no.
- ✓ Identificar quien es la entidad que expresa la opinión.
- ✓ Identificar el objeto o tema de la opinión.
- ✓ Identificar la polaridad de la opinión positiva o negativa.

Para Kevvit, Partridge & Wilks (1992), definen las diferentes teorías y modelos de PLN argumentan que "la coherencia de un discurso puede medirse en tres temas principales, el significado, la estructura y la intención" (p. 12). Zadeh (2004) determina que la función primaria del PLN "es proporcionar un marco computacional para servir como medio de compresión del significado y la representación del significado, entre lenguaje computadora y lenguaje humano" (p. 75).

Reshamwala, Misrha & Pawar (2013, p. 113) determinan que el PLN es "un área de investigación y aplicación que explora como se pueden usar las computadoras para comprender y manipular el texto o el habla del lenguaje natural para hacer cosas útiles". La Tabla 2.7 demuestra la clasificación del lenguaje natural propuesta por Reshamwala, Misrha & Pawar (2013, p. 113). De acuerdo a la síntesis del habla o al reconocimiento de voz.

Tabla 2.7.

Clasificación del PLN por método

Métodos para la clasificación del lenguaje natural	
Procesamiento del lenguaje natural para la síntesis del habla:	Basado en la velocidad de la conversación observando cual es primer dato en entrar al sistema, utiliza la segmentación de oraciones que trata con los signos de puntuación con un árbol de decisión.
Procesamiento del lenguaje natural para el reconocimiento de voz:	El sistema automático de reconocimiento utiliza técnicas del PLN basadas en gramática. Utiliza las gramáticas libres de contexto para representar sintaxis de ese lenguaje, presenta un medio para trata lo espontaneo a través de la adición de resúmenes automáticos, incluida la indexación que extrae la esencia de las transcripciones del habla para

	tratar la recuperación de información IR y los problemas del sistema de dialogo.

Fuente: Reshamwala, Misrha y Pawar (2013, p. 113).

La Tabla 2.8 explica los diferentes niveles de aplicación en el PLN de acuerdo a fonología y morfología además demuestra los diferentes tipos de análisis.

Tabla 2.8.

Niveles de PLN.

<table>
<tr><td>Fonología:
Es aquella que trata con la interpretación de los sonidos del habla dentro y entre las palabras.</td><td colspan="2">• Reglas de fonética.
• Reglas de fonémicas.
• Reglas prosódicas.</td></tr>
<tr><td rowspan="2">Morfología:
Es la primera etapa donde la información ha sido recibida, es útil para identificar las partes del discurso en una oración y las palabras que interactúan juntas.</td><td>Sintaxis:
Es la aplicación de reglas de gramática del idioma destino, su función es determinar el papel de cada una de las palabras en una oración y organizar sus datos en una estructura</td><td>• Gramática.
• Análisis sintáctico.
• Análisis semántico.
• Análisis pragmático</td></tr>
<tr><td colspan="2">Ambigüedad:
Explicado como el de que un enunciado en un lenguaje humano puede tener más de un significado.</td></tr>
</table>

Fuente: Reshamwala, Misrha y Pawar (2013, p. 113).

Cortez et al. (2009) determinan que el PLN puede ser aplicado en: traducción automática de textos, recuperación de la información, extracción de información y resúmenes, resolución cooperativa de problemas, tutores inteligentes y reconocimiento de voz. Por lo tanto, su aplicación puede ser variada siendo un beneficio para la investigación el poder contar con una herramienta tan útil para la obtención del conocimiento. El mayor reto para PLN es el poder entender la solicitud del usuario, en donde cada cliente tiene necesidades diferentes,

por lo tanto, la resolución para cada uno será diferente. La formulación de su respuesta es la que determina la utilización de los distintos procesos que se servirán para resolverlo.

2.7. Extracción de la información (EI)

Para Lathila & Meenakshi (2014) definen la EI como "las tareas relacionadas con en localizar ítems específicos en documentos del lenguaje natural" (p. 433), el gran problema radica en transformar los documentos en más estructurados. Wilks (1997) aclara que la tecnología de EI tiene gran significancia en la información derivada de los usuarios finales, especialmente en compañías de finanzas, bancos, revistas y en gobiernos. Para Cunningham (1997) lo describe como "el proceso de que toma textos no vistos como entrada y produce datos de formato fijo e inequívocos como salida". La Tabla 2.9 muestra los diferentes tipos de EI clasificados por categoría.

Tabla 2.9.

Tipos de EI

Nombre de EI	Función
Reconocimiento de entidad nombrada (REN)	Busca y encuentra todos los nombres, lugares, organizaciones, fechas, y montos de dinero.
Resolución de la conferencia (RC)	Identifica las relaciones entre entidades en textos, las entidades son ambas identificadas por el reconocimiento REN y referencias anafóricas a esas entidades.
Construcción de elemento de plantilla (CEL)	Agrega información descriptiva a los resultados de REN. La información descriptiva asociada con las entidades.
Producción de plantilla de escenario (PPE)	Ajusta los resultados CEL en escenarios de eventos especificados.

Fuente: Elaboración propia basado en Cunningham (1997, p. 2).

Desde el punto de vista del usuario final *REN*, *RC* y *PPE* son los más relevantes para realizar la tarea de EI, además de proveer un nivel elevado acerca de los textos. Abdelmagid et al (2015, p. 1068), afirman que la *EI* es el campo de extracción útil que utiliza diferentes métodos y enfoques mediante el *PLN*. En la actualidad los investigadores siguen en la

búsqueda continua de buscar métodos que ayuden en la maximización de tiempo para llevar a cabo la *EI*, tratando de acotar los procesos simplificándolos eliminando, examinando los métodos actuales para innovarlos. Su principal función es localizar información específica en documentos textuales, videos, audios etc. La Tabla 2.10 muestra los procesos internos de la *EI*, a través de análisis de texto.

Tabla 2.10.

Clasificación de los procesos de la EI

Número de proceso	Función
1	El sistema extrae hechos individuales del documento a través de análisis de texto local.
2	Integra los hechos para producir un hecho más grande o inferir en un hecho nuevo.
3	Toma lugar después de la integración de hechos pertinentes se traducen en formato de salida.

Fuente: Abdelmagid et al (2015, p. 1068).

La tabla 2.11 muestra las diferentes tareas que se realizan cuando se realiza el proceso de la EI.

Tabla 2.11.

Clasificación de las tareas de la EI

Nombre de la tarea	Función
1	Reconocimiento de la entidad (REN), es el proceso de encontrar cosas específicas en el texto (personas, ubicación y organización.
2	Resolución de la correferencia de frase nominal que es el "proceso de verificar si dos expresiones en el lenguaje natural se refieren a la misma entidad y resolver referencias anafóricas por pronombres y sintagmas nominales definidos.
3	La resolución de correferencia entre documentos que se usa cuando se discute el mismo nombre de una entidad, persona, organización o ubicación en más de una fuente de texto.
4	Reconocimiento de roles semánticos que es un conjunto de roles que van desde significado restringido "especifico" al amplio significado "general".

Fuente: Abdelmagid et al (2015, p. 1069).

Cimmiano et al. (2005, p. 60) declaran que la *EI* es la tarea de identificar, recopilar y normalizar información relevante de textos en lenguaje natural y producir un conjunto de estructuras de conocimiento como resultado. Varsha & Khandewal (2016, p. 16) manifiestan que la EI es una técnica que tiene como objetivo de extraer los nombres de entidades y objetos del texto e identificar los caracteres que interpretan descripciones inventivas. La Tabla 2.12 muestra los factores que afectan el desempeño de *EI*.

Tabla 2.12.

Factores que afectan el desempeño del EI

Factor	Función
Tipo de texto	Todos los textos con los que se está trabajando.
Dominio	El amplio tema de los textos.
Escenario	Tipo particular de evento en el que EI está interesado en trabajar.

Fuente: Cunningham (1997, p. 4).

Un proceso típico de *EI* únicamente analizar textos y presenta información sobre la cual el usuario esté interesado. Grishman (1997,) declara que la *EI es la identificación de una clase particular de eventos o relaciones en un texto de lenguaje natural y la extracción de argumentos relevantes del evento o relación* (p. 1). La EI involucra la creación de bases de datos provenientes de los textos analizados. Sarawagi (2007, p. 263) ratifica que el *EI se refiere a la extracción automática de información estructurada como entidades a partir de fuentes no estructuradas*. La Tabla 2.13 muestra las diferentes aplicaciones de la EI en la industria.

Tabla 2.13.

Aplicaciones de la EI

Nombre	Función
Aplicaciones empresariales	Seguimiento de noticias: Rastrea automáticamente tipos de eventos específicos de fuentes de noticias.

	Atención a clientes: recopilan información no estructurada de la interacción con el cliente, se encuentran ligadas estrechamente con las bases de datos estructuradas y comerciales de la empresa. El amplio tema de los textos. Limpieza de datos: es aquella que se encarga de organizar, limpiar y estructurar los datos, provenientes de formularios estructurados.
Gestión de información personal	Buscar organizar datos personales como documentos, emails, proyectos, personas, en un formato estructurado entre enlaces.
Aplicación científica	Aquella encargada de buscar entidades u objetos en proteínas y genes.
Aplicaciones orientadas a la web	Estructuras de datos de muchos niveles, para segmentar la información, por ejemplo, bases de datos para citas textuales.
Bases de datos de opinión	Opiniones en texto libre, se pueden mejorar las bases de datos al largo de campos estructurados.
Sitios web comunitarios	Creación de bases de datos a partir de documentos web, comunitarios.
Comparación de compras	Se utilizar para comparación de compras que rastrean automáticamente los sitios web.
Colocación de anuncios de páginas web	Extracción de menciones de productos y el tipo de opinión del producto, ejemplo: google.
Búsqueda de web estructuradas	Búsqueda de palabras clave para obtener información de entidades que son típicamente sustantivos o frases nominales.

Fuente: Sarawagi (2007, p. 268).

Chaix et al (2018) mencionan que la EI es la encargada en reconocer piezas específicas de información previamente pre-definidas, estas entidades incluyen términos de EI de particular interés y la relación entre los mismos. Jiang (2012) enfatiza que "la EI consiste en descubrir información estructurada de texto no estructurado o semi-estructurado" (p. 11). La Tabla 2.14 muestra los tipos de estructuras extraídas por medio de la *EI* y la clasificación de las mismas.

Tabla 2.14.

Tipos de estructuras extraídas.

Nombre	Descripción

Entidades	Son frases nominales y comprenden de uno a algunos tokens en el texto no estructurado. Contienen un sustantivo o un adjetivo. Las más populares son llamadas (Named entities) o entidades de nombre que pueden ser, personas, ubicaciones, etc.
Relaciones	Las relaciones se definen sobre dos o más entidades relacionadas de una manera predefinida.
Adjetivos que describen las entidades	Es necesario asociar una entidad dada, con el valor de un adjetivo que describa la entidad, el adjetivo se deriva al combinar pistas blandas repartidas en muchas palabras diferentes alrededores de la entidad.
Estructuras como listas, tablas y ontologías	El alcance de los sistemas de extracción ahora se ha expandido para incluir la extracción de entidades atómicas y registros planos, pero también de estructuras más ricas como tablas, listas y árboles de varios tipos de documentos.
Tipos de fuentes no estructuradas.	
Granularidad de la extracción	Registro de sentencias: a partir de fragmentos de texto que son registros de texto no estructurado, como citas, anuncios clasificados, frases extraídas de un párrafo en el lenguaje natural. Párrafos y documentos: muchas tareas hacen necesario considerar el contexto de oraciones múltiples o documento completo para extracciones significativas, como extracciones de eventos periodísticos, extracción de números de parte, descripción de problemas de correo electrónico, etc.
Heterogeneidad de fuentes no estructuradas:	Páginas generadas por maquina: aquellas generadas por máquina altamente personalizada. Fuentes específicas del dominio parcialmente estructurado: son las que se encuentran dentro de un alcance bien definido, como las noticias, citas, currículos. Fuentes abiertas recientemente: existen gracias a extraer instancias de relaciones y entidades de dominios abiertos como las páginas web.
Recursos de entrada para la *EI*.	
Bases de datos estructuradas	Las bases de datos estructuradas existentes de entidades y relaciones son un recurso valioso para mejorar la precisión de la extracción.
Texto desestructurado etiquetado	Esta recopilación de texto no estructurado etiquetado requiere un tedioso esfuerzo de etiquetado. Este tipo de fuentes es más valioso que una base de datos estructurada ya que proporciona información contextual sobre una entidad y también porque la forma en que la entidad aparece en los datos no estructurados suele ser una forma muy ruidosa de su aparición en las bases de datos.

Fuente: Sarawagi (2008, pp. 272-275).

2.8. Análisis crítico del discurso (ACD)

Se cree que sus inicios comenzaron en la teoría crítica Frankfurt antes de la segunda guerra mundial, en los 70´s se conoció el surgimiento de un tipo de análisis de discurso y tenía como rol el juego del poder en la sociedad. Algunos otros investigadores solo se centraban en los aspectos formales del lenguaje, ahí es donde se forma la relación entre el lenguaje y el contexto centrado en la sociolingüística de los habitantes, pero se formalizó en Ámsterdam en enero de 1991. Desde entonces han nacido y se han planteado innumerables propuestas. Además, el *ACD* es considerado como un paradigma establecido en el campo de la lingüística. Dijk (1999, p. 2) afirma que el *ACD "*es el uso del lenguaje, los discursos y la comunicación entre gente real además poseen dimensiones intrínsecamente cognitivas, emocionales, sociales, políticas, culturales e históricas. Por otro lado, contribuye al entendimiento de las relaciones entre el discurso y la sociedad en general".

En el contexto actual las necesidades humanas basadas en una sociedad en desarrollo continuo se han dado cuenta que la información que trasmiten al comunicarse, permite generar conocimiento no solo en lenguaje hablado, sino también escrito. Para poder crearlo, es necesario generar un análisis del discurso, ya que los científicos se enfrentan a retos como a textos escritos, entrevistas, diálogos y a lenguaje. Por tal motivo, es necesaria una recolección de datos y antecedentes para que sean sometidos al análisis en forma de acción social ya que de cada discurso quedan huellas, pistas que se deben describir e interpretar.

Sayago (mencionado en Santander 2011, p. 207) señala que el *ACD* lo considera como "campo de estudio por su multidisciplinariedad por las diferentes ciencias del cual se constituye como por ejemplo lingüística, sociología, antropología, psicología social, cognitiva, ciencias políticas, ciencias de la comunicación, pedagogía etc. Por tanto, es una técnica de análisis potente y precisa*"*.

En el análisis del discurso existe la opacidad del discurso que tiene como significado la existencia de lenguaje oculto a través de los signos, que a veces expresa indirectamente lo que se quiere decir. El *ACD* estudia el lenguaje como practica social y considera que el uso del lenguaje es crucial y se interesa de modo particular en la relación entre el lenguaje y el

poder y como manifestación a través del lenguaje. Wodak & Meyer (2003, p. 19) determinan que el *ACD* "propone investigar de forma critica la desigualdad social tal como viene expresada, señalada, constituida, etc., por los usos del lenguaje No solo se centra en textos hablados o escritos considerados como objetos de investigación, comprende tres conceptos básicos poder, historia e ideología".

Para Wodak mencionado en (Van Dirjk, 1999, p. 3). Resume los principios básicos del *ACD* como: El *ACD* trata los problemas sociales. Las relaciones del poder son discursivas. El discurso constituye la sociedad y la cultura. El discurso es histórico. Es un enlace entre el texto y la sociedad de inmediato. El análisis del discurso es interpretativo y explicativo. El discurso es una forma de acción social.

2.9. Investigación social

El Análisis Cualitativo de Contenido *ACC* (Qualitative Analysis of Content) es uno de los métodos más utilizados desde los años 50s para analizar datos de texto. Para Downe-Wamboldt mencionado en Hsieh & Shannon, (2005, p.1278), explican que el objetivo del análisis de contenido es "proporcionar conocimiento y comprensión del fenómeno objeto de estudio, también el *ACC* es el método de investigación para la interpretación subjetiva de los contenidos de texto y datos a través del proceso de clasificación de sistemática de codificación e identificación de temas o patrones" los datos deben de ser almacenados en bases de datos u hojas de cálculo, para su análisis posterior.

Un reto para el *ACC* es que se puede confundir fácilmente con otros métodos cualitativos. Para Hsiu-Fang & Shannon (2005), el *ACC* convencional es "limitado en el desarrollo de la teoría y de la descripción de la experiencia vivida porque los procedimientos de muestreo y el análisis de la relación teórica entre los conceptos es difícil de inferir de los resultados" (p. 1278). Otro enfoque de *ACC* es el análisis cualitativo de contenido directo (ACCD) y puede ser definido como la forma de validar o extender conceptualmente un marco teórico. Su uso dependerá directamente de la estrategia que vaya a utilizar el investigador. Existe también la *Sumativa ACC* (SACC). Babbie et al mencionados en (Hsieh & Shannon 2005, p. 1285) definen el ACCD como el enfoque consistente en "descubrir subyacentes

significados de las palabras o el contenido y puede ser utilizado para analizar específicamente periódicos o contenido de libros de texto".

Este enfoque tiene la ventaja de que, estudiando el fenómeno de interés en una forma discreta, se pueden proporcionar ideas básicas de cómo se usan las palabras en realidad, pero los resultados están limitados por la falta de atención a los significados presentes en los datos. La Tabla 2.11 se muestran las diferencias entre 3 enfoques de ACC desde su estudio, definición y origen.

Tabla 2.15.

Diferencias entre 3 enfoques de ACC.

Tipo de ACC:	El estudio comienza con:	Momento de la definición:	Origen de los códigos.
Análisis de cualitativo de contenido convencional. (ACCC)	Observación.	Los códigos son definidos durante el análisis.	Los códigos derivan de los datos.
Análisis cualitativo de contenido directo. (ACCD)	Teoría.	Los códigos son definidos antes y durante el análisis de datos.	Los códigos se derivan de la teoría o los resultados de la investigación.
Sumativa de análisis cualitativo de contenido. (SACC)	Palabras Clave	Se identifican las palabras clave antes y durante los datos de análisis.	Las palabras clave se derivan del interés de los investigadores o revisión de la literatura.

Fuente: Hsieh y Shannon (2005, p. 1286).

Sandelowski (1995, p. 373-375). Define el análisis ACC en 5 etapas como sigue:

Preparación: los datos requieren ser preparados para lograr cumplir los objetivos, es decir una entrevista necesita ser trascrita de manera que conserve la preservación de los elementos de la entrevista inclusive no verbales como gestos o movimientos.

Análisis: continúa con el esfuerzo de entender el enfoque que tiene la entrevista. Sentido: el investigador debe estar enfocado y entender antes de hacer cualquier comparación con otras entrevistas, además debe leer las veces necesarias para no perder la esencia y características de las mismas.

Desarrollar un sistema: después de entender la entrevista, es necesario descartar los diferentes enfoques hasta encontrar el más productivo, de cada uno de esos enfoques.

Extracción de hechos: Se deben extraer todas las entrevistas en objetivos de lo más importante. De esta manera al investigador le pueden servir para un nuevo análisis de datos verbales.

2.10. Resumen en textos Automático (RTA) o Text Summarization (TSUM)

En la actualidad algunas personas prefieren leer resúmenes de textos en lugar de leerlo completo. La finalidad es presentar una herramienta que permita reducir tiempo y facilitar al lector una interpretación rápida resumiendo textos. Mani et al (2002), concluyeron que los lectores pueden leer resúmenes de textos más rápido que leer el texto completo. Por su parte Marbury (1995), define RTA como el proceso de extracción o destilación de la información más importante de un conjunto de textos, para producir una versión abreviada para usuarios en particulares tareas. El proceso de RTA puede ser aplicado para cualquier tipo de texto, puede ser utilizado para distinguir entre resúmenes de textos generando nuevo texto. Hans, Agus & Suhartono (2016), manifiestan que TS reduce el texto eliminando oraciones sin valor, ayudando al lector a obtener información más relevante de una manera rápida. También Jezek & Steinberger (2008), señalan que los RTA puede contener secuencias de palabras que no están en el texto original. Por su parte Babar (2013), concluye que la parte más importante de usar el TS es la reducción del tiempo de lectura. La tabla 2.12 muestra los diferentes enfoques de los TS.

Tabla 2.16.

Diferentes tipos de RTA

Basados en su enfoque

Extractivo	Abstractivo
los RTAE extractivos (EXT) funcionan seleccionando un subconjunto de palabras, frases u oraciones existentes del original texto para formar nuevos RTAE.	Hans, Agus & Suhartono (2016), ultiman que los TSUM o RTA puede realizarse por el método abstractivo (RTAA) que consiste en generar un enunciado de la representación semántica, y utiliza la técnica de PLN para crear un TS parecido a lo que un humano generaría. Por su parte Nallapati et al. (2016), Concluyen que el enfoque de RTAA es la tarea de generar un resumen breve que consta de pocas oraciones, que captura las ideas más destacadas de un artículo o un pasaje.
Basados en tipo de detalles	
Informativos	Indicativos
Gholamrezazadeh, Salehi, & Gholamzadeh (2009), determinan que el informativo sirve como una constitución original del documento, proporciona información concisa sobre el documento original para el usuario.	Gholamrezazadeh, Salehi, & Gholamzadeh (2009), concluyen que se utiliza para una vista rápida de un documento extenso y proporciona solo la idea principal del texto.
Basados en tipo de contenido	
Resumen Genérico	Basado en consultas
Gholamrezazadeh, Salehi, & Gholamzadeh (2009), declaran que, de ser utilizado por cualquier tipo de usuario, y toda la información tienen el mismo nivel de importancia.	Es aquel que se basa en una pregunta con respuesta, donde el resumen es resultado de una consulta.
Basados en evaluación	
Intrínsecamente	Extrínsecamente
Jing et al (1998), proponen que las evaluaciones intrínsecas se basan en la coherencia y la información. Bharti & Babu (2017), destacan que juzga la calidad del resumen directamente basada en el análisis en términos de algún conjunto de normas.	Mani (2001), define que los RTA extrínsecos son aquellos que se concentran en el uso de resúmenes en una tarea específica, por ejemplo, ejecución de instrucciones, recuperación de información, repuesta a preguntas y relevancia. Bharti & Babu (2017), proponen que es aquella que juzga calidad del resumen en función de cómo afecta la finalización de alguna otra tarea.
Predicción	Consultas
Dorr et al (2005), exponen que la RTA puede ser evaluada por la relevancia de	Goldstein et al (1999), determinan RTA basados en consultas para extraer oraciones,

predicción la cual compara los juicios de los sujetos en resúmenes con sus propios juicios en documentos de textos completos	relevantes en un documento. Pei-ying & Cun-he (2009), destacan las oraciones se agrupen de acuerdo a su distancia semántica, entre oraciones y luego calcula la similitud de oración acumulativa, entre todo el documento y finalmente se elijan oraciones basadas en reglas de extracción.

Fuente: Elaboración propia.

Joachims (1996), determina que para poder realizar el proceso de RTA se utiliza la representación vectorial *Frecuencia de Termino-Frecuencia inversa* de documento (TF-FID) por sus siglas en inglés *(TF-IDF).* Cuando se evalúan varios documentos es necesario utilizar *TF-FID*. El término frecuencia de una palabra w en un documento d, denotado *tf (w, d),* es el número de veces que la palabra *w* aparece es un documento *d.* Entre mayor sea el *tf* será más representativo. La frecuencia de un documento denotada *df(w)* es las veces en las que aparece w en un documento. La frecuencia inversa del documento *fid* expresa la importancia de la palabra en el documento y se calcula por la siguiente formula:

$$fid(w)=log(\frac{|d|}{df(w)}) \quad (1)$$

Si el *fid* de una palabra es bajo y si ocurre en varios documentos, quiere decir que tiene un poca representativo, pero si el *fid* es alto y aparece en pocos documentos, quiere decir que tiene gran nivel de representación. Por lo tanto, si el *td* y el *fid* son altos tendrán mayor representación y son los que se deben de considerar. Para calcular el *tf-fid* se utiliza la siguiente formula:

$$tf\text{-}fid(w,d)=tf(w,d)*fid(w). \quad (2)$$

Por el contrario, cuando se evalúa un solo documento es necesario realizar un cambio a la formula Tf-Idf quedando de la siguiente manera:

$$tf\text{-}isf(w,s)=tf(w,s)*fis(w). \quad (3)$$

La fórmula de *Idf* cambiara a *Isf* y quedara de la siguiente manera:

$$Isf(W)=Log(\frac{|s|}{sf(w)}) \quad (4)$$

Dónde: *sf* = número de oraciones en las que aparece la palabra *w* por sus siglas en inglés (sentence frequency). *s* = número total de oraciones, por sus siglas en inglés

(sentences). |s| =valor absoluto de *s*, se utiliza el valor absoluto para obtener la magnitud numérica sin importar su signo positivo o negativo. Log =Logaritmo, es el exponente que se debe elevar un número, llamado base para obtener otro número determinado. Para cada uno de los enunciados el promedio o (average) de *tfIsf* (w, s) es el peso del enunciado. El cálculo del promedio se realiza con la siguiente formula:

$$promedioTfIsf(s) = \sum_{i=1}^{W(S)} TfIsf(i,s)/w(s) \quad (5)$$

Dónde: *w(s)*=número de palabras en el enunciado. Cuando se terminan los cálculos correspondientes a cada una de las oraciones, a continuación, se selecciona los enunciados con más relevancia, los que tengan mayor promedio. El cálculo se realiza con el Máximo de promedioTfIsf.

$$MaxpromedioTfIsf = max(promedioTfIsf) \quad (6)$$

El límite o (Threshold) por sus siglas en inglés, es el porcentaje de promedio que asigna el usuario para realizar la comparación para realizar RTAE. Para calcular el *LimiteTiIsf* se utiliza la siguiente formula:

$$LímiteTsIsf = \%límite * max(promedioTfIsf) \quad (7)$$

Por lo tanto, el resultado de RTAE serán las oraciones que tengan un valor mayor o igual al promedio *TfIsf*. Tanto *TfIdf* y *TsIsf* tienen la misma función solo es cuestión de nomenclatura.

3. Modelos /Heurísticos

3.1. Modelo de espacio vectorial (MEV)

La forma de tratar con documentos textuales no estructurados es a través utilizar el modelo MEV ya que proporciona eficacia de análisis en grandes cantidades de textos. Allahyari et al. (2017), sugieren que la mejor forma de representar los documentos textuales es convertirlos en vectores numéricos, a este procedimiento se denomina *modelo del espacio vectorial*. Para lograr una mejor eficiencia en el manejo de grandes volúmenes de datos es necesario realizar una reducción de dimensión.

Para Salton et al. (1975), la forma más sencilla de representar el MEV es a través de la definición de las variables a utilizar: D= son las colecciones de documentos. D= $\{d_1, d_2, d_3 ...d_D\}$. V= también llamada vocabulario, es la colección de las diferentes palabras del texto en cada colección. V= $\{ w_1, w_2, w_3...w_v\}$

Hotho & Nürnberger (2005), admiten que la función principal del MEV es encontrar una codificación adecuada del vector de características. *La frecuencia del término w* $\in$ *(pertenecen) en el documento D* $\in$*(pertenece*) y es mostrada en la función fd*(w) y el número de documentos (w) es representada por fd(w).* El tamaño del vector se define por el número de palabras o grupos contenidos en la colección de textos.

El termino vector para un documento es d y se representa por: d: (fd(w1) , fd(w2) ... fd(wv). En el MEV cada *palabra* se representa por una *variable* con un valor numérico, dicho valor representa la importancia de la palabra en el texto, para calcular el peso de ese valor numérico se utiliza la frecuencia del término *tf*, y es considerada como la frecuencia inversa del documento *Idf (Tf-Idf).* Siendo *q* el termino de peso después el peso de cada palabra *w* $\in$ *d: q(w)=fd*(*w*)**log* $\frac{|D|}{fD(w)}$. |D|= es el número de documentos en la colección.

Tf-Idf = Esta normalización produce un decremento en el peso de los términos que ocurren con más frecuencia en la colección, consiguiendo que el emparejamiento de los términos se efectué por el distintivo y que sean palabras que se encuentren con menor

frecuencia en la colección, por consiguiente, cualquier documento textual puede ser comparado usando el MEV con operaciones simples.

Basado en el esquema del peso de los términos, cada documento es representado por un vector de termino de peso, $w(d) = (w(d, w1), w(d, w2) \dots w(d, wn)$.

La fórmula más similar de medida es la de coseno:

$$s\ (d1,d2) = cos(\theta) = \frac{d1.d2}{\sqrt{\sum_{i=1}^{n} w_{1i}^2} \cdot \sqrt{\sum_{i=1}^{n} w_{2i}^2}}$$

Se debe considerar que no todas las colecciones de texto que sean analizadas por el MEV no pueden descubrir entidades de relacionales entre sí por lo tanto no pueden ser determinas por el MEV. Sin embargo, si un documento contiene información desestructurada es necesario estructurarla para su posterior manipulación con herramientas de minería de textos.

3.2. Clasificación o categorización de texto (CT)

La categorización de los textos es una actividad que se realiza de manera automatizada. Para Sebastiani (2002) la CT consiste en etiquetar textos provenientes del lenguaje natural con categorías temáticas predefinidas de un conjunto de textos como una tarea programada. Consiste en asignar un valor booleano a cada par *de* $<d_j, c_i> \in DxC$ donde D es el dominio de documentos C es igual a $\{c_{1.....}c_{|c|}\}$ como un conjunto predefinido de categorías. Un valor de T asignado para $<d_j, c_i>$ indica la decisión para el archivo d_j sobre c_i mientras un valor de F indica la decisión de no archivo d_j *sobre* c_i.

De forma más natural consiste en aproximar el objetivo desconocido mediante la función $ø: D\ x\ 0 \rightarrow \{T, F\}$ llamando el clasificar la regla debe coincidir lo máximo posible con ø. En la actualidad la CT es utilizada principalmente por entidades de inteligencia de artificial. CT es una ayuda esencial para la MT. Niharika et al (2012), determinan que la CT consiste en determinar los temas principales de un documento, al categorizarlos, por lo tanto, consiste en clasificar un conjunto de documentos en un número fijo de categorías predefinidas.

Se ha convertido en un problema basado en técnicas de ML que inician como un proceso inductivo que construye automáticamente un clasificador aprendiendo de un conjunto de

documentos preclasificados en diferentes categorías y además comparte características con recuperación de la información (RI).

Mitchell (1997) declara que la clasificación de textos es aquella cuyo objetivo es asignar clases predeterminadas a documentos textuales. La clasificación es definida como un conjunto $D = \{d_1, d_2, d_3 \ldots d_n\}$ de documentos en donde cada documento d_i es etiquetado como l_i del conjunto $L = \{l_1, l_2, l_3, \ldots l_k\}$. La labor de encontrar un modelo de clasificador donde $f: d \rightarrow l \; f(d) = l$ Rao & Vemuri (2017) concluyen que la función del clasificador es fusionar documentos de texto en uno o categorías más predefinidas en función de su contenido. Wang & Xia (2010) declaran que el algoritmo KNN es uno de los algoritmos de aprendizaje más simple, con el objetivo de crear clases predefinidas de un grupo muestra.

McCallum & Nigam (2001) determinan que el clasificador Näive Bayes es un clasificador famoso por su simplicidad, además de asumir que todos los atributos son independientes entre sí dando el contexto entre sí. Generalmente se utiliza como base en ML como un condicional de probabilidad, basado en clases que son seleccionadas mediante métodos.

3.3. Análisis de sentimientos

Bolande & Olumide (2012) concluye que la minería de opinión (MDP) consiste en determinar la polaridad positiva o negativa de una opiniones o comentarios, acerca de un producto o servicio, articulo o evento con el propósito de tomar decisiones basadas en dichas opiniones. Wei & Hung (2009) declaran que la MP se puede enfocar en dos direcciones en MDP en documento y MDP en características.

4. Aspectos Metodológicos

4.1. Definición del problema

Para identificar con claridad el problema a resolver, se realizó una entrevista a estudiantes del Doctorado en Tecnologías Educativas que realizan investigaciones cualitativas o mixtas y que han utilizado el Atlas.ti como herramienta de análisis de entrevistas. Los estudiantes manifestaron que el contador de palabras, contabiliza palabras simples y no compuestas, tampoco hace diferencia entre palabras con la misma raíz como por ejemplo clase o clases, las cuenta como palabras diferentes, por otro lado, no distingue sinónimos, así mismo no existe un diccionario de palabras por categoría y por último no hay un clasificador de palabras. También es importante mencionar que al clasificar la información por categorías que el investigador definió con anterioridad, el software no realiza ninguna interpretación y es el investigador quien explica subjetivamente el significado de datos con base a la información que le fue proporcionada y de acuerdo a su experiencia

4.2. Objetivos generales y específicos

Objetivo general: Proponer un algoritmo para el proceso de extracción del conocimiento (*Knowledge Discovery in Data Bases*) a través de algoritmos de *Minería de Texto (MT)* y *Análisis Crítico del Discurso (ACD)* para facilitar la interpretación objetiva de los datos generados por la investigación social.

Objetivos específicos:

- ✓ Identificar los algoritmos de MT que permitan la extracción del conocimiento.
- ✓ Analizar metodologías de análisis del discurso.
- ✓ Identificar algoritmos que coadyuven en el análisis del discurso
- ✓ Realizar pruebas y presentar resultados del funcionamiento del algoritmo.
- ✓ Fundamentar el uso de dicho algoritmo.

4.3. Justificación

En los programas doctorales de la Facultad de Informática de la Universidad Autónoma de Querétaro se realizan investigaciones de corte cualitativo, cuantitativo y de investigación

aplicada. Esta propuesta va a contribuir a que los investigadores, tanto alumnos como docentes, puedan extraer el conocimiento con mayor facilidad de la información procedente de investigaciones sociales o cualitativas en las que la información procede del análisis de entrevistas. Al aplicar el algoritmo propuesto con un conocimiento filtrado por categorías de análisis permitirá a los investigadores una visión panorámica de los datos para que puedan encontrar con mayor facilidad el significado de lo que quisieron decir los entrevistados ayudando a encontrar patrones. Con la propuesta se contribuye a la conservación de la consistencia de la información logrando una interpretación objetiva y homogénea.

4.4. Metodología para la investigación cualitativa, estudio de caso, transversal

4.4.1. Fuentes de información.

Documentos generados por el Atlas.ti

4.4.2. Método aplicado.

Para poder generar una solución práctica al problema propuesto en esta tesis fue necesario generar un proceso general que permita la obtención del conocimiento, se creó un algoritmo y se convertio en aplicación para la implementación del método RTAE provenientes de RI de la MT para que pudiera ser compatible con la información proveniente de entrevistas derivadas del Atlas.ti.

Se consideró el comportamiento de la integración de un corrector ortográfico como un pre-proceso extra durante la etapa 1 de la MT, se analizaron diferentes modelos y propuestas que ejemplificaron un modelo de las instancias de la MT y los pre-procesos como un elemento particular, indicando la importancia y la necesidad de ejecución además resumiendo las herramientas de código abierto para la evaluación. Se analizaron y describieron las técnicas para realizar la MT; el modelo es completo, con un algoritmo de reglas y técnicas de uso orientada a la web aplicadas a RTAE. Se determinó la necesidad del tiempo que consumen los pre-procesos en MT, a través de un modelo secuencial con el objetivo de limpiar e identificar patrones.

Se analizaron y describieron las áreas y el modelo general de la MT y realizo una comparación de las herramientas disponibles. Se describen las reglas para evaluar la gramática para que pueda ser procesada por la computadora a través del PLN. Se adaptaron los procesos con el enfoque de descripción el modelo de pre-procesos y modelo de procesos para la aplicación de la MT.

Se creó una notación capaz de eliminar caracteres especiales y no deseados. Se desarrolló un algoritmo que fuera capaz de separar el conjunto de palabras que se analizaron en elementos independientes (tokenización). Se adaptó el algoritmo snowball (AS) (2003) procedente del algoritmo Porter (AP) (1980) como algoritmo stemming, que fue desarrollado para el idioma inglés, además se creó un procedimiento para la eliminación de Stop Words. Se creó un conjunto de procedimientos para crear matrices dentadas en las cuales se utilizaron para análisis, almacenamiento de palabras y de resultados de operaciones estadísticas provenientes del modelo *Tf-Idf* para determinar la relevancia de las palabras en el texto analizado. El resultado del método aplicado se puede observar en la creación de la aplicación llamada "Resúmenes de textos J Quillo-Espino".

Hipótesis de la Investigación

Si se utiliza un algoritmo para el proceso de extracción del conocimiento (KDD) con base en algoritmos de minería de textos y de análisis del discurso entonces se logrará una interpretación objetiva de los datos generados por la investigación Social

4.4.3. Listado de indicadores.

Medidores de software.

Cuando se produce un sistema, se deben de emplear métodos o herramientas que colaboren a medir el proceso del funcionamiento, por tal motivo es necesario utilizar métricas que permitan realizar una evaluación cuantitativa del software. Mah (1999, p. 7) define el concepto de métrica como "mediciones basadas en técnicas para el proceso de desarrollo de software y sus productos, y conjunto de revisiones que mejoran los procesos y sus productos". De acuerdo con Pressman (2016, p. 357) las revisiones que se pueden llevar a

cabo para la medición del cumplimiento de los objetivos de la tesis de acuerdo a la evaluación del algoritmo propuesto.

La Tabla 4 muestra la clasificación de esfuerzo en horas hombre para la evaluación de un programa y da una descripción de cada uno de los términos utilizados en la evaluación de esfuerzo.

Tabla 4.

Clasificación de esfuerzo en horas hombre para la revisión de un programa

Clasificación	Descripción
Esfuerzo de preparación E_p	Esfuerzo (en horas hombre), para revisar un producto del trabajo antes de la revisión real.
Esfuerzo de evaluación E_a	Esfuerzo requerido (horas hombre), que se dedica a la revisión real.
Esfuerzo de la repetición E_r	Esfuerzo requerido (horas hombre) que se dedica a la corrección de errores descubiertos durante la revisión.
Tamaño del producto del trabajo TPT	Medición del tamaño del producto del trabajo que se ha revisado (por ejemplo, número de páginas del documento o de líneas de código).
Errores menores detectados $Err_{menores}$	Numero de errores detectados que pueden clasificarse como menores (requieren menos de algún esfuerzo especificado para corregirse.
Errores mayores detectados $Err_{mayores}$	Números de errores detectados que puedan clasificarse como mayores (requieren más que un error especificado para corregirse).

Fuente: Pressman (2016, p. 357)

Se puede calcular las siguientes fórmulas: Esfuerzo total de la revisión = $E_{revisión} = E_p + E_a + E_r$. Errores totales= $Err_{tot} = Err_{menores} + Err_{mayores}$

La densidad del error representa los errores encontrados por unidad del producto del trabajo revisada. Densidad del error = Err_{tot} / TPT

Las Revisiones Técnicas Formales (RTF): son aquellas encargadas de la calidad del software y sus funciones serán: descubrir errores en el funcionamiento, lógica o implementación de cualquier representación de software, verificar que software cumpla con los requerimientos. En la tabla 4.1 muestra el formato de hoja para reporte de revisión.

Tabla 4.1

Reporte y registro de la revisión

Fecha: _____	
Objetivo de la Revisión: _____ _____	Nombre del revisor: _____ _____
Descubrimiento de la revisión: _____ _____	Conclusiones: _____ _____

Fuente: Elaboración propia basada en Pressman (2016, p. 357).

Se puede anexar una hoja de pendientes con la finalidad identificar problemas en los productos y funciona como lista de verificación de acciones. En la tabla 4.2 se muestran los lineamientos a seguir para llevar a cabo la RTF.

Tabla 4.2.

Consideraciones o lineamientos para la revisión:

Consideraciones	Descripción
Revisar el producto y no al productor.	Sin importar el ego y sentimientos personales la RTF debe dejar los participantes con una sensación de logro. Los errores se deben de señalar de forma amable. Debe de ser constructivo.
Se debe establecer una agenda y seguirla.	Una RTF debe de mantenerse activa durante el desarrollo del software.
Enunciar áreas de problemas, pero no intentar resolver cada uno.	La solución de problemas debe posponerse para después de la reunión.
Tomar notas por escrito.	Pueden tomarse en cualquier medio ya sea físico o digital.
Asignar tiempo de la RTF cortos	Dar capacitación a todos los revisores constantemente.

Fuente: Elaboración propia basado en Pressman (2016, p. 365).

Gracias a las RTF se considera que el software que se desarrollara tiene una mayor probabilidad de producir software de mayor calidad por ser estructuradas y además ayudan a reducir los errores en el desarrollo de software de manera notable.

4.4.4. Estrategias de pruebas de software.

Son aquellas que proporciona los pasos que deben realizarse como parte de la prueba ayudan a determinar el tiempo que se empleara para la ejecución y la recolección y evaluación de los resultados. La Tabla 4.3. demuestra la definición de verificación y validación.

Tabla 4.3.

Definiciones de verificación y validación.

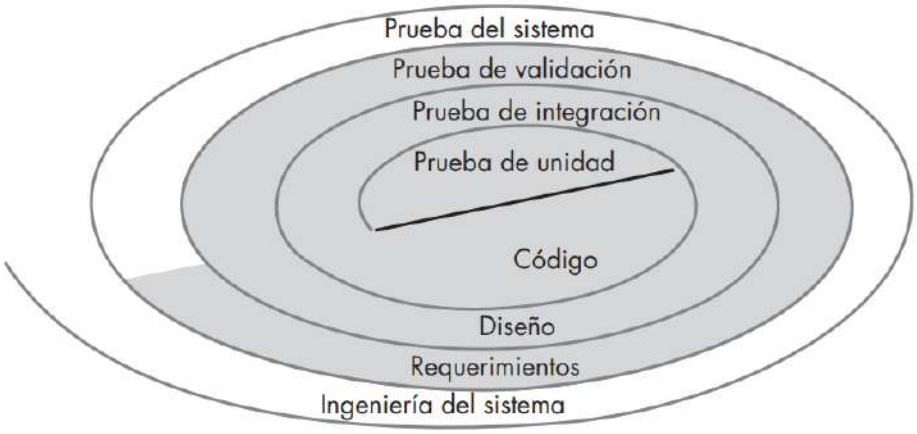
Nombre	Descripción
Verificación	Conjunto de tareas que garantizan que el software se implementara para cumplir una función específica (construido correctamente).
Validación	Conjunto de tareas que aseguran que el software que se construye sigue los requerimientos del cliente (construimos el producto correcto).

Fuente: Elaboración propia basada en (Boehm, p. 103).

Ambos procesos son los nombres técnicos que se les da a los procesos de comprobación y análisis que aseguran que el software que se desarrolla esta acorde a su especificación y cumple con las necesidades requeridas.

Las inspecciones de software son el análisis de representaciones del sistema como diagramas de flujo, código fuente etc. Se deben realizan durante todo el ciclo de desarrollo. La Figura 4.1 muestra el esquema de visión general de pruebas de software.

Figura 4.1 Esquema visión general de pruebas de software.



Fuente: Pressman (2010, p. 386).

La Figura 4.1 hace una muestra clara de cómo se debe comenzar la verificación y validación de software. El proceso comienza en prueba de unidad enfocándose en cada componente de manera individual, a continuación, los componentes deben de integrarse para formar el paquete o software, después la validación proporciona una la certeza de que el software cumple con todos los requerimientos y el rendimiento necesario. Las pruebas se deben de establecer de manera explícita. En la tabla 4.4 se muestra las diferentes pruebas que se le pueden realizar a un software orientado a objetos.

Tabla 4.4.

Tipos de pruebas para software orientado a objetos

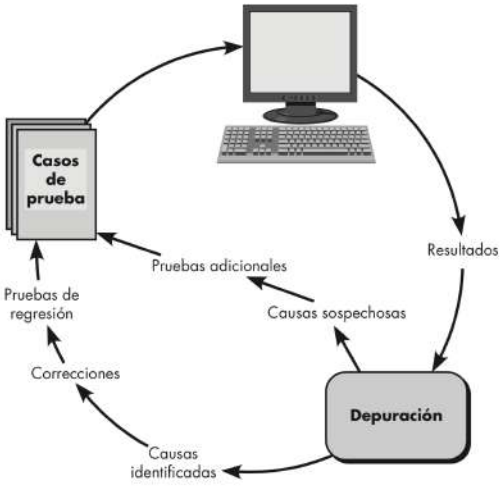
Tipo de prueba	Descripción
Prueba de unidad	es la que se enfoca en la verificación de la unidad más pequeña, y se aplica a un módulo de software o un componente.
Pruebas de integración:	son las que se efectúan para descubrir errores asociados con la interfaz.
Pruebas de validación:	la validación es correcta cuando el software funciona de tal manera que se cumplan las expectativas del cliente.
Revisión de la configuración:	son los que garantizan que todos los elementos de la configuración del software se desarrollaron de manera adecuada.
Pruebas alfa:	aquellas sé que llevan a cabo con una muestra representativa de usuarios se realizan en un ambiente controlado.
Pruebas beta	son aquellas que se realiza con uno o más usuario final y por lo general el desarrollador no se encuentra presente.

Pruebas de seguridad:	Intenta verificar los mecanismos de protección que se construyen en un sistema, en realidad lo protegerán de cualquier penetración impropia.
Pruebas de esfuerzo	son las que me miden el funcionamiento y rendimiento del sistema.
Pruebas de rendimiento	son las que se diseñan para poner a prueba el rendimiento del software dentro de un sistema integrado.

Fuente: Elaboración propia basada en Pressman (2010, p. 401).

La Figura 4.2 demuestra el proceso general de depuración del desarrollo de un software.

Figura 4.2. Proceso general de depuración



Fuente: Pressman (2010, p. 405).

El proceso de depuración comienza identificando las posibles causas de la problemática, a continuación, se aplican las posibles correcciones y terminan con pruebas de regresión, si en caso de dado que se encuentren otros posibles fallos se debe de aplicar de nuevo la depuración buscando las causas sospechosas y se le aplican pruebas adicionales.

Las pruebas se deben de diseñar de tal manera que tengan la facilidad de encontrar los mayores errores posibles con el mínimo esfuerzo dentro del código de tal manera que las pruebas sean factibles y comprobables. La tabla 4.5 muestra las características que debe poseer un diseño de pruebas de software.

Tabla 4.5

Características de las pruebas de software.

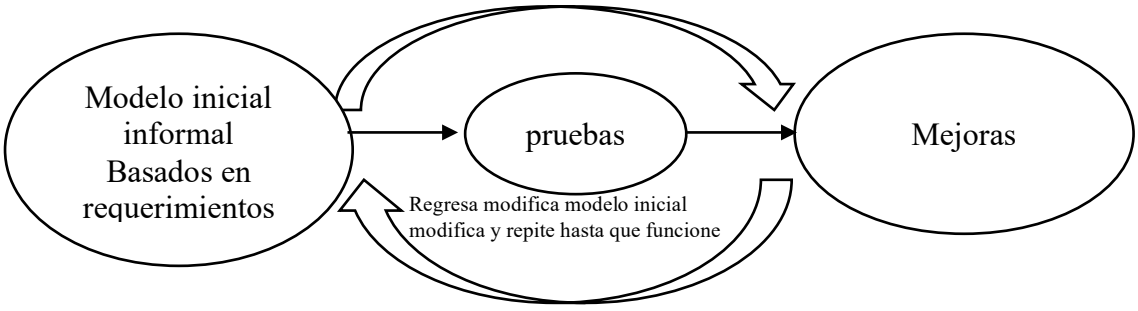
Nombre	Descripción
Comprobabilidad	La facilidad de que un software pueda ser comprobado.
Operatividad	Mientras más eficiente sea más puede probarse.
Observabilidad	Su ejecución permite ver los estados del sistema y las variables son visibles durante su ejecución.
Controlabilidad	Entre más se pueda controlar el software más se podrá automatizar las pruebas.
Descomponibilidad	La capacidad de aislar más rápidamente los problemas para realizar pruebas nuevas y más inteligentes.
Simplicidad	En cuanto menor sea el número de cosas a probar rápidamente se le puede probar.
Estabilidad	Entre menos cambios, menos perturbaciones para probar.
Comprensibilidad	Entre más información contenga se probará con mayor inteligencia.

Fuente: Elaboración propia basada en Pressman (2010, p. 412).

Métricas orientadas a objetos (MOO)

Las pruebas orientadas a objetos son aquellas que están orientadas a encontrar el número máximo de errores con una cantidad mínima de esfuerzo en un tiempo determinado. La Figura 4.3 muestra el diagrama de ciclo de métricas orientadas a objetos comienza por un bosquejo informal de la representación de los requisitos del sistema se realizan pruebas se realizan y las modificaciones y se repite hasta que se obtenga los objetivos esperados.

Figura 4.3 Diagrama ciclo de métricas orientadas a objetos.



Fuente: Elaboración propia basada en Pressman (2010, p. 438).

5. Algoritmo Propuesto

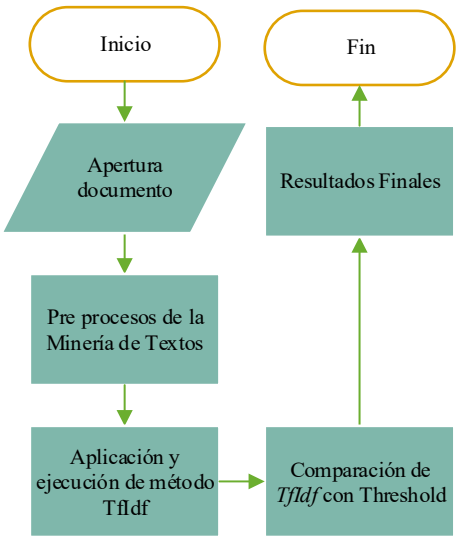
Para poder llevar a cabo este proyecto fue necesario desarrollar e implementar un algoritmo que permitiera desarrollar y aplicar la minería de textos que permitiera el análisis de las entrevistas.

El algoritmo Quillo Espino consta de las siguientes fases:

- Primera fase: abre, guarda y cierra el archivo de texto.
- Segunda fase: realiza preprocesos de la MT (dividir, transformar y estructurar los datos).
- Tercera fase: aplicación y ejecución del método extractivo *Tfldf* (realiza operaciones *Tf*, *Idf* y obtiene resultados)
- Cuarta Fase: Comparación de valores de *Tfldf* con valor *threshold* (se comparan los resultados obtenidos de *Tfldf* con el valor de *threshold* ingresado por el usuario).
- Quinta Fase: Obtención de resultados finales (el algoritmo presenta los resultados finales).

El algoritmo se presenta por medio de una aplicación, La Figura 5.0 muestra el diagrama de flujo del algoritmo Quillo Espino.

Figura 5.0. Diagrama de flujo Algoritmo Quillo Espino.

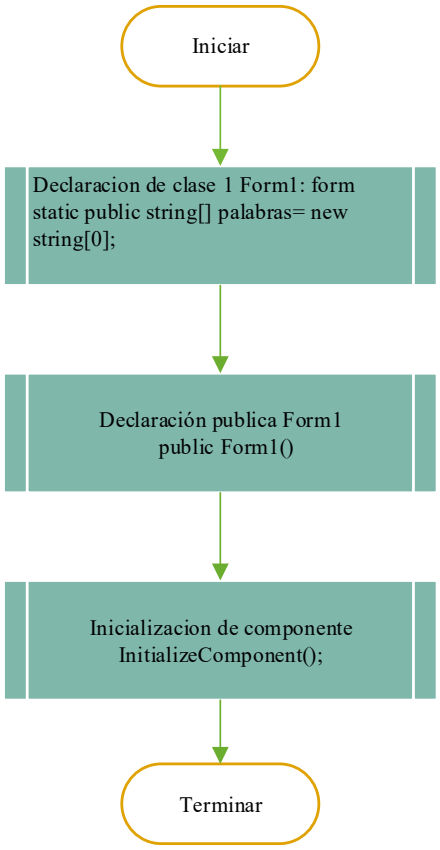


Fuente: Elaboración propia.

5.1. Creación de forma

Para poder manipular la información fue necearía la aplicación de programación orientada a objetos a través de una forma de trabajo llamada *Form1*. La figura 5.1 muestra el diagrama de flujo para función de Form1.

Figura 5.1. Diagrama de flujo para inicializar la forma1.

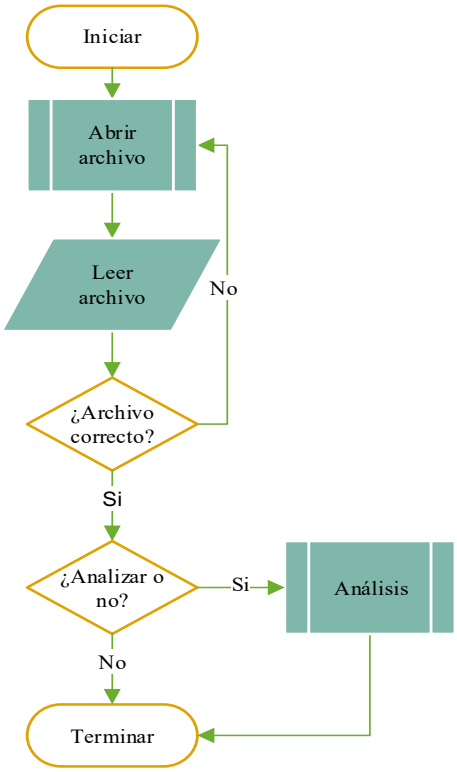


Fuente: Elaboración propia.

5.2. Abrir y Cerrar archivo

El análisis de un texto requiere un archivo de entrada. Es primer paso para el desarrollo del algoritmo. La Figura 5.2 muestra el algoritmo para abrir el archivo.

Figura 5.2. Diagrama de flujo para abrir un archivo.



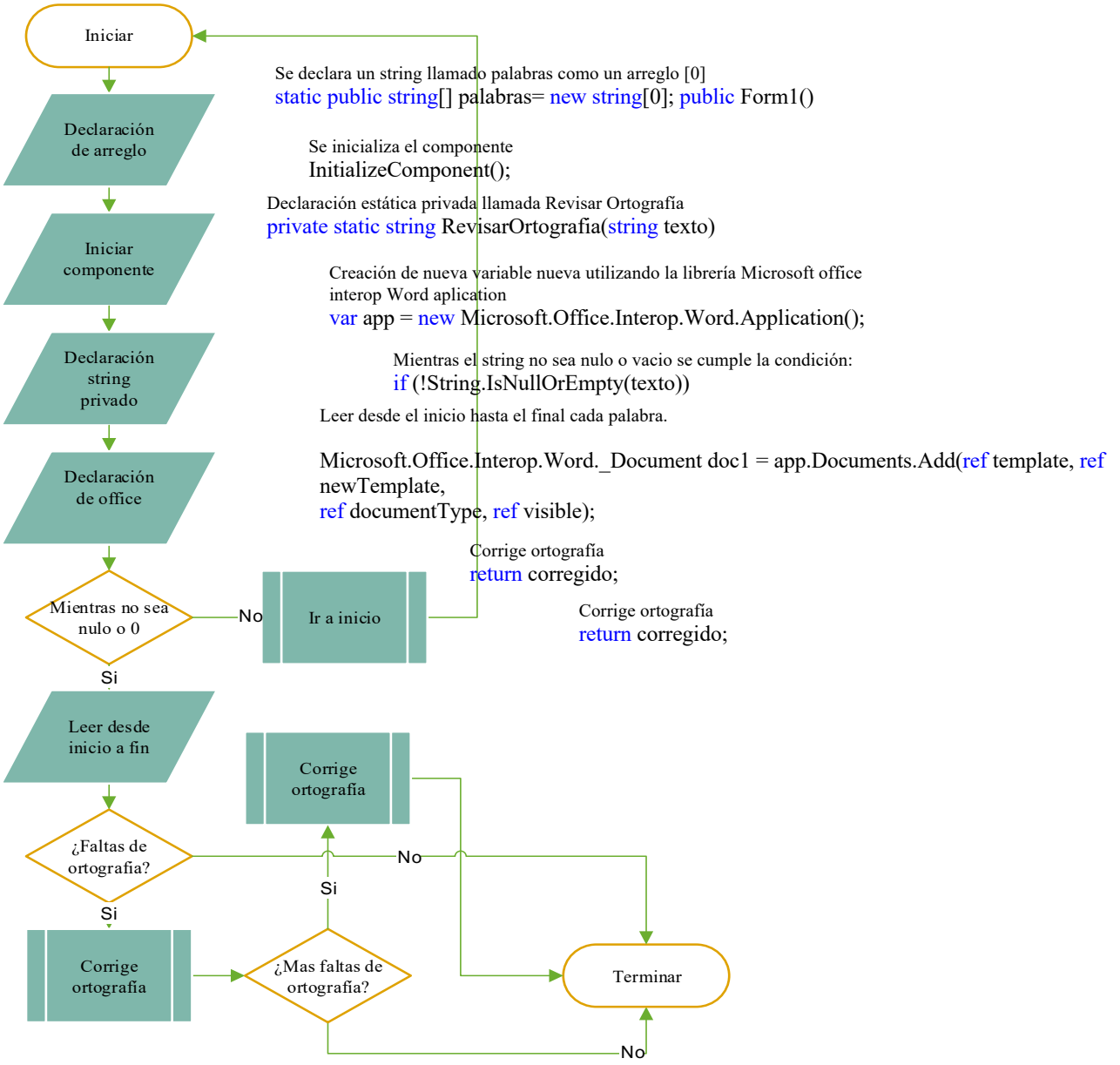
Fuente: Elaboración propia.

El diagrama de flujo de un algoritmo que demuestra el procedimiento para abrir un archivo de texto que será analizado y le da la posibilidad de verificar si el archivo que escogió fue el adecuado si no también le da la oportunidad de corregirlo de nuevo, si fue el correcto se pasa a la siguiente fase que consiste en comenzar a con el procedimiento de la MT.

5.3. Corrección ortográfica

El origen de los documentos textuales con los que se ha realizado esta investigación son resultado de entrevistas que han sido transcritas de manera manual. Los scripts de dichas entrevistas contienen faltas de ortografía. Es necesario corregir la ortografía para mejorar la eficiencia en los pre-procesos de la MT disminuyendo el tiempo para cada proceso. La Figura 5.3 muestra el diagrama de flujo para llevar a cabo la corrección ortografía.

Figura 5.3. Diagrama de flujo para corrección ortográfica.



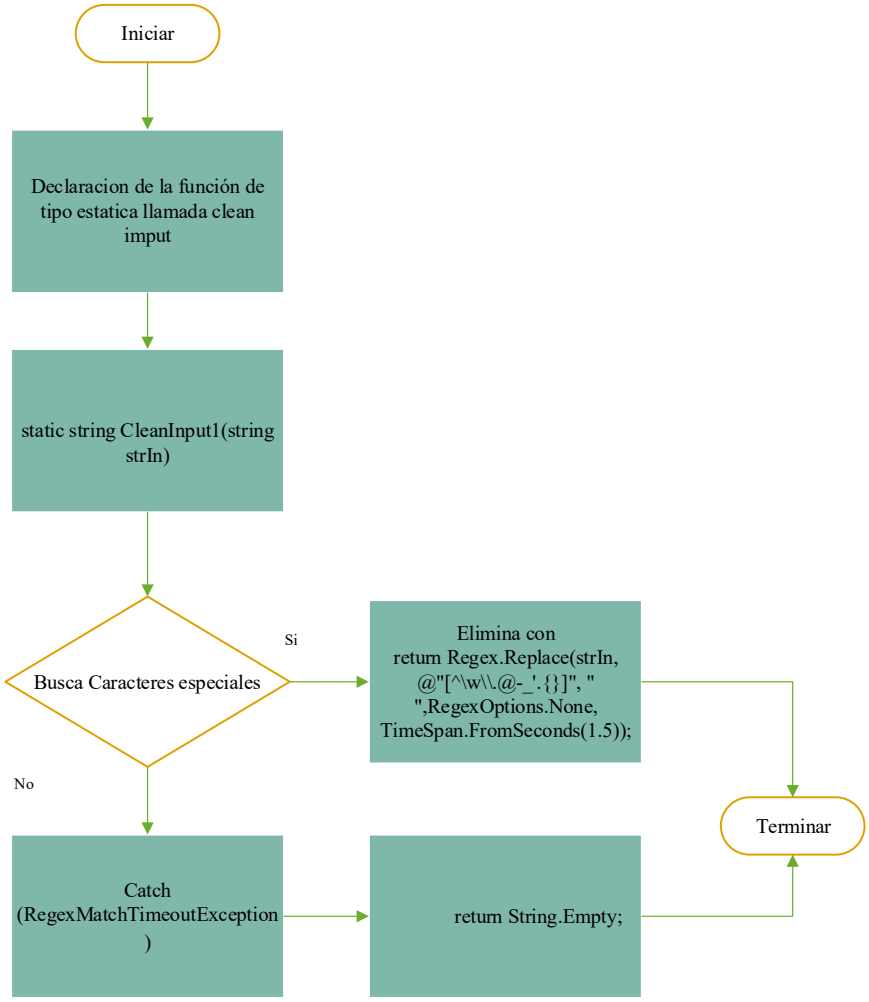
Fuente: Elaboración propia.

La búsqueda de errores ortográficos es necesaria, buscar en cada una de las palabras que se encuentren dentro del contenido textual satisfaciendo las necesidades del usuario.

5.4. Eliminación de caracteres especiales

En la MT existen elementos llamados caracteres especiales que al no ofrecer ningún tipo de valor para el proceso de obtención de conocimiento por lo tanto es necesario eliminarlos, algunos ejemplos son: ()´*+{}-., etc. No se consideran requeridos en la MT. La Figura 5.4 muestra el diagrama de flujo de algoritmo para la eliminación de caracteres especiales.

Figura 5.4. Diagrama de flujo de algoritmo para la eliminación de caracteres especiales.



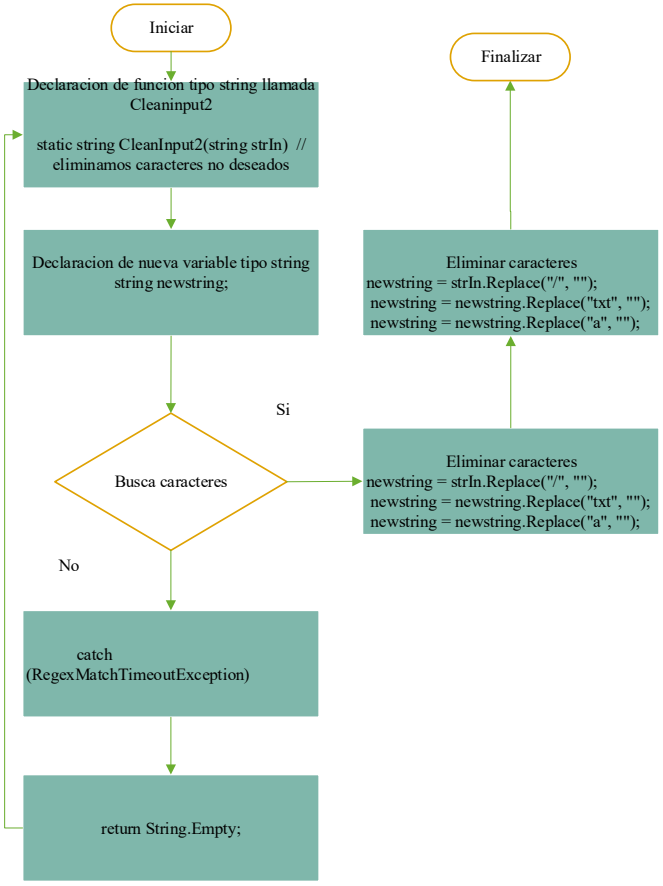
Fuente: Elaboración propia.

5.5. Eliminación de caracteres no deseados

Los caracteres no deseados con aquellos que no proporcionan o representan algún interés para la MT, pueden ser las comas, números, signos de admiración e interrogación, paréntesis,

diagonales etc. En esta parte del proceso el contenido textual ya se encuentra perfectamente corregido ortográficamente, a continuación, se procede a eliminar todos los caracteres no deseados. La Figura 5.5 muestra el diagrama de flujo del algoritmo para la eliminación de caracteres no deseados.

Figura 5.5. Diagrama de flujo eliminación de caracteres especiales.

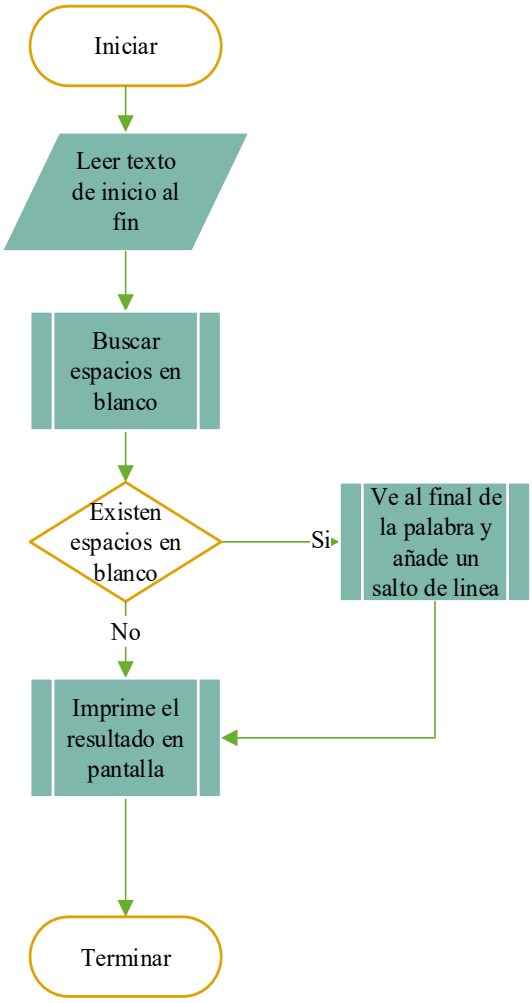


Fuente: Elaboración propia.

5.6. Tokenizador

La tokenización consiste en buscar y separar cada uno de los elementos del conjunto de la colección de datos textuales en un elemento independiente generalmente busca mediante un indicador que puede ser un espacio en blanco. La Figura 5.6 muestra el diagrama de flujo del algoritmo para realizar la tokenización.

Figura 5.6. Diagrama de flujo para tokenización.



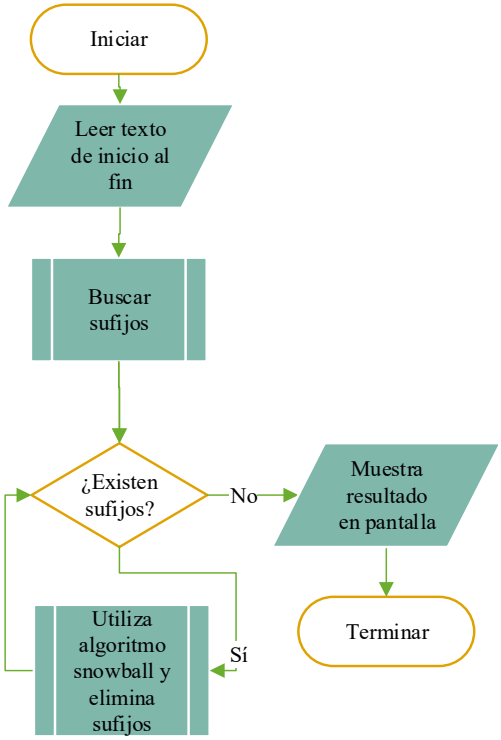
Fuente: Elaboración propia.

Se observa en la Figura 5.6 que el ciclo se repetirá hasta que no encuentre espacios en blanco, a partir de ese momento se ira al fin.

5.7. Stemming

Durante la etapa 1 de la MT, es necesario realizar el proceso de las palabras llamado *stemming* que consiste intentar que cada una de las palabras del conjunto o colección de datos textuales sea convertida o llevada a su raíz. La Figura 5.7 demuestra el diagrama de flujo del algoritmo *stemming.*

Figura 5.7. Diagrama de flujo de algoritmo stemming.

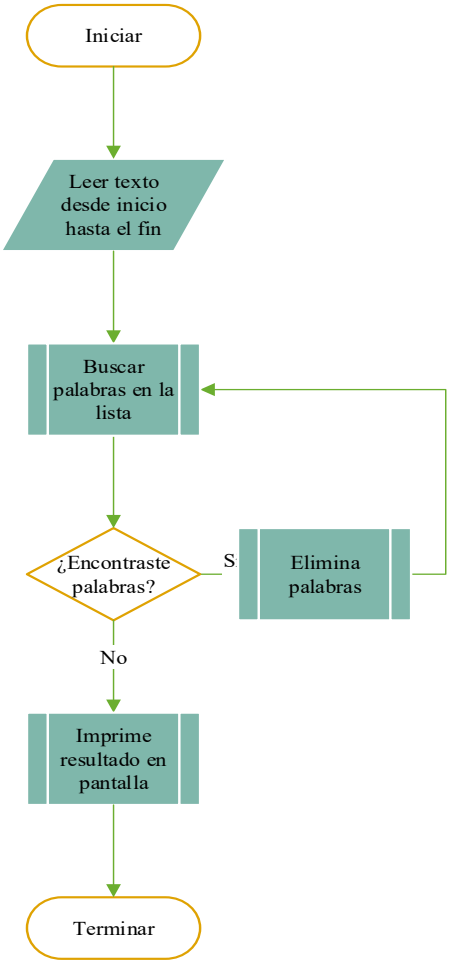


Fuente: Elaboración propia.

5.8. Stop words removal tool

Existe un diccionario de palabras predefinidas desde el inicio de PLN, con la finalidad de promover con el funcionamiento de MT, por lo tanto, se eliminan palabras que no proporcionan utilidad en la MT. La Figura 5.8 demuestra el diagrama de flujo que permite eliminar las palabras establecidas en stop Word removal tool.

Figura 5.8. Diagrama de flujo stop removal tool.

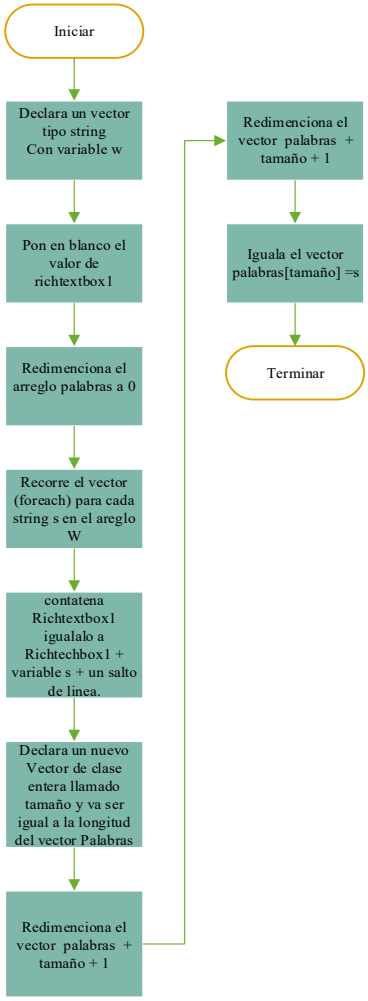


Fuente: Elaboración propia.

5.9. Separar strings

Para conocer el valor del vector de palabras se separan y realiza operación permitiendo convertir el valor que se encuentra en richtextbox1 a un string y editar los valores. La Figura 5.9 demuestra el diagrama de separar string.

Figura 5.9. Diagrama de flujo separar string.

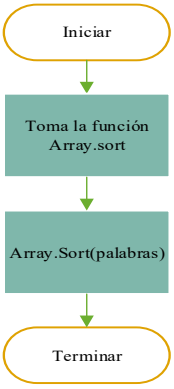


Fuente: Elaboración propia.

5.10. Ordenar string alfabéticamente

Para la representación organizada es necesario ordenarlos alfabéticamente por lo tanto se realiza dicha operación. La Figura 5.10 muestra diagrama de flujo para ordenar alfabéticamente un vector.

Figura 5.10. Diagrama de flujo para ordenar un arreglo alfabéticamente.

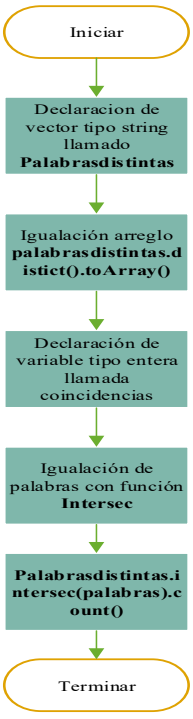


Fuente: Elaboración propia.

5.11. Selección de elementos distintos en el vector palabras

Selecciona elementos que sean diferentes al vector, y los guarda para su posterior comparación. En la Figura 5.11 se muestra el diagrama de flujo para poder realizarlo.

Figura 5.11. Diagrama de flujo Selección de elementos distintos en vector.

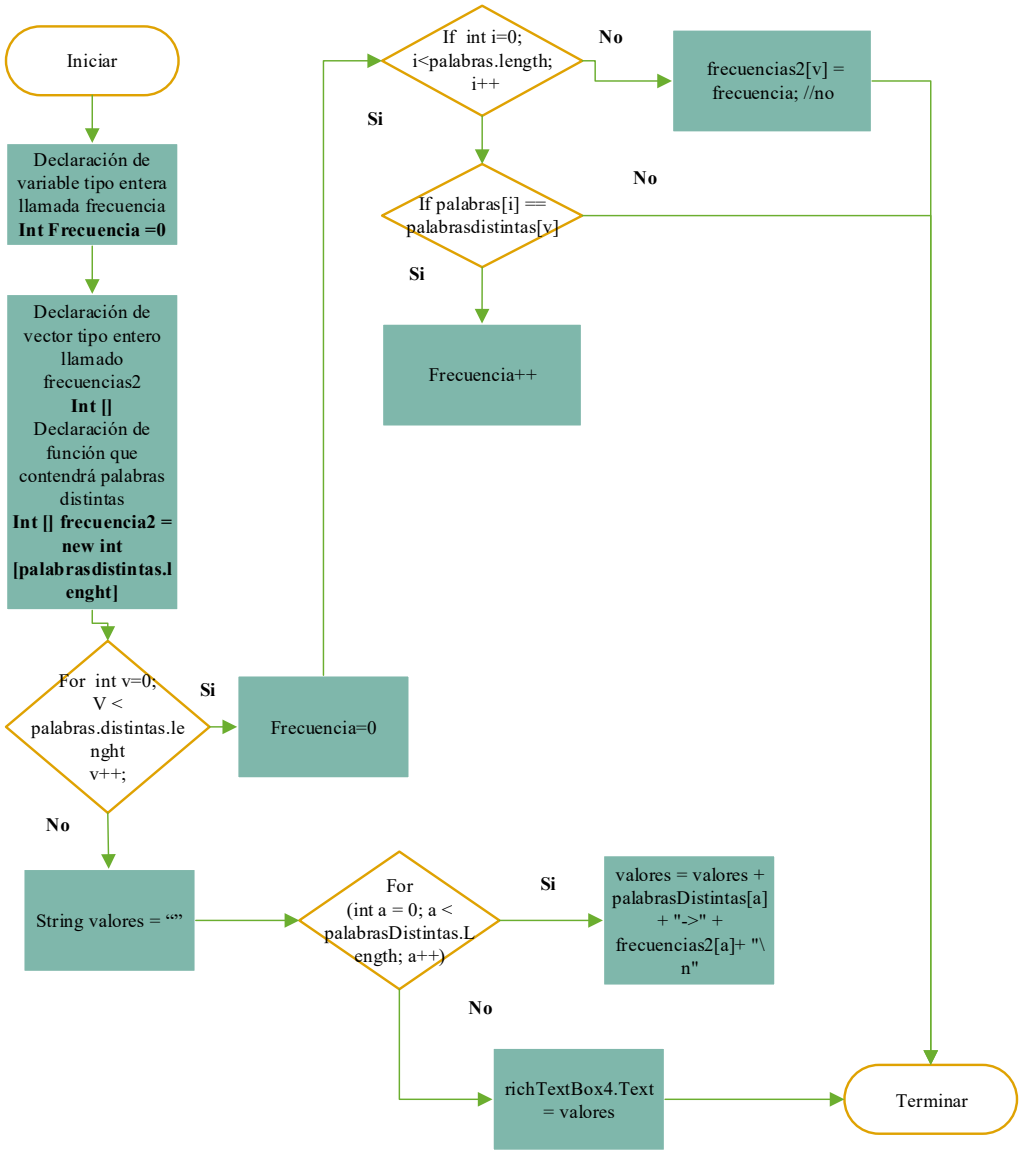


Fuente: Elaboración propia.

5.12. Coincidencias

Se almacena una vector que contiene la cantidad total de coincidencias entre el vector palabras y palabras distintas, es un procedimiento para obtener la frecuencia de las palabras. La Figura 5.12 el diagrama de flujo para obtener coincidencias.

Figura 5.12. Diagrama de flujo de coincidencias.

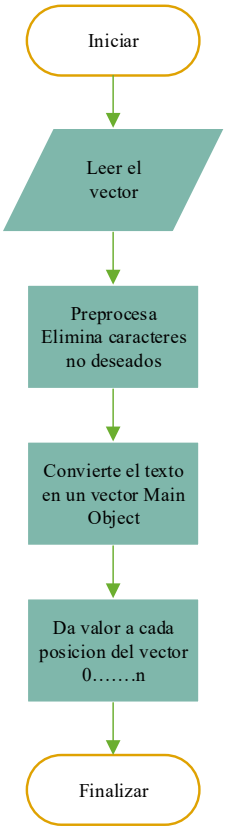


Fuente: Elaboración propia.

5.13. Posición del vector

Para realizar la categorización y búsqueda de palabras se eliminan los caracteres no deseados se convierte en main object y asignar una posición dentro del vector. La Figura 5.13 muestra el proceso para asignar posición en el vector.

Figura 5.13. Diagrama de flujo posición del vector.

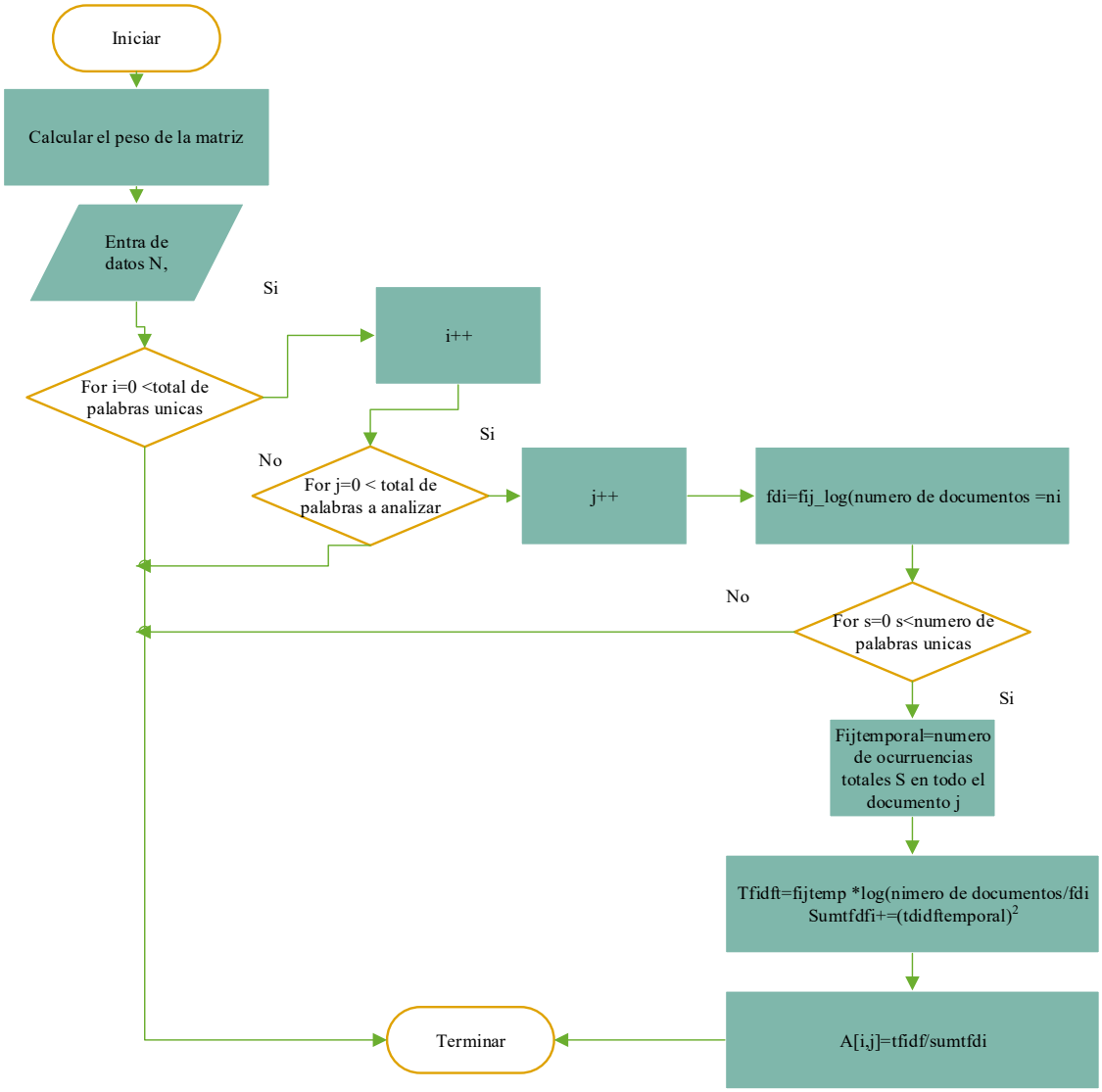


Fuente: Elaboración propia.

5.14. Cálculo de matriz de palabras

El cálculo del peso de la matriz se necesita ya que esta contiene información única de cada una de las palabras en los documentos. Cada uno de los documentos representa un vector en de n dimensiones en el espacio vectorial. La Figura 5.14 muestra el diagrama de flujo para el cálculo de peso de la matriz.

Figura 5.14. Diagrama de flujo de cálculo de matriz de palabras.

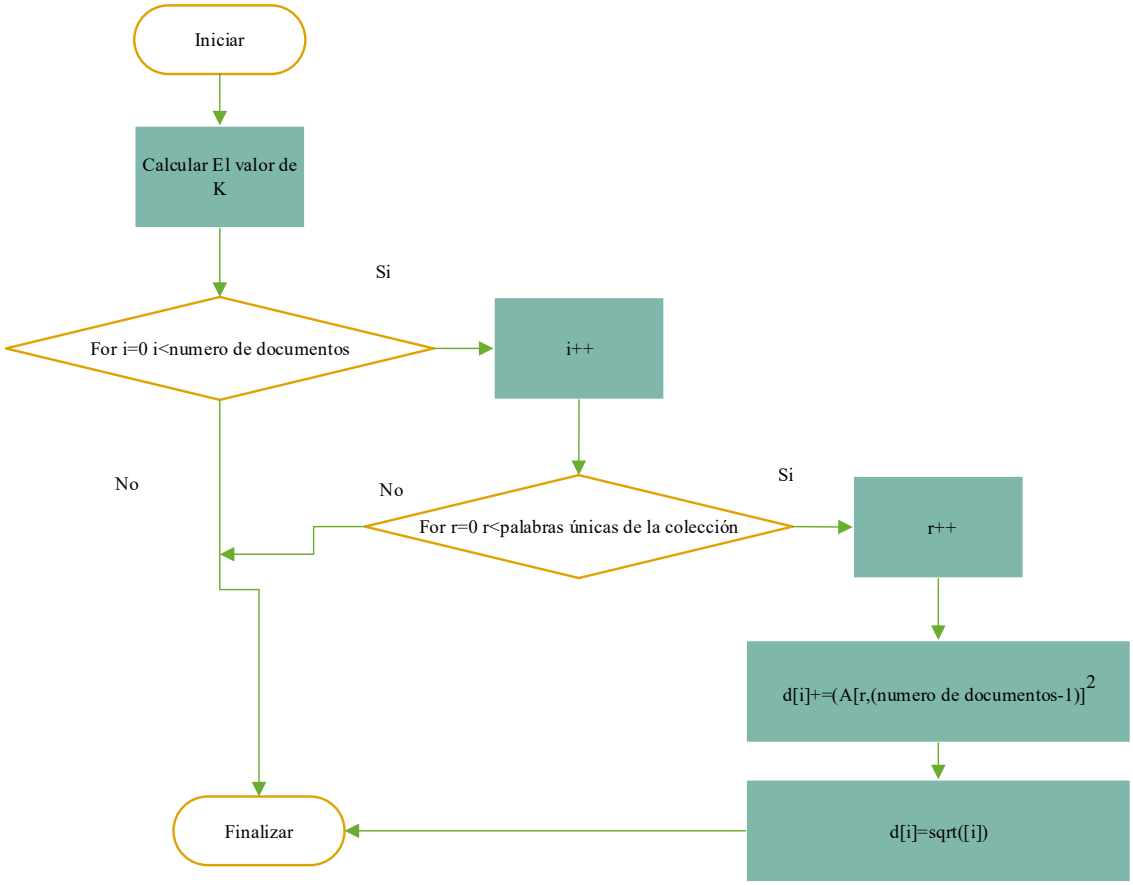


Fuente: elaboración propia.

5.15. Cálculo de K

El cálculo de K ya que es el factor que representa el número requerido de documentos de la colección que está cerca al documento seleccionado. La Figura 5.15. muestra el proceso para ejecutar el cálculo de K.

Figura 5.15 Diagrama de flujo para calcular K.

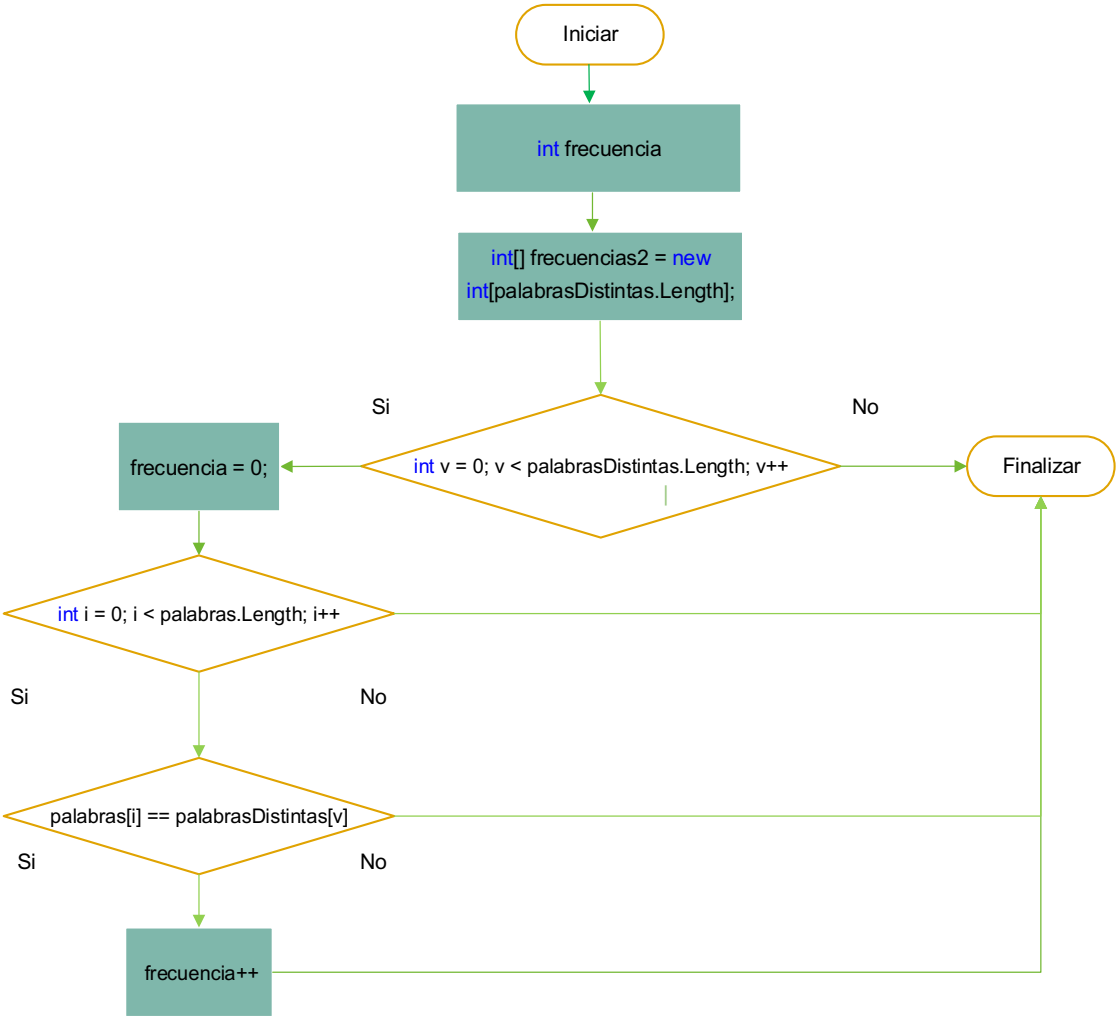


Fuente: Elaboración propia.

5.16. Cálculo de frecuencias generales totales

Se solicita conocer las frecuencias generales de todo el documento por tal motivo se requiere calcular peso de cada una de las palabras existentes en el documento. La Figura 5.16 demuestra el diagrama de flujo de cálculo de frecuencias generales.

Figura 5.16. Diagrama de flujo para cálculo de frecuencias generales.

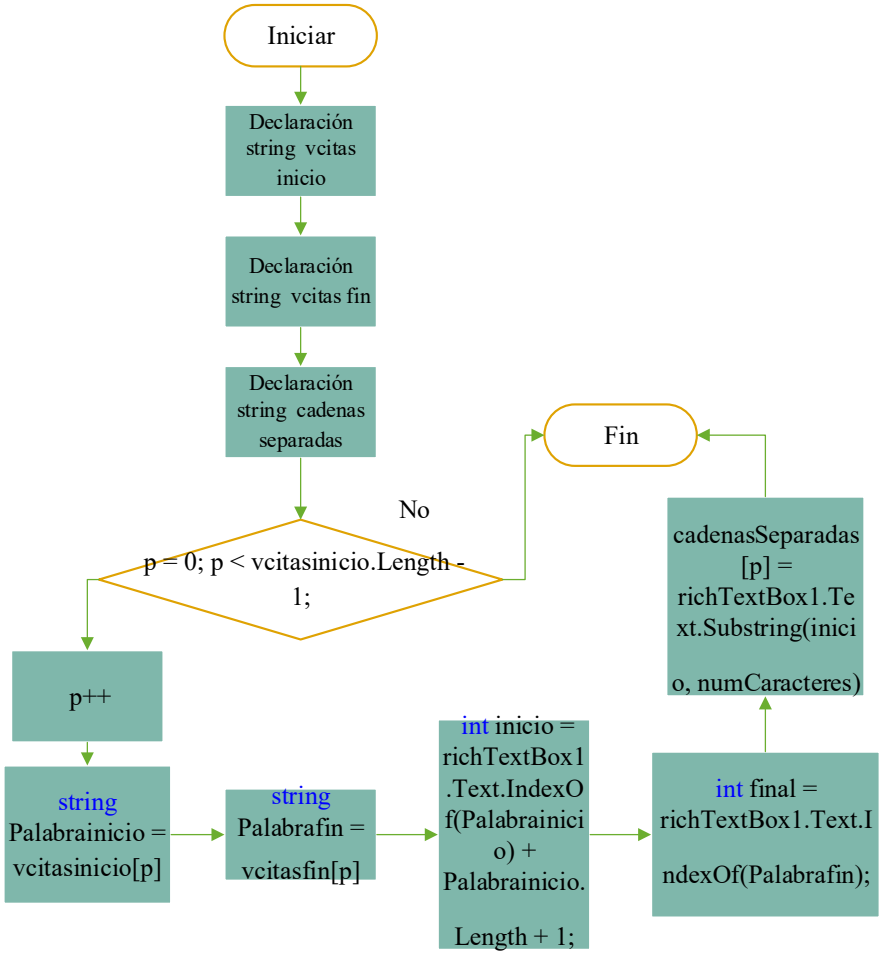


Fuente: Elaboración propia.

5.17. Separador de citas textuales

Para poder realizar el conteo de la frecuencia se separa las citas textuales en elementos independientes. Como resultado arroja un conjunto de vectores independientes los cuales se recorren para generar diferentes cálculos. La Figura 5.17 muestra el diagrama de flujo de separación de citas textuales.

Figura 5.17. Diagrama de flujo separador de citas textuales.

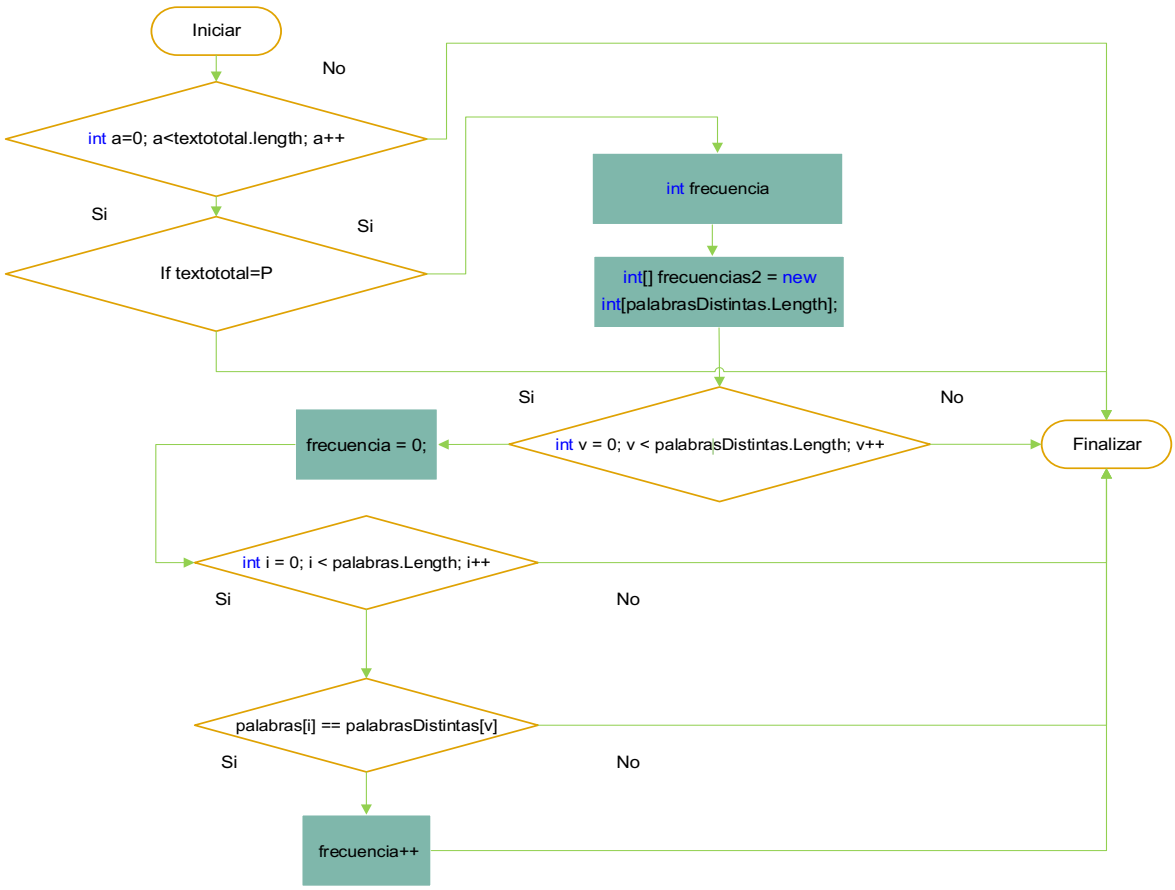


Fuente: Elaboración propia.

5.18. Cálculo de frecuencia de palabras por cita

Es imprescindible conocer las frecuencias correspondientes por cita textual para obtener el peso de cada uno de las palabras contenidas en la cita. La Figura 5.18 demuestra el diagrama de flujo para el cálculo de frecuencias independientes por cita textual.

Figura 5.18. Diagrama de flujo de frecuencias independientes.

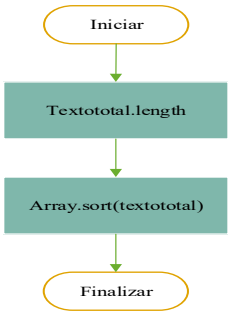


Fuente: Elaboración propia.

5.19. Ordenamiento de frecuencias

Las frecuencias se ordenarlas de manera descendente. La Figura 5.19 muestra el diagrama de flujo para ordenarla de manera descendente.

Figura 5.19. Diagrama de flujo de ordenación de frecuencias.

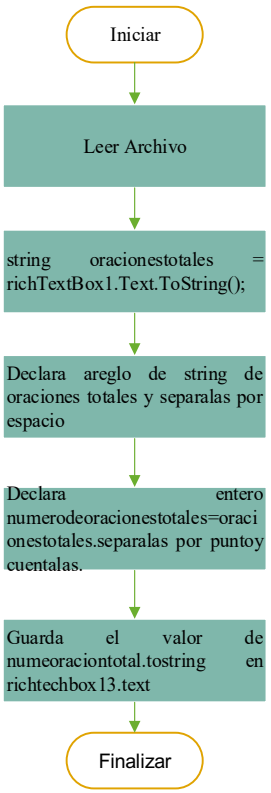


Fuente: Elaboración propia.

5.20. Conversión de conjunto de palabras a String array

Es inexcusable convertir el contenido de richtextbox1.text en un arreglo de string. La Figura 5.20 muestra el Diagrama de flujo del proceso de conversión de string array.

Figura 5.20. Diagrama de flujo para convertir conjunto de palabras en string array.



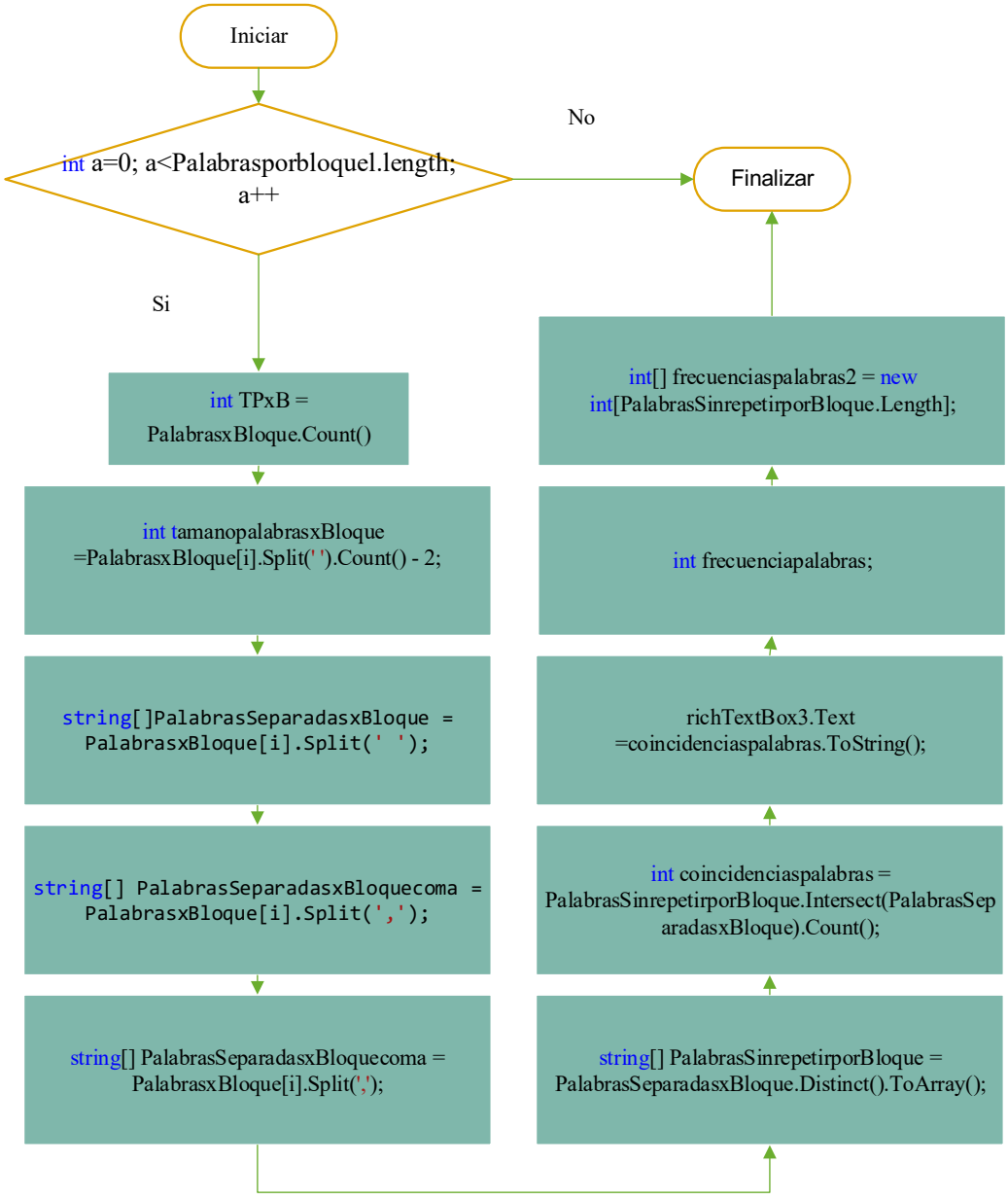
Fuente: Elaboración propia.

5.21. Declaración de ciclo y obtener datos de oraciones totales

Para poder acceder a los datos de palabras por bloque es necesario declarar y realizar un conjunto de operaciones para su ejecución, se declara un arreglo de palabras separadas por espacio vacío en el índice i. Se declara un arreglo llamado palabras separadas por coma y se toman los datos de palabras por bloque índice i separadas por una coma. Se declara un arreglo de palabras sin repetir por bloque y se toman los datos de palabras separadas por bloque se

utiliza el *método distintc* y se guarda como tipo *array.* La Figura 5.21 demuestra el diagrama de flujo para llevar a cabo la declaración del ciclo y obtención de datos.

Figura 5.21. Diagrama de flujo de declaración de ciclo y obtención de datos.

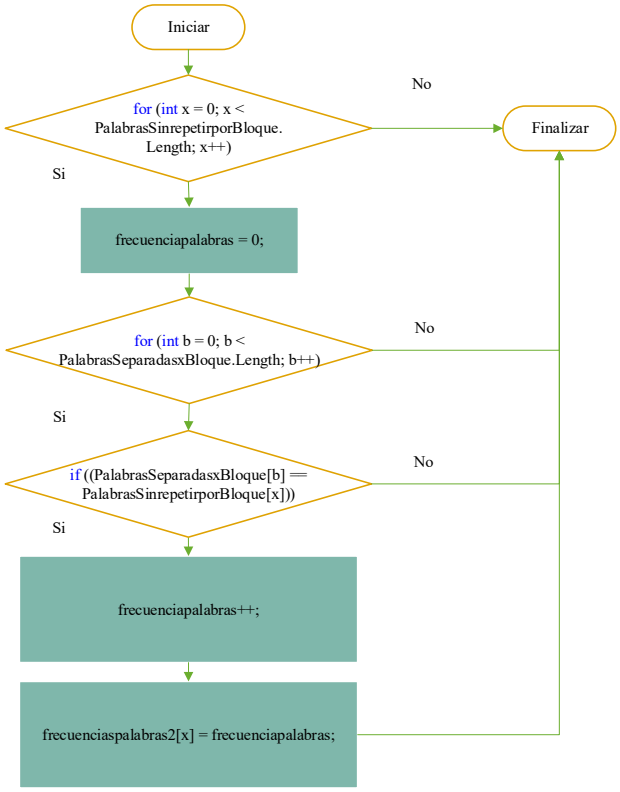


Fuente: Elaboración propia.

5.22. Declaración de ciclo y obtener frecuencias separadas por bloque

Para la obtención de palabras repetidas del arreglo se recorre el arreglo de palabras *sin repetir* y después se busca dentro del arreglo de *palabras separadas por bloque* y comparar cuando palabras *separadas por bloque índice b* sea igual a *palabras sin repetir por bloque índice x* entonces se aumentará la frecuencia de tal manera que cuando se recorra todo el arreglo palabras sin repetir por bloque se obtendrán todos los datos correspondientes a las frecuencias de cada una de las palabras. La Figura 5.22 muestra el desarrollo del ciclo para obtener las frecuencias.

Figura 5.22. Diagrama de flujo para cálculo de frecuencias

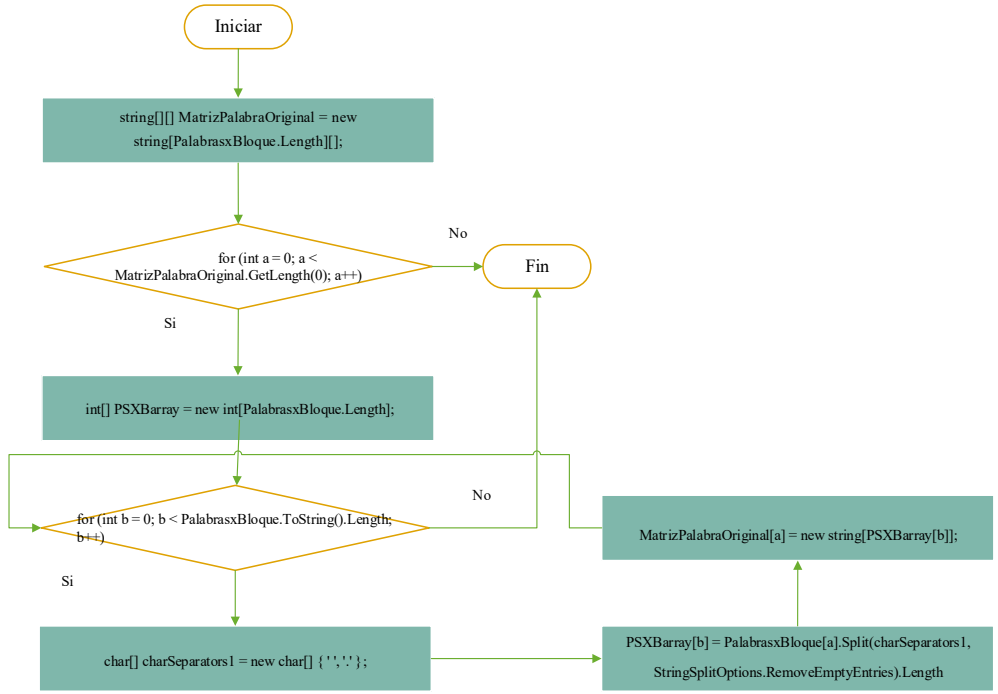


Fuente: Elaboración propia.

5.23. Declaración e inicialización de Matriz de palabras original

El almacenamiento de los datos textuales es necesario por tal motivo se declara una matriz llamada matriz original. La Figura 5.23 muestra el diagrama de flujo para su creación.

Figura 5.23. Diagrama de flujo declaración de matriz.

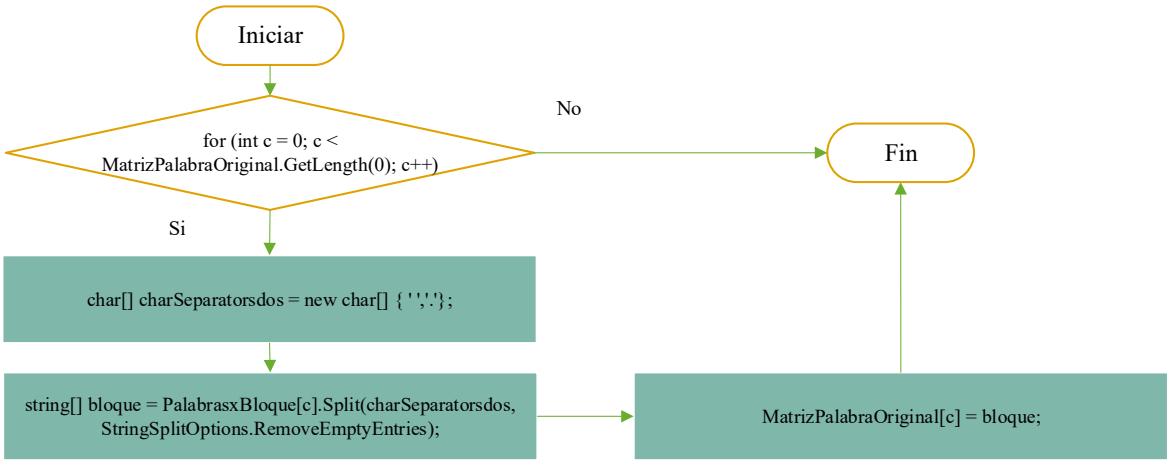


Fuente: Elaboración propia.

5.24. Inclusión de datos a Matriz de palabras original

Creada la matriz se procede a insertar los datos. La Figura 5.24 muestra el diagrama de flujo para insertar los datos.

Figura 5.24. Diagrama de flujo inclusión de datos en Matriz

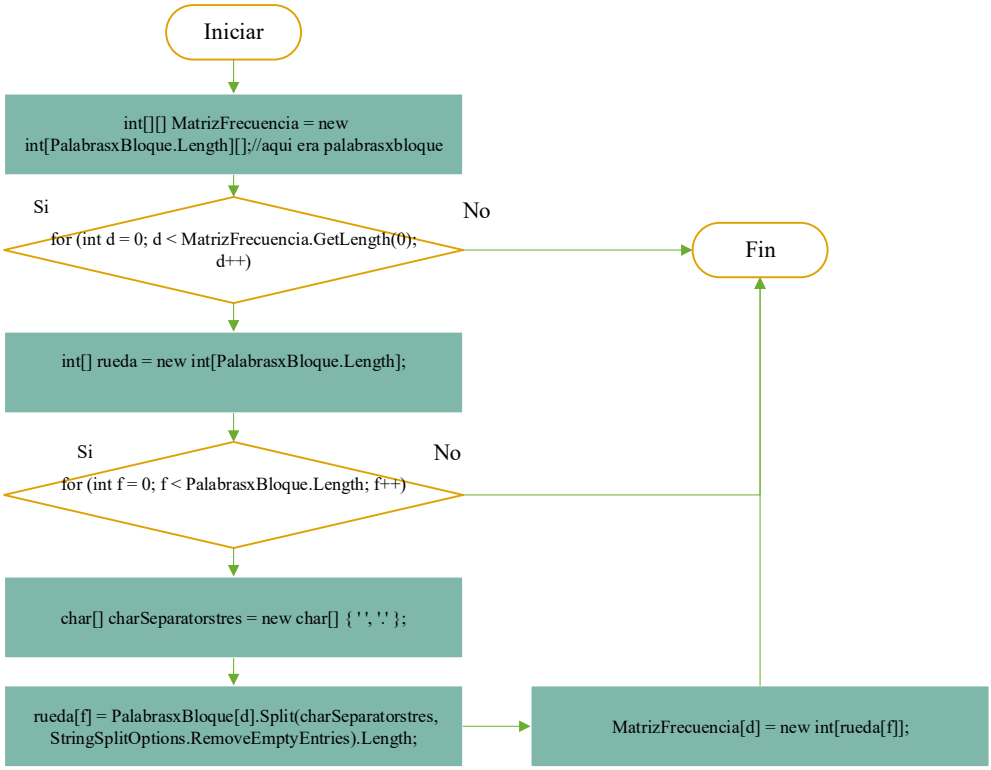


Fuente: Elaboración propia.

5.25. Declaración de Matriz de Frecuencia de palabras

Se declara e inicia una matriz que contendrá los valores de las frecuencias separadas por bloque. La Figura 5.25 muestra el diagrama de flujo declaración de matriz frecuencia de palabras.

Figura 5.25. Diagrama de flujo declaración de inicialización de Matriz de frecuencias.

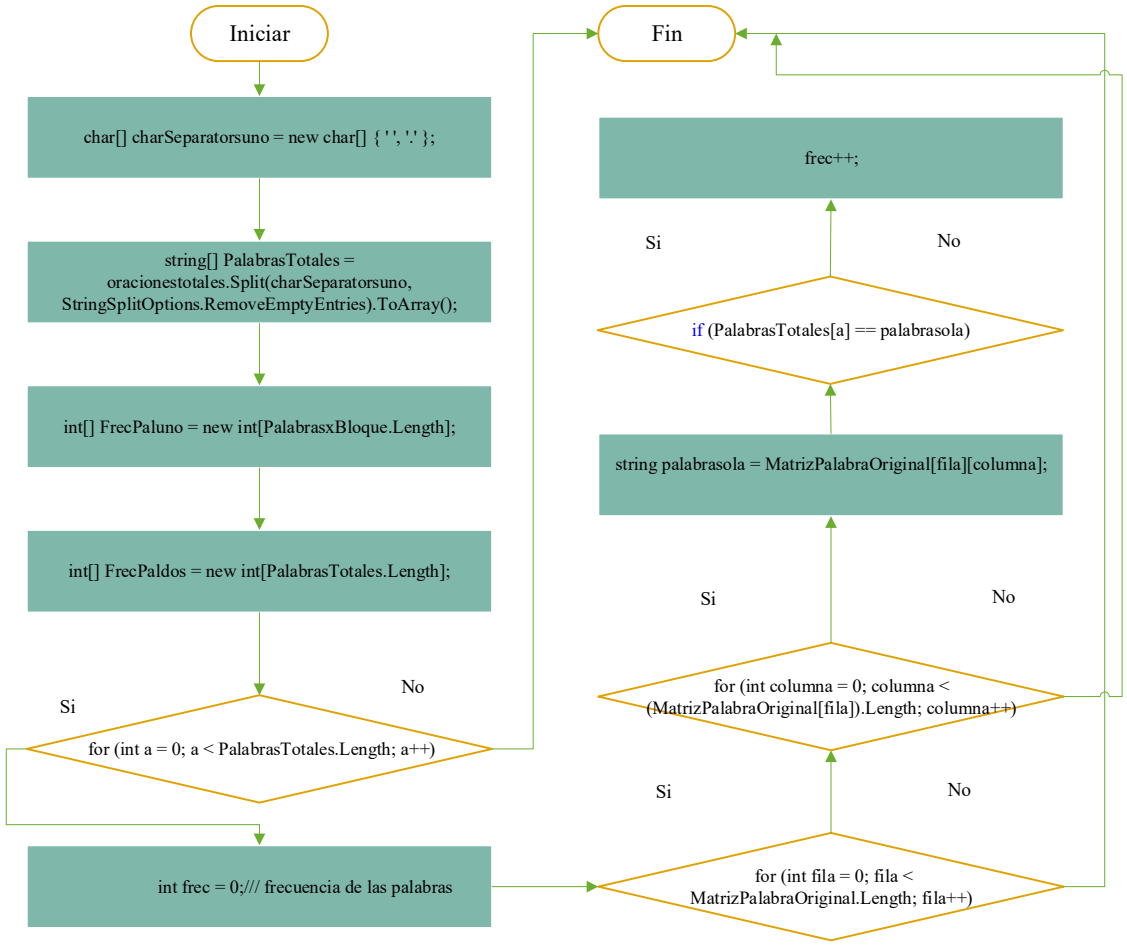


Fuente: Elaboración propia.

5.26. Declaración vector contenedor de palabras totales

Para poder declarar un vector que contendrá las frecuencias se requiere realizar una conversión del total de las palabras, y separarlas por limitador que es un espacio vacío. La Figura 5.26 muestra el diagrama de flujo para crear el vector contenedor de palabras totales.

Figura 5.26. Diagrama de flujo declaración de inicialización de vector contenedor de palabras totales.

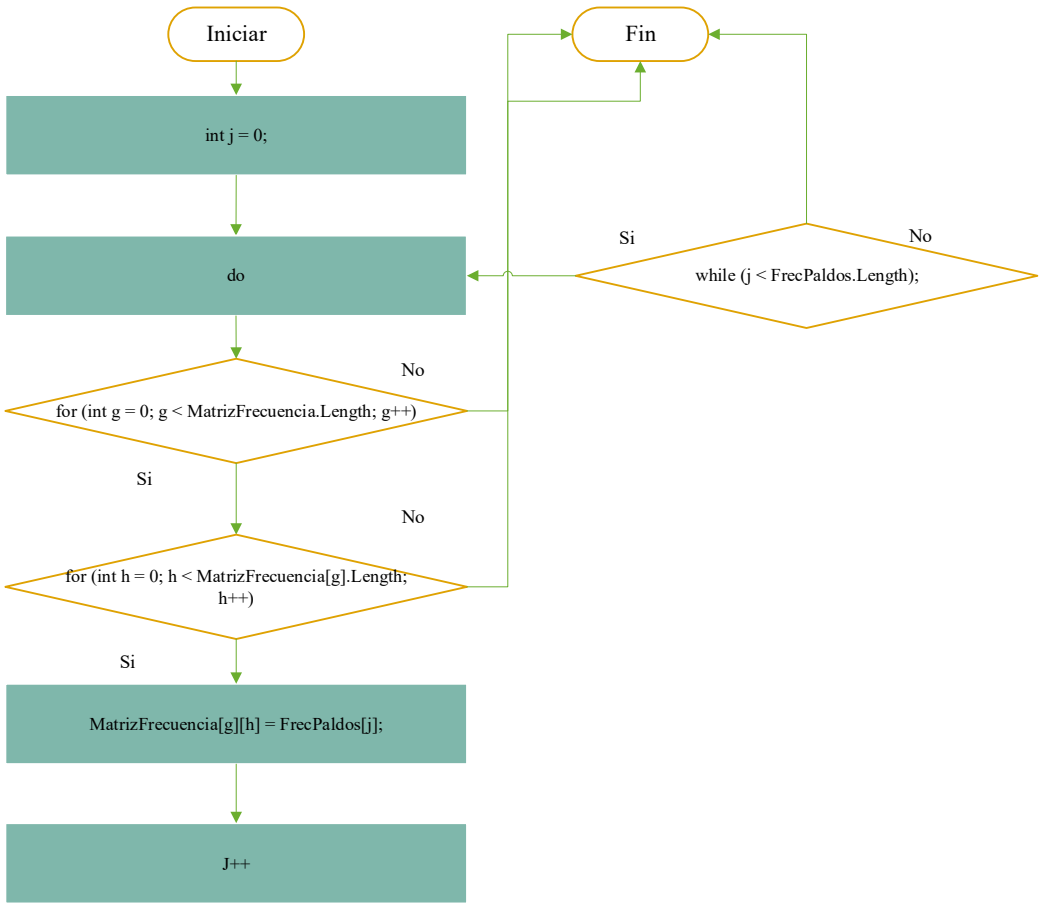


Fuente: Elaboración propia.

5.27. Inclusión de datos a Matriz de frecuencias de palabras

Para mantener la preservación de los datos se requiere introducirlos en la matriz de frecuencia de palabras. La Figura 5.27 muestra el diagrama de flujo para la inclusión de los datos en la matriz.

Figura 5.27. Diagrama de flujo inclusión de datos en la matriz de frecuencia de palabras.

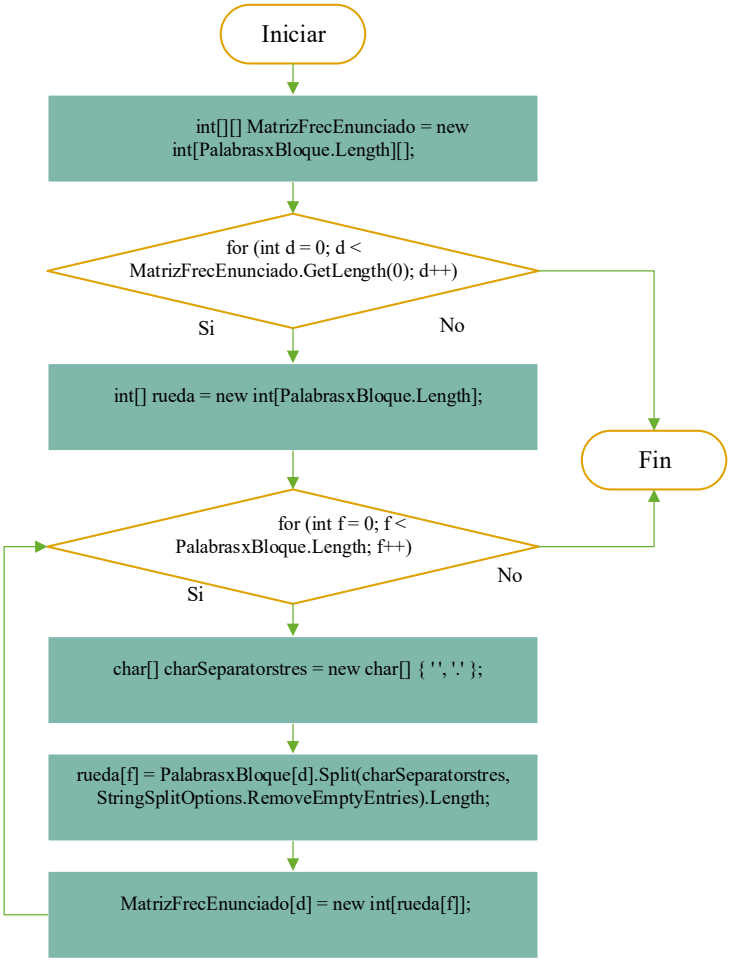


Fuente: Elaboración propia.

5.28. Creación e iniciación de Matriz de frecuencias por oración

Se requiere llevar un conteo para conocer el número de veces aparece una palabra por oración por tal motivo es necesario crear la matriz para contenerlas. La Figura 5.28 muestra el diagrama de flujo de creación de la matriz de frecuencia de palabra por enunciado.

Figura 5.28. Diagrama de flujo de creación de la matriz de frecuencia de palabra por enunciado.

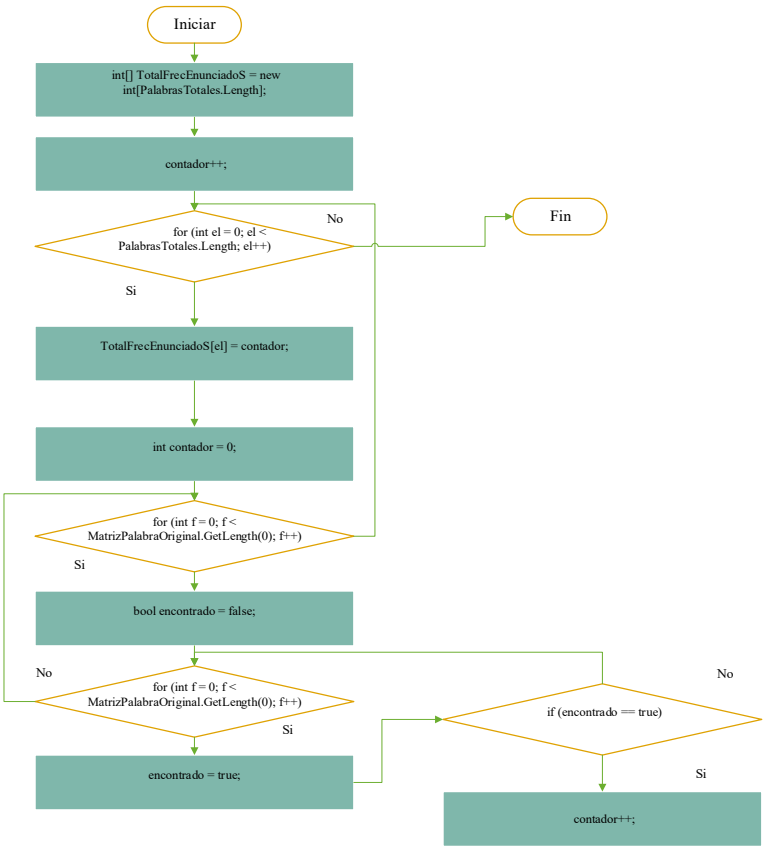


Fuente: Elaboración propia.

5.29. Ciclo para la obtener frecuencias de oración por palabra

La necesidad de obtener las frecuencias de las veces que las palabras se repiten por oración obliga a que se recorra los bloques de palabras. La Figura 5.29 muestra el diagrama de flujo del ciclo para obtener las frecuencias de oración por palabra se guarda todo el un vector para su posterior inserción en la matriz.

Figura 5.29. Diagrama de flujo ciclo de frecuencias de oración por palabra.

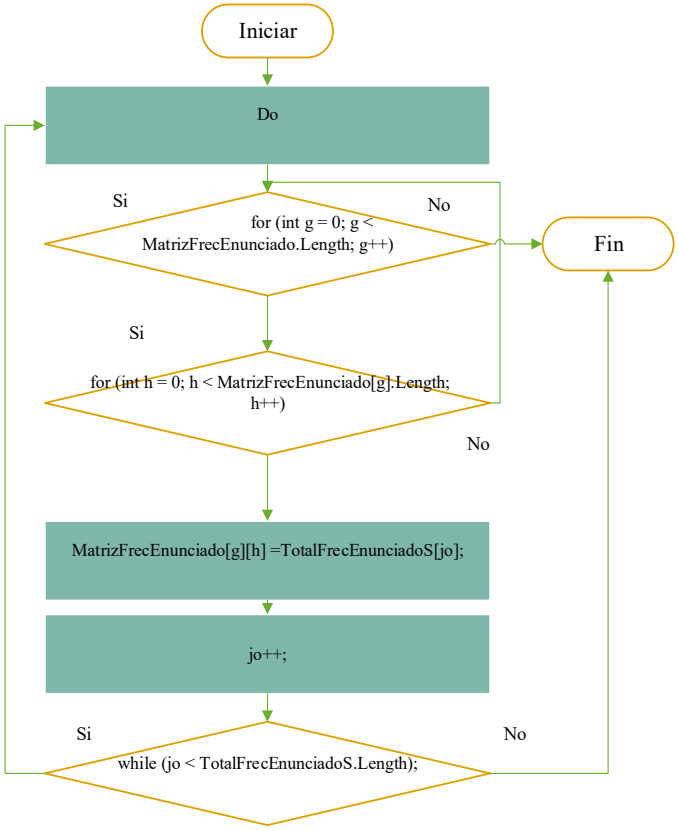


Fuente: Elaboración propia.

5.30. Inserción de elementos en Matriz de frecuencias por oración

Insertar los elementos obtenidos del ciclo anterior es requerido. La Figura 5.30 muestra el diagrama de flujo para la inserción de los datos en la matriz frecuencias de oración por palabra.

Figura 5.30. Diagrama de flujo inserción de valores en matriz frecuencias de oración por palabra.

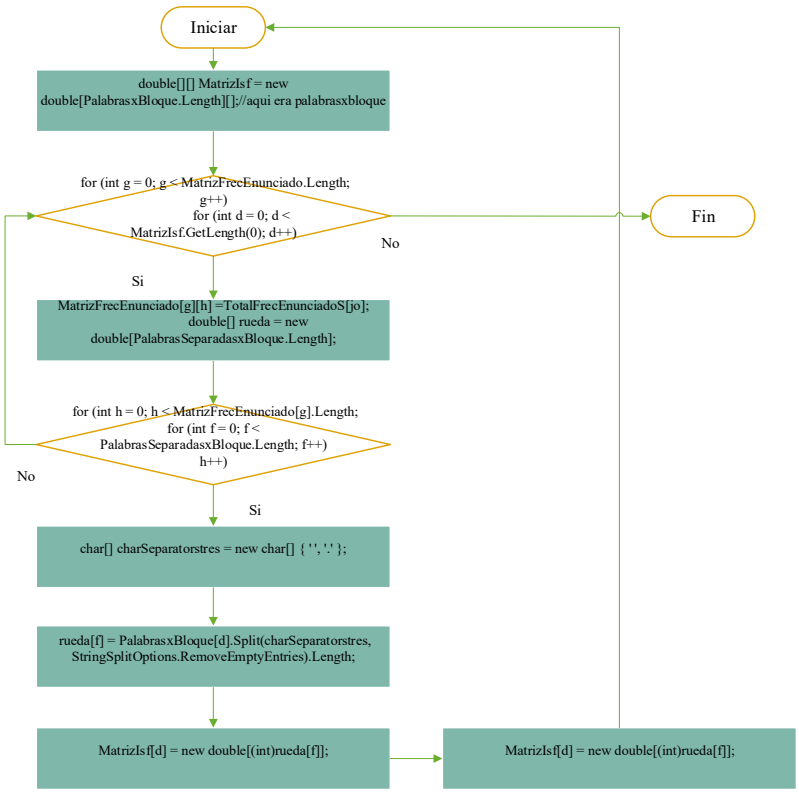


Fuente: Elaboración propia.

5.31. Declaración de Matriz de ISF

La frecuencia inversa del enunciado requiere guardarse en una matriz por tal motivo se declara la matriz con nombre *Isf*. La Figura 5.31 muestra el diagrama de flujo para declarar la matriz e inicializarla.

Figura 5.31. Diagrama de flujo declaración de matriz *Isf.*

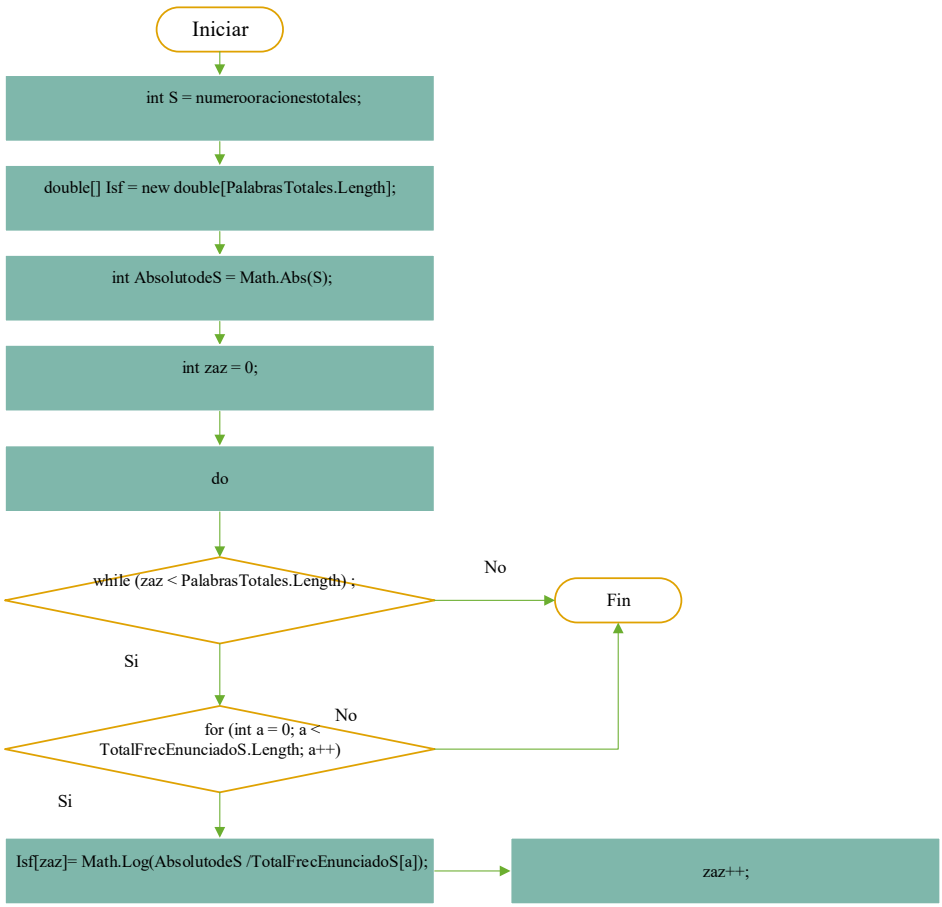


Fuente: Elaboración propia.

5.32. Ciclo para la obtención de valores a Matriz de ISF

Obtener los valores a la matriz *Isf* se realiza mediante un ciclo. La Figura 5.32 muestra el diagrama de flujo del ciclo de obtención de valores de matriz *Isf*.

Figura 5.32. Diagrama de flujo del ciclo para la obtención de valores de Matriz *Isf.*

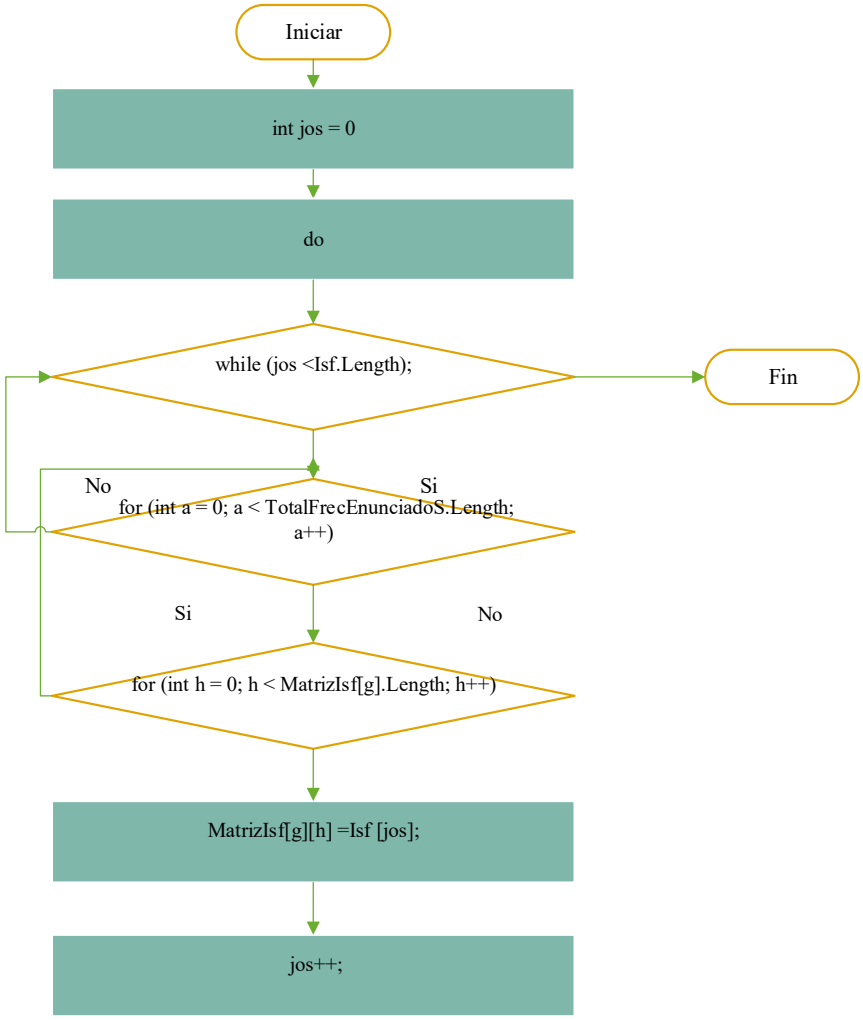


Fuente: Elaboración propia.

5.33. Ciclo para la obtención de valores a Matriz de *Isf*

Para la conservación de los valores obtenidos de la *Isf* se pretende guardar en la matriz. La Figura 5.33 muestra el diagrama de flujo para insertar los valores en la matriz.

Figura 5.33. Diagrama de flujo del ciclo para la inserción de valores de Matriz *Isf.*

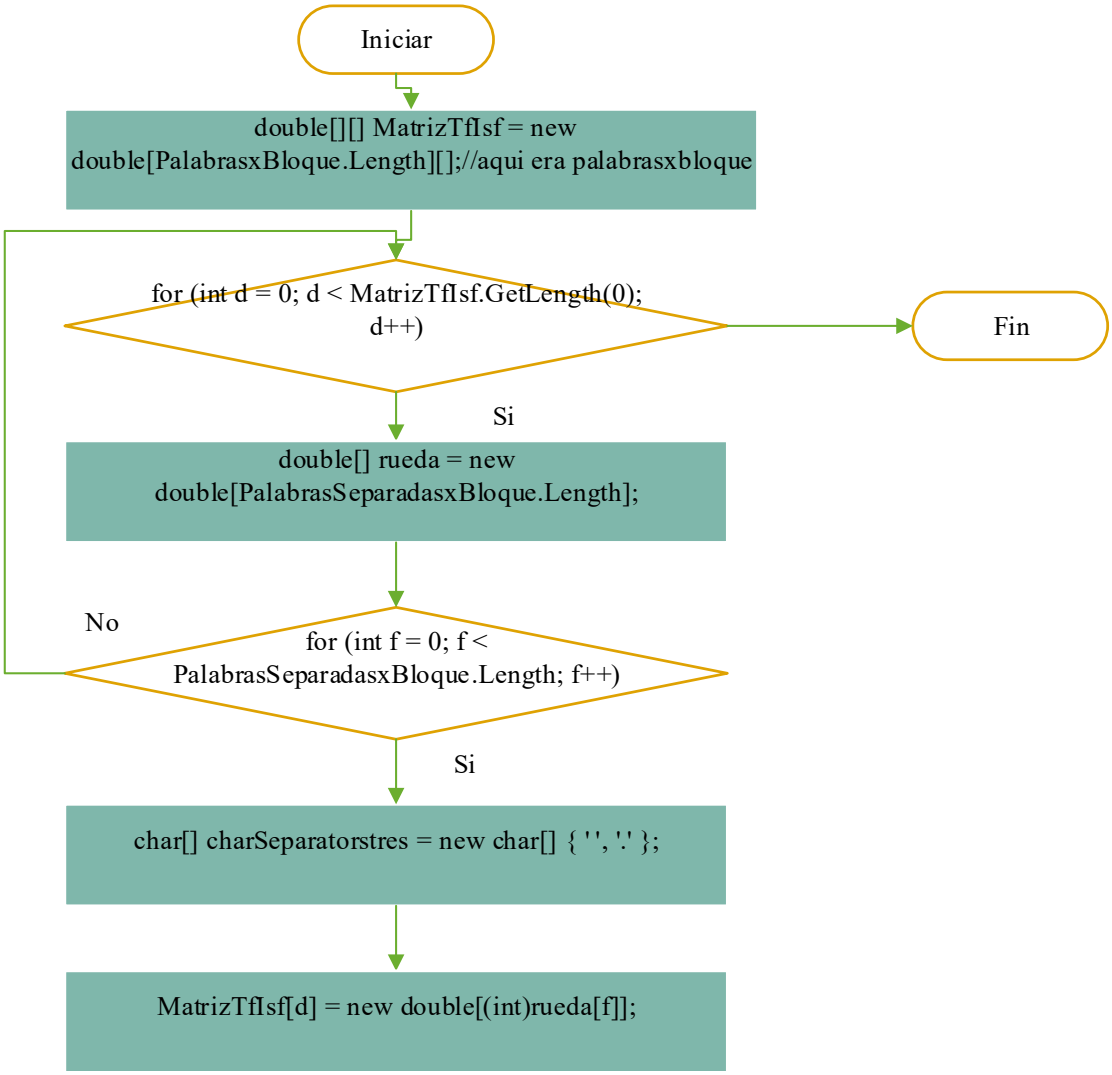


Fuente: Elaboración propia.

5.34. Declaración e iniciación de matriz *TfIsf*

Es determinante la declaración de una matriz que permita guardar los valores obtenidos de la operación *TfIsf*. La Figura 5.34 muestra el diagrama de flujo para la declaración de la matriz *TfIsf*.

Figura 5.34. Diagrama de flujo para la declaración de la Matriz *TfIsf*.

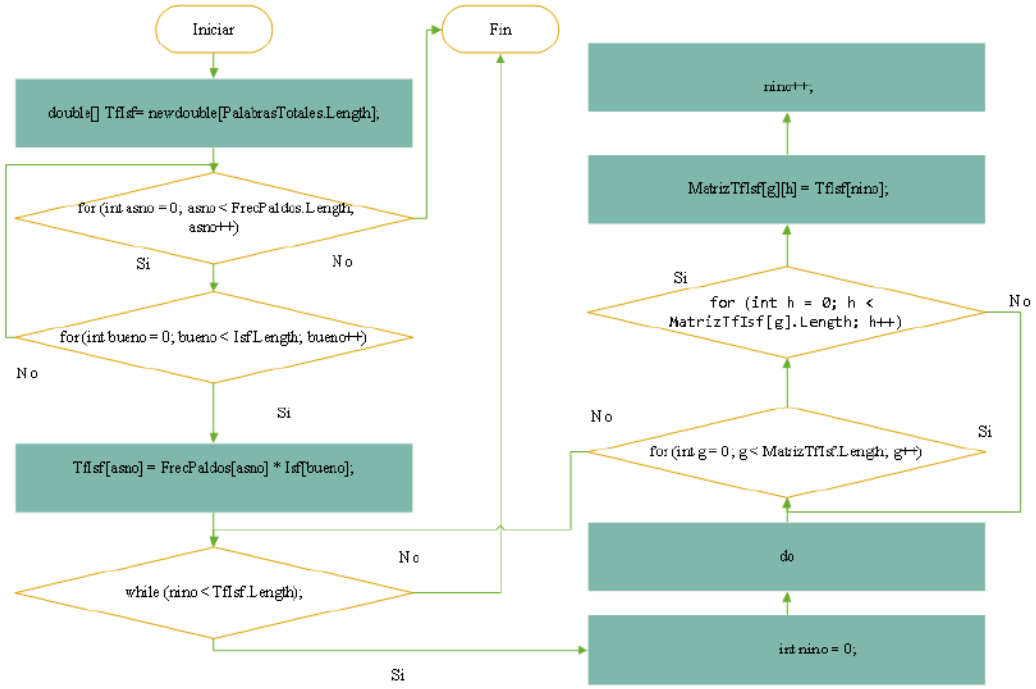


Fuente: Elaboración propia.

5.35. Inserción de datos a matriz TfIsf

Los datos necesitan ser depositados en la matriz para poder guardarlos y además para poder realizar operaciones de búsqueda. La Figura 5.35 muestra el diagrama de flujo de inserción de datos en la matriz *TfIsf*.

Figura 5.35. Diagrama de flujo para la inserción de datos en la Matriz *TfIsf*.

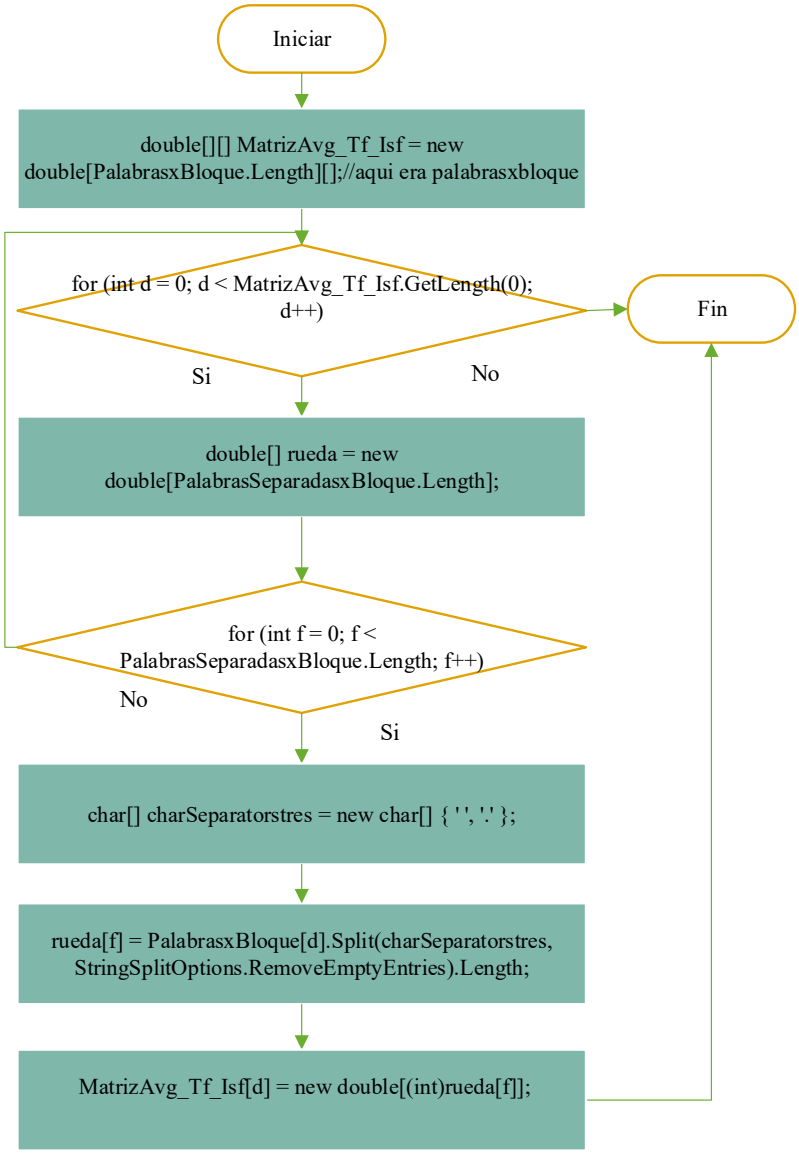


Fuente: Elaboración propia.

5.36. Declaración de matriz promedioTfIsf

La creación de una matriz para el almacenamiento es preciso. La Figura 5.36 muestra el diagrama de flujo para llevar la declaración de la matriz e inicialización.

Figura 5.36. Diagrama de flujo para declaración e iniciación de datos en la Matriz promedioTfIsf.

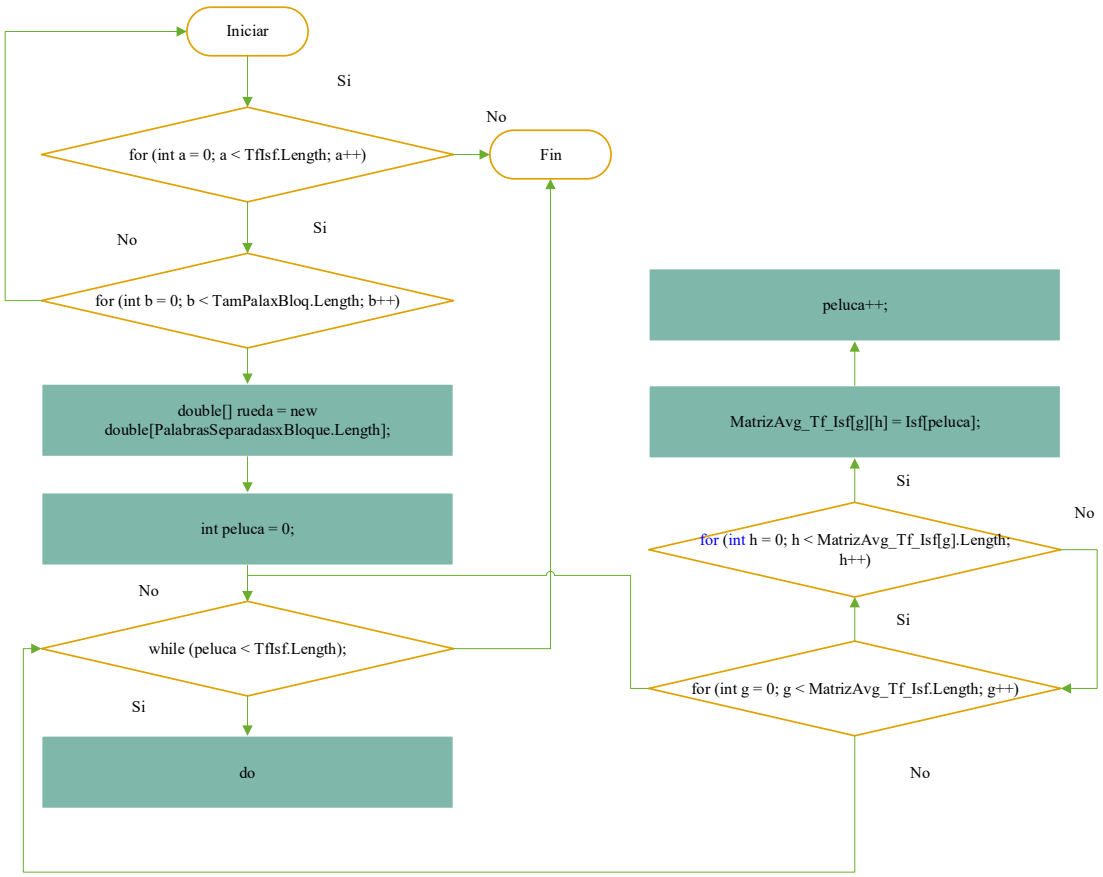


Fuente: Elaboración propia.

5.37. Inserción de datos a matriz *promedioTfIsf*

Realizar los cálculos de la operación para obtener los promedios *promedioTfIsf* requiere que los resultados sean guardados en un matriz la Figura 5.37 muestra el diagrama de flujo para la inserción de los datos en la matriz *promedioTfIsf*.

Figura 5.37. Diagrama de flujo para la inserción de datos en la Matriz *promedioTfIsf*.

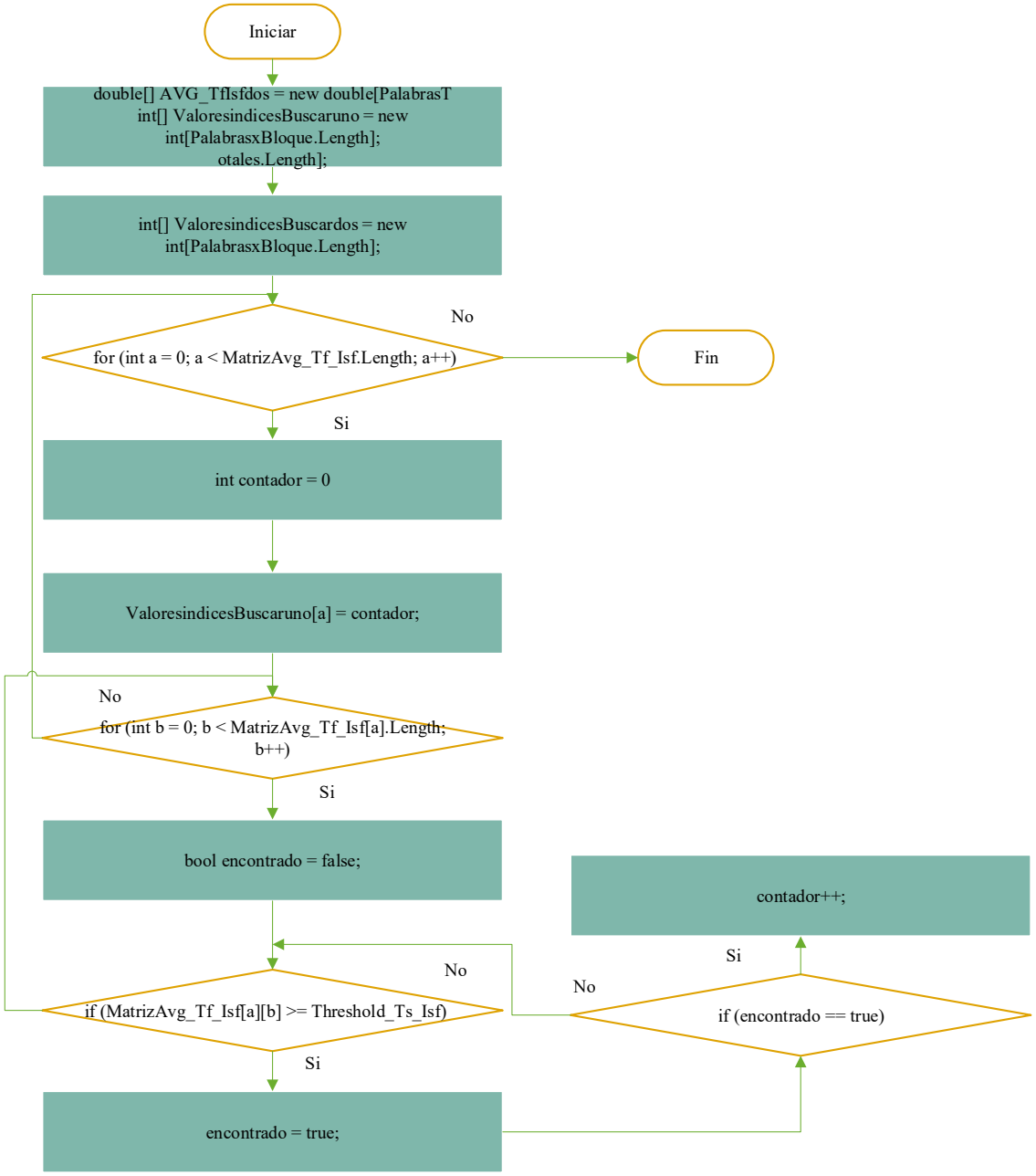


Fuente: Elaboración propia.

5.38. Declaración e inserción de vector para guardar valores de búsqueda en matriz Isf para obtener el máximo

Se requiere buscar el valor obtenido del máximo de *promedioTfIsf* en la matriz *Isf* cuyos resultados serán los valores son depositados en el vector *valorindicebuscado*. La Figura 5.38 muestra el diagrama de flujo de declaración e inserción en vector *valoríndicebuscado*.

Figura 5.38. Diagrama de flujo declaración e inserción en vector *valoríndicebuscado*.

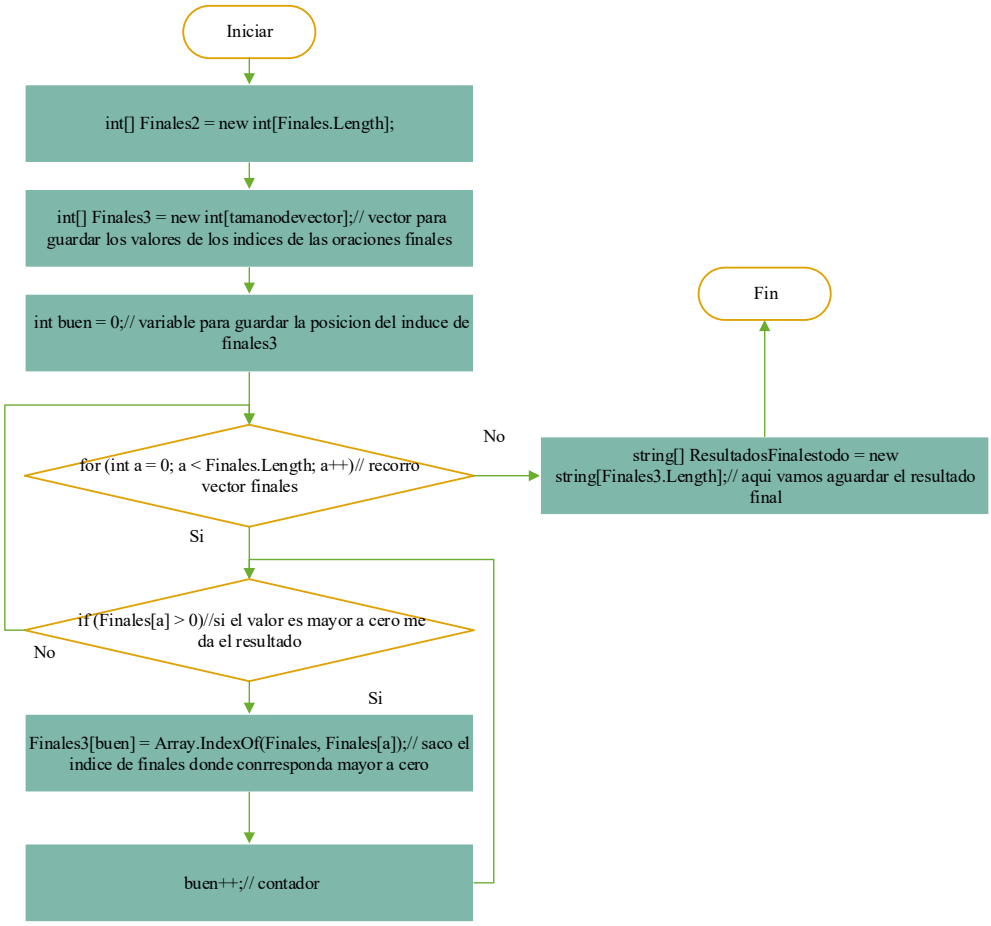


Fuente: Elaboración propia.

5.39. Búsqueda de índice para obtener cita

La búsqueda del valor correspondiente al índice es necesaria para identificar las citas que correspondan a ese índice. La Figura 5.39 muestra el diagrama de flujo correspondiente a la obtención del número de cita.

Figura 5.39. Diagrama de flujo búsqueda de índice para obtener cita o citas.

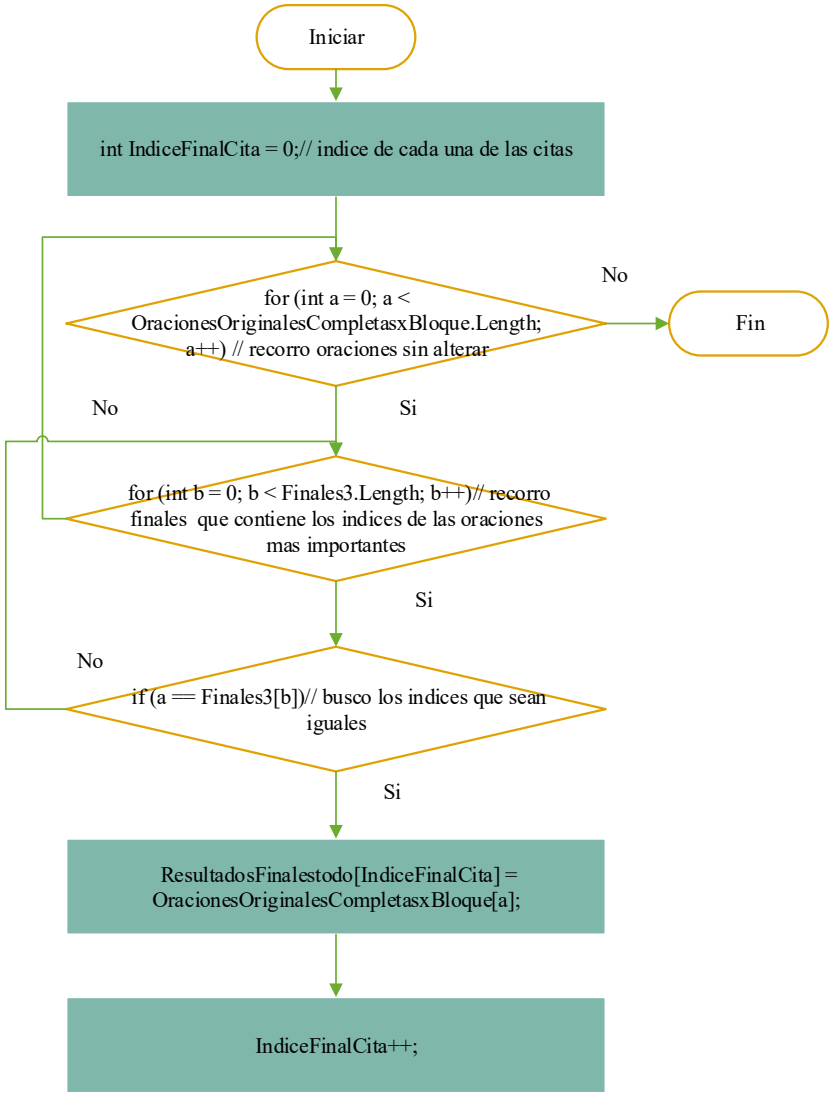


Fuente: Elaboración propia.

5.40. Búsqueda de incide en citas originales y creación de vector

Cuando se obtiene los números de citas la última operación consiste en realizar la búsqueda en las en el conjunto de citas originales para presentaras en un richtexbox que contendrá los resultados finales. La Figura 5.40 muestra el diagrama de flujo para realizar la búsqueda de las citas.

Figura 5.40. Diagrama de flujo de búsqueda de citas en citas originales



Fuente: Elaboración propia.

6. Resultados

6.1. Resultados de los pre-procesos.

La Tabla 6.1 muestra los resultados de los pre-procesos. Al analizar la evaluación se concluyó que fue el mismo tiempo de ejecución con o sin corrector ortográfico.

Tabla 6.1.

Comparación de evaluación de tiempo con o sin corrector ortográfico.

Pre-proceso	Tiempo	Pre-proceso	Tiempo
Tokenizador sin corrector ortográfico	1.90s.	Stemming sin corrector ortográfico	1.28s.
Tokenizador con corrector ortográfico	1.90s.	Stemming con corrector ortográfico	1.25s.
Stop removal word sin corrector ortográfico	1.22s.	Tiempo total sin corrector ortográfico	4.40s.
Stop removal word con corrector ortográfico	1.13s	Tiempo total con corrector ortográfico	4.28s.

Fuente: Elaboración propia.

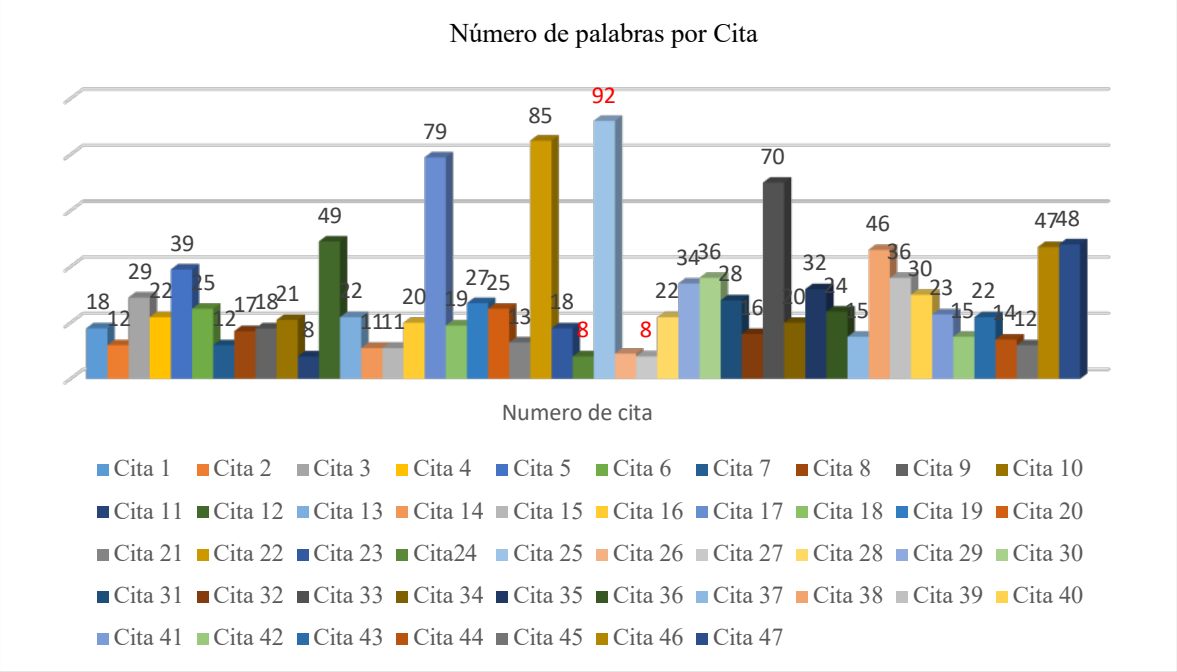
Por lo tanto, la primera suposición no pudo ser probada ya que no se mostró ser significativa la diferencia de tiempos durante el pre-proceso. La segunda suposición se considera comprobada y correcta, ya que para el pre-proceso de *stemming* con corrector ortográfico hubo un ahorro de 30 milisegundos respecto a la opción sin corrector ortográfico.

Para el tercer pre-proceso de *stop removal word* la diferencia fue de 90 milisegundos, lo cual quiere decir que la tercera suposición se cumple correctamente y da mejor beneficio en relación tiempo cuando las palabras están ortográficamente bien escritas. Observando los resultados detalladamente el tiempo total sin corrector fue de 4.40 segundos y con corrector ortográfico fue de 4.28 segundos por lo tanto se ha comprobado el beneficio de utilizar un corrector ortográfico en los pre-procesos de la MT, es evidente que entre más grande sea el conjunto de palabras a analizar mayor será el tiempo que se tardara en realizar el conjunto de pre-procesos.

6.2. Resultados de Conteo de palabras por cita

Los resultados obtenidos del conteo de palabras son mostrados a continuación. La Figura 6.1 muestra el resultado de conteo de palabras por cita. Es notable que la cita con más palabras fue el número 25 con un total de 92 palabras por cita y las menores fueron el número con 24 y 27 con un total de 8 palabras por cita. Es importante destacar que las palabras que se muestran son palabras que han llevado el pre-proceso de *stemming* y han sido llevadas a su forma raíz.

Figura 6.1. Ejemplo de número de palabras por cita.



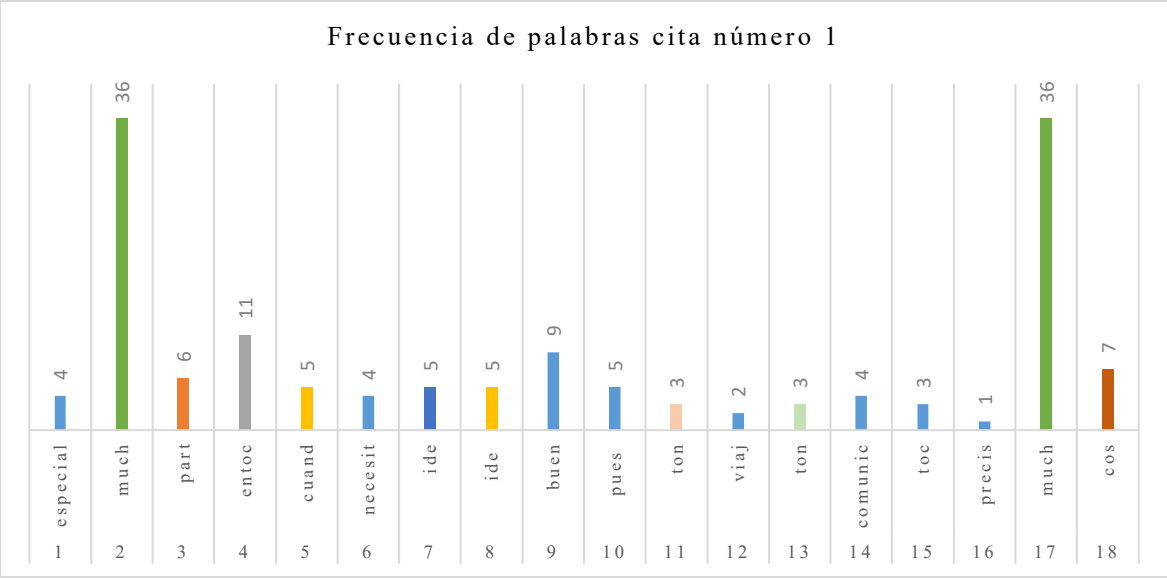
Fuente: Elaboración propia.

6.3. Resultados de Conteo de frecuencias de palabras

Al realizar el conteo de cada una de las palabras se obtuvo el resultado las frecuencias, la Figura 6.2 muestra un ejemplo de la primera cita. La palabra *much* para la cita 1 se repite un total de 2 veces en la palabra 2 y 17 respectivamente, pero se repite en todas las citas con

un valor máximo de 36 veces del total de las 47 citas. La Figura 6.2 muestra la tabla de frecuencias correspondientes a la cita número 1.

Figura 6.2. Frecuencias de palabras cita número 1.

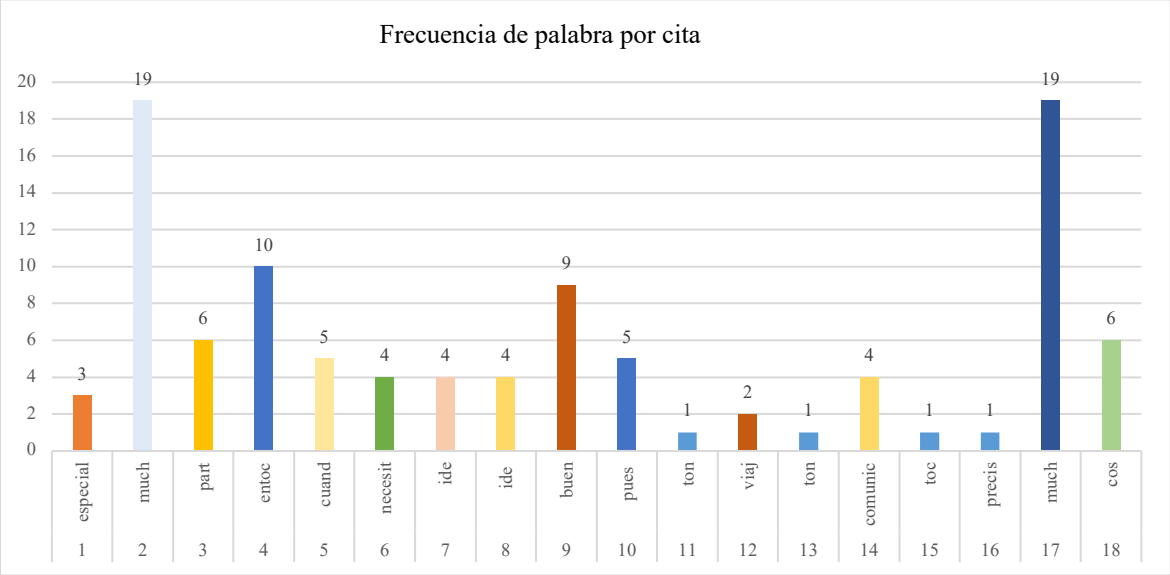


Fuente: Elaboración propia.

6.4. Resultados de Frecuencia de palabra por cita

Fue necesario contar el número de citas en las que una palabra aparece, para poder realizar las operaciones necesarias para el cálculo de *TfIsf*, se observa que la palabra *much* correspondiente a la palabra número 2 y el número 17 de la cita 1, pero se repite un total de 19 veces en las 47 citas, la Figura 6.3 muestra un ejemplo claro del resultado del conteo de frecuencia de palabra por cita en la cita número 1.

Figura 6.3. Frecuencia de palabra por cita.

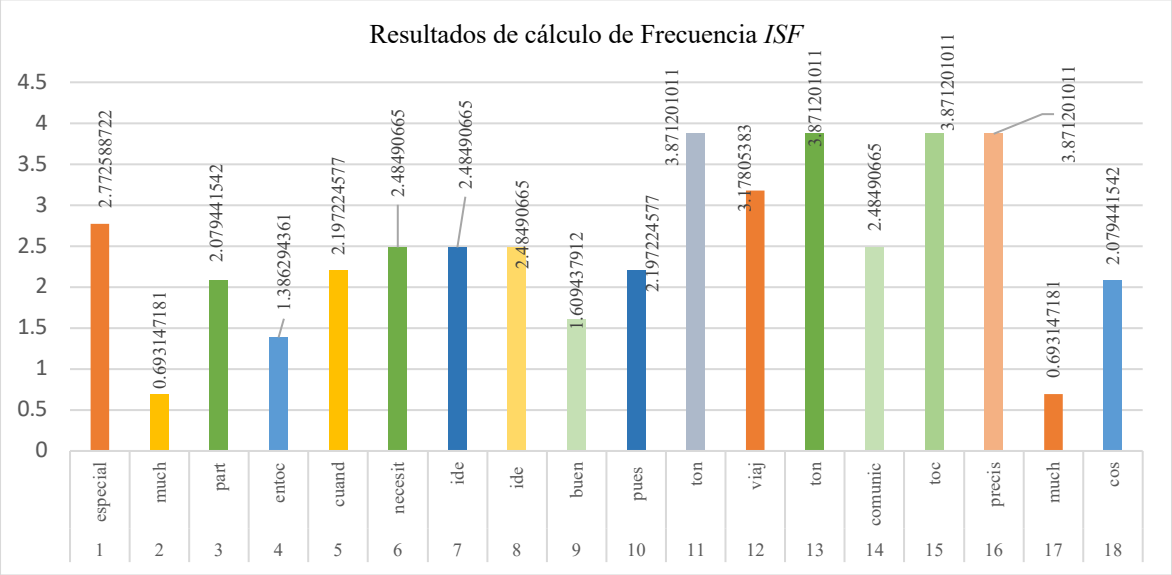


Fuente: Elaboración propia.

6.5. Resultados de Frecuencia Inversa del enunciado *Isf*

La Frecuencia Inversa del Enunciado (Isf) determina la ocurrencia de la frecuencia en las citas, La Figura 6.4 muestra los resultados obtenidos para la cita número 1, se observa que la frecuencia número 2 y 17 correspondientes a la palabra *much* son las más pequeñas con el mismo valor de .693147 y la frecuencia número 15 y 16 correspondientes a la palabra *ton, toc, percis* son las que tienen valor mayor de un 3.7120 respectivamente.

Figura 6.4. Resultados de cálculo de Frecuencia *Isf*

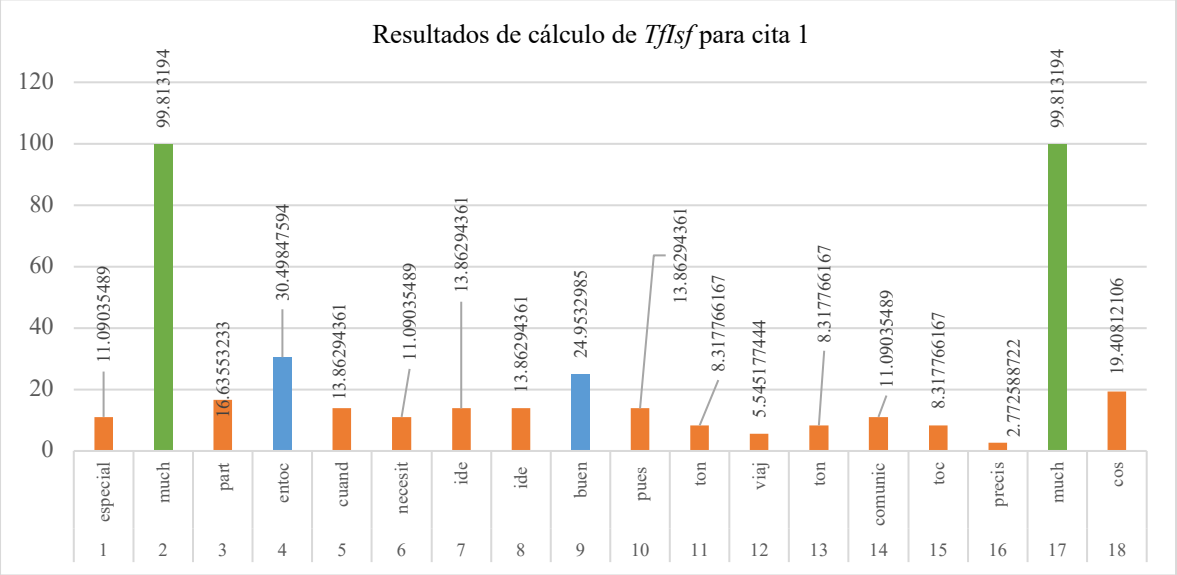


Fuente: Elaboración propia.

6.6. Resultados de *TfIsf*

Los resultados del cálculo *TfIsf* para la cita número 1 el valor más bajo fue la frecuencia número 16 con un valor de 2.77 y la más altas la frecuencia número 2 y 17 siendo la misma palabra con un valor de 99.81. La Figura 6.5 muestra el resultado del cálculo de *TfIsf* de la cita número 1.

Figura 6.5. Resultados de cálculo de *TfIsf* para cita 1.



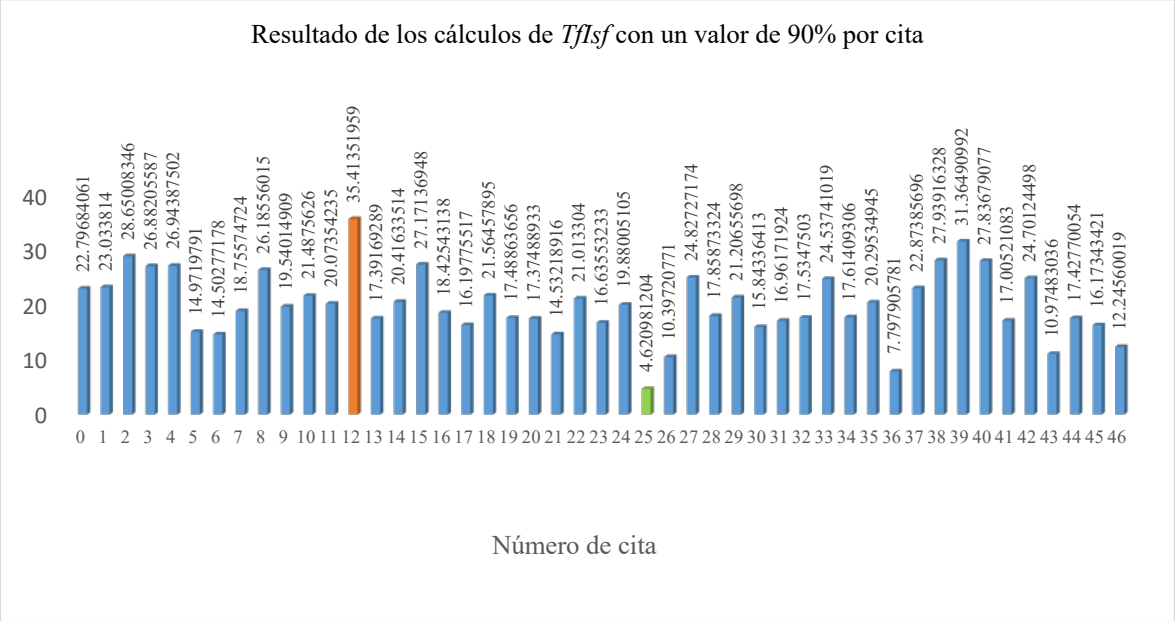
Fuente: Elaboración propia.

Los Figuras 6.2, 6.3, 6.4, 6.5, muestran los resultados obtenidos de realizar el análisis de cálculos de un total de 2431 palabras, correspondientes a un total de 47 diferentes citas procedentes del software de investigación Atlas.ti, de la categoría de conocimiento únicamente se ejemplificaron los resultados de la cita número 1. Para mostrar el funcionamiento del programa resumen de textos se realizará un ejemplo para resumir un 90% de las 47% citas, los resultados se muestras a continuación.

6.7. Resultados de Cálculo de términos con un Threshold de 90%

El resultado de este tipo de cálculo de resumen de textos también llamado Resumen de Texto Automático Extractivo debido que permite extraer un subconjunto del texto siendo el texto más representativo contenido en las citas textuales. La Figura 6.6 muestra los resultados de los cálculos con un Threshold de 90%, se puede observar que el valor menor fue la cita número 25 con un valor de 4.62 y el valor máximo fue la cita número 12 con un valor de 35.41.

Figura 6.6. Resultado de los cálculos de *TfIsf* con un valor de 90% por cita.

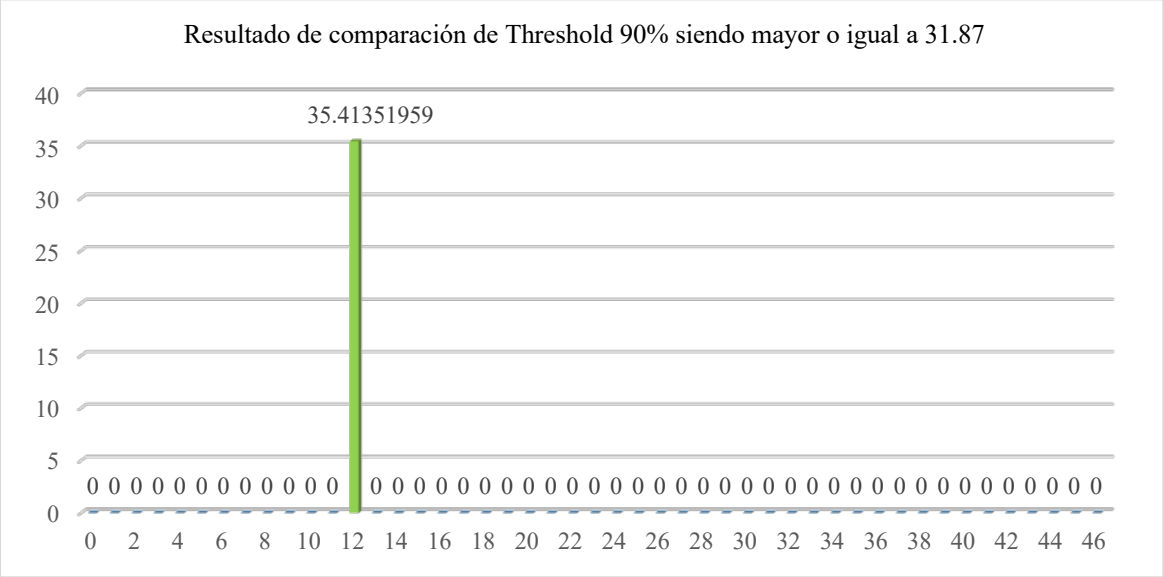


Fuente: Elaboración propia.

6.8. Resultados de valores iguales o mayores a un Threshold de 90%

Al realizar el cálculo con un Threshold de 90% se obtuvo el valor de 31.87 ese valor fue comparado con los valores de la Figura 6.7 siendo la cita número 13 con un valor 35.41 por lo tanto esta es la cita más representativa. La Figura 6.7 muestra el resultado de la comparación con Threshold de 90%.

Figura 6.7. Resultado de comparación de Threshold 90% siendo mayor o igual a 31.87.



Fuente: Elaboración propia.

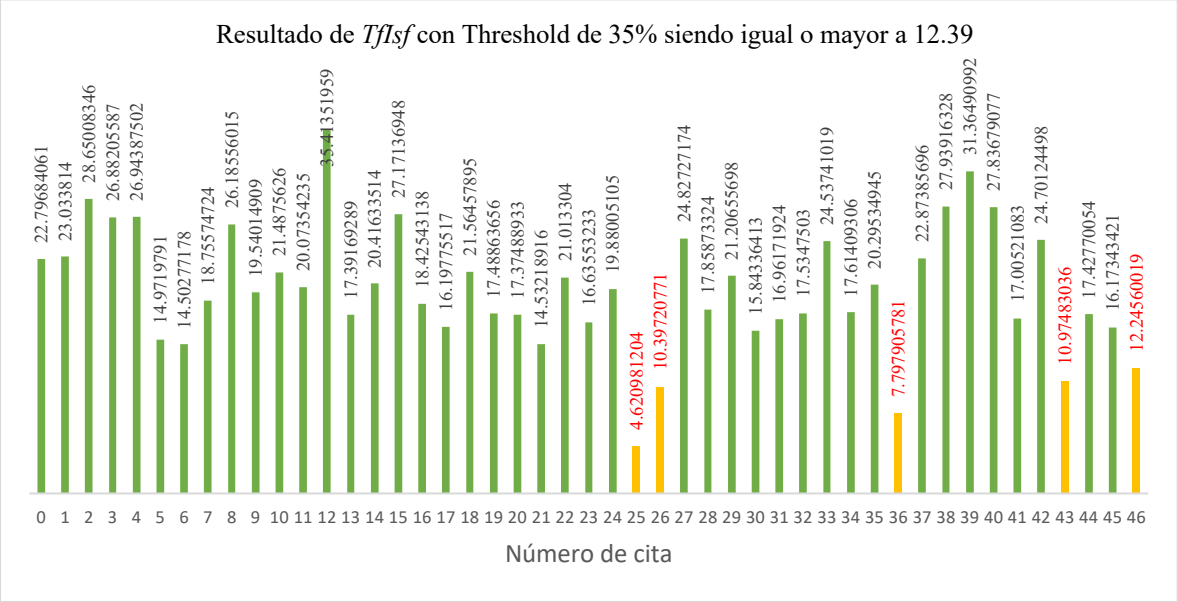
6.9 Resultados de citas con valor de Threshold mayor o igual a 31.87.

La Tabla A1.1 muestra los resultados con un Threshold de 90% mayor o igual a 31.87. Anexo A1

6.10. Resultados de valores iguales o mayores a un Threshold de 35%

Al realizar el cálculo con un Threshold de 35% se obtuvo el valor 12.39 ese valor fue comparado con los valores de la Figura 6.6 siendo las citas número 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 27, 28, 29, 30, 31, 32, 33, 34, 35, 37, 38, 39, 40, 41, 42, 44, 45 fueron bastantes citas debido que el valor de Threshold fue pequeño por lo tanto, caen dentro del rango mayor o igual al Threshold. La Figura 6.8 muestra el resultado del Threshold 35%.

Figura 6.8. Resultados de *TfIsf* con Threshold de 35% siendo igual o mayor a 12.39.



Fuente: Elaboración propia.

6.11. Resultados de citas con un valor 35% de Threshold mayor o igual a 12.39

Las citas que se presentan a continuación provienen de una entrevista realizada a estudiantes del Doctorado en Tecnologías Educativas que a su vez realizan investigaciones cualitativas o mixtas y que han utilizado el Atlas.ti como herramienta de análisis de entrevistas. Se presentan los resultados correspondientes con un Threshold de 35%. La Tabla A2.1 muestra los resultados con un Threshold de 35% mayor o igual a 12.39.
Anexo A2.

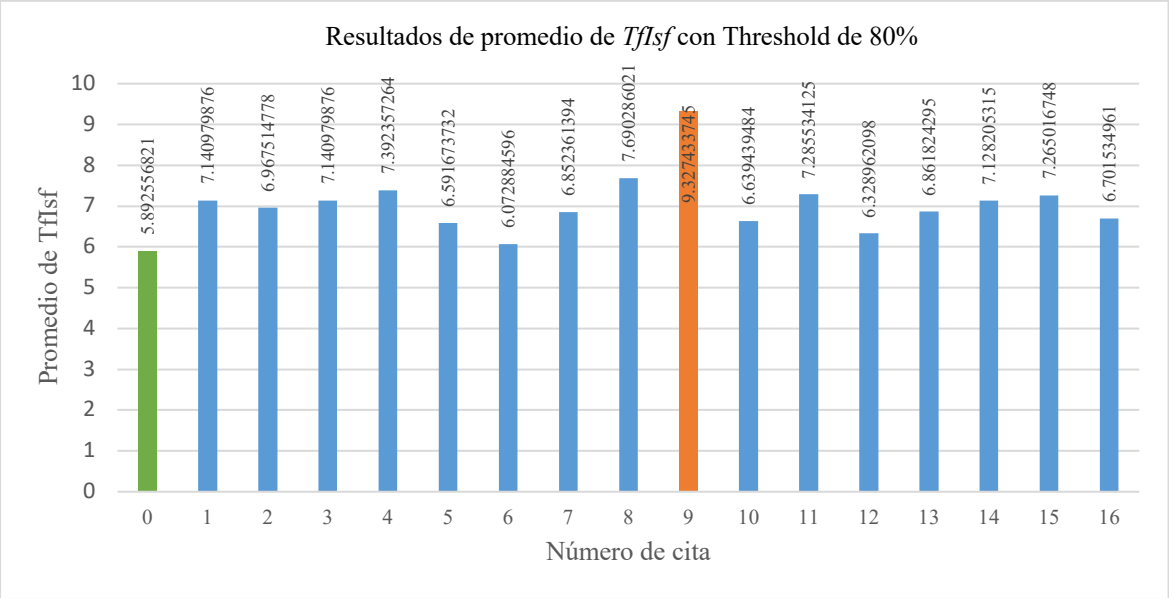
Las citas anteriores fueron el resultado de las citas que corresponden a un resumen con un Threshold al 35%, el número de citas dependerá del porcentaje de Threshold que el usuario introduzca cuando realice un resumen, entre más se eleve el porcentaje del mayor será la importancia de la cita por lo tanto entre mayor el número de Threshold menores citas aparecerán como es el primer ejemplo con un valor del 90%.

Con estos resultados se pretende demostrar la eficiencia de resumen de textos *Q&E*, el uso del algoritmo permite resumir textos de una manera eficaz y eficiente, además de coadyuva con trabajo de los investigadores, al realizar resúmenes de las citas colabora en la reducción de tiempo de lectura de la información, sin que se pierda la integridad de la misma.

6.12. Resultados de Cálculo de términos con un Threshold de 80%

Existe un nuevo formato del documento entregado por Atlas.ti en este año por lo tanto fue necesario adaptar el Resumen de textos a formato 2021 quedando de la siguiente manera. El documento formato 2021 contiene 17 citas textuales. La Figura 6.9 muestra el resultado del cálculo del promedio de *TfIsf* con un Threshold de 80%.

Figura 6.9. Resultados de promedio de *TfIsf* con Threshold de 80%



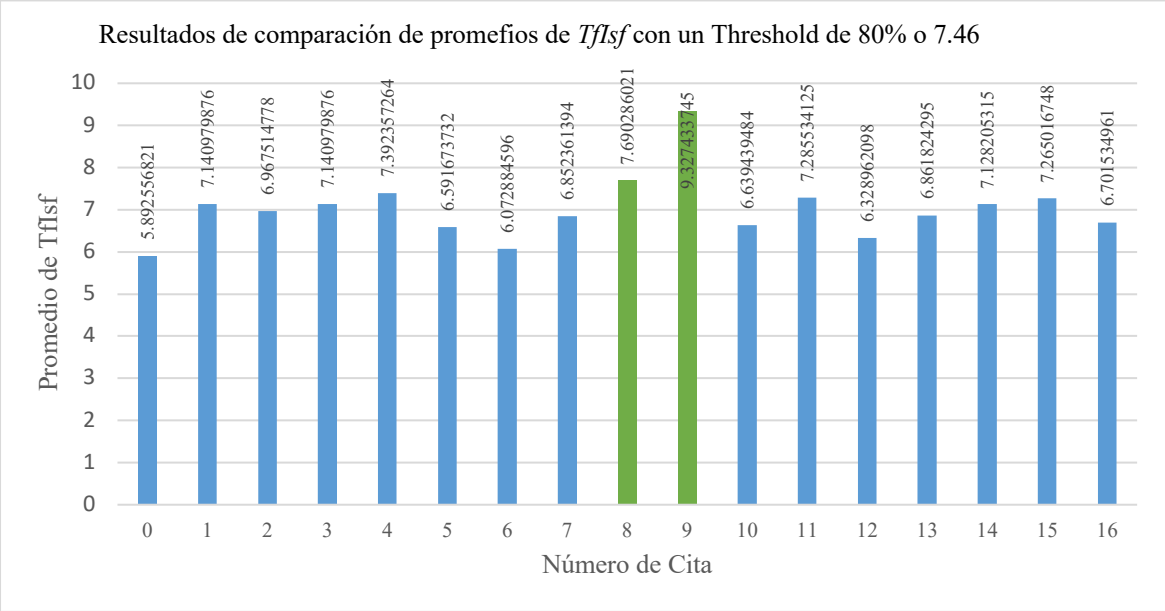
Elaboración: Propia.

Se observa en la Figura 6.11 que la cita con el valor más alto es la cita número 9 con un valor de 9.32 y la cita número 0 con un valor de 5.89.

6.13. Resultados de comparación de promedios de *TfIsf* con un Threshold de 7.46 correspondiente a un *Threshold* de 80%

El valor de Threshold obtenido para esta operación es de 7.46 por lo tanto se comparan y se buscan valores iguales o mayores a ese valor dando como resultado la cita número nueve con un valor de promedio *TfIsf* de 9.32 y la cita número 8 con un valor de promedio *TfIsf* de 7.69. La Figura 6.10 muestra los resultados de comparación con un Threshold de 80% o 7.46.

Figura 6.10. Resultados de comparación de promedios de *TfIsf* con un Threshold de 80% o 7.46.



Fuente: Elaboración propia.

6.14. Resultados Finales con un valor de Threshold de 7.46 correspondiente al 80%

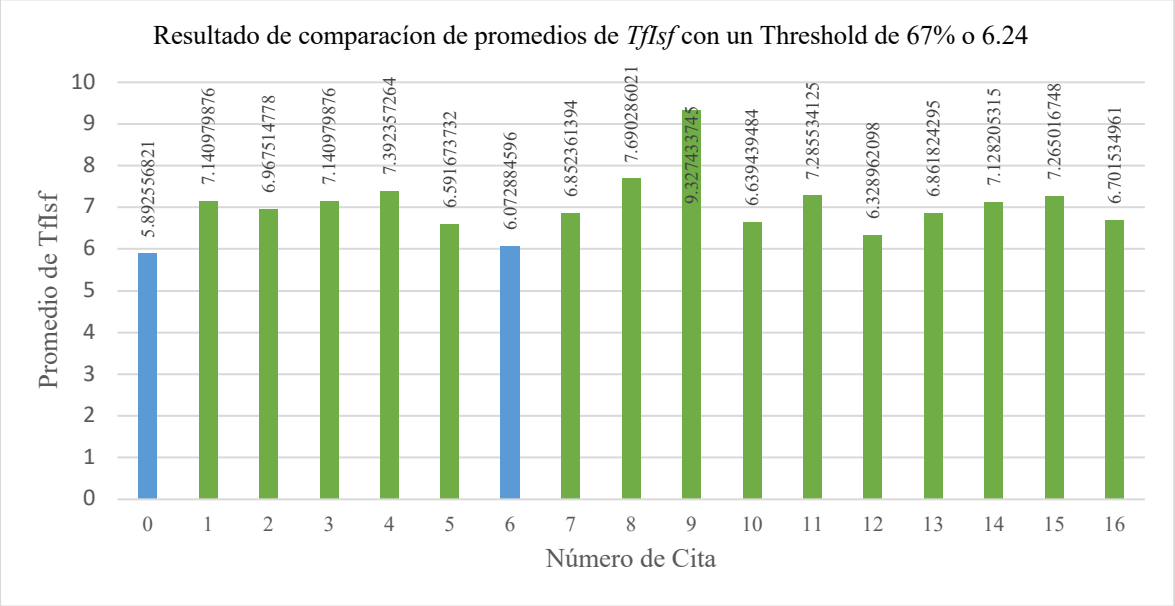
La Tabla A3.1 muestra los resultados Finales con un valor de Threshold mayor o igual a 7.46 correspondiente al 80%.

Anexo A3.

6.15. Resultados de comparación de promedios de *TfIsf* con un Threhold de 6.24 correspondiente a un *Threshold* de 67%

La Figura 6.11 muestra el resultado de la comparación de promedios de con un Threshold de 6.24 correspondiente de 67%

Figura 6.11. Resultado de comparación de promedios de *TfIsf* con un Threshold de 67% o 6.24.



Fuente: Elaboración propia.

Se observa en la Figura 6.11 que la cita número 1 con un valor de 5.89 y la cita número 6 con un valor de 6.07 son las únicas dos citas que no son ni mayores ni iguales al valor del Threshold de 67% se descartan siendo la citas número: 1, 2, 3,4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16 las mostradas.

6.16. Resultados Finales con un valor de 67% deThreshold mayor i igual a 6.24

La Tabla A4.1. muestra los resultados finales con un valor de 67 % Threshold mayor o igual a 6.24.

Anexo A4

A continuación, se muestras las conclusiones finales de esta investigación.

6.17 Resultados de opinión de funcionamiento final del algoritmo

Para conocer y validar el funcionamiento del algoritmo Quillo Espino y comprobar la hipótesis de esta investigación se realizó una encuesta 3 a investigadores de área cualitativa de la universidad Autónoma de Querétaro, a continuación, se reportan los resultados obtenidos:

6.18. Resultados de encuesta

1. ¿Existieron problemas durante la instalación del programa?

✓ Los encuestados reportaron que no existió ningún problema durante la instalación.

2. ¿Consideras que el tamaño de la aplicación es adecuado?

✓ Los encuestados reportaron que el tamaño de la aplicación fue adecuado.

3. ¿Consideras adecuado el tamaño de la fuente de los títulos de los botones de la aplicación?

✓ Los encuestados reportaron que fue adecuado el tamaño de la fuente del algoritmo.

4. ¿Consideras adecuado el tamaño de los cuadros de texto?

✓ Los encuestados reportaron que fue adecuado el tamaño de los cuadros de texto, uno de ellos comento *Considero que se podría modificar el tamaño para que la información de visualice de mejor forma y sobre todo se entienda el seguimiento que se está llevando a cabo*

5. ¿El programa se mantiene estable durante su funcionamiento?

✓ Los encuestados respondieron que fue estable su funcionamiento durante su ejecución.

6. ¿El algoritmo responde con velocidad adecuada?

Los encuestados reportaron:

✓ *El programa es bastante rápido para su procesamiento. Quizás lo más tardado del sistema es la secuencia de botones que se deben de seleccionar.*

✓ *Sentí que tardo un poco en procesar el texto debido a la cantidad del análisis que se tiene que hacer*

7. ¿El uso del algoritmo es?

✓ *Es relativamente complejo por la secuencia de botones que se tienen que seleccionar. Un aspecto que también impacta en la complejidad es la falta de ayuda dentro del sistema. Si se pudieran reducir los botones de la secuencia o resaltar el siguiente botón a presionar creo que se simplificaría mucho el uso.*

✓ *Considero que le falta ser más intuitivo, a primera vista no se entiende cómo es que debes usarlo ya después de analizarlo si lo comprendes*

8. ¿Consideras que el programa debería tener por default un porcentaje de resumen preestablecido?

✓ 2 reportaron que debería ser el 80% y uno menciono que debería ser entre 50 y 60%

9. ¿El programa realiza el trabajo deseado?

✓ Los encuestados reportaron que si realiza el trabajo de manera adecuada.

- ✓ *Uno de ellos menciono: Sí, en general, encontré que realizó los resúmenes de manera adecuada y fidedigna. Hubo casos en los que probé el programa y obtuve malos resultados, pero esto era por la calidad de la transcripción original de la entrevista.*

10. ¿Consideras que el programa cumple con el objetivo de ayudar al investigador a reducir el tiempo de lectura presentando la información más importante?

Los encuestados reportaron:

- ✓ *Sí, ya que resumir una entrevista es un trabajo que sí requiere de bastante tiempo y atención.*
- ✓ *Considero que es una buena herramienta para el análisis de este tipo de textos para obtener ideas más claras acerca de la investigación*

11. De acuerdo a tu experiencia ¿Consideras funcional el algoritmo para tu trabajo de investigación?

- ✓ Los 3 encuestados reportaron: que si es funcional el desarrollo del algoritmo.

12. ¿Cuál es tu opinión general del programa?

- ✓ *Me gustó bastante la propuesta y la idea. Al final del día, lo importante de una investigación son los resultados que se obtienen, al agilizar y facilitar el proceso el programa me ayudó mucho.*
- ✓ *Es demasiado útil y sobre todo reduce el tiempo para los investigadores, no tienen que estar en distintas plataformas, teniendo tu software tienen todo en un mismo lugar*
- ✓ *Es un programa muy útil para el análisis de investigaciones cualitativas. Me gustaría aprender un poco más sobre el tema y por supuesto, tener la oportunidad de utilizarlo.*

13. ¿Es útil el programa para la interpretación objetiva de la información generada a través de atlas.ti?

- ✓ Los 3 encuestados reportaron que si es útil.

14. ¿Consideras que el programa contribuye a obtener conocimiento nuevo favoreciendo la conservación de la información?

- ✓ *Los 3 encuestados respondieron que si ayuda a contribuir a obtener el conocimiento nuevo favoreciendo la conservación de la información.*

¿Consideras que los resultados del programa presentan la información más Importante del texto?

Los encuestados reportaron que si presenta información más importante de acuerdo a su análisis además uno de ellos comento:

- ✓ *Sí, el análisis que presenta el sistema es acertado en cuanto al producto que entrega.*

7. Discusión

El objetivo general de esta investigación fue proponer un algoritmo para el proceso de extracción de conocimiento a través de la minería de textos y análisis crítico del discurso, por lo tanto, se analizaron diferentes enfoques de RTA.

El enfoque abstractivo cuyo resultado es crear nuevas oraciones a partir de la información analizada, contiene métodos como ontologías, redes neuronales, basado en plantillas, basado en regla, se descartó por completo debido a que sus resultados generarían cambios notables en el texto nuevo produciendo perdida de la consistencia de la información y la homogeneidad.

Por su parte el enfoque extractivo cuyo objetivo es encontrar oraciones más relevantes dentro del texto analizado sin cambiar su idea por esta razón fue el enfoque más adecuado, en consecuencia se realizó el análisis de sus diferentes métodos extractivos para generar resúmenes de texto automático, por el ejemplo *método de localización* Allahyari, Pouriyeh & Assefi (2017), indican que se concentra en el origen de la posición de la información importante de una oración, puede ser en el inicio o el final de la oración, sin embargo este método no puede ser ocupado por que únicamente se concentra en una oración y no se aplica a documento o multidocumentos, a su vez el *método Cue,* Abdi et al (2016), determinan que es el que se basa en la hipótesis que la relevancia de una oración es calculada por la presencia o ausencia de palabras claves en específico, provenientes del diccionario de palabras, sin embargo es un método diseñado para multidocumento, pero el impedimento que no se contaba con un diccionario de palabras preestablecido, por el número limitado de palabras analizadas fue imposible utilizarlo, también el *método de Titulo,* Xiao & Munro (2019), declaran que es el que calcula el peso de la oración se calcula sumando las palabras contenidas en el encabezado y son comparadas con los demás títulos, pero en esta investigación no se enfocó en analizar los títulos de las entrevistas, se centró en el análisis del contenido de las entrevistas por tal motivo el método de títulos tampoco fue empleado, el método de *palabras temáticas,* Rahman & Borah (2020), refieren que son aquellos en los que una palabra aparece más en el texto, pero la repetición de las palabras por si solas en un texto no puede determinar si la palabra tiene relevancia dentro del texto, además solo se aplica a un solo documento, por tal motivo tampoco se consideró utilizarlo, el *método por grafos*

Khan et al (2018), deducen que representan la oración como un conjunto de palabras y utilizan una medida de similitud de contenido que puede detectar oraciones redundantes semánticamente o equivalentes, sin embargo es un método con mayor complejidad debido a la redundancia de la información de por esta razón no se utilizó.

Se concluyó que el método que más apropiado para esta investigación fue el *método extractivo TfIdf,* este utiliza dos métricas implícitamente para su funcionamiento la primera *es el termino de frecuencia* (TF) por sus siglas en inglés term frequency Qaiser & Ali (2018), la definen *como la medida que se utiliza para medir cuantas veces hay un término presente en un documento*, y la segunda *es la frecuencia inversa del documento* por sus siglas en inglés inverse document frequency (IDF) Kim & Gil (2019), *concluyen que es la métrica que se utiliza para medir la importancia de las palabras en todos los documentos analizados*. Gracias a esas dos métricas se puede obtener la relevancia de las palabras clave en o en los documentos con las cuales pueden para ser identificados o categorizados. Hans, Pramodana & Suhartono (2016), concluyen que TfIdf proporciona el peso de basado en la frecuencia de uno cada elemento de la oración, la puntuación aumenta o disminuye de acuerdo al número de la oración. Al ser un método estadístico numérico refleja la importancia de una palabra para cada documento de la colección, además puede ser aplicado a documento independiente o multidocumento. De acuerdo con Zhou & Salvendy (2018), determinan que el 83% de los sistemas de recomendación y resúmenes basados en texto digital, utilizan el método TfIdf como medida de ponderación y medida de calificación de la correlación entre documentos y las consultas de los usuarios, consecuentemente, es el método más adecuado para esta investigación siendo su confiabilidad alta en cualquier cantidad de porcentaje de threshold de resumen seleccionada por el usuario, para obtener resultados acertados.

Después del análisis de los resultados de la encuesta realizada a los investigadores del área de la facultad de informática de la Universidad Autónoma de Querétaro se concluye que el principal aporte para esta investigación fue que: *el algoritmo presenta y facilita una visión panorámica para la interpretación objetiva proveniente de la investigación social, debido a que el algoritmo muestra las oraciones más significativas de acuerdo al threshold seleccionado por el usuario por lo tanto, su funcionamiento y objetivo se cumplen correctamente, además contribuye a la conservación de la consistencia de la información*

logrando una interpretación objetiva y homogénea, debido a que el sistema es acertado en cuanto al producto que entrega, reduciendo el tiempo de análisis para los investigadores, además de agilizar y facilitar el proceso de investigación, ya que los investigadores no tienen que estar en diversas plataformas para realizar su trabajo.

Una de las limitaciones para el funcionamiento correcto del algoritmo fue la mala calidad proveniente del texto que se analizó, debido a que los textos utilizados en esta investigación fueron capturados manualmente por lo tanto es recomendable utilizar el corrector ortográfico que se encuentra en el mismo algoritmo para tratar de mejorar la calidad del texto y obtener resultados satisfactorios. Es necesario mencionar que el algoritmo fue programado para realizar análisis de entrevistas provenientes de atlas.ti, con características de formato de texto muy particulares, no obstante, se realizó otra prueba con otro tipo de texto con un total de 5000 palabras y el algoritmo funciono correctamente.

8. Conclusiones

Después de haber realizado esta investigación se realizan las siguientes conclusiones:

1. Existen diferentes tipos de formatos de texto que complican los procesos de análisis: el objetivo de los pre procesos fue brindar soporte para diferentes formatos analizando, borrando, limpiando y organizando el texto, cuyo fin es ayudar a la estructuración del texto de tal manera que se puedan encontrar los patrones que permitan la extracción de información importante.
2. Se sugiere la integración de un corrector ortográfico: debido a las faltas de ortografía que pueden presentar en el documento original procedente del software Atlas.ti fue necesario la aplicación de un corrector ortográfico para mejorar la ortografía de las palabras y se puede considerar como un pre-proceso extra durante la etapa 1 de la MT, con la finalidad de reducir el tiempo de ejecución del pre proceso stemming.
3. La tokenización es un elemento necesario para RTAE: gracias a la tokenización se pudo conocer el número de términos iniciales en el desarrollo del proceso de RTAE.
4. Se recomienda la aplicación del algoritmo Snowball: fue necesario adaptar el algoritmo snowball derivado del algoritmo Porter, con el propósito de eliminar sufijos de cada uno de los términos analizados, llevándolos a su raíz, a través de este proceso se puede contabilizar de frecuencia de términos de una manera perfecta.
5. El pre proceso de Stop words colaboran con la reducción de términos en el análisis: son elementos que no aportan ningún significado durante el proceso de análisis del texto por lo tanto fue necesario su aplicación y reducir el número de términos.
6. El espacio vectorial fue forzoso para el desarrollo de los RTAE: durante la ejecución del proceso general de RTAE la separación del documento en subconjuntos de documentos independientes fue llevada a cabo, para poder realizar el análisis de cada uno de los documentos fue necesario convertir cada uno de esos documentos en vectores de n-tamaños, cabe señalar que cada documento cuenta con tamaños diferentes.

7. El uso de matrices dentadas fue necesario para el almacenamiento de vectores: las características de filas y columnas desiguales permite almacenar los vectores de forma simple y sencilla además trabajar con matrices facilita la rapidez para su recorrido. Las matrices, asimismo contienen la relación entre términos y documentos ya que permiten realizar el cálculo de peso de cada documento(vector).
8. El cálculo la relevancia de cada uno de los vectores se ejecutó mediante la aplicación del método *TfIsf:* el método *TfIdf* forma parte de la recuperación de la información procedente de técnica de minería de textos.
9. El método también puede ser utilizado en la categorización de textos: de hecho, desde el momento en el que el texto es estructurado en vectores permite la ejecución de categorización de forma implícita ya que se asignan valores a los términos analizados.
10. No fue necesario el conocimiento y aplicación de semántica en los vectores: gracias a la aplicación del método *TfIsf* únicamente se aboco en realización de cálculos estadísticos tales como la frecuencia y la frecuencia inversa de documentos debido a las características del método, permitiendo que la ejecución del método sea de manera rápida. El objetivo fue buscar encontrar los valores más altos de *TfIdf* ya que representan mayor índice de discriminación en el cálculo de *TfIdf.*
11. Las operaciones implícitas dentro de *TfIdf* fueron ejecutadas de manera automática buscando que los valores que se obtuvieron fueran claros: el resultado proporciono un resumen de texto rápido y poderoso ya que al ser un resumen extractivo proporciona los vectores con mayor similitud o igualdad al Threshold solicitado por el usuario, denotando los resultados más relevantes del documento completo.
12. El objetivo de esta investigación fue demostrar los resultados y lograr la obtención de información importante en un texto, promoviendo el análisis y reducción de tamaño de grandes volúmenes de texto, y puede ser aplicado a un documento o a varios documentos.
13. Los resultados finales demuestran el texto organizado y categorizado de forma proporcional al Threshold: la aplicación del algoritmo promovió la reducción de tiempo de lectura de información relevante del documento, permitiendo un entendimiento de la investigación de manera sencilla con la finalidad de mantener la

conservación de la información lo máximo posible, además presenta de manera organizada y separada los documentos.

Referencias

1. Aghila, G., & Vidhya, K. A. (2010). Text Mining Process, Techniques and Tools : an Overview. *International Journal of Information Technology and Knowledge Management, 2*(2), 613-622. Obtenido el 10 septiembre de 2017 desde: http://csjournals.com/IJITKM/PDF%203-1/86.pdf
2. Aluja, T. (2001). La minería de datos, entre la estadística y la inteligencia artificial. *25(3).* pp. 479-498. *Sort international journal published by the Statistical Institute of Catalonia.* Obtenido el 11 de septiembre de 2017, desde: https://w ww.idescat.cat/sort/questiio/questiiopdf/25.3.4.Aluja.pdf
3. Bender, E. M. (2013). *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Washington Graeme Hirst. https://doi.org/10.2200/S00493ED1V01Y201303HLT020.
4. Bharati, A., Chaitanya, V., & Sangal, R. (1996). *Natural Languaje Processing: a paninian perspective*. New Delhi: Prentice-Hall. https://cdn.iiit.ac.in/cdn/ltrc.iiit.ac.in/downloads/nlpbook/nlp-panini.pdf.
5. Cambria, E., & White, B. (2014). Jumping NPL curves: A review of natural language Processing Research*. IEEE.Computacional intelligence magazine.* Obtenido desde 12 de septiembre de 2017, desde: https://w.sentic.net/jumping-nlp-curves.pdf.
6. Cano, J. R., & Herrera, F. (2006). Técnicas de reducción de datos en KDD. El uso de Algoritmos Evolutivos para la Selección de Instancias. *Actas del 1 seminario sobre sistemas inteligentes.* pp.165-181. Madrid: Universidad Rey Juan Carlos, Madrid.
7. Chakraborty, G. Pagolu, M., & Garla, S. (2013). Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*.* USA:SAS.
8. Dijk, V. & Teun, A. (1999). El análsis crítico del discurso*. Anthropos.* 186. pp. 22-36.
9. Dooley, A. R., & Levinsohn, S. H. (2007). *Análisis del discurso Manual de conceptos básicos.* 1ra Ed. Lima: Instituto Lingüístico de Verano.

10. Fayyad, U. Shampiro-Piatetsky. G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *AAAI Press.* Proceeding of Second International Conference on Knowledge Discovery and Data Mining. 1(1). pp. 2-4.
11. Fayyad, U. Shampiro-Piatetsky, G., & Smyth, P. (1997). Data Mining and Knowledge Discovery in Databases. *AAAI.* 17(3). pp. 27-31.
12. Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. 349(6245). *Science*. pp.261-267. DOI: 10.1126/science.aaa8685
13. Hotho, A. Nürnberger, A. & Paas, G. (2005). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*. 20 (1). pp 19-62. Obtenido el 15 de septiembre de 2017, desde:
14. https://www.semanticscholar.org/paper/A-Brief-Survey-of-Text-Mining-Hotho-N%C3%BCrnberger/8f74f5623c4e5c5931641a264cfd7c02097e1e22
15. Kao, A. & Poteet, S.R. (2007). *Natural Languaje processing and text mining.* London: Springer .
16. Kevvit, P. Mc. Patridge, D., & Wilks, Y. (1992). Approaches to natural language discourse processing. 6(4). pp. 333-364. Artificial Intelligence Review. Obtenido el 16 de semptiembre de 2017, desde:
17. https://www.researchgate.net/publication/226896389_Approaches_to_natural_language_discourse_processing
18. Leiva, I. G., & Rodríguez, M. J. V. (1996). El procesamiento del lenguaje natural aplicado al análisis del contenido de los documentos*. Revista general de informacion y documentacion. 6(2). p. 205.* Obtenido el 17 de septiembre de 2017, desde: https://dialnet.unirioja.es/servlet/articulo?codigo=169971
19. Leskovec, J. Rajaraman, A., & Ulman, J.D. (2010). *Mining of Massive Datasets.* California: Stanford University.
20. Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook.* New York, New York, USA: Springer.
21. North, M. (2012). *Data Mining for the Masses .* USA: Global Text Project.

22. Panessi, W., & Bordignon, F. (2017). Procesamiento de Variantes Morfológicas en Búsquedas de Textos en Castellano. *Revista interamericana de bibliotecología*. 24(1). pp. 69-88. Obtenido el 9 de octubre de 2017, desde http://www.tyr.unlu.edu.ar/TYR-publica/debaja-Varia-Morfo.pdf
23. Porter, M. (1979). Algoritmo para la extracción de raiz de palabra. Obtenido el 7 de agosto de 2017 desde https://tartarus.org/martin/PorterStemmer/index-old
24. Pushpa, S., & Balamurugan, R. (2015). A review on Various text mining techniques and algorithms. *International journal of advance research in science and engineering.* 4(11). pp. 288-299.
25. Saggion, H. (2010). Procesamiento del lenguaje natural para el análisis del lenguaje subjetivo. 14(2). *Subjetividady procesos cognitivos*. Obtenido el 16 de octubre de 2017, desde: http://dspace.uces.edu.ar:8180/xmlui/bitstream/handle/123456789/970/Procesamiento_de_lenguaje_Saggion.pdf?sequence=1
26. Sandelowski, M. (1995). *Focus on Qualitative methods qualitative analysis: What it is and how to begin.* SNRS. 18(4). pp. 371-375. Obtenido el 17 de octubre de 2017, desde: https://onlinelibrary.wiley.com/doi/epdf/10.1002/nur.4770180411.
27. Santander, P. (2011). Por qué y cómo hacer análisis de discurso. *Revista de Epistemología de Ciencias Sociales.* 41(1). pp. 207-224. Obtenido el 28 de septiembre de 2017, desde: https://www.moebio.uchile.cl/41/santander.html
28. Sayago, S. (2014). El análisis del discurso como técnica de investigación cualitativa y cuantitativa en las ciencias sociales. *Revista de Epistemología de Ciencias Sociales. 49(1).* pp. 1-10. Obtenido el 29 de septiembre de 2017 desde: https://www.moebio.uchile.cl/49/sayago.htmlHsien, H. F., & Shannon, S. (2005).
29. Hsiu-Fang, H., & Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative health Research*. 15(9). pp. 1277-1288. Obtenido el 30 de septiembre de 2017, desde:

https://www.researchgate.net/publication/7561647_Three_Approaches_to_Qualitative_Content_Analysis

30. Stifelman, L. J. (1995). A Discourse Analysis Approach to Structured Speech. Cambridge: *MIT Media Laboratory*. Obtenido el 30 de septiembre de 2017, desde: https://www.media.mit.edu/speech/people/lisa/aaai.html.
31. Usama, M. F., Piatesky-Shapiro G. Smyth P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining.* USA: MIT Press.
32. Valcárcel, A. V. (2004). Data Mining y el descubrimiento del conocimiento. *Industrial Data.* Lima, Lima, Perú: Industrial Data Revista de Investigación. 7(2). pp. 83-86. Obtenido el 1 de octubre de 2017, desde: https://revistasinvestigacion.unmsm.edu.pe/index.php/idata/article/view/6140/5331.
33. Wodak, R., & Meyer, M. (2003). *Métodos de análisis crítico del discurso .* España: Gedisa .
34. Zadeh, L. A. (2004). Precinsiated Natural Languaje (PNL). *AI Magazine.* 25(3). p.74. Obteido el 6 de octubre de 2017, desde: https://doi.org/10.1609/aimag.v25i3.1778
35. Nigro, O., Xodo, D., Cordi, G., & Terren, D. (2004). Kdd (Knoledge discovery in databases): un proceso centrado en el usuario. *VI Workshop de Investigadores en Ciencias de la Computación.* pp.53-58. Obtenido el 6 de Octubre de 2017 desde: http://sedici.unlp.edu.ar/handle/10915/21220
36. Troche, C. A. (2014). Aplicación de la minería de datos sobre bases de datos transaccionales. *Fides y Ratio.* 7(7). pp.58-66. Obtenido el 6 diciembre de 2017 desde http://www.scielo.org.bo/pdf/rfer/v7n7/v7n7_a05.pdf .
37. Han, J. Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techiques.* USA: Elsiever.
38. Maracano, A.Y. J., & Talavera, P. R. (2007). Minería de datos como soporte a la toma de decisiones emperesariales. *Produccion cientifica luz*. 23(52). pp.104-118. Universidad de Zulia Maracaibo, Venezuela. Obtenido el 14 de diciembre de 2017 desde: http://www.redalyc.org/articulo.oa?id=31005208

39. Hand, D.J. (2007). Principles of Data Mining. *Springer link.* Obtenido el 15 de diciembre de 2017 desde: https& Hall, M. A. E. (2011*). Data Mining: Practical Machine Learning Tools and Techniques*. USA: Elsevier. Obtenido el 17 de diciembre de 2017 desde https://books.google.es/books?id=bDtLM8CODsQC&printsec=frontcover&hl=es&source=gbs_atb#v=onepage&q&f=false
40. Quiroz, G. N. L. Valencia, C. A. (2012). Aplicación del proceso de KDD en el contexto de bibliomining: El caso Elogim. *Revista Interamericana de bibliotecología*. 35(1). pp. 97-108. Obtenido el 18 de diciembre de 2017 desde http://www.scielo.org.co/pdf/rib/v35n1/v35n1a9.pdf
41. Contreras, B. M. (2014) Minería de texto: Una visión actual. *Sistema de Información Científica Redalyc Red de Revistas Científicas.* 17(2). pp.129-138. Obtenido del 18 de diciembre de 2017 desde: https://www.redalyc.org/articulo.oa?id=28540279005
42. Sukanya, M. Biruntha, S. (2012). Techniques on Text Mining. *ICACCCT*. pp. 269-271. Obtenido el 19 diciembre de 2017 http://ieeexplore.ieee.org/document/6320784/?reload=true
43. Eíto, B.R. Senso, A. J. (2004). Minería textual. *En el Profesional de la información.* 13(1). p.11-27. Obtenido el 5 de enero de 2018, desde: http://profesionaldelainformacion.com/contenidos/2004/enero/2.pdf
44. Reyes, S. J. F., & García. F.R. (2005). El proceso de descubrimiento de conocimiento en bases de datos. *Ingenierias.* 3(26). pp.37-47. Obtenido el 10 de enero desde: https://www.google.com.mx/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&cad=rja&uact=8&ved=0ahUKEwj7iKWj5vvYAhVGb60KHfdSC94QFgg7MAM&url=http%3A%2F%2Fingenierias.uanl.mx%2F26%2Fpdfs%2F26_el_proceso.pdf&usg=AOvVaw1xZmlgMWEaCL7aauc-eoRV.pdf
45. Vijayarani, S., Llamathi, J., & Nithya, S. (2015). *Preprocessing Techniques for Text Mining*. *International Journal of Computer Science & Communication Networks.* 5 (1). pp. 7-16. Obtenido el 11 de enero de 2018, desde:

https://www.researchgate.net/profile/Vijayarani-Mohan/publication/339529230_Preprocessing_Techniques_for_Text_Mining_-_An_Overview/links/5e57a59f299bf1bdb83e7972/Preprocessing-Techniques-for-Text-Mining-An-Overview.pdf

46. Dursun, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications.* pp. 1707-1720. Obtenido el 11 de enero de 2018, desde: https://scholars.okstate.edu/en/publications/seeding-the-survey-and-analysis-of-research-literature-with-text-
47. Vijay, G. S., & Chaugule, A. Patil, P. (2014). Text Mining Methods and Techniques*. International Journal of Computer Applications*. 85(17). p. 42-45. Obtenido el 12 de enero de 2018, desde: https://www.ijcaonline.org/archives/volume85/number17/14937-3507
48. Cortez, V. A., Vega, H. H., & Pariona, Q. J. (2009). Procesamiento del Lenguaje Natural. *Revista de Ingeniería de sistemas e Informática*. 6(2). p. 48. Obtenido el 13 de enero de 2018, desde: https://core.ac.uk/download/pdf/304898423.pdf
49. Vijayarani, S. Janani, R. (2016). Text Mining Open-Source Tokenization Toll-An Analysis. *Advanced Computational Intelligence: An International journal* (ACII), 3(1). p. 38. Obtenido el 14 de enero de 2018, desde: https://aircconline.com/acii/V3N1/3116acii04.pdf
50. Sigh, J. Gupta, V. (2016). *Approaches, Applications and Challenges. ACM Computing Surveys.* 4(3). Obtenido el 14 de enero de 2018, desde: https://dl.acm.org/doi/10.1145/2975608
51. Nadkarni, P. M. Ohno-Machado, L., & Chapman, W. W. (2011). Natural Language Processing: an introduction. J*ournal of the American Medical Informatics Association: JAMIA,* 18(5). pp. 544-551. Obtenido el 15 de enero de 2018, desde: http://doi.org/10.1136/amiajnl-2011-000464 .
52. Lalitha, T., & Meenakshi, S. (2014). Text mining algorithm discotext (discovery from text extraction with information extraction). J*ournal of*

theoretical and applied information technology. 64(2). p. 433-445. Obtenido el 15 de enero de 2018, desde: http://www.jatit.org/volumes/Vol64No2/17Vol64No2.pdf

53. Cunningham, H. (1997). Information extraction a user guide. *Institute for language, speech and hearing (ILASH)*. Obtenido el 15 de agosto de 2018 desde: http://home.mit.bme.hu/~dezsenyi/research/cikkek/cunningham97information.pdf
54. Abdelmagid, E. M., Ahmed, A., & Himmat, M. (2015). Information extraction methods and extraction techniques in the chemical document´s contents: survey. *ARPN Journal of engineering and applied sciences*. 10(3), p. 1068-1073. Obtenido el 16 de agosto de 2018, desde: https://www.researchgate.net/publication/282382161_Information_Extraction_methods_and_extraction_techniques_in_the_chemical_document%27s_contents_Survey
55. Cimmano, P., Reyle, U., & Saric, J. (2005). Ontology- driven discourse analysis for information extraction. *Data & knowledge engineering*. 55(1). pp. 59-83. Obtenido el 17 de agosto de 2018, desde: https://www.sciencedirect.com/science/article/abs/pii/S0169023X04002228
56. Varsha, P., & Khandelwal, S. A. (2016). Information extraction technique: a review. *IOSR Journal of Computer Engineering*. pp. 16-20. Obtenido el 21 de agosto de 2018 desde: http://www.iosrjournals.org/iosr-jce/papers/conf.15013/Volume%209/4.%2016-20.pdf
57. Grishman, R. (1997). Information extraction: techniques and challenges. *Computer science department.* Obtenido el 22 de agosto de 2018, desde: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.6923&rep=rep1&type=pdf
58. Sarawagi, S. (2008). Information extraction. *Foundation and trends in databases*. 1(3), pp. 261-377. Obtenido el 23 de agosto de 2018, desde: https://dl.acm.org/doi/10.1561/1900000003

59. McCallum, A. & Nigam, K. (2001). A comparison of event models for Näive Bayes text classification. *Work Learn Text Categ*. pp. 752. Obtenido del 23 de agosto de 2018 desde: https://www.researchgate.net/publication/2408220_A_Comparison_of_Event_Models_for_Naive_Bayes_Text_Classification
60. Chaix, E., Deléger, L., Bossy, R., & Nédellec, C. (2018). Text mining tools for extracting information about microbial biodiversity in food. *Elsevier ltd*. 81. pp.1-13. Obtenido el 24 de agosto de 2018, desde: https://www.researchgate.net/publication/324682808_Text-mining_tools_for_extracting_information_about_microbial_biodiversity_in_food
61. Mitchell, T. M. (1997). Machine Learning. Burr Ridge, McGraw Hill. 45. pp.180-185. Obtenido el 20 de agosto de 2018 desde: http://www.cs.cmu.edu/%7Etom/NewChapters.html
62. Jiang, J. (2012). Information extraction from text. *Research Collection School Of Computing and Information Systems.* pp. 11-41 Obtenido el 21 de agosto de 2018 desde: https://ink.library.smu.edu.sg/sis_research/1711/
63. Singapure Magament University. pp. 11-41. ISBN. 9781461432227 Obtenido el 27 de agosto de 2018 desde: https://ink.library.smu.edu.sg/sis_research/1711/
64. Pressman, R. S. (2010). *Ingeniería de software*. *Un enfoque práctico*. USA: McGrawHill. Obtenido el 10 de noviembre de 2018 desde: http://cotana.informatica.edu.bo/downloads/ld-Ingenieria.de.software.enfoque.practico.7ed.Pressman.PDF
65. Bryant, A., & Kirkham, J. A. (1983). Software Engineering Economics: a review essay. *ACM SIGSOFT Software Engineering Notes*. 8(3). pp. 201-217. Obtenido el 11 de noviembre de 2018 desde: http://csse.usc.edu/TECHRPTS/1984/usccse84-500/usccse84-500s.pdf
66. Mah, M. (1999). High-definition software measurement. *Software Development.* 7 (5). pp. 14-15. Obtenido el 12 de noviembre de 2018, desde: https://dl.acm.org/doi/abs/10.5555/315382.315389
67. Salton, G. Wong, A. & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM.* 18(11). pp. 613-618. Obtenido el 30 de enero de 2019, desde: https://dl.acm.org/citation.cfm?id=361220
68. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Guitierrez, J. B., & Kochut, K. A brief survey of text mining: classification, clustering and

extraction techniques. (2017). *In procceding of KDD bigdas*. Obtenido el 30 de enero de 2019, desde: https://arxiv.org/pdf/1707.02919.pdf
69. Salton, G. Allan, J. (1994). Automatic text decomposition and structuring. *Conference on research and development in information retrieval*. pp. 21-30. Obtenido el 30 de enero de 2019, desde: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.5398&rep=rep1&type=pdf
70. Wilks, Y. (1997). Information extraction as a core language technology. *In M-T pzienza, editor*. Springer, Berlin. Obtenido el día 1 enero de febrero de 2019, desde: https://pdfs.semanticscholar.org/2c90/fa59c6d9beed8dcb0e844725b872d3f33a35.pdf
71. Sebastiani, F. (2002). Machine Learning in automated text categorization. *ACM Computing Surveys*. 34(1). pp 1-47. Obtenido el 25 de agosto de 2019 desde: https://dl.acm.org/citation.cfm?id=505283
72. Niharika, S. Latha, S. & Lavanya, D. R. (2012). A survey on text categorization. *International Journal of computer Trends and technology*. Obtenido el 8 de agosto de 2019, desde: http://www.ijcttjournal.org/Volume3/issue-1/IJCTT-V3I1P108.pdf
73. Wang, J., & Li, X. (2010). An Improved KNN algorithm for text classification. *International Conference of Information*. Networking and Automation (ICINA). pp. 436-439. Obtenido el 8 de agosto de 2019, desde: https://ieeexplore.ieee.org/document/5636476/figures#figures
74. Salton, G., & Yang, C. S. (1973). On the specification of term values in automatic indexing. *Department of Computer Science*. (29), pp. 351-372. Obtenido el 10 de septiembre de 2019, desde: https://ecommons.cornell.edu/bitstream/handle/1813/6016/73-173.pdf?sequence=1&isAllowed=y
75. Ming-Yang, S. (2011). Using clustering to improve the KNN – based classifier for online anomaly network traffic identification. *Elsevier.* (34)2. pp. 722-730. Obtenido el 11 de septiembre de 2019, desde: https://www.sciencedirect.com/science/article/pii/S1084804510001785?via%3Dihub
76. Robertson, S. (2004). Understanding inverse document frequency on theoretical arguments for IDF. *Journal of Documentation.* (60), pp. 503-520. Obtenido el 14 de septiembre de 2019 desde: https://pdfs.semanticscholar.org/8397/ab573dd6c97a39ff4feb9c2d9b3c1e16c705.pdf
77. Mikawa, K. Ishida, T., & Goto, M. (2011). A proporsal of extended cosine measure for distance metric learning in text classification. *IEEE.* Obtenido el

14 de septiembre de 2019 desde: https://ieeexplore.ieee.org/document/6083923

78. Bolande, O., & Olumide, K. (2012). A feature opinion extraction approach to opinion mining. *Journal of web engineering*. 11(1). pp.51-66 Obtenido el 20 de octubre de 2019 desde: https://www.researchgate.net/publication/220538260_A_Feature-Opinion_Extraction_Approach_to_Opinion_Mining
79. Wei, J., & Hung, H. H. (2009). A novel lexicalized HMM-based learning framework for web opinion mining. *Appearing in proceeding of the 26th international conference on machine learning*. Obtenido el 21 de octubre de 2019, desde: http://people.cs.pitt.edu/~huynv/research/aspect-sentiment/A%20novel%20lexicalized%20HMM-based%20learning%20framework%20for%20web%20opinion%20mining.pdf
80. Karanokas, H. Tjortjis, C., & Theodoulidis, B. (2000). An approach of text mining using information extraction. *Centre for research in information Management*. Obtenido el 23 de octubre de 2019, desde: https://s3.amazonaws.com/academia.edu.documents/41359587/WS5_13.pdf?response-content-disposition=inline%3B%20filename%3DAn_approach_to_text_mining_using_informa.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20191110%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20191110T234658Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=c60ec3650c333a691bb7ab7e3860fa746d5b1130c6b3449013ca6848cf84c62d
81. Nayak, T. Prasad, S., & Senepati. (2015). A survey on web text information retrieval in text mining. *Research journal of applied sciences, Engineering and technology*. 10(10). pp.1164-1174. Obtenido el 23 de octubre de 2019, desde: https://www.researchgate.net/publication/283225863_A_Survey_on_Web_Text_Information_Retrieval_in_Text_Mining
82. Dang, S., & Ahmad, P.H. (2013). A review of text mining techniques associated with various application areas. *International journal of science and research*. 4(2). pp. 2461-2466. Obtenido el 24 de octubre de 2019, desde: https://pdfs.semanticscholar.org/58be/9b174785c74444beea35c944b4aa57ce23f7.pdf
83. Ahmad, P. H. & Dang, S. (2014). A Comparative Study on Text Mining Techniques. *International Journal of Science and Research*. 3(4). pp.2222-2226. ISSN: 2319-7064. Obtenido el 24 de octubre de 2019 desde: https://pdfs.semanticscholar.org/7cb6/fbc1d2abc7e63d2faa399406a2199ae80881.pdf?_ga=2.128559438.390162867.1573435999-1310702993.1573435999

84. Liao, Y. & Vemuri, R. V. (2017). Use of k-nearest neighbor classifier for instruction detection. *Computers & Security*. 21(5). pp.439-448. Obtenido el 25 de octubre de 2019 desde: https://www.researchgate.net/publication/220614594_Use_of_K-Nearest_Neighbor_classifier_for_intrusion_detection
85. Wang, J. & Xia, L. (2010). An improved algorithm for text classification. *International conference on information, networking and automation (ICINA).* pp.436-438. Obtenido el 25 de octubre de 2019, desde: https://ieeexplore.ieee.org/document/5636476?reload=true&arnumber=5636476&contentType=Conference%20Publications
86. Mani, I. Klein, G. House, D., & Hirshchman, L. (2002). Summac: a text summarization evaluation. Natural Language Engineering. 8(1). pp.43-68. Obtenido el 18 de marzo de 2020 desde: https://pdfs.semanticscholar.org/d61c/0b9d86f60ce2220be9ee2baf5009c5ce8841.pdf
87. Maybury, M. T. (1995). Generating summaries from event data. Information processing and management. *Elsevier.* 31(5). pp. 735-751. Obtenido el 20 de marzo de 2020 desde: https://www.sciencedirect.com/science/article/abs/pii/030645739500025C
88. Jing, H., Barzilay, R., McKeown, K., & Elhadad, M. (1998). Summarization Evaluation Methods Experiments and Analysis. *AAAI Technical Report*. Obtenido el 21 de marzo de 2020 desde: https://www.aaai.org/Papers/Symposia/Spring/1998/SS-98-06/SS98-06-007.pdf
89. Mani, I. (2001). Summarization Evaluation: An Overview. *In Proceedings of the NAACL 2001* Workshop on Automatic Summarization. Obtenido el 22 de marzo de 2020 desde: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/sum-mani.pdf
90. Dorr, J. B., Monz, C., President, S., Schawartz., & Zajic, D. (2005). A methodology for extrinsic evaluation of text summarization does rouge correlate. *Proceeding of the ACL workshop on intrinsic evaluation measures for machine translation*. pp.1-8. Obtenido el 23 de marzo de 2020 desde: https://www.aclweb.org/anthology/W05-0901.pdf
91. Khan, R. Qian, Y., & Naeem, S. (2019). Extractive based text summarization using K means and TF-IDF. *International Journal of Information Engineering and Electronic Business*. 3(33). pp.33-44. Obtenido el 24 de marzo de 2020 desde: https://www.researchgate.net/publication/333081743_Extractive_based_Text_Summarization_Using_KMeans_and_TF-IDF.
92. Hans, C. Agus. P. A. M., & Suhartono, D. (2016). ComTech: Computer, Mathematics and Engineering Applications. *Computer and technology*.

7(285). pp.285-294. Obtenido el 25 de marzo de 2020 desde: https://www.semanticscholar.org/paper/Single-Document-Automatic-Text-Summarization-using-Christian-Agus/b61e1d017eb3c1b2a00e7d0f1230a07397376aa5

93. Jezek, K., & Steinberger, J. (2008). Automatic Text summarization (The estate of art 2007 and new challenges). *Znalosti*. 30(2), pp.1-12. Obtenido el 26 de marzo de 2020 desde: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.456.58&rep=rep1&type=pdf
94. Golstein, J. Krantrowitz, M. Mittal, V., & Carbonell, J. (1999). Summarizing text documents: sentence selection and evaluation metrics. *SIGIR.* Obtenido el 26 de marzo de 2020 desde: https://www.cs.cmu.edu/~jgc/publication/Summarizing_Text_Documents_Sentence_SIGIR_1999.pdf
95. Pei-Ying, Z., & Cun-he, L. (2009). Automatic text summarization based on sentences clustering and extraction. 8(11), pp.8-11. *2nd IEEE International Conference on Computer Science and Information Technology*. Obtenido el 27 de marzo de 2019 desde: https://ieeexplore.ieee.org/document/5234971
96. Gholamrezazadeh, S. Salehi, A, M., & Gholamzadeh, B. (2009). A Comprehensive Survey on Text Summarization Systems. 2009 *2nd International Conference on Computer Science and its Applications.* Obtenido el 27 de marzo de 2020 desde: https://ieeexplore.ieee.org/document/5404226
97. Babar, S. (2013). Text Summarization: An Overview. Obtenido el 28 de marzo de 2020 desde:https://www.researchgate.net/publication/257947528_Text_SummarizationAn_Overview
98. Nallapati, R. Zhou, B. Dos Santos, C. Gulcehre, C., & Xiang, B. (2016). Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. *The SIGNLL Conference on Computational Natural Language Learning (CoNLL).* Obtenido el 29 de marzo de 2020 desde: https://arxiv.org/pdf/1602.06023.pdf
99. Bharti, D., & Babu, K. (2017). Automatic Keyword Extraction for Text Summarization: A Survey. *National Institute of technology.* Obtenido el 30 de marzo de 2020 desde: https://www.researchgate.net/publication/316028143_Automatic_Keyword_Extraction_for_Text_Summarization_A_Survey
100. Joachims, T. (1996). A probabilistic analysis of the rocchio algorithm with Tf-Idf for text categorization. *Technical Report CMUS-CS-96-118*. Obtenido el 3 de mayo de 2020 desde: https://www.cs.cornell.edu/people/tj/publications/joachims_97a.pdf

101. Allahyari, M., Pouriyeh, S., & Assefi, M. (2017). Text summarization techniques a brief survey. arXiv:1707.02268
102. Abdi, A., Idris, N., Alguliyev, R. M., Alguliyev, R. M. (2016). An automated summarization assessment algorithm for identifying strategies. doi:10.1371/journal.pone.0145809
103. Xiao, J., & Munro, R. (2019). Text Summarization of Product Titles. Proceedings of the SIGIR 2019 eCom workshop. Obtenido el 20 de mayo de 2020 desde: http://ceur-ws.org/Vol-2410/paper36.pdf
104. Rahman, N., & Borah, B. (2018). Improvement of query-based text summarization using word sense disambiguation. Vol (6), pp (75-85). https://doi.org/10.1007/s40747-019-0115-2
105. Khan, A., Salim, N., Farman, H., Khan, M., Jan, B., Ahmad, A., Ahmed, I., & Paul A. (2018). Abstractive Text Summarization based on Improved Semantic Graph Approach. International journal of parallel programming. DOI: 10.1007/s10766-018-0560-3
106. Hans, C., Pramodana, A., M., & Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). Obtenido el 22 de mayo de 2021 desde: https://doi.org/10.21512/comtech.v7i4.3746.
107. Qaiser, S., & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of computer applications. 181(1), pp.25-29. DOI:10.5120/ijca2018917395.
108. Kim, S. W., & Gil, J. M. (2019). Research paper classification systems based on TF-IDF and LDA scheme. Hum. Cent. Comput. Inf. Sci. 9, 30 (2019). https://doi.org/10.1186/s13673-019-0192-7
109. Zhou, J., & Salvendy, G. (2018). Human aspects of IT for the aged population. Applications in Health, assistance and etertaiment. 4th International Conference, ITAP 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part II. Las Vegas, NV, USA. ISBN978-3-319-92036-8

Anexos

Anexo A1

Tabla A1.1.

Resultados con un Threshold de 90% siendo mayor o igual a 31.87.

Cita #12 (nadie tiene el poder pero adems como somos colegas y pares trabajamos muy horizontalmente eso ha generado mucha comunidad mucha relacin entre los profesores entre todas las universidades que forman parte de la red.)

Fuente: Elaboración propia.

Anexo A2

Tabla A2.1.

Resultados con un Threshold de 35% mayor o igual a 12.39.

Cita #45 (hay mucho apoyo mira mucha gente me dice oye cmo le haces para que vengan desde canad a dar una pltica nada mas y tiempo y esto existe espritu de cooperativista sabes que necesitas ayudas o vas a hacer un evento y bueno ok voy implica horas de viaje sacrificar una semana de tu vida por una pltica estn presentes hay cooperacin entonces a lo mejor no si tu no quieres con todos pero en la gran parte si hay ese compromiso .)
Cita #44 (entrevista a amalia rico hernandeztxt 1736 a travs de la red y a travs 453455 super a travs de la red y a travs de todos estos eventos me he involucrado con la gente y lo he vivido de una manera pues mas por decirlo estoy mas presente.)
Cita #42 (conoces a muchas personas a nivel internacional y por ejemplo cuando haces congresos entonces tu ya tienes a quien acudir tu puedes comunicarte con ellos tienes la confianza para mandarles un correo alguna bibliografa.)
Cita #41 (la red permiti a estos investigadores de la universidad trabajar juntos y nunca hubiera sido posible escribir y pensar sobre la tica sin la re.)
Cita #40 *(*es una red de gente tambin mucha gente que ahora se conocen y que venimos aqu a quertaro es porque hay una red de personas eso a nivel humano es un valor que no tiene precio y tenemos el planteamiento para seguir adelante con ms seguridad.)
Cita #39 *(*he tenido un crecimiento personal importante he madurado mucho en muchas cosas mucho ms rpido y mucho ms fcil porque he tenido muchos colegas que generosamente han compartido conocimientos y experiencias yo estoy muy satisfecho porque ha valido la pena y a seguir colaborando y aportando todo lo que podamos aportar.)
Cita #38 (atlas 15 ahora somos redes annimas de gentes con muchas virtudes y con muchos defectos pero como cuerpos orgnicos entonces uno tiene que apostarle a eso el investigador tambin a trabajar por el todo ser parte de un todo y a sacarle provecho de ese todo y aportarle a ese todo lo que uno tenga yo creo que la red unircoop ha creado comunidad cientfica.)
Cita #37 (digo que trabajo en red lleg para quedarse porque hoy en dia no se puede investigar solo ya no hay investigadores como haban en el siglo diecinueve que saban de todo eran grandes cientficos grandes pensadores universalistas abarcaban todos los temas ellos solos se hacan bolas ahora ya no el mundo es ms complejo mucho ms difcil de entender de comprender hay que estar actualizado hay un bombardeo informtico mucha informacin que circula y uno no tiene como procesarlo.)
Cita #35

(hemos creado una comunidad acadmica donde hay muchos puntos de inters nos conocemos hemos creado relaciones de amistad y de colaboracin ms o menos slida ya nos conocemos nuestras virtudes y defectos y eso es una ganancia siempre que se trabaja con personas.)
Cita #34 (sobre todo y para todos el proyecto fue una oportunidad de salir del aislamiento relativo ligado a un tema muy especializado cooperativas y asociaciones para insertarse en equipos de trabajo a nivel nacional y internacional tambin les permiti tener una visibilidad ms grande frente a los organismos nacionales internacionales de prctica cooperativa y de investigación.)
Cita #33 (contribuy a crear grupos de profesores interesados en los mismos temas en compartir sus experiencias pedaggicas para poder eventualmente ofrecer formaciones conjuntas para distintas universidades de la red.)
Cita #31 (siento que me reconocen las otras personas como un ser humano que est metido dentro de las practicas de la solidaridad me siento muy compartido de tener amigos y amigas solidarias .)
Cita #30 (lo que si tengo claro es que la red ha sido fuente de satisfaccin tanto emocional como productiva acadmicamente que me ha permitido posesionar mejor a uni san gil en el medio que he conseguido y he compartido con muchsimas personas de las amricas mi propia experiencia de trabajo en san gil .)
Cita #29 *(*algunas decisiones que requeriran nuestra presencia fsica pero ya sabemos que para negocios para documentos para la investigacin no es estrictamente necesario la presencia fsica sera ms para estrechar los lazos emocionales para poder convivir y compartir y debatir cara a cara algunos asuntos que si interesan debatirlos cara a cara .)
Cita #28 (he podido abrir perspectivas he podido conocer experiencias que me han ayudado mucho a interiorizar mis conceptos de economa solidaria y bueno estoy siempre pendiente de poder cubrir el proceso como estudiante y tal vez dar el paso ya de estudiante a investigadora me gustara muchsimo poder dar ese salto y me parecera importante mas que nada .)
Cita #27 (la red es un espacio en donde puedes compartir puedes vivir y puedes conocer y a mi me resulta importantsimo el tema porque yo como estudiante he aprendido muchsimo de la experiencia del resto de personas como investigadores .)
Cita #24 (tenemos tres aos de haber hecho investigaciones y nstor rodrguez de la javeriana se suma por primera vez a nuestra investigacin lo recibimos como si fuese el primer da y como te recibimos a vos lgicamente uno tiene mas vnculos y charla con uno con otro porque es la ley de la vida pero ac hay un nivel de compaerismo bsicamente en todo el sentido de compaero y eso me lo puedes decir vos o me equivoco ac nadie viene a viene para

compartir lo poco o mucho que tiene para compartir de l yo me espero llevar despus de dar lo mejor de m siempre y espero llevarme mucho de los dems pero no en el sentido interesado si no en el sentido cuando uno viene a dar algo siempre se lleva mucho mas y ac hay cien investigadores no se cuantos hay sesenta investigadores creo que hay sesenta investigadores que cada uno tiene una visin distinta particular y nos da otra visin otra experiencia otros vnculos un montn de cosas .)
Cita #23 (unos lleva un montn de amigos no amigos en el sentido de compaeros de militancia de trabajo .)
Cita #22 (es la satisfaccin ms grande que nosotros nos podemos llevar como docentes como investigadores no el recurso econmico sino la satisfaccin de que contribuimos von un aporte a la economa social a la red .)
Cita #21 (somos de mucho debatir y se logra mucho en consenso sabemos a dnde queremos llegar en la investigacin siempre el objetivo est a largo plazo pensando en donde queremos llegar entonces fruto de eso se arma un montn de dilogo de debate y dems tenemos diferencia tambin con el nivel con el tipo de educacin formal que tiene cada uno bsicamente en un lado uno es profesor de lingstica en el otro economista o contador exclusivamente en el caso de la javeriana en el otro un profesor de literatura o de historia lo que sea otro tcnico cooperativo entonces a veces cada uno socilogos administradores contadores bueno abogados lo que vos quieras eso enriquece mucho a la investigacin pero adems se arma un enriquecedor debate que se necesita subsanarlo con jornadas de intercambio porque cada uno tiene una visin smale a eso el pas smale a eso las realidades legales realidades coyunturales realidades econmicas infinidad de circunstancias .)
Cita #20 (tambin dira que una de las fortalezas de la red sera que existe una interrelacin poderosa en el rea acadmica es pues esta interrelacin ms acadmica que poltica .)
Cita #19 (es agradable cuando tu consigues un espacio donde poder expresarte con toda libertad y sabes que tienes interlocutores que estn interesados en tu propio tema o te prestan una atencin porque humanamente y por delicadeza lo hacen pero al mismo tiempo hay un feedback .)
Cita #18 (tienes espacio donde presentar tus inquietudes tus conclusiones tienes auditorio con quien discutir en forma tus ideas cuestin que muchas veces en tu propio pas en tu propia universidad no tienes pues porque no hay pares que sientan ese en lo tuyo por ejemplo .)
Cita #17 (tengo confianza con la gente nos valoramos sobre elementos positivos los elementos de confianza y amistad y hasta de complicidad como te deca hace rato en el sentido positivo son muy muy buenos .)
Cita #16 (hay una interrelacin bien interesante entre nosotros hay amistad hay confianza nos sentimos a gusto nos picamos el ojo agarramos la sea como se dice hay hasta cierta

complicidad positiva en el sentido de que al uno conocerse con una sea ya es suficiente para saber muchas cosas porque son aos de trabajo como consecuencia de esa interrelacin que la red impuso para nosotros es muy positivo entonces tenemos ya una cultura organizacional una forma de reunirnos una manera de montar la agenda hasta en la bsqueda de locales para el sitio de la reunin estamos siempre en una forma de hacerlo muy agradable y bueno funcionamos y esa misma relacin de confianza de complicidad si se quiere de tener cdigos que nos permitan comunicarnos entender rpidamente las cosas tambin la tuvimos con el comit cientfico son cositas que te dicen vaya t te sientes bien .)
Cita #15 (una red de investigacin como un espacio de encuentro entre acadmicos de distintas universidades o centros que se encuentran para compartir para intercambiar conocimiento metodologas experiencias de manera de enriquecerse mutuamente .)
Cita #14 (no ha sido un trabajo pesado sino ha sido algo gratificante en la medida que se trabaja con confianza hay buen nivel de compromiso .)
Cita#13 (una relacin de confianza de de solidaridad de trabajar muy bien todos y tuvimos la posibilidad la facilidad de juntarse de encontrarse .)
Cita #12 (nadie tiene el poder pero adems como somos colegas y pares trabajamos muy horizontalmente eso ha generado mucha comunidad mucha relacin entre los profesores entre todas las universidades que forman parte de la red .
Cita #11 (efinira una red de investigacin como un grupo de personas interesadas en investigar un tema especfico que con la posibilidad de producir conocimiento pero al mismo tiempo fundamentalmente los componentes de esta red va a ser la discusin la promocin buscando siempre conocimiento de frontera sera pues aprovechar la capacidad de produccin colectiva de un grupo de pares que estn en temas parecidos a partir de un debate serio riguroso acadmico pero al mismo tiempo la red permita consolidar los resultados .)
Cita #10 (unidad siempre nos definimos como una unidad autogestionaria o sea un equipo de trabajo .)
Cita #9 (apertura mental deseo de intercambiar deseo de discutir con sus pares tener un espacio donde debatir el intercambio la discusin el aporte todo eso la vida acadmica uno cuestiona a la comunidad en una cosa pero esto)
Cita #8 (el inters de intercambiar de conocer de debatir de interactuar porque si yo simplemente me ligo a una red buscando plata que es lo que mucha gente hace ah muere la idea de una red .)
Cita #7 (lo que tenemos que hacer es relacin debatir buscar que vengan los profesores que vayan es muy bueno circular que nos fomentemos que vengan para mi comit cientfico es la actividad acadmica .)

Cita #6 (las fortalezas que esos otros tienen entonces fueron dndose una serie de relaciones entre las universidades entre nodos el nodo andino es muy unido otros nodos tambin .)
Cita #5 (con los canadienses hay que reconocer ellos siempre han tenido una visin de horizontalidad en las relaciones siendo realmente pares no el hecho en el caso de colombia estoy aportando el recurso entonces imponerlo se hace bien social en el caso de canad han sido muy horizontales eso hay que resaltar .)
Cita #4 (esto involucra muchos valores muchos principios que hace que haya algo especial cuando esas personas tienen la misma forma de comunidad se renen y comparten la misma visin de la sociedad entonces para mi es algo muy especial porque pienso que el hecho mismo que las personas que creen en el cooperativismo mediante cooperacin hacen que hay muchos intereses en comn manera de pensar de funcionar que se renen de una red .)
Cita #3 (a nivel de las relaciones personales yo pienso que es un tipo de persona especial que est en la red porque son todas personas que se interesan en el cooperativismo entonces son todas personas que tienen una meta objetivo y tienen la misma cosa .)
Cita #2 (hace 5 aos que estn desarrollando conocimiento en comn al inicio no saban trabajar juntos porque trabajar muchos pases es difcil hay veces hay muchos idiomas hay muchos paradigmas que trabajar no es fcil pero poco a poco a travs de la red y de los comits acadmicos.)
Cita #1 (la red me permite a m al menos yo creo que es as por muchos me permite reconocer el ideal que es un ideal que hay que definir siempre .)
Cita #0 (hay especialistas en muchas partes entonces cuando se necesita la idea de uno o la idea de otro bueno pues nos toca viajar nos toca comunicarnos nos toca precisar muchas cosas .)

Fuente: Elaboración propia.

Anexo A3

Tabla A3.1.

Resultados Finales con un valor de Threshold mayor o igual a 7.46 correspondiente al 80%

Número de Cita: 9 compartir recursos o material o actividades de una manera mucho más sencilla y también por la parte siento yo que los alumnos es que yo ahí también les dejo pues ciertas cómo como recursos que ellos pueden utilizar para más adelante y trata siempre de dividirse los por temas o por o por actividades buscando que sea sencillo para ellos entender qué es lo que hace ese material ahí para que lo pueden utilizar entonces creo que eso es lo que me puede beneficiar a mis clases 220 es que de repente que siento que a veces los alumnos no le ven como la no la seriedad que tampoco ex .
Número de Cita: 8 así como las notificaciones de las actividades que había de repente llega la hora se me llega a pasar hay una tarea y nunca me daba cuenta y entonces classroom por otra parte como que si lo tienes instalado qué es lo que yo les recomiendo a mis alumnos pues te muestra las notificaciones de tienes una tarea ahorita y la tienes que entregar para esta hora inclusive desde el lado como del maestro ahí me aparece porque sincronizo con mi calendario me aparece los eventos que tengo y cuando van a esperar entonces cómo puede estar un poco más atenta a ver quienes entregaron quienes todavía no pues empezar a ver si si se les olv si se le está complicando si necesitan más tiempo o cualquier otra cosa entonces el virtual como tal sí así es de tipo pero a mí se me hace muy confuso muy no sé cómo muy complicado de utilizar como que no no puedes entrar y hacer lo que tienes que hacer como que tienes que entrar a tu perfil buscar la materia buscar la actividad y no sé eso no me gusta sí 218 compartir recursos o material o actividades de una manera mucho más sencilla y también por la parte .

Fuente: Elaboración propia.

Anexo A4

Tabla A4.1

Resultados Finales con un valor de 67 % Threshold mayor o igual a 6.24.

Número de Cita: 16 el maestro para precisamente poder ya sea mejorar inclusive más su clase o generar diferentes materiales nuevos ya que al tener todo centralizado también pues hace que sea más ráp tanto calificar enviarles las calificaciones asignar trabajos inclusive darles avisos hace que todos sean pues muy sencillo .
Número de Cita: 15 al hecho de la grabación de clase el poderles compartir la misma grabación el pizarrón antes estaba más orientado a la gestión de la clase dividir actividades temas el poder hacer publicaciones compartirles material 425 el maestro para precisamente poder ya sea mejorar inclusive más su clase o generar diferentes materi .
Número de Cita: 14 un poco el hecho de que las calificaciones son en base 100 y pues no puedo poner calificaciones con punto decimal y luego pues también a la hora de pasar las por ejemplo en excel para quedar mis promedios o así tengo que hacer ese match entre base 100 y base 10 para poner con punto decimal a lo mejor 421 al hecho de la grabación de clase el poderles compartir la misma grabación el pizarrón antes estaba .
Número de Cita: 13 entonces gran parte de esto es para empezar que es específico para clases y segundo esa integración que tiene con otras herramientas de google no si por ejemplo pon forms ahorita acaban de lanzar una integración con blackboard que me permite generar un pizarrón virtual en el cual puedo interactuar con los alumnos explicando a manera de dibujo y pues ellos mismos se pueden quedar con esa herramienta aparte de que por ejemplo hoy en día que son las clases virtuales grabó mi clase y los alumnos en caso de dudas pueden consultar la grabación de la clase 418 un poco el hecho de que las calificaciones son en base 100 y pues no puedo poner calificaciones con .
Número de Cita: 12 utilizarse al menos en este entorno en este contexto en que estamos donde todo es a distancia porque utilizaron claro en lugar de utilizar un grupo de whatsapp o un este grupo de correo electrónico por ejemplo creo que algo de lo que lo que más escuchado de mis estudiantes ahorita es que están muy fastidiados están exhaustos de las clases en línea y como mi experiencia en otras áreas administrativas los grupos de whatsapp no siempre se toman con seriedad 413 entonces gran parte de esto es para empezar que es específico para clases y segundo esa integración .
Número de Cita: 11 si tuviera que señalar una característica específico dices por eso utilizo classroom sería creo que no tiene tal vez tantas opciones como otras plataformas entonces a mi parecer y de lo que he escuchado de mis alumnos también es que es más fácil de utilizar cómo eso mismo que no tiene tantas funciones como otras herramientas todo se puede encontrar un poquito más fácil porque es un poco menos lo que te deja hacer entonces las fases más coherente es más legible y por lo tanto es más amigable o sea es más fácil de utilizar pues

39 utilizarse al menos en este entorno en este contexto en que estamos donde todo es a distancia porque .
Número de Cita: 10 es que de repente que siento que a veces los alumnos no le ven como la no la seriedad que tampoco exijo como que sean completamente serio con esto pero la como la importancia de estos recursos y simplemente pues optan una o por no poner nada o no contestar o no hacer nada de esto o por escribir cualquier tipo de comentario que nada que ver con la materia con la clase con todo lo que tiene que ver ahí 35 si tuviera que señalar una característica específico dices por eso utilizo classroom sería creo que .
Número de Cita: 9 compartir recursos o material o actividades de una manera mucho más sencilla y también por la parte siento yo que los alumnos es que yo ahí también les dejo pues ciertas cómo como recursos que ellos pueden utilizar para más adelante y trata siempre de dividirse los por temas o por o por actividades buscando que sea sencillo para ellos entender qué es lo que hace ese material ahí para que lo pueden utilizar entonces creo que eso es lo que me puede beneficiar a mis clases 220 es que de repente que siento que a veces los alumnos no le ven como la no la seriedad que tampoco ex .
Número de Cita: 8 así como las notificaciones de las actividades que había de repente llega la hora se me llega a pasar hay una tarea y nunca me daba cuenta y entonces classroom por otra parte como que si lo tienes instalado qué es lo que yo les recomiendo a mis alumnos pues te muestra las notificaciones de tienes una tarea ahorita y la tienes que entregar para esta hora inclusive desde el lado como del maestro ahí me aparece porque sincronizo con mi calendario me aparece los eventos que tengo y cuando van a esperar entonces cómo puede estar un poco más atenta a ver quienes entregaron quienes todavía no pues empezar a ver si si se les olv si se le está complicando si necesitan más tiempo o cualquier otra cosa entonces el virtual como tal sí así es de tipo pero a mí se me hace muy confuso muy no sé cómo muy complicado de utilizar como que no no puedes entrar y hacer lo que tienes que hacer como que tienes que entrar a tu perfil buscar la materia buscar la actividad y no sé eso no me gusta sí 218 compartir recursos o material o actividades de una manera mucho más sencilla y también por la parte .
Número de Cita: 7 porque me facilita muchísimo el llevar el control de las actividades y el tener como la evidencia de los trabajos que se realizaban eh simplemente no me imagino con mi correo lleno de tareas y actividades y no era la versión le va a esta otra o sea no me imagino así creo que no me organizaría nisiquiera me gustaría y entonces conclusión pues ya te tienes una materia en esa materia solamente recibes cosas de ese material y puedes dividir como actividades no entonces también en cierta actividad solamente ves cosas actividad 213 así como las notificaciones de las actividades que había de repente llega la hora se me llega a pasa .
Número de Cita: 5 para mí fue mucho más sencillo porque pone mi cuenta de email porque no tuve que hacer mucho para poder administrar mis clases simplemente das de alta una clase compartes un código los alumnos se unen ella más que una clase no y tiene la facilidad de llevar un

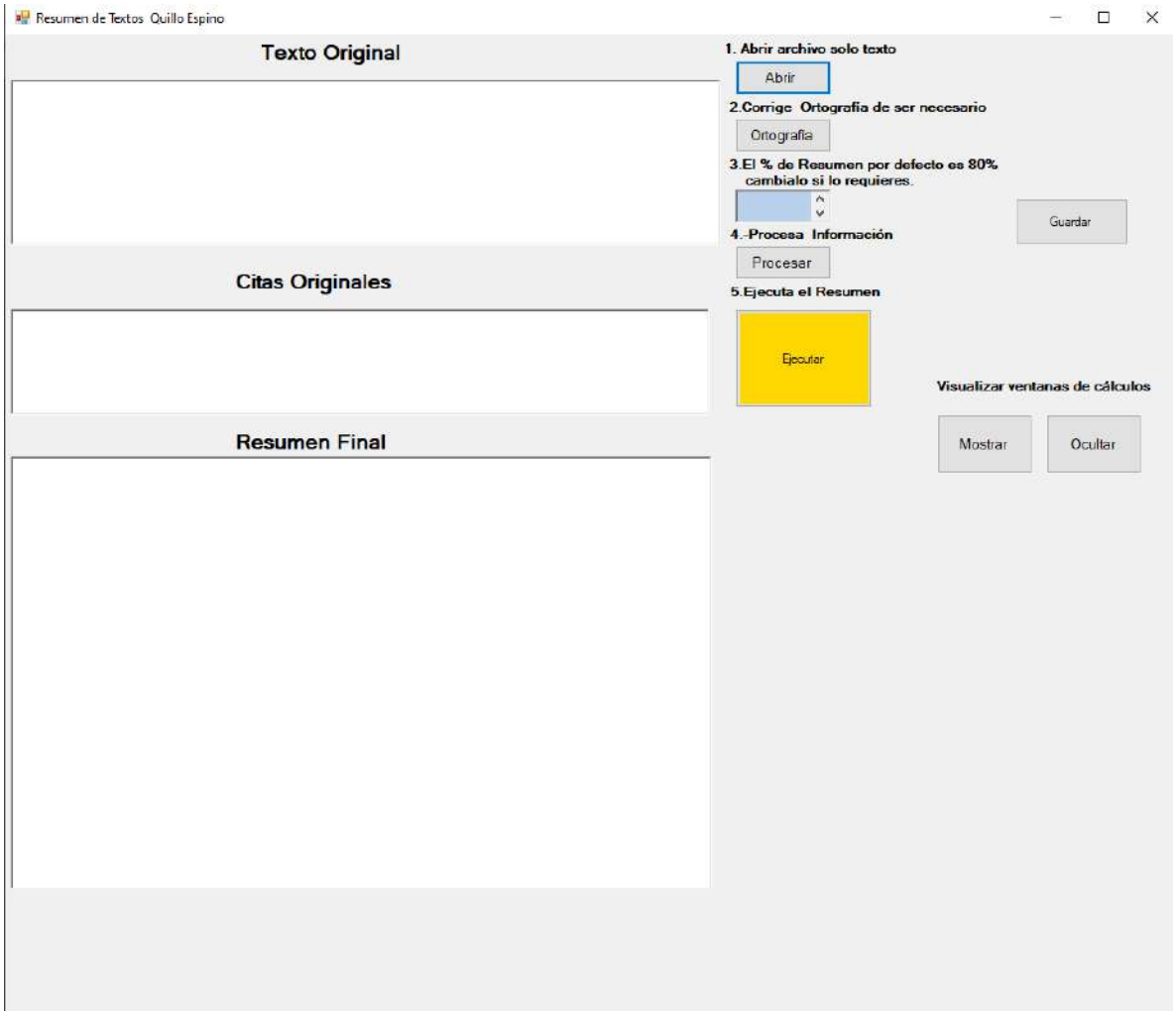
control de fichas actividades anuncios y barato que segu comentando hasta aquí por ejemplo 27 que darle click luego irse a los puntitos luego seleccionar editar y como que siento que es mucha vu .
Número de Cita: 4 que las pueda o sport perdón sin importar de excel o exportarlos y compartirlas con otros este pues eso ayuda bastante no creo que para efectos los documentos pues no tengo ningún problema porque los cargo en el draft de los comparto digo eso es ilógico pero pues me gustaría poder consultar o visualizar todos los tipos de archivos pero bueno es ilógico porque no tenía que ser compatible con algunos simuladores que también utilizó para otras materias 122 para mí fue mucho más sencillo porque pone mi cuenta de email porque no tuve que hacer mucho para po .
Número de Cita: 3 permite utilizar un fue un pizarrón virtual donde también utilizó para explicar algunas otras cosas y darle un poquito el cambio a que no solo sea compartir mi pantalla y que hago algo sino poder dibujarle sí que tengan esa como alternativa similar a lo que sería un pizarrón 121 que las pueda o sport perdón sin importar de excel o exportarlos y compartirlas con otros este pues .
Número de Cita: 2 hay ocasiones donde cuestiones muy específicas tal vez solo para mí si se repite la plataforma y hay otras ocasiones que para los alumnos con base a la experiencia de ese mes anterior puede que exista una combinación de de varias plataformas o sí sí hubo un éxito pues por qué no repetir 117 permite utilizar un fue un pizarrón virtual donde también utilizó para explicar algunas otras cosa .
Número de Cita: 1 sí depende la materia y también pues depende los alumnos 111 hay ocasiones donde cuestiones muy específicas tal vez solo para mí si se repite la plataforma y hay .

Fuente: Elaboración propia.

Anexo A5

La Figura A 4.1 muestra ejemplo de la aplicación Resumen de Textos Quillo Espino.

Figura A4.1. Resumen de Textos Quillo Espino



Fuente: Elaboración propia.

Anexo A6

Productividad académica:

Participación en Congresos:

- Primer Congreso Nacional en Computación y Tecnología Educativa "Extracción de conocimiento a través de texto".
- Segundo Congreso Nacional de Computación y Tecnología Educativa "Reducción de tiempo utilizando corrector ortográfico en la minería de textos".
- Tercer Congreso Nacional de Computación y Tecnología Educativa "Visión de los pre procesos de la minería de textos en C#".
- Memorias en 2do. Congreso Nacional de Computación y Tecnología Educativa "Optimización de los pre procesos de la Minería de Textos".

Publicaciones:

- Quillo-Espino. J., Romero-González, R. M. & Lara-Guevara. A. (2018). Advantages of Using a Spell Checker in Text Mining Pre-Processes. Journal of Computer and Communications. 6(11), pp.43.54. https://doi.org/10.4236/jcc.2018.611004
- Quillo-Espino. J., Romero-González, Paulin-Martinez. F. J. (2019). Text Mining Preprocessing in Times of Python vs MVCS. International Journal of Computer Science and Software Engineering (IJCSSE). 8(11), pp. 266-275.
- Quillo-Espino. J., Romero-González. (2021). Where are the Automatic Text Summaries Located in the 2021? A review. International Journal of Advanced Research in Computer and Communication Engineering. 10(4), pp.11-16. DOI 10.17148/IJARCCE.2021.10402
- Quillo-Espino. J., Romero-González. (2021). Text mining and its Techniques, Applications: An Overview. International Journal of Advanced Research in Computer and Communication Engineering. 10(5), pp. 49-53. DOI 10.17148/IJARCCE.2021.10507
- Quillo-Espino, J., Romero-González, R.M. & Herrera-Navarro, A. M. (2021). A Deep Look into Extractive Text Summarization. Journal of Computer and Communications. 9(6), pp. 24-37. https://doi.org/10.4236/jcc.2021.96002