

PROPUESTA DE UNA TÉCNICA PARA LA CLASIFICACIÓN DE TEXTOS RESPECTO

A UNA TEMÁTICA ESPECIFICA

FABIAN ARTURO RIAÑO SANTIESTEBAN

Autor

EZEQUIEL GLINSKY

Tutor

MAESTRIA EN TECNOLOGÍAS DE LA INFORMACIÓN

Buenos Aires

2013

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tutor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Presidente del jurado

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Jurado

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Jurado Externo

Defensa de Tesis

Buenos Aires, Capital Federal, A los \_\_\_ días del mes \_\_\_\_\_\_\_\_\_\_\_\_\_ del \_\_\_\_\_\_\_\_ Agradezco a todas las personas que me ayudaron en este largo camino de construcción de conocimiento en donde me encontré con dificultades que tuve que hallar la forma de superar y que me hicieron crecer como persona y como profesional.

Gracias a mis padres, a mi familia, amigos, y sobre todo gracias a Dios que sin su ayuda no hubiera sido posible.

"Hay una fuerza motriz más poderosa que el vapor, la electricidad y la energía atómica: la voluntad" Albert Einstein

# RESUMEN

Gracias al desarrollo tecnológico hoy en día se cuenta con amplios volúmenes de información en formato digital. Dicha información es improductiva si no se tienen los métodos necesarios para que sea analizada. Se presenta como una oportunidad explotar las fuentes de información digital disponibles para diferentes usos, ya sea para investigación científica, análisis de opiniones o marketing, entre otros.

En la presente tesis se propone un método de clasificación, para una etapa inicial de selección de textos por temática para Opinion Mining. La técnica consiste en escoger tres textos que se refieran a un mismo tema, los cuales son procesados para obtener las keywords que representan la temática que se trata en estos.

Luego de obtenidas las keywords se realiza una encuesta para validar y asignar un valor en peso o ponderación (relevancia) a cada palabra obtenida. El análisis se hace definiendo un rango de keywords que deben tener los textos analizados y haciendo una agrupación según la distancia definida entre las palabras.

Posterior al análisis, obtención de keywords y definición de rango de distancias se procede a analizar 12 nuevos textos que se refieren a diversos temas, luego mediante el análisis y comparación se seleccionan los textos que coinciden en la temática correspondiente a las keywords y los otros son descartados.

# TABLA DE CONTENIDO

[RESUMEN iii](#_Toc125489)

[TABLA DE CONTENIDO iv](#_Toc125490)

[LISTA DE TABLAS viii](#_Toc125491)

[LISTA DE FIGURAS xi](#_Toc125492)

[LISTA DE ECUACIONES xi](#_Toc125493)

[CAPÍTULO 1 1](#_Toc125494)

[INTRODUCCIÓN 1](#_Toc125495)

[1.1 PLANTEAMIENTO DEL PROBLEMA 3](#_Toc125496)

[1.2 JUSTIFICACIÓN 4](#_Toc125497)

[1.3 OBJETIVOS DE LA TESIS 4](#_Toc125498)

[1.4 ALCANCE 5](#_Toc125499)

[1.5 LIMITACIONES 5](#_Toc125500)

[1.6 METODOLOGÍA Y HERRAMIENTAS 6](#_Toc125501)

[CAPÍTULO 2 9](#_Toc125502)

[ESTUDIOS PRELIMINARES SOBRE DATA MINING 9](#_Toc125503)

[2.1 MARCO TEÓRICO 9](#_Toc125504)

[2.2 ¿QUÉ ES DATA MINING? 10](#_Toc125505)

[2.3 CLASIFICACIÓN DE LAS TÉCNICAS DE DATA MINING 14](#_Toc125506)

[2.4 TÉCNICAS Y APLICACIONES PARA DATA MINING 16](#_Toc125507)

[2.5 OPINION MINING Y REDES SOCIALES 19](#_Toc125508)

[CAPÍTULO 3TECNICA PROPUESTA PARA EL ANALISIS Y CLASIFICACIÓN DE TEXTOS 22](#_Toc125509)

[3.1 ANALISIS Y OBTECIÓN DE KEYWORDS 22](#_Toc125510)

[3.1.1 Frecuencias del primer texto 22](#_Toc125511)

[3.1.2 Frecuencias del segundo texto 23](#_Toc125512)

[3.1.3 Frecuencias del tercer texto 25](#_Toc125513)

[3.1.4 Promedio ponderado de los tres textos y keywords 31](#_Toc125514)

[3.2 ENCUESTA Y VALIDACIÓN DE KEYWORDS OBTENIDAS 36](#_Toc125515)

[3.2.3 Análisis de promedios ponderados y resultados de la encuesta 39](#_Toc125516)

[3.3 DISTANCIA ENTRE TEXTOS SEGÚN KEWORDS 40](#_Toc125517)

[CAPÍTULO 4 47](#_Toc125518)

[IMPLEMENTACIÓN DE LA TECNICA PROPUESTA PARA LA CLASIFICACIÓN DE TEXTOS 47](#_Toc125519)

[4.1 PRUEBA Y ANÁLISIS PARA DETERMINAR SI UN TEXTO HACE REFERENCIA A UNA TEMÁTICA ESPECÍFICA 47](#_Toc125520)

[4.1.1 Prueba con el texto (a), título: Datacenters de Nube Argentina de Telecom, socios estratégicos del resguardo de información de las empresas 48](#_Toc125521)

[4.1.2 Prueba con el texto (b), título: Cloud Computing 51](#_Toc125522)

[4.1.3 Prueba con el texto (c), título: Tendencias de los servicios de Cloud Computing en el mercado Español 53](#_Toc125523)

[4.1.4 Prueba con el texto (d), título: Educación y Computadoras – Enseñanzas de América Latina 55](#_Toc125524)

[4.1.5 Prueba con el texto (e), título: Comida Saludable 57](#_Toc125525)

[4.1.6 Prueba con el texto (f), título: ¿Qué el Cloud Computing y qué beneficios aporta a la empresa? 58](#_Toc125526)

[4.1.7 Prueba con el texto (g), título: Calentamiento Global 60](#_Toc125527)

[4.1.8 Prueba con el texto (h), título: Cuentos Vargas Llosa 62](#_Toc125528)

[4.1.9 Prueba con el texto (i), título: Espantos De Agosto- Garcia Marquez 64](#_Toc125529)

[4.1.10 Prueba con el texto (j), título: Las empresas hablan sobre La Nube 66](#_Toc125530)

[4.1.11 Prueba con el texto (k), título: El Tiempo En Campana 68](#_Toc125531)

[4.1.12 Prueba con el texto (L), título: Formacion de las Nubes 70](#_Toc125532)

[4.2 ANALISIS DE RESULTADOS 72](#_Toc125533)

[CAPÍTULO 5 75](#_Toc125534)

[CONCLUSIONES Y TRABAJOS FUTUROS 75](#_Toc125535)

[5.1 CONCLUSIONES 75](#_Toc125536)

[5.2 TRABAJOS FUTUROS 77](#_Toc125537)

[REFERENCIAS 79](#_Toc125538)

[ANEXOS 81](#_Toc125539)

22

# LISTA DE TABLAS

*Tabla 1. . Frecuencias del texto número 1, Resultado obtenido con el contador de palabras*

*desarrollado en java. .............................................................................................................................................................. 23*

*Tabla 2. Frecuencias del texto número 2, Resultado obtenido con el contador de palabras desarrollado en java. .......................................................................................................................................................................................... 25*

*Tabla 3. Frecuencias del texto número 3, Resultado obtenido con el contador de palabras desarrollado en java. .......................................................................................................................................................................................... 29*

*Tabla 4. Promedios ponderados de los tres textos y análisis de Pareto tomando únicamente el 81% del*

*total de palabras de los tres textos. ................................................................................................................................... 32*

*Tabla 5. En esta tabla se pueden ver las palabras de la encuesta con sus respectivos promedios y están organizadas de mayor a menor según el nivel de importancia que los encuestados les dieron. .............. 36*

*Tabla 6. Resumen de los datos relevante de la encuesta, palabra con mayor promedio, promedio de los puntajes, Media aritmética y medina................................................................................................................................ 36*

*Tabla 7. Keywords con sus respectivos pesos obtenidos de la encuesta. Se observa el rango en donde se debe encontrar la suma de keywords halladas en los textos que son analizados (PT y K). ................. 40*

*Tabla 8. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del texto (a). Las palabras sombreadas son las keywords de la temática de Cloud Computing que se encontraron en este texto. .. 47*

*Tabla 9. Keywords identificadas en el texto (a) con sus respetivos pesos y el peso total w. ...................... 47*

*Tabla 10. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del texto (b). Las palabras sombreadas son las keywords de Cloud Computing que se encontraron en esté texto. ............................... 49*

*Tabla 11. Palabras identificadas en el texto (b) con sus respectivos pesos y peso total w. ........................ 49*

*Tabla 12. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del texto (c). Las palabras*

*sombreadas son las keywords que se encontraron en este texto. .......................................................................... 51*

*Tabla 13. Tabla 20. Palabras identificadas en el texto (c) con sus respectivos pesos y peso total w. .... 51*

*Tabla 14. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del texto (d). Las palabras*

*sombreadas son las keywords que se encontraron en este texto. .......................................................................... 53*

*Tabla 15. Palabras identificadas en el texto (d) con sus respectivos pesos y peso total W. ....................... 53*

*Tabla 16. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del texto (e). ....................... 55*

*Tabla 17. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del Texto (f) de prueba.*

*Las palabras sombreadas son las keywords que se encontraron en este texto. ............................................... 57*

*Tabla 18. Palabras identificadas en el texto (f) con sus respectivos pesos y peso total W. ........................ 57*

*Tabla 19. Primeras 15 palabras y sus ocurrencias en orden mayor a menor del texto (f). ........................ 59 Tabla 20. Primeras 15 palabras y sus ocurrencias en orden mayor a menor del texto (h). ....................... 60*

*Tabla 21. Primeras 15 palabras y sus ocurrencias en orden mayor a menor del texto (i). ........................ 62*

*Tabla 22. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del Texto (j). Las palabras*

*sombreadas son las keywords que se encontraron en esté texto. .......................................................................... 64*

*Tabla 23. Palabras identificadas en el texto (j) con sus respectivos pesos y peso total w. ......................... 64*

*Tabla 24. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del Texto (k). La palabra*

*que aparece sombreada es una keyword que se encontró en esté texto l w. ..................................................... 66*

*Tabla 25. Palabras identificadas en el texto (k) con sus respectivos pesos y peso total w. ........................ 66*

*Tabla 26. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del Texto (L). La palabra*

*que aparece sombreada es una keyword que se encontró en esté texto. ............................................................ 68*

*Tabla 27. Palabras identificadas en el texto (L) con sus respectivos pesos y peso total w. ....................... 68*

*Tabla 28. Después del análisis se obtuvo que los textos que se muestran sombreados (a, b, c,f y j)*

*hacen referencia a temática de Cloud Computing). ................................................................................................... 69*

# LISTA DE FIGURAS

*Figura 1. Técnicas de Data Mining y un ejemplo de posible resultado. ............................................................ 14*

*Figura 2. Flujo de datos y técnicas aplicadas sobre una base de datos Oracle 11g con tecnología*

*OLAP para Data Mining. ...................................................................................................................................................... 15*

*Figura 3. Proceso estadístico que se realizó sobre los tres textos que tratan el tema de Cloud*

*Computing, luego se obtuvieron los promedios ponderados y finalmente se aplicó la técnica de Pareto*

*para descartar palabras que hacían ruido al análisis. .............................................................................................. 29*

*Figura 4. Sobre la tabla de promedios ponderados se hace el proceso de Pareto para establecer las*

*palabras más relevantes. ....................................................................................................................................................... 33*

*Figura 5. Primeras 30 palabras del promedio ponderado de todas las palabras contenidas en los tres*

*textos analizados. ...................................................................................................................................................................... 33*

*Figura 6. Las 10 Keywords referentes a la temática de Cloud Computing obtenidas después del*

*análisis de Pareto. .................................................................................................................................................................... 37*

*Figura 7. Con las Keywords obtenidas sobre la temática de Cloud Computing que se pueden analizar nuevos textos x, y, z, etc de temas desconocidos y posteriormente inferir si también pertenecen a esta*

*misma temática. ......................................................................................................................................................................... 39*

*Figura 8. Triángulo formado por las distancias: d(W\_1,T),d(W\_2,T),d(W\_1,W\_2) ..................................... 44*

*Figura 9. Proceso de la técnica propuesta para la clasificación de textos. ...................................................... 46*

*Figura 10. Proceso de prueba de distancia con 12 textos, de los cuales se obtuvo que 5 hacen*

*referencia la temática de Cloud Computing. ................................................................................................................. 70*

*Figura 11. Proceso de prueba de distancia con 12 textos, de los cuales se determinó que 5 pertenece a*

*la temática de Cloud Computing. ....................................................................................................................................... 71*

# LISTA DE ECUACIONES

*Ecuación 1 Promedios ponderados ................................................................................................................................... 30*

*Ecuación 2. Distancia euclidea .......................................................................................................................................... 38*

*Ecuación 3. Distancia propuesta para determinar cercanía de un texto respecto a la temática de Cloud*

*Computing. .................................................................................................................................................................................. 41*

*Ecuación 4. Rango distancia que determina si un texto hace referencia a la temática de Cloud*

*Computing o no. ........................................................................................................................................................................ 42*

*Ecuación 5. Muestra la no negatividad de la distancia propuesta. ................................................................... 43*

*Ecuación 6. Simetría de la distancia propuesta. .......................................................................................................... 43*

*Ecuación 7. Desigualdad triangular de la distancia propuesta. ........................................................................... 44*

# CAPÍTULO 1

# INTRODUCCIÓN

Con los volúmenes de información que se manejan hoy en día y que están disponibles ya sea de forma privada para una empresa o pública a través de la web, se puede pensar en diferentes estrategias de análisis que permitan beneficiar la ciencia y otras áreas de conocimiento. Un ejemplo de esto son las grandes bases de datos denominadas Big Data, bases de datos del orden los Peta Bytes (Google, Facebook,

Twitter , Amazon, entre otras).

Se presenta como una oportunidad explotar las fuentes de información que se encuentran en bases de datos normalizadas y des normalizadas[[1]](#footnote-1). Se debe hacer un análisis de la información contenida en dichas bases de datos, logrando encontrar tendencias de consumo respecto a algún producto o servicio, así como también es posible analizar la popularidad de un candidato a la presidencia en un país.

En la web encontramos fuentes de información de todo tipo, por lo cual se hace indispensable una buena clasificación de documentos para su posterior análisis.

En la presente tesis se propone una técnica con una distancia propia para comparar textos con palabras claves. Esta técnica puede ser utilizada para una etapa inicial de Opinion Mining[[2]](#footnote-2) en la que se clasifiquen textos de interés.

Para hacer Opinion Mining se identifica inicialmente la pertinencia del análisis de algún texto, pues se debe tener certeza de que los textos se refieran a un mismo tema, dicha etapa de clasificación puede resultar engorrosa si se quiere clasificar una alto volumen de textos.

La técnica propuesta consiste en tomar de la web tres textos que tratan un mismo tema de interés, estos son analizados para obtener las palabras de mayor peso o keywords, las cuales representan conjunto de textos analizados, por lo cual son usadas para procesar otros textos de los cuales se desconozca el tema que tratan.

Con la ayuda de las keywords y la técnica planteada en esta tesis se logra identificar textos que hacen referencia a una misma temática, esto es posible haciendo una agrupación según la distancia definida entre las palabras. Esta propuesta puede ser usada para investigaciones posteriores en las que se continúe con el análisis de

Opinion Mining de los textos ya clasificados por temática de interés.

## 1.1 PLANTEAMIENTO DEL PROBLEMA

Los amplios volúmenes de información digitalizada que contienen registros de diversa índole, han sido almacenados a través del tiempo de manera manual por una persona o de forma automática por un sistema[[3]](#footnote-3). Una falencia que se tiene es el desaprovechamiento de esta información, pues al no ser analizada, solo queda un registro de datos a los cuales no se les está sacando utilidad. En su libro sobre generación de conocimiento, Davenport y Prusack resaltan que: "El conocimiento no son los datos ni la información, aunque esté relacionado con ambos y las diferencias entre aquellos términos sea frecuentemente una cuestión de grado” [1]. Un estudio de EMC Corporation revela que solo un 0,5% de la información almacenada en el mundo está siendo analizada [2].

En el caso de los sitios de internet a los que se tiene acceso de forma libre y de los cuales se desea analizar la información que allí se encuentra, se tiene la dificultad de que no siempre están bien indexados respecto a la temática, género, o autor, entre otros. Este es uno de los principales problemas a los que se enfrentan quienes desean realizar Opinión Mining puesto que primero se debe tener certeza de que los documentos o textos con los que se va a trabajar se refieran a una misma temática.

## 1.2 JUSTIFICACIÓN

Para hacer un análisis de opinión y ver tendencias de consumo en el mercado, opiniones sobre libros, películas, favoritismo sobre algún candidato político, o cualquier otra aplicación que se quiera hacer con Opinion Mining, se debe llevar a cabo como paso inicial una selección de los textos que se van a analizar, pues si se desea conocer la opinión alrededor de un mismo tópico es necesario asegurarse de que la información que esta descrita en los textos trate de una misma temática. Dichos textos pueden ser datos digitales que sean tomados de una página de internet o directamente de una base de datos.

Es pertinente proponer una técnica de análisis mediante la cual se pueda clasificar textos por temáticas, para que en trabajos futuros sea posible diseñar un sistema autómata en el cual la información sea tomada directamente de la web y se pueda hacer un proceso de clasificación y selección de textos de interés. De esta manera sería posible analizar grandes volúmenes de información que se refieran a un mismo tema.

## 1.3 OBJETIVOS DE LA TESIS

#### OBJETIVO GENERAL

Proponer una técnica con una distancia propia entre textos, que funcione para una etapa inicial de Opinion Mining, con la que se determine si un texto hace referencia a un tema específico o no.

#### OBJETIVOS ESPECÍFICOS

1. Realizar una encuesta y un análisis estadístico para determinar las keywords que van a ser utilizadas en el análisis de nuevos textos de los cuales se desconoce su temática.
2. Realizar pruebas con 12 textos diferentes para determinar si la técnica propuesta es eficiente para la clasificación de textos en una etapa inicial para opinion-mining.

## 1.4 ALCANCE

Se presenta una nueva forma de identificar textos referentes a una misma temática mediante el análisis con una distancia propia. Inicialmente se estudian tres textos que referentes a un mismo tema conocido. El estudio se hace con un análisis estadístico de los textos y con una encuesta que permite conocer la percepción de las personas respecto a las temáticas de los textos.

Con base en los resultados del análisis se determinan keywords de la temática tratada en los textos, luego se propone una distancia para poder analizar nuevos textos de temas desconocidos y determinar si hacen referencia o no al mismo tema que los de las keywords. Finalmente se hace evaluación de la técnica propuesta, presentando los resultados y las conclusiones. Se abre un camino de investigación para Opinion

Mining, el cual se puede seguir depurando en posteriores trabajos.

## 1.5 LIMITACIONES

Durante el desarrollo del estudio se presentan las siguientes limitaciones:

* Los textos escogidos de la web no son descargados automáticamente con ningún tipo de programa. Este proceso se hace manualmente y el contenido textual de la URL es copiado y pegado en un archivo de extensión *texto.txt* para poder ser analizados.
* Se hace una etapa de conteo de palabras con un programa desarrollado en Java y el resultado es exportado a un archivo de Excel.
* El análisis estadístico mediante promedios ponderados de la frecuencia de las palabras y la comparación con los resultados se hace con Microsoft Excel

2010.

## 1.6 METODOLOGÍA Y HERRAMIENTAS

El estudio se realiza con una metodología de investigación definida en 7 etapas que se explican a continuación.

Etapa 1: Se seleccionaron tres textos alojados en URL distintas que se refieren a un mismo tema. Para el caso específico mostrado en el presente documento se escogió el tema de Cloud Computing y los textos fueron tomados de la web de forma no automática. Dichos textos son extraídos limpios manualmente sin ningún tipo de información publicitaria que pueda hacer ruido en el análisis.

Etapa 2: Se toman los tres textos limpios en un formato .txt (EjemploTexto1.txt) y son usados como datos de entrada a un programa de java que hace un conteo de frecuencias de las palabras y el resultado se exporta a un Excel.

Etapa 3: Se hace un análisis de las frecuencias mediante un promedio ponderado para determinar las keywords que representan los tres textos en Microsoft

Excel 2010.

Etapa 4: Se realiza una encuesta para conocer la percepción de personas respecto al contenido temático de los tres textos. Para este caso la muestra seleccionada para responder la encuesta son personas que tienen conocimientos de tecnología y manejan el tema desde una posición técnica, el contenido de los textos que para este caso en particular es Cloud Computing.

Etapa 5: En esta etapa se hace una comparación de los resultados obtenidos en la encuesta y se compara con los que se obtuvieron con el análisis estadístico, determinando de esta forma las keywords y sus pesos (relevancia) correspondientes a esta temática específica (Cloud Computing). Se propone una distancia para analizar temáticas de cualquier texto desconocido respecto a las Keywords determinadas

previamente.

Etapa 6: En esta etapa se evalúa la técnica propuesta con 12 textos, cinco de estos tratan el mismo tema de los representados por las keywords (Cloud Computing) y los restantes se refieren a temas diferentes.

Etapa 7: En esta última etapa se presentan las conclusiones obtenidas luego de las pruebas y los resultados obtenidos. Se proponen trabajos futuros para dar continuidad a la investigación.

A continuación se puede observar la lista de herramientas que fueron

utilizadas para el análisis:

* Laptop: ACER Aspire serie 4738
  + CPU: Intel Core i5
  + Sistema Operativo: Microsoft Windows 7
* Programas utilizados:
  + Microsoft Excel 2010
  + IDE Eclipse Estándar para Java
  + Google Docs

# CAPÍTULO 2

# ESTUDIOS PRELIMINARES SOBRE DATA MINING

## 2.1 MARCO TEÓRICO

La minería de datos es un área interdisciplinaria que está relacionada con la gestión de bases de datos y la manera como dichos datos son tratados para lograr la optimización de análisis de la información.

En un comienzo grandes conjuntos de datos se fueron agrupando y comenzaron a crecer los volúmenes de información, es ahí como surgen los sistemas de bases de datos relacionales con gestor de base de datos que permiten llevar un control y hacer diferentes tipos de acciones sobre los datos: select, update, insert y delete. De la misma forma se pueden hacer join entre tablas y definir alguna estructura relacional entre las tablas de acuerdo a sus claves primarias. El trabajo de base de datos relacional comenzó a ser desarrollado por Codd, E.F y siguió creciendo en la década de 1970 [3,4], permitiendo analizar información, ayudando en estrategias empresariales y desarrollo de diferentes campos del conocimiento, al permitir una mejor gestión de la información.

Internet como red global de información y con su crecimiento a pasos agigantados en los últimos años gracias a los avances e infraestructuras de alto rendimiento, estables y robustas que se pueden utilizar hoy en día, hacen posibles que el almacenamiento sea mayor. A diario se crean registros de los comentarios hechos en foros o en diferentes sitios en la web. La dificultad de analizar grandes volúmenes de información que no siempre se generan o son almacenados en bases de datos relacionales, abre paso a la necesidad de desarrollo e investigación de nuevas herramientas de análisis de información. Entre las cuales se encuentra Data Mining[[4]](#footnote-4).

## 2.2 ¿QUÉ ES DATA MINING?

En los últimos años aparece una tendencia al uso e implementación de técnicas de aprendizaje automático de sistemas y la implementación de diferentes técnicas que permiten analizar la información de la Web5

Existe información que se encuentra implícita dentro los textos que están almacenados en las bases de datos con las que contamos hoy en día, no sólo las de tipo relacional sino también todas las nuevas fuentes de información que están en la red como son; foros de opinión, redes sociales,Twitter, Facebook, Instagram, Google, Yahoo, entre otras. Es pertinente hacer un análisis de opinión aplicado sobre fuentes de información de este estilo para extraer la información y transformarla en conocimiento, determinando patrones y tendencias.

Con Data Mining se puede pensar en aplicaciones que brinden soluciones como las que se mencionan a continuación:

* Saber si una película va a ser exitosa en cines según comentarios sobre un tráiler en un foro de opinión.

* Se puede determinar el posible ganador de las elecciones presidenciales sabiendo la opinión de un gran número de personas.

* Una empresa puede hacer un análisis de mercado y saber qué opinión tienen sus clientes sobre los productos o servicios que ofrece y con base en esto pensar tomar decisiones sobre la forma de innovar en el mercado.

Para la implementación del análisis con Data Mining a nivel empresarial se debe tener comprensión del negocio y saber el problema que se quiere resolver o el proceso que se quiere optimizar. Por tal razón es de vital importancia vincular a los stakeholders que tengan participación activa en un proceso de para tomar decisiones concretas, Data Mining [5].

Las posibilidades de implementación en la industria son muchas y se pueden escoger las técnicas más convenientes para un determinado proceso inclusive se pueden llegar a combinar. El hecho de conocer los diferentes circuitos funcionales de un negocio, permite evaluar y definir puntos en los cuales se puede implementar el uso de Data mining para la mejora de procesos y toma de decisiones más acertadas.

Cuando se cuenta con una base de datos existe información que se encuentra implícita en los registros almacenados. Dicha información puede ser valiosa para el negocio y se puede obtener teniendo en cuenta las siguientes técnicas de Data Mining [6]:

1. Multilevel data generalization (Datos multinivel generalización, síntesis y caracterización): Estas técnicas tienen como objetivo analizar los datos almacenados en una base de datos dándole una vista de alto nivel mediante la generalización, síntesis y caracterización. Estas técnicas son usadas sobre todo en Bases de datos

relacionales.

1. Reglas de La asociación Minera: Estas reglas están basadas en encontrar conocimiento mediante el uso de técnicas de asociación entre ítems dentro de un conjunto de transacciones, un ejemplo es analizar el comportamiento de los clientes respecto a las facturaciones de determinado producto.

1. Clasificación de Datos: Como su nombre lo indica está técnica se usa para hacer clasificaciones de un grupo de datos. Esto se logra hacer con el uso de técnicas como: árbol de decisiones, casos basados en razonamiento y redes neuronales. En estas técnicas unos atributos son vistos como un atributo target de clasificación (salida) y los otros como patrón de los datos.

1. Agrupación o Clustering: Estas técnicas están basadas en patrones de similitudes en las que registros con patrones similares se agrupan en un clúster. Para este tipo de técnicas se usan herramientas estadísticas para determinar similitudes y agrupar por patrones, K-means es un ejemplo.

1. Patrón de búsqueda basado en similitudes: Esta técnica puede ser dividida en dos tipos: búsqueda de texto y similitud de series de tiempo por similitud de búsqueda.

Los tipos de datos más adecuados para estas técnicas son: los datos textuales para buscar texto de similitud y formularios de datos temporales.

1. Descubrimiento y análisis de series temporales o tendencias: Esta técnica está basada en el cambio de los datos en diferentes instantes de tiempo, la tendencia de series cronológicas pueden encontrarse y utilizarse para predecir la posibilidad de un cambio a futuro.

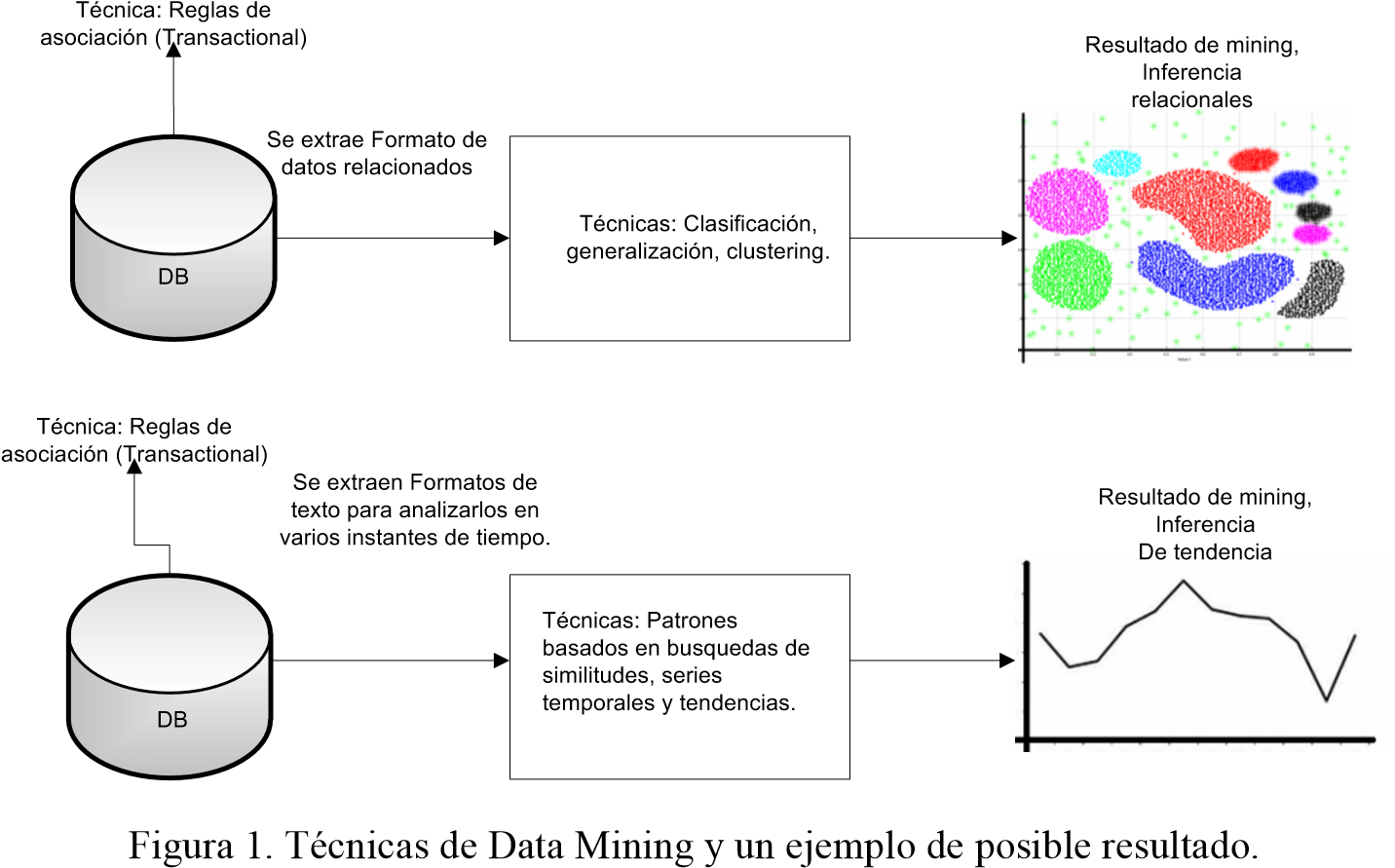
Con la implementación de una técnica de Data Mining a nivel empresarial se podría procesar información sobre las personas que compran algún producto o solicitan un servicio frecuentemente y de esta manera poder direccionar nuevas ofertas de productos con un potencial grupo de clientes.

Desde otro punto de vista pensando en aplicaciones del tipo web 2.0 se pueden hacer sondeos de opinión mediante el uso de las redes sociales y analizar la respuesta de visitantes a un sitio web y la opinión (Opinion Mining) sobre un tema en específico. De esta manera una empresa puede tomar decisiones y hacer mejoras de servicios o productos entendiendo y satisfaciendo al consumidor o usuario.

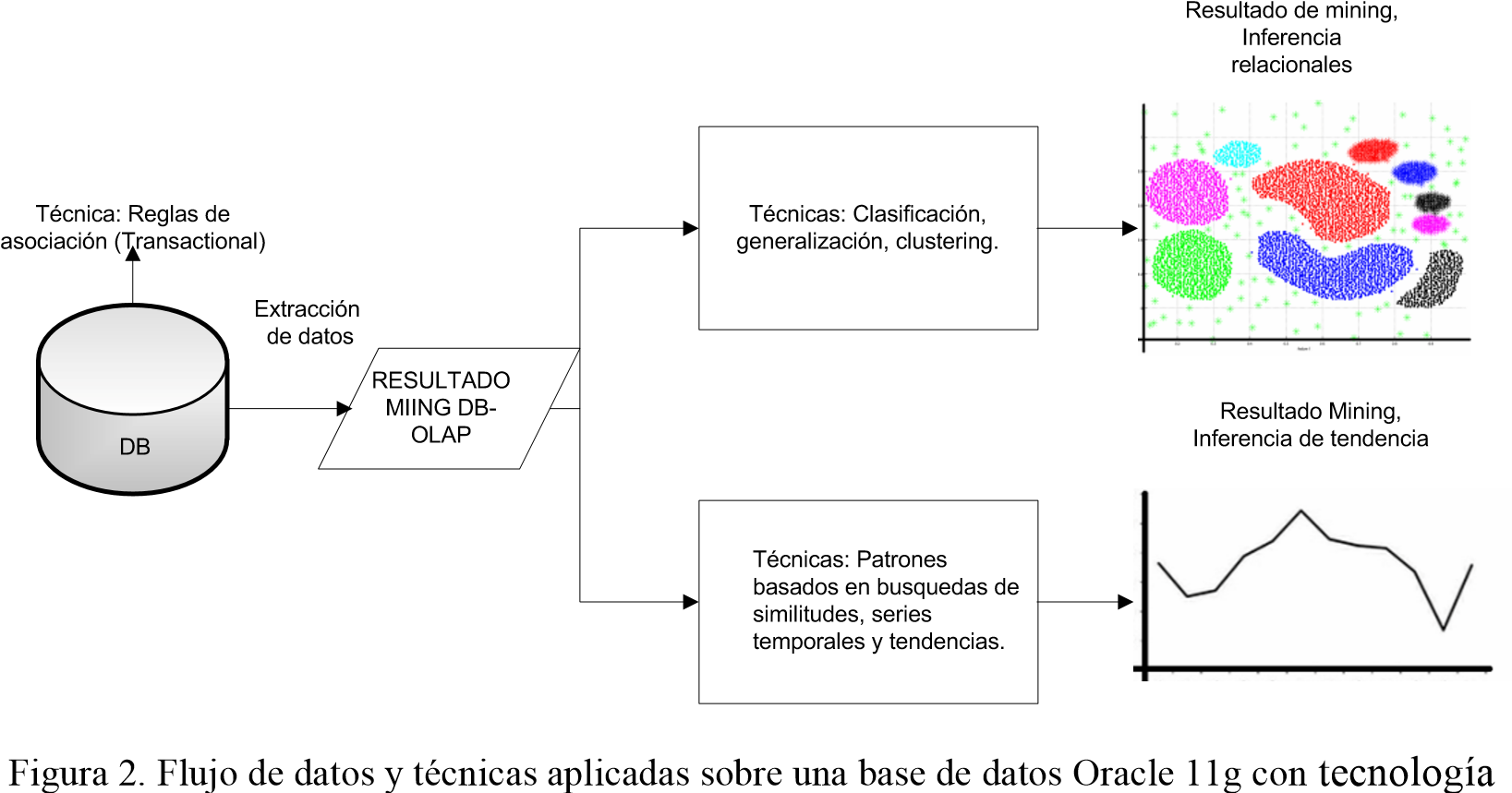
En términos generales Data Mining abre paso a la exploración de nuevas lecturas e interpretaciones de la información que se tiene almacenada digitalmente, determinando similitudes, tendencias, opiniones y así entendiendo como se interrelacionan unas variables con otras, extrayendo información útil para diferentes campos de interés.

## 2.3 CLASIFICACIÓN DE LAS TÉCNICAS DE DATA MINING

Para el desarrollo de aplicaciones para Data Mining se deben tener en cuenta técnicas que permitan la inferencia estadística o heurística, logrando extraer conclusiones de los datos en base a una muestra tomada. Un query que se hace sobre una base de datos no se puede considerar una técnica de Data Mining si no se le aplican técnicas estadísticas o de minería con algoritmos matemáticos que permitan obtener inferencias, en la figura 1 se ilustra el proceso de Minería de datos y sus posibles resultados.



Por otro lado la tecnología OLAP de Oracle DB 11g y posteriores son un ejemplo de una DB que incorpora lógica de Data Mining como señala su documentación: “Oracle Database brinda un enfoque completamente diferente. Oracle ha implementado OLAP, Data Mining y estadísticas dentro de su motor de base de datos. En lugar de mover los datos de un data warehouse a otros motores analíticos para un mayor análisis, Oracle incorporó los algoritmos avanzados de análisis en su base de datos, donde los datos residen. Asimismo, Oracle brinda acceso SQL a todo su análisis, de manera que puedan implementarse con cualquier herramienta basada en SQL o cualquier entorno de aplicación.”[7,8], esta tecnología tiene técnicas de aprendizaje automático para analizar datos y descubrir patrones. Si se obtiene un resultado con procesamiento de minería en la base de datos se obtiene un grupo de datos más refinados a los cuales luego se le pueden seguir aplicando técnicas Data Mining. En la figura 2 se ilustra el proceso de minería de datos con una etapa inicial dentro de la misma base de datos.



No necesariamente se debe contar con una tecnología del tipo OLAP [7,8], se puede hacer una extracción de la información de una DB común y luego de esto aplicarle alguna técnica de Data Mining.

Tanto en la figura 1 como en la figura 2 se puede apreciar que las técnicas se pueden dividir en dos grupos (Inferencia Relacionales e Inferencia de tendencias) según si genera una inferencia de relaciones que permite hacer un análisis de la relación que puede tener un determinado archivo con otro, determinando de esta manera su cercanía según patrones en común y agrupándolos mediante una distancia como se hace por ejemplo con la técnica de clustering [9,10].

Por otro lado el grupo de técnicas está dedicado a encontrar patrones mediante la búsqueda de similitudes y permite predecir posibles comportamientos con el uso de métodos estadísticos analizando resultados en el tiempo. En la siguiente sección del presente documento se describirán algunas aplicaciones de las diferentes técnicas de

Data Mining.

## 2.4 TÉCNICAS Y APLICACIONES PARA DATA MINING

En esta sección se presentan algunos ejemplos de implementación de las diferentes técnicas de Data Mining y en unos casos se puede observar la combinación de algunas de estas.

• Reglas de Asociación:

Una de las técnicas de Data Mining que se comenzó a usar en warehouse es la de Reglas de asociación. Ad-Hoc Association-Rule Mining dentro de Data Warehouses es un Framework que incorpora el Data Mining en la tecnología del propio warehouse, teniendo de esta manera una reducción de posibles issues de privacidad y permitiendo en un query una alta capacidad de procesamiento de datos y permitiendo mediante las reglas de asociación el descubrimiento de patrones que se encuentran implícitos en las bases de datos.

Un análisis de Data Mining en una base de datos es denominado KDD ("Knowledge Discovery in Databases"), Se basa en el descubrimiento de

conocimiento en las bases de datos, se hace principalmente con reglas de asociación y estableciendo patrones en grandes volúmenes de información. Este modelo basado en división simple y compleja de datos, sirve de marco de trabajo de Data Mining [11], se clasifican los datos en tres grupos asociándolos dependiendo el análisis que se desee hacer sobre los atributos. Se obtienen tres grupos; uno con algoritmos de lógica espacial de la DB completa, lógica espacial sobre una parte de la DB y uno de datos generalizados, los cuales son pasados por etapas separadas y al final del proceso de minería por cada etapa se obtiene un modelo de conocimiento.

Las reglas de asociación también han sido útiles para plantear nuevos formas de implementar el Data Mining en lenguajes como el DMQL (Data Mining Query Lenguage) que permite un análisis común de minería y a partir de este surgen propuestas de lenguajes como el TDML, (Transaction Data Mining Lenguage) el cual propone una minería con reglas de asociación y patrones secuenciales, permitiendo almacenar en buffers los resultados, logrando hacer minería online, reduciendo las transacciones con una minería distribuida, logrando incrementar el procesamiento de los datos en menor tiempo [12].

* Reglas de Clasificación y asociación y clustering.

Se pueden encontrar implementaciones donde se combinan diferentes técnicas de Data Mining, un ejemplo de esto es el uso en una Data Mining Distribuida (DDM) con un sistema multiagente como presenta el autor Khan [13], CAKE (Classifying, Associating & Knowledge DiscovEry). Esta implementación está inmersa en una arquitectura llamada Agentes en Paralelo de Data Mining (PADMAs). Esta es una arquitectura distribuida que permite la orquestación de diferentes módulos y aprovechando esto, se usan las técnicas de clasificación, asociación y Clustering para lograr un procesamiento óptimo de los datos en un Data warehouse.

Clustering using k-Neighbor y reglas de inducción usando árboles de decisión. Troubleshooting Distributed [14]. Es un sistema que propone mezclar técnicas como clustering usando K-Neighbor y de árboles de inducción como una buena combinación para el aprovechamiento de sistemas en paralelo en ambientes distribuidos permitiendo una ejecución con mejor performance, y adecuado análisis de los datos. Por un lado el clustering permite el almacenamiento más conveniente de la información para búsquedas y ejecuciones. Por otro lado las reglas que se implementan en el árbol de decisiones permiten estimar el gasto de recursos, de memoria, dispositivos de almacenamiento y procesadores que requiere determinado trabajo. Luego con la información obtenida, se determina algún conjunto de clusters de servidores de servidores o alguno en específico está en la capacidad de hacer un trabajo Determinado.

* Redes neuronales

ANN (Artificial Neural Network) [15] es una aplicación de Data Mining que se basa en la implementación de sistema en paralelo con la utilización de redes neuronales, con lo cual en sistemas paralelizados permite mejorar la robustez, adaptabilidad, procesamiento y almacenamiento en paralelo ofreciendo un alto grado de tolerancia.

## 2.5 OPINION MINING Y REDES SOCIALES

Los avances tecnológicos han permitido el almacenamiento de grandes volúmenes de información a lo que se le ha dado el nombre de Big data. Esto ha cambiado la forma mediante la cual se han venido haciendo análisis de Data Mining y de estadística en general.

Es una oportunidad interesante poder extraer información valiosa de repositorios digitales de forma automática y entender la opinión que se tiene sobre algo específico. Un ejemplo interesante de esto es la solución planteada en el artículo; [16] Análisis automático de emociones en la Red Internacional e-Culturas (2011), en el que se describen las actividades que se hicieron para la creación del corpus sobre emociones de la edición 2009 de la Red Internacional e-Culturas, en la que participaron alumnos de 10 y 11 años de diferentes países. Los autores hacen un repaso de la clasificación de emociones. Posteriormente se prevé integrar tal categorizador dentro de la plataforma e-Culturas, donde el análisis de sentimientos puede resultar de gran ayuda al poder detectar automáticamente una posible reacción negativa del alumno frente a determinadas situaciones.

Twitter proporciona un caudal de opinión totalmente espontáneo y al alcance del que quiera analizarlo, a diario unos 500 millones de usuarios comparten sus opiniones en esta aplicación. Estudios han permitido conocer por ejemplo la opinión de las personas referente a las grandes compañías de petróleo, lo cual brinda información útil y aplicable en la toma de decisiones de los altos mandos de dichas compañías para mejorar la imagen que tienen ante la comunidad y mejorar la relación con los stakeholders [17]. Uno de los aspectos interesantes de almacenar información como la almacena Twitter es que el volumen de su sondeo se hace de forma contínua, pero por otro lado una desventaja que se presenta es que la información se tiene de forma desestructurada, limitada a 140 caracteres, por lo cual presenta dificultades para análisis y segmentar por edad, sexo, nivel educativo, capacidad adquisitiva, entre otros. Otra dificultad que puede surgir al querer analizar las opiniones que las personas dejan en Twitter es la clasificación de la información, es muy común que los usuarios hagan clasificación mediante el uso de hash tags para definir el tema del que están opinando o hacen referencia. Dichos hash tags no siempre son una buena clasificación de las temáticas a las que se están refiriendo las personas [18].

Por otro lado las diferentes compañías de tecnología (IBM, Oracle, Microsoft, Hewlett Packard) están al margen del crecimiento de la información y se preocupan por ofrecer soluciones tecnológicas tanto de software como de hardware para lograr hacer análisis a cantidades del orden de los peta bytes [19]

Hoy en día al tener tan altos volúmenes de información se requieren métodos refinados que permitan extraer información de los datos almacenados para luego ser procesada y obtener beneficios en diferentes áreas del conocimiento.

En el siguiente capítulo se muestra el análisis propuesto de textos para una primera etapa de clasificación de Opinion Mining que podría ser de gran ayuda para el análisis de grandes volúmenes de información.

# CAPÍTULO 3 TECNICA PROPUESTA PARA EL ANALISIS Y CLASIFICACIÓN DE TEXTOS

## 3.1 ANALISIS Y OBTECIÓN DE KEYWORDS

Para este trabajo se seleccionaron tres textos de la web que tratan un misma temática (Cloud Computing) contenidos en diferentes URLs (Ver Anexo 1).

Se quitó la publicidad que tenían los sitios web y se tomaron únicamente los tres textos limpios, esté proceso se hizo de forma manual, luego los contenidos textuales fueron guardados en un archivo de extensión *texto.txt*. Posteriormente se procedió a analizar cada texto con un programa desarrollado en java para el conteo de frecuencias de palabras (Ver Anexo 2).

El contador de palabras desarrollado en java permite identificar las ocurrencias de las palabras de cada uno de los textos, también discrimina palabras como conectores o sujetos con lo que se logra quitar ruido al análisis, puesto que estas palabras generalmente tienen una alta frecuencia y no son de ayuda para identificar temáticas de un texto en particular.

A continuación se muestran los resultados obtenidos al pasar cada uno de los textos por el contador de palabras y las respectivas graficas de frecuencias sin tener en cuenta sujetos ni conectores.

### 3.1.1 Frecuencias del primer texto

Título del texto: Definición de Cloud Computing

##### Fuente: http://www.e-conomic.es/programa/glosario/definicion-cloud-computing

Texto1: Cloud Computing

|  |  |  |
| --- | --- | --- |
| Palabra | Frecuencia | Frecuencia Relativa |
| acceso | 2 | 0,033333333 |
| bajo | 2 | 0,033333333 |
| capacidad | 2 | 0,033333333 |
| cloud | 3 | 0,05 |
| computación | 3 | 0,05 |
| computing | 3 | 0,05 |
| conocido | 2 | 0,033333333 |
| consiste | 2 | 0,033333333 |
| coste | 2 | 0,033333333 |
| forma | 5 | 0,083333333 |
| información | 4 | 0,066666667 |
| internet | 7 | 0,116666667 |
| negocio | 3 | 0,05 |
| nube | 5 | 0,083333333 |
| nueva | 2 | 0,033333333 |
| poseer | 2 | 0,033333333 |
| posibilidad | 2 | 0,033333333 |
| quiera | 2 | 0,033333333 |
| servicios | 2 | 0,033333333 |
| software | 3 | 0,05 |
| utilización | 2 | 0,033333333 |
| Total de palabras | 60 | 1 |

Tabla 1. . Frecuencias del texto número 1, Resultado obtenido con el contador de palabras desarrollado

en java.

### 3.1.2 Frecuencias del segundo texto

Título del texto: Cloud computing

##### Fuente:http://www.tecnologiahechapalabra.com/tecnologia/glosario\_tecnico/articulo.a sp?i=3024

Texto2: Cloud Computing

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Palabra | | Frecuencia | | Frecuencia  Realativa | | |
| computadoras | 2 | | 0,008474576 | |
| informáticos | 2 | | 0,008474576 | |
| accede | 3 | | 0,012711864 | |
| acceder | 2 | | 0,008474576 | |
| actualidad | 2 | | 0,008474576 | |
| almacenamiento | 2 | | 0,008474576 | |
| aplica | 3 | | 0,012711864 | |
| aplicaciones | 2 | | 0,008474576 | |
| capacidad | 2 | | 0,008474576 | |
| centros | 4 | | 0,016949153 | |
| cloud | 15 | | 0,063559322 | |
| computación | 15 | | 0,063559322 | |
| computing | 17 | | 0,072033898 | |
| cómputo | 3 | | 0,012711864 | |
| conocido | 2 | | 0,008474576 | |
| conocimiento | 2 | | 0,008474576 | |
| contar | 5 | | 0,021186441 | |
| cual | 2 | | 0,008474576 | |
| cuando | 2 | | 0,008474576 | |
| datos | 7 | | 0,029661017 | |
| desde | 2 | | 0,008474576 | |
| donde | 2 | | 0,008474576 | |
| ejemplo | 3 | | 0,012711864 | |
| escritorio | 2 | | 0,008474576 | |
| forma | 6 | | 0,025423729 | |
| google | 4 | | 0,016949153 | |
| grandes | 4 | | 0,016949153 | |
| grid | 2 | | 0,008474576 | |
| grilla | 2 | | 0,008474576 | |
| incluyen | 2 | | 0,008474576 | |
| infraestructura | 2 | | 0,008474576 | |
| linea | 2 | | 0,008474576 | |
| localmente | 2 | | 0,008474576 | |
| más | 2 | | 0,008474576 | |
| mientras | 2 | | 0,008474576 | |
| muchos | 3 | | 0,012711864 | |
| navegador | 2 | | 0,008474576 | |
| necesidad | 2 | | 0,008474576 | |
| negocios | 2 | | 0,008474576 | |
| normalmente | 2 | | 0,008474576 | |
| nue | 8 | | 0,033898305 | |
| ofrece | 5 | | 0,021186441 | |
| ofrecer | 2 | | 0,008474576 | |
| presenta | 2 | | 0,008474576 | |
| reciente | 2 | | 0,008474576 | |
| redes | 3 | | 0,012711864 | |
| refiere | 2 | | 0,008474576 | |
| servicio | 18 | | 0,076271186 | |
| servicios | | | 12 | | 0,050847458 | |
| servidores | | | 3 | | 0,012711864 | |
| software | | | 8 | | 0,033898305 | |
| tecnología | | | 4 | | 0,016949153 | |
| tecnologías | | | 2 | | 0,008474576 | |
| tema | | | 3 | | 0,012711864 | |
| tendencia | | | 3 | | 0,012711864 | |
| uso | | | 3 | | 0,012711864 | |
| usuario | | | 5 | | 0,021186441 | |
| virtual | | | 2 | | 0,008474576 | |
| web | | | 5 | | 0,021186441 | |
| Total de palabras | | | 236 | | 1 | |

Tabla 2. Frecuencias del texto número 2, Resultado obtenido con el contador de palabras desarrollado en java.

### 3.1.3 Frecuencias del tercer texto

Título del texto: Cloud computing

Fuente:http://www.ceditec.etsit.upm.es/index.php?option=com\_content&view=article

##### &id=21808&Itemid=1439&lang=en

Texto3: Cloud computing

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Palabra | | Frecuencia | Frecuencia Relativa | |
| almacenamiento | 6 | 0,019169329 |
| fin | 2 | 0,006389776 |
| accesible | 3 | 0,009584665 |
| actores | 2 | 0,006389776 |
| ahora | 2 | 0,006389776 |
| ahorro | 2 | 0,006389776 |
| alojadas | 2 | 0,006389776 |
| amazon | 7 | 0,022364217 |
| analizar | 2 | 0,006389776 |
| aplicciones | 13 | 0,041533546 |
| aspecto | 3 | 0,009584665 |
| automática | 3 | 0,009584665 |
| bajo | 3 | 0,009584665 |
| basa | 3 | 0,009584665 |
| basado | 2 | 0,006389776 |
| base | 2 | 0,006389776 |
| bases | 5 | 0,015974441 |
| caso | 2 | 0,006389776 |
| cliente | 5 | 0,015974441 |
| cloud | 28 | 0,089456869 |
| computacional | 3 | 0,009584665 |
| computing | 25 | 0,079872204 |
| comunidad | 14 | 0,044728435 |

|  |  |  |
| --- | --- | --- |
| concepto | 9 | 0,028753994 |
| confianza | 4 | 0,012779553 |
| conjunto | 2 | 0,006389776 |
| construcción | 2 | 0,006389776 |
| construido | 11 | 0,03514377 |
| consumido | 2 | 0,006389776 |
| consumo | 9 | 0,028753994 |
| contenido | 2 | 0,006389776 |
| convertirse | 2 | 0,006389776 |
| coste | 3 | 0,009584665 |
| costes | 4 | 0,012779553 |
| cuando | 4 | 0,012779553 |
| datos | 14 | 0,044728435 |
| definición | 3 | 0,009584665 |
| definir | 3 | 0,009584665 |
| definitiva | 2 | 0,006389776 |
| demanda | 3 | 0,009584665 |
| denominación | 2 | 0,006389776 |
| desarrollador | 2 | 0,006389776 |
| desarrolladores | 5 | 0,015974441 |
| desarrollo | 9 | 0,028753994 |
| desde | 2 | 0,006389776 |
| dicho | 2 | 0,006389776 |
| diferentes | 8 | 0,025559105 |
| dispositivos | 2 | 0,006389776 |
| Distribución | 2 | 0,006389776 |
| documento | 2 | 0,006389776 |
| donde | 3 | 0,009584665 |
| ejecución | 2 | 0,006389776 |
| embargo | 5 | 0,015974441 |
| empresa | 9 | 0,028753994 |
| empresas | 4 | 0,012779553 |
| escalabilidad | 3 | 0,009584665 |
| escalables | 15 | 0,047923323 |
| escogido | 2 | 0,006389776 |
| estamos | 2 | 0,006389776 |
| externa | 6 | 0,019169329 |
| externalización | 2 | 0,006389776 |
| externalizar | 2 | 0,006389776 |
| favor | 2 | 0,006389776 |
| final | 4 | 0,012779553 |
| forma | 29 | 0,092651757 |
| función | 2 | 0,006389776 |
| fundamentalmente | 2 | 0,006389776 |
| fundamentos | 2 | 0,006389776 |
| garantizada | 4 | 0,012779553 |
| google | 4 | 0,012779553 |
| grandes | 4 | 0,012779553 |
| hardware | 4 | 0,012779553 |
| herramientas | 3 | 0,009584665 |
| hosting | 4 | 0,012779553 |
| IaaS importante | 6 | 0,019169329 |

|  |  |  |
| --- | --- | --- |
| incluyendo | 2 | 0,006389776 |
| información | 7 | 0,022364217 |
| infraestructura | 7 | 0,022364217 |
| integración | 3 | 0,009584665 |
| intereses | 3 | 0,009584665 |
| internet | 2 | 0,006389776 |
| Inversión | 3 | 0,009584665 |
| largo | 2 | 0,006389776 |
| lenguaje | 3 | 0,009584665 |
| llegada | 2 | 0,006389776 |
| mantener | 3 | 0,009584665 |
| mantenimiento | 3 | 0,009584665 |
| máquinas | 3 | 0,009584665 |
| marco | 2 | 0,006389776 |
| máximo | 2 | 0,006389776 |
| mercado | 5 | 0,015974441 |
| meses | 3 | 0,009584665 |
| Microsoft | 4 | 0,012779553 |
| misma | 2 | 0,006389776 |
| mismos | 4 | 0,012779553 |
| modelo | 17 | 0,054313099 |
| modelos | 3 | 0,009584665 |
| móviles | 4 | 0,012779553 |
| mundo | 2 | 0,006389776 |
| necesario | 3 | 0,009584665 |
| necesidad | 2 | 0,006389776 |
| nueva | 4 | 0,012779553 |
| nuevo | 9 | 0,028753994 |
| oferta | 2 | 0,006389776 |
| ofrecer | 3 | 0,009584665 |
| operativo | 2 | 0,006389776 |
| orientada | 2 | 0,006389776 |
| PaaS | 8 | 0,025559105 |
| pago | 3 | 0,009584665 |
| paradigma | 2 | 0,006389776 |
| persigue | 2 | 0,006389776 |
| pionera | 2 | 0,006389776 |
| pionera | 2 | 0,006389776 |
| plataforma | 17 | 0,054313099 |
| podemos | 2 | 0,006389776 |
| posibilidad | 4 | 0,012779553 |
| pretende | 5 | 0,015974441 |
| pretenden | 2 | 0,006389776 |
| principios | 6 | 0,019169329 |
| problema | 6 | 0,019169329 |
| problemas | 2 | 0,006389776 |
| procesamiento | 2 | 0,006389776 |
| proceso | 3 | 0,009584665 |
| productos | 6 | 0,019169329 |
| programación | 3 | 0,009584665 |
| proveedor | 6 | 0,019169329 |
| proveedores puedan | 8 | 0,025559105 |

|  |  |  |
| --- | --- | --- |
| puede | 4 | 0,012779553 |
| puesto | 2 | 0,006389776 |
| punto | 2 | 0,006389776 |
| realmente | 3 | 0,009584665 |
| recientemente | 2 | 0,006389776 |
| recursos | 13 | 0,041533546 |
| Red | 11 | 0,03514377 |
| referencia | 2 | 0,006389776 |
| remota | 2 | 0,006389776 |
| respuesta | 4 | 0,012779553 |
| resto | 2 | 0,006389776 |
| saas | 7 | 0,022364217 |
| seguridad | 2 | 0,006389776 |
| serie | 2 | 0,006389776 |
| server | 2 | 0,006389776 |
| services | 7 | 0,022364217 |
| servicio | 39 | 0,124600639 |
| servicios | 15 | 0,047923323 |
| servidores | 5 | 0,015974441 |
| sistema | 5 | 0,015974441 |
| sistemas | 2 | 0,006389776 |
| sobre | 13 | 0,041533546 |
| software | 16 | 0,051118211 |
| sólo | 2 | 0,006389776 |
| solución | 5 | 0,015974441 |
| soluciones | 12 | 0,038338658 |
| soporte | 2 | 0,006389776 |
| sql | 2 | 0,006389776 |
| sun | 5 | 0,015974441 |
| supone | 2 | 0,006389776 |
| también | 3 | 0,009584665 |
| tanto | 2 | 0,006389776 |
| tecnologías | 3 | 0,009584665 |
| tendencia | 9 | 0,028753994 |
| tendencias | 6 | 0,019169329 |
| tercero | 2 | 0,006389776 |
| término | 2 | 0,006389776 |
| the | 2 | 0,006389776 |
| tiempo | 2 | 0,006389776 |
| tiempos | 2 | 0,006389776 |
| tipo | 2 | 0,006389776 |
| toda | 2 | 0,006389776 |
| todo | 3 | 0,009584665 |
| todos | 3 | 0,009584665 |
| transferencia | 2 | 0,006389776 |
| transparente | 3 | 0,009584665 |
| trata | 7 | 0,022364217 |
| bicuidad | 3 | 0,009584665 |
| único | 3 | 0,009584665 |
| usuario | 25 | 0,079872204 |
| usuarios | 6 | 0,019169329 |
| ventajas virtual | 3 | 0,009584665 |

web 10 0,031948882 windows 4 0,012779553

|  |  |  |
| --- | --- | --- |
| Total de palabras | 313 | 1 |

Tabla 3. Frecuencias del texto número 3, Resultado obtenido con el contador de palabras desarrollado en java.

### 3.1.4 Promedio ponderado de los tres textos y keywords

El proceso desarrollado en esta sección se hizo con el fin de ponderar los resultados de los tres textos y de esta manera tener una mejor vista de las palabras más relevantes que más adelante se van a tomar como keywords, el proceso se ilustra en la figura 3.

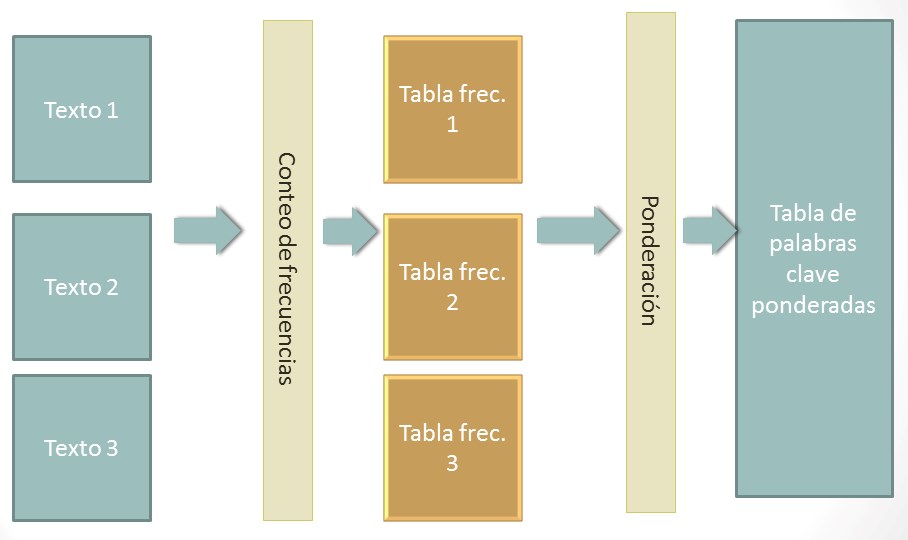


Figura 3. Proceso estadístico que se realizó sobre los tres textos que tratan el tema de Cloud

Computing, luego se obtuvieron los promedios ponderados y finalmente se aplicó la técnica de Pareto para descartar palabras que hacían ruido al análisis.

Para este análisis se tomaron los promedios ponderados de las frecuencias relativas de todas las palabras contenidas en los tres textos, con el fin de obtener las palabras más importantes y de forma igualitaria y representativa en los tres textos.

Se usa la ecuación 1:

𝑝&𝑥& + 𝑝)𝑥) + 𝑝\*𝑥\*+ … … … + 𝑝,𝑥, ,.0&(𝑝. ∗ 𝑥.)



𝑝&  … … … +𝑝, ,.0& 𝑝.

Ecuación 1 Promedios ponderados

Para el presente análisis las variables que se usan son las siguientes:

𝑝& = 𝑓𝑟𝑒𝑐𝑢𝑒𝑛𝑐𝑖𝑎 𝑟𝑒𝑙𝑎𝑡𝑖𝑣𝑎 𝑑𝑒 𝑢𝑛𝑎 𝑝𝑎𝑙𝑎𝑏𝑟𝑎 𝑒𝑛 𝑒𝑙 𝑝𝑟𝑖𝑚𝑒r 𝑡𝑒𝑥𝑡𝑜.

𝑥& = 𝐼𝑚𝑝𝑜𝑟𝑡𝑎𝑛𝑐𝑖𝑎 𝑞𝑢𝑒 𝑠𝑒 𝑙𝑒 𝑑𝑎 𝑎 𝑙𝑎 𝑓𝑟𝑒𝑐𝑢𝑒𝑛𝑐𝑖𝑎 𝑟𝑒𝑙𝑎𝑡𝑖𝑣𝑎

𝑑𝑒 𝑒𝑠𝑡𝑎 𝑝𝑎𝑙𝑎𝑏𝑟𝑎 𝑒𝑛 𝑒𝑙 𝑡𝑒𝑥𝑡𝑜 1

𝑝) = 𝑓𝑟𝑒𝑐𝑢𝑒𝑛𝑐𝑖𝑎 𝑟𝑒𝑙𝑎𝑡𝑖𝑣𝑎 𝑑𝑒 𝑢𝑛𝑎 𝑝𝑎𝑙𝑎𝑏𝑟𝑎 𝑒𝑛 𝑒𝑙 𝑠𝑒𝑔𝑢𝑛𝑑𝑜 𝑡𝑒𝑥𝑡𝑜.

𝑥) = 𝐼𝑚𝑝𝑜𝑟𝑡𝑎𝑛𝑐𝑖𝑎 𝑞𝑢𝑒 𝑠𝑒 𝑙𝑒 𝑑𝑎 𝑎 𝑙𝑎 𝑓𝑟𝑒𝑐𝑢𝑒𝑛𝑐𝑖𝑎 𝑟𝑒𝑙𝑎𝑡𝑖𝑣𝑎

𝑑𝑒 𝑒𝑠𝑡𝑎 𝑝𝑎𝑙𝑎𝑏𝑟𝑎 𝑒𝑛 𝑒𝑙 𝑡𝑒𝑥𝑡𝑜 2

𝑝\* = 𝑓𝑟𝑒𝑐𝑢𝑒𝑛𝑐𝑖𝑎 𝑟𝑒𝑙𝑎𝑡𝑖𝑣𝑎 𝑑𝑒 𝑢𝑛𝑎 𝑝𝑎𝑙𝑎𝑏𝑟𝑎 𝑒𝑛 𝑒𝑙 𝑡𝑒𝑟𝑐𝑒𝑟 𝑡𝑒𝑥𝑡𝑜.

𝑥\* = 𝐼𝑚𝑝𝑜𝑟𝑡𝑎𝑛𝑐𝑖𝑎 𝑞𝑢𝑒 𝑠𝑒 𝑙𝑒 𝑑𝑎 𝑎 𝑙𝑎 𝑓𝑟𝑒𝑐𝑢𝑒𝑛𝑐𝑖𝑎 𝑟𝑒𝑙𝑎𝑡𝑖𝑣𝑎

𝑑𝑒 𝑒𝑠𝑡𝑎 𝑝𝑎𝑙𝑎𝑏𝑟𝑎 𝑒𝑛 𝑒𝑙 𝑡𝑒𝑥𝑡𝑜 3

Para el análisis de los tres textos se quiere considerar el mismo nivel de importancia a cada uno de estos, por un valor constante a las variables:

𝑥& = 𝑥) = 𝑥\* = 1

Al tener igualadas las variables 𝑥&, 𝑥) 𝑦 𝑥\* se está dando la misma relevancia a todos los textos sin importar si alguno es más extenso que otros. Luego de tener el resultado de los promedios ponderados se realizó un análisis estadístico de Pareto con el cual se toma solo el 80% de las palabras que tienen mayor valor en su promedio ponderado y se desprecia al otro 20% que son datos que no aportan al análisis (para este estudio se tomó el 81%).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Palabra | | Promedio Ponderado | Pareto | Porcentaje % | |
| computing | 0,14865797 | 5% | 5,30% |
| cloud | 0,14337828 | 10% | 5,11% |
| forma | 0,13964098 | 15% | 4,98% |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| internet | 0,11879659 | | 20% | | 4,24% |
| computación | 0,11355932 | | 24% | | 4,05% |
| software | 0,10093771 | | 27% | | 3,60% |
| servicios | 0,10015523 | | 31% | | 3,57% |
| nube | 0,08333333 | | 34% | | 2,97% |
| información | 0,07412141 | | 36% | | 2,64% |
| negocio | 0,05 | | 38% | | 1,78% |
| usuario | 0,04781051 | | 40% | | 1,70% |
| datos | 0,0445705 | | 42% | | 1,59% |
| conocido | 0,04180791 | | 43% | | 1,49% |
| capacidad | 0,04180791 | | 45% | | 1,49% |
| servicio | 0,04153355 | | 46% | | 1,48% |
| coste | 0,03652822 | | 47% | | 1,30% |
| bajo | 0,03652822 | | 49% | | 1,30% |
| nube | 0,03389831 | | 50% | | 1,21% |
| utilización | 0,03333333 | | 51% | | 1,19% |
| quiera | 0,03333333 | | 52% | | 1,19% |
| poseer | 0,03333333 | | 53% | | 1,19% |
| consiste | 0,03333333 | | 55% | | 1,19% |
| acceso | 0,03333333 | | 56% | | 1,19% |
| web | 0,03183607 | | 57% | | 1,13% |
| aplicaciones | 0,02231909 | | 58% | | 0,80% |
| tendencia | 0,02229653 | | 58% | | 0,79% |
| grandes | 0,021209 | | 59% | | 0,76% |
| google | 0,021209 | | 60% | | 0,76% |
| Ofrece | 0,02118644 | | 61% | | 0,76% |
| contar | 0,02118644 | | 62% | | 0,76% |
| plataforma | 0,01810437 | | 63% | | 0,65% |
| podelo | 0,01810437 | | 64% | | 0,65% |
| servidores | 0,01803668 | | 64% | | 0,64% |
| tecnología | 0,01694915 | | 65% | | 0,60% |
| centros | 0,01694915 | | 65% | | 0,60% |
| escalables | 0,01597444 | | 66% | | 0,57% |
| infraestructura | 0,01592932 | | 67% | | 0,57% |
| comunidad | 0,01490948 | | 67% | | 0,53% |
| sobre | 0,01384452 | | 68% | | 0,49% |
| recursos | 0,01384452 | | 68% | | 0,49% |
| soluciones | 0,01277955 | | 69% | | 0,46% |
| cuando | 0,01273443 | | 69% | | 0,45% |
| uso | 0,01271186 | | 69% | | 0,45% |
| tema | 0,01271186 | | 70% | | 0,45% |
| muchos | 0,01271186 | | 70% | | 0,45% |
| ejemplo | 0,01271186 | | 71% | | 0,45% |
| cómputo | 0,01271186 | | 71% | | 0,45% |
| aplica | 0,01271186 | | 72% | | 0,45% |
| accede | 0,01271186 | 72% | | 0,45% | |
| red | 0,01171459 | 73% | | 0,42% | |
| construido | 0,01171459 | 73% | | 0,42% | |
| tecnologías | 0,01166946 | 73% | | 0,42% | |
| ofrecer | 0,01166946 | 74% | | 0,42% | |
| dónde | 0,01166946 | 74% | | 0,42% | |
| donde | 0,01166946 | 75% | | 0,42% | |
| virtual | 0,0106045 | 75% | | 0,38% | |
| necesidad | 0,0106045 | 75% | | 0,38% | |
| desde | 0,0106045 | 76% | | 0,38% | |
| nuevo | 0,00958466 | 76% | | 0,34% | |
| empresa | 0,00958466 | 76% | | 0,34% | |
| desarrollo | 0,00958466 | 77% | | 0,34% | |
| consumo | 0,00958466 | 77% | | 0,34% | |
| concepto | 0,00958466 | 77% | | 0,34% | |
| proveedores | 0,0085197 | 78% | | 0,30% | |
| paas | 0,0085197 | 78% | | 0,30% | |
| diferentes | 0,0085197 | 78% | | 0,30% | |
| refiere | 0,00847458 | 79% | | 0,30% | |
| presenta | 0,00847458 | 79% | | 0,30% | |
| normalmente | 0,00847458 | 79% | | 0,30% | |
| navegador | 0,00847458 | 80% | | 0,30% | |
| mientras | 0,00847458 | 80% | | 0,30% | |
| más | 0,00847458 | 80% | | 0,30% | |
| localmente | 0,00847458 | 81% | | 0,30% | |
| línea | 0,00847458 | 81% | | 0,30% | |
| demás | 0,53841988 | 100% | | 19,20% | |
| Total | | 2,80494666 | | | | | |

Tabla 4. Promedios ponderados de los tres textos y análisis de Pareto tomando únicamente el 81% del total de palabras de los tres textos.

Del análisis de Pareto se obtiene un gráfica con el 81% de la totalidad de las palabras más relevantes, los resultados se pueden apreciar en las figuras 4 y 5.

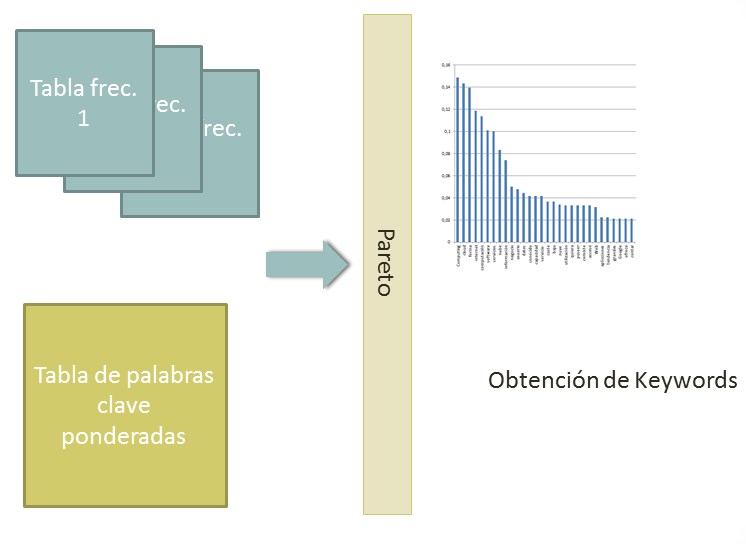


Figura 4. Sobre la tabla de promedios ponderados se hace el proceso de Pareto para establecer las palabras más relevantes.

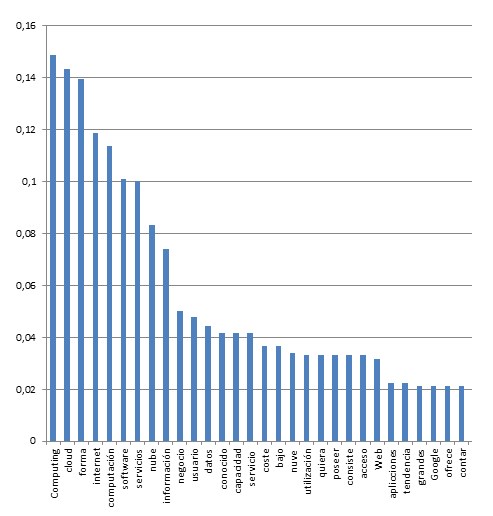


Figura 5. Primeras 30 palabras del promedio ponderado de todas las palabras contenidas en los tres textos analizados.

De la figura 5 Se ve que las primeras 30 palabras están con un promedio ponderado por encima de (0,02) lo cual indica una relevancia importante en los tres textos (Ver Anexo 3).

El resultado del análisis estadístico resumido en la tabla 4 permite identificar las palabras más relevantes de la temática tratada en los tres textos (Cloud Computing), este resultado será analizado junto con los resultados de una encuesta en la siguiente sección.

## 3.2 ENCUESTA Y VALIDACIÓN DE KEYWORDS OBTENIDAS

La idea principal de desarrollar una encuesta sobre los textos, es poder relacionar las percepciones de las personas y los resultados de análisis estadísticos sobre los mismos textos. Posteriormente se van a validar que las keywords halladas de forma estadística con la frecuencia corresponden a una temática en particular (Cloud

Computing).

##### 3.2.1 DESCRIPCIÓN DE LA ENCUESTA

La encuesta fue realizada con el fin de entender la percepción de las personas respecto al contenido temático de tres textos que se refieren a Cloud Computing. Los tres textos fueron presentados a las personas sin título puesto que en una de las preguntas de la encuesta tenían que seleccionar un posible título para cada texto (Ver

Anexo 4).

##### 3.2.2 RESULTADOS DE LA ENCUESTA

* Primera pregunta

Para la primera pregunta el 100% (50 personas) de los encuestados coincidieron en su respuesta al asignar un título a los textos.

Todos respondieron: Cloud Computing

* Segunda Pregunta

En la tabla 5 se muestran los promedios de cada una de las 20 palabras seleccionadas en la encuesta. El promedió corresponde a la importancia que tiene cada palabra en los tres textos de Cloud Computing según los encuestados. Este promedio en otras palabras es el Peso en importancia que tiene cada palabra por la temática tratada.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Palabra** | | **Peso de keywords** | | |
| Cloud | | 4,759259259 | | |
| Nube | | 4,740740741 | | |
| Computing | | 4,555555556 | | |
| Internet | | 4,555555556 | | |
| Red | | 4,407407407 | | |
| Servicio | | 4,37037037 | | |
| Información | | 4,351851852 | | |
| Web | | 4,333333333 | | |
| Datos | | 4,277777778 | | |
| Usuario | | 4,222222222 | | |
| Aplicaciones | 4,111111111 | |
| Software | 4,018518519 | |
| Plataforma | 3,925925926 | |
| Computación | 3,87037037 | |
| Coste | 3,648148148 | |
| Escalables | 3,333333333 | |
| Negocio | 3,074074074 | |
| Tendencia | | 3,037037037 |
| Nueva | | 2,87037037 |
| Modelo | | 2,685185185 |

Tabla 5. En esta tabla se pueden ver las palabras de la encuesta con sus respectivos promedios y están organizadas de mayor a menor según el nivel de importancia que los encuestados les dieron.

**Resumen estadístico**

Palabra de Mayor promedio: Cloud 4,759259259

Promedio 3,957407407

Media Aritmética 3,957407407

Mediana 4,166666667

Tabla 6. Resumen de los datos relevante de la encuesta, palabra con mayor promedio, promedio de los puntajes, Media aritmética y medina.

En el siguiente apartado se hace el análisis de los promedios ponderados y del resultado de la encuesta para establecer palabras relevantes entre los tres textos procesados.

### 3.2.3 Análisis de promedios ponderados y resultados de la encuesta

Como resultado de la primera pregunta de la encuesta el 100% de encuestados seleccionaron de forma correcta el título para los textos (Cloud Computing), por lo que se puede ver que son personas que conocían el concepto y dieron una buena

interpretación a los textos.

Por otro lado, al analizar los resultados obtenidos del análisis de promedios ponderados y de Pareto que se hizo en el apartado 3.1.4 de esta tesis, se obtuvo un 81 % de palabras relevantes. Gracias al análisis de Pareto se identificaron las 20 palabras de mayor relevancia y se usaron para incluir en la encuesta. Para cada una de estas palabras los encuestados asignaron un valor según su importancia y relevancia en el contenido temático de los textos.

De acuerdo a los puntajes asignados por los encuestados se determinaron las diez palabras más relevantes para la temática de Cloud Computing:

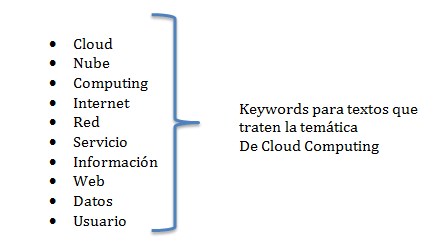


Figura 6. Las 10 Keywords referentes a la temática de Cloud Computing obtenidas después del análisis de Pareto.

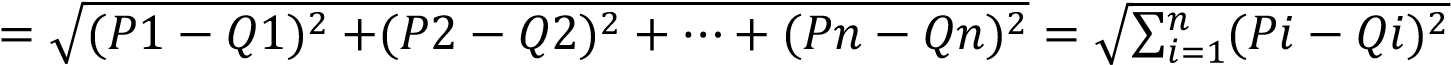
## 3.3 DISTANCIA ENTRE TEXTOS SEGÚN KEWORDS

Las técnicas usadas para la clasificación de textos son Native Bayes classifier, J48 Decision Trees y Support Vector Machines [20, 21, 22], estas son adaptaciones del método K-means Clustering. Para determinar cercanía entre patrones de información el método de Clustering usa la distancia euclidea, haciendo posible la clasificación de datos que tengan similitud, esta distancia es aplicable en un espacio euclideo donde se puedan identificar diferentes objetos coordenados y hallar las distancias entre ellos atreves de métricas[[5]](#footnote-5).

La fórmula canónica de la distancia euclidea par puntos:

𝑃 =𝑃1. 𝑃2 … 𝑃𝑛 𝑦 𝑄(𝑄1. 𝑄2 … 𝑄𝑁)

En un espacio de n dimensiones se define en la ecuación 2:

(2) 𝑑𝐸 𝑃, 𝑄 

Ecuación 2. Distancia euclidea

La distancia que se propone en la presente tesis para identificar la pertenencia o no de un texto a una temática específica está basada en un análisis estadístico. El procesamiento de un texto referente a una temática desconocida se realiza determinando las ocurrencias de cada palabra en el texto, luego se puede analizar según el peso de sus palabras con mayor frecuencia y medir[[6]](#footnote-6) en relación a las keywords halladas previamente de una temática conocida. Este proceso permite determinar si un texto desconocido pertenece al mismo grupo o trata la misma temática de los 3 textos de los cuales se obtuvieron las keywords, el proceso para determinar las keywords está ilustrado en la figura 7 y el de clasificación de textos en la figura 9.

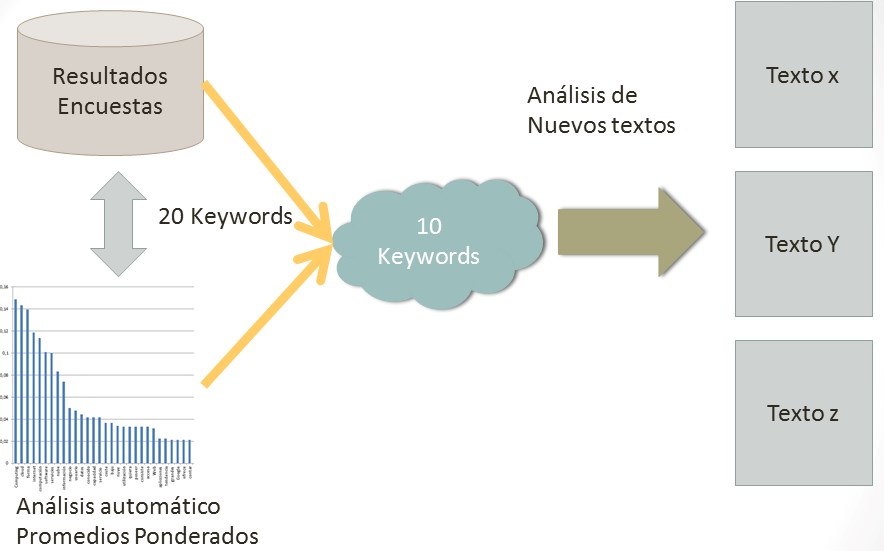


Figura 7. Con las Keywords obtenidas sobre la temática de Cloud Computing que se pueden analizar nuevos textos x, y, z, etc de temas desconocidos y posteriormente inferir si también pertenecen a esta misma temática.

Dentro del análisis de los promedios ponderados se habían seleccionado las 20 primeras palabras más relevantes, y en base a ellas se hizo la encuesta en donde los encuestados dieron un puntaje en nivel de importancia a cada palabra. De esto se dedujeron 10 keywords que son representativas para la temática de Cloud Computing.

Para poder analizar textos nuevos se define un rango entre las palabras más relevantes halladas previamente. De las primeras 20 keywords que fueron valoradas en nivel de importancia por los encuestados corresponden a un 100% de peso en relevancia en los textos (Promedio total = *PT*).

Se toma *PT* como límite superior y cómo límite inferior el 60% de la suma de pesos de las 10 keywords definidas en la encuesta, puesto que estas son las de mayor importancia (60% de la suma de 10 keywords de mayor relevancia = *K*).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Palabra** | | **Peso De keywords** |  | | |
| Cloud | | 4,75 |
| Nube | | 4,74 |  | | |
| Computing | | 4,556 |  | | |
| Internet | | 4,55 |  | | |
| Red | | 4,40 |  | | |
| Servicio | | 4,37 |  | | |
| Información | | 4,35 |  | | |
| Web | | 4,33 |  | | |
| Datos | | 4,27 | **Suma de Promedio keywords** | | |
| Usuario | | 4,22 | (44,57 X 50%) = 22,35 *K*=22,35 | | |
| Aplicaciones | 4,11 | |  |
| Software | 4,01 | |  |
| Plataforma | 3,92 | |  |
| Computación | 3,87 | |  |
| Coste | 3,64 | |  |
| Escalables | 3,33 | |  |
| Negocio | 3,07 | |  |
| Tendencia | 3,03 | |  |
| Nueva | 2,87 | | **Suma total de los promedios** |
| Modelo | 2,68 | | *PT*= 79,14 |

Tabla 7. Keywords con sus respectivos pesos obtenidos de la encuesta. Se observa el rango en donde se debe encontrar la suma de keywords halladas en los textos que son analizados (*PT* y *K*).

Para saber si un texto cualquiera habla o no de una temática en particular se debe procesar con el contador de frecuencias/ocurrencias y se tendrán en cuenta sus primeras 20 palabras en orden de mayor a menor.

Dentro de las primeras 20 palabras el texto puede tener algunas, todas o ninguna palabra de las 20 keywords halladas previamente para la temática de Cloud

Computing,

De las palabras que se encuentren en el texto se sumarán sus respectivos pesos y esto va a dar un valor *w* que debe estar entre el rango de valores posibles entre *K* y *PT*. Se puede inferir entonces que si para un texto x cualquiera su peso W se encuentra en el siguiente rango:

𝐾 ≤ 𝑊 ≤ 𝑃𝑇



Por tanto un texto cualquiera *texto x* pertenece a la temática Cloud Computing si tiene una distancia DT cercana o igual cero[[7]](#footnote-7):

(3) 𝐷𝑇 = (𝑃𝑇 − 𝑊))

Ecuación 3. Distancia propuesta para determinar cercanía de un texto respecto a la temática de Cloud Computing.

La menor Distancia que se puede presentar es cero, que sería un caso en que un *texto x* tiene las 20 palabras claves, ósea cuando *W=PT*.

𝑊 = 𝑃𝑇

𝐷𝑇 = (𝑃𝑇 − 𝑊))

𝐷𝑇 = (79,14 − 79,14))

𝐷𝑇 =0

Por otro lado la distancia mayor que puede tener un texto perteneciente a la temática de Cloud Computing, se presenta cuando *W=K.*

𝑊 = 𝐾

𝐷𝑇 = (𝑃𝑇 − 𝑊))

𝐷𝑇 = (79,14 − 22,35))

𝐷𝑇 = 3225,10

Se define entonces que para un *texto x* pertenece a la temática de Cloud

Computing siempre y cuando su distancia *DTx* se encuentre entre el siguiente rango:

 𝐷𝑇𝑥 ≤ 3225,10

Ecuación 4. Rango distancia que determina si un texto hace referencia a la temática de Cloud Computing o no.

La ecuación propuesta de distancia se puede reconocer como una distancia valida puesto que cumple con las tres propiedades básicas que se validan a

continuación:

* No negatividad:

(5) 𝑑 𝑇, 𝑊 

Ecuación 5. Muestra la no negatividad de la distancia propuesta.

Esta condición se cumple por construcción puesto que todo número elevado al cuadrado es positivo y por otro lado *W* siembre va a ser menor o igual que *PT*.

* Simetría

(6) 𝑑 𝑇, 𝑊 = 𝑑(𝑊, 𝑇)

Ecuación 6. Simetría de la distancia propuesta.

𝑑 𝑇, 𝑊 = (𝑃𝑇 − 𝑊))

𝑑 𝑇, 𝑊 = (𝑊 − 𝑃𝑇))

Son simétricas las distancias de *T* a *W* y de *W* a *T* puesto que se cumple la igualdad tanto en magnitud como en signo.

* Desigualdad triangular

Se puede construir un triángulo de lados:

𝑑 𝑊&, 𝑇 , 𝑑 𝑊), 𝑇 , 𝑑(𝑊&, 𝑊))

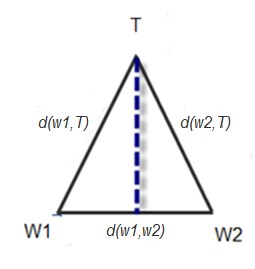


Figura 8. Triángulo formado por las distancias: d(W\_1,T),d(W\_2,T),d(W\_1,W\_2)

Y por propiedades de los triángulos, se tiene que la suma de dos de sus lados es mayor que el tercero.

(7) 𝑑 𝑊&, 𝑇 + 𝑑(𝑊), 𝑇) ≥ 𝑑(𝑊&, 𝑊))

Ecuación 7. Desigualdad triangular de la distancia propuesta.

Para verificar la distancia propuesta se procede a hacer una prueba con tres textos en el siguiente capítulo.

# CAPÍTULO 4

# IMPLEMENTACIÓN DE LA TECNICA PROPUESTA PARA LA CLASIFICACIÓN DE TEXTOS

## 4.1 PRUEBA Y ANÁLISIS PARA DETERMINAR SI UN TEXTO HACE REFERENCIA A UNA TEMÁTICA ESPECÍFICA.

Para la realización de las pruebas de la técnica de clasificación propuesta se usaron 12 textos que tratan diversos temas, dentro de los cuales 5 hacen referencia a la temática de Cloud Computing. A continuación se ven los títulos da cada uno de los textos y en la Figura 9 se ilustra el proceso de clasificación.

* *Texto a:* El deporte y sus beneficios en la salud física, mental y psicológica.
* *Texto b:* Cloud Computing.
* *Texto c:* Educación y computadoras – Enseñanzas de América Latina.
* *Texto d:* El deporte y sus beneficios en la salud física, mental y psicológica.
* *Texto e:* Cloud Computing.
* *Texto f:* Educación y computadoras – Enseñanzas de América Latina.
* *Texto g*: El deporte y sus beneficios en la salud física, mental y psicológica.
* *Texto h:* Cloud Computing.
* *Texto i:* Educación y computadoras – Enseñanzas de América Latina.
* *Texto j:* El deporte y sus beneficios en la salud física, mental y psicológica.
* *Texto k:* El Tiempo En Campana.
* *Texto l:* Formación de las Nubes

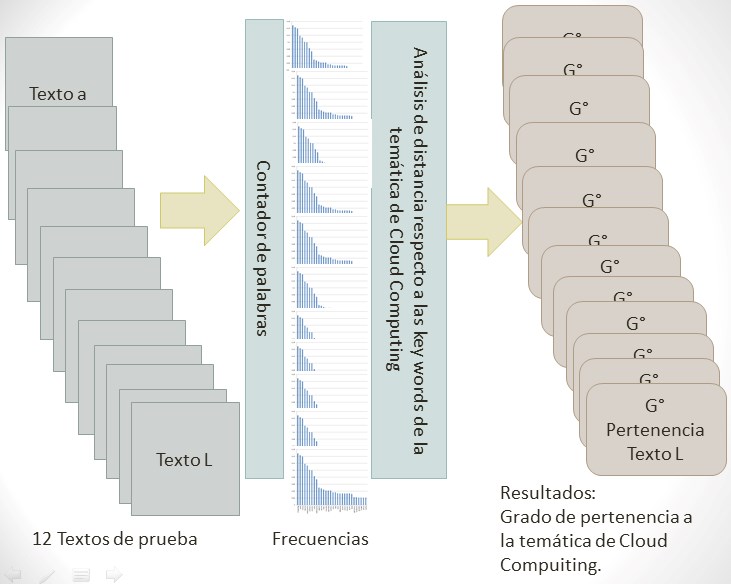


Figura 9. Proceso de la técnica propuesta para la clasificación de textos.

### 4.1.1 Prueba con el texto (a), título: Datacenters de Nube Argentina de Telecom, socios estratégicos del resguardo de información de las empresas.

El texto (a) fue seleccionado de una URL de la web (Ver Anexo 5) y se procesó con un contador de frecuencias (Ver Anexo 2). Para la medida de la distancia respecto a las keywords de Cloud Computing se toman las primeras 20 palabras obtenidas en orden descendente comenzando por la palabra con mayor frecuencia.

|  |  |
| --- | --- |
| Palabra | Frecuencia |
| telecom | 9 |
| servicios | 9 |

infraestructura 7 datacenters 7 argentina 6 sistemas 5

|  |  |
| --- | --- |
| datos | 5 |
| seguridad | 4 |
| resguardo | 4 |
| equipos | 4 |
| cloud | 4 |
| cliente | 4 |
| tecnología | 3 |
| servicio | 3 |
| nube | 3 |
| negocio | 3 |
| información | 3 |

como 3 todos 2 soluciones 2

Tabla 8. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del texto (a). Las palabras sombreadas son las keywords de la temática de Cloud Computing que se encontraron en este texto.

El Texto (a) tiene keywords de Cloud Computing que están resumidas con su respectivo peso en la Tabla 9.

|  |  |
| --- | --- |
| Palabras | Peso |
| datos | 4,27777778 |
| cloud | 4,75925926 |
| servicio | 4,37037037 |
| nube | 4,74074074 |
| negocio | 3,07407407 |
| información | 4,35185185 |
| *W*= | 25,5740741 |

Tabla 9. Keywords identificadas en el texto (a) con sus respetivos pesos y el peso total w.

El texto (a) hace referencia a la implementación de Datacenters en Argentina para Cloud Computing. Luego del análisis estadístico de ocurrencias de las palabras se obtiene el siguiente peso total representativo por las keywords:

𝑊 = 25,57

Al obtener el peso del texto (a) se puede calcular la distancia que tiene respecto a la Temática de Cloud Computing usando la ecuación (3), el proceso se puede ver a continuación.

(3) 𝐷𝑇𝑎 = (𝑃𝑇 − 𝑊))

𝐷𝑇𝑎 = (79,14 − 25,57))

𝐷𝑇𝑎 = (53,57))

𝐷𝑇𝑎 = 2869,74

Según el resultado de *DTa* obtenido anteriormente se puede observar que el texto (a) pertenece a la temática de Cloud Computing puesto que la distancia está dentro del rango definido anteriormente en la ecuación 4, como se ve a continuación:

 𝐷𝑇𝑎 ≤ 3225,10

### 4.1.2 Prueba con el texto (b), título: Cloud Computing

El texto (b) se seleccionado de una URL de la web (Ver Anexo 5) y se procesó con el contador de frecuencias (Ver Anexo 2).

Para la medida de la distancia respecto a las keywords de Cloud Computing se toman las primeras 20 en orden de mayor a menor obtenidas del procesamiento del texto y se obtuvo el siguiente resultado:

|  |  |
| --- | --- |
| Palabra | Ocurrencia |
| datos | 16 |
| cloud | 16 |
| computing | 12 |
| computadora | 11 |
| Red | 10 |
| usted | 9 |
| computación | 9 |
| Nube | 8 |
| aplicaciones | 7 |
| modelo | 6 |

tampoco 5 programas 5 archivos 5 trabajo 4 tiene 4

|  |  |
| --- | --- |
| internet | 4 |

correo 4 acceso 4

trabajar 3

Tabla 10. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del texto (b). Las palabras sombreadas son las keywords de Cloud Computing que se encontraron en esté texto.

A continuación en la Tabla 11 se puede observar el resumen de las keywords de Cloud Computing obtenidas en el texto (b) con sus respectivos pesos.

|  |  |
| --- | --- |
| Palabras | Peso |
| datos | 4,27777778 |
| cloud | 4,75925926 |
| computing | 4,55555556 |
| Red | 4,40740741 |
| computación | 3,87037037 |
| Nube | 4,74074074 |
| aplicaciones | 4,11111111 |
| modelo internet | 2,68518519 |
| *W=* |  |

Tabla 11. Palabras identificadas en el texto (b) con sus respectivos pesos y peso total w.

Al obtener el peso del texto (b) se puede calcular la distancia que tiene respecto a la Temática de Cloud Computing usando la ecuación (3), el proceso se puede ver a continuación en la ecuación.

𝑊 = 37,96

(3) 𝐷𝑇𝑏 = (𝑃𝑇 − 𝑊))

𝐷𝑇𝑏 = (79,14 − 37,96)) 𝐷𝑇𝑏 = (41,18))

𝐷𝑇𝑏 = 1695,79

Según el resultado de *DTb* obtenido anteriormente se puede observar que el texto (b) pertenece a la temática de Cloud Computing puesto que la distancia está dentro del rango definido anteriormente en la ecuación (4), como se ve a continuación:

 𝐷𝑇𝑏 ≤ 3225,10

### 4.1.3 Prueba con el texto (c), título: Tendencias de los servicios de Cloud Computing en el mercado Español.

El texto (c) fue seleccionado de una URL de la web (Ver Anexo 5) y se procesó con el contador de frecuencias (Ver Anexo 2).

Para la medida de la distancia respecto a las keywords de Cloud Computing se toman las primeras 20 palabras de forma descendente comenzando con la palabra de mayor frecuencia como se ve en la Tabla 12.

|  |  |
| --- | --- |
| Palabra | Ocurrencia |
| cloud | 35 |
| servicios | 11 |
| modelo | 11 |
| infraestructura | 9 |

empresas 8 mercado 7

|  |  |
| --- | --- |
| computing | 7 |
| red | 6 |
| proveedores | 6 |
| hecho | 6 |
| proveedor | 5 |
| nivel | 5 |
| mayor | 5 |
| acceso | 5 |
| servicio | 4 |
| seguridad | 4 |
| privado | 4 |
| ofrece | 4 |
| garantizar | 4 |
| aplicaciones | 4 |

Tabla 12. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del texto (c). Las palabras sombreadas son las keywords que se encontraron en este texto.

El texto (c) tiene keywords de Cloud Computing que están resumidas con su respectivo peso en la tabla 20.

|  |  |
| --- | --- |
| Palabras | Peso |
| cloud | 4,75925926 |
| modelo | 2,68518519 |
| computing | 4,55555556 |
| red | 4,40740741 |
| servicio | 4,37037037 |
| aplicaciones | 4,11111111 |
| W= | 24,8888889 |

Tabla 13. Tabla 20. Palabras identificadas en el texto (c) con sus respectivos pesos y peso total w.

Al obtener el peso del texto (c) se puede calcular la distancia que tiene respecto a la Temática de Cloud Computing usando la ecuación (3), el proceso se puede ver a continuación en la ecuación (25).

W = 24,88

1. 𝐷𝑇𝑐 = (𝑃𝑇 − 𝑊))

𝐷𝑇𝑐 = (79,14 − 24,88))

𝐷𝑇𝑐 = (54,26))

𝐷𝑇𝑐 =2944,14

Según el resultado de *DTc* obtenido anteriormente se puede observar que el texto (c) pertenece a la temática de Cloud Computing puesto que la distancia está dentro del rango definido anteriormente en la ecuación (4), como se ve a continuación:

1. 0 ≤ 𝐷𝑇𝑐 ≤ 3225,10

### 4.1.4 Prueba con el texto (d), título: Educación y Computadoras – Enseñanzas de América Latina

El texto (d) fue seleccionado de una URL de la web (Ver Anexo 5) y fue procesado con el contador de frecuencias (Ver Anexo 2).

Para la medir la distancia del texto (d) respecto a las keywords de Cloud Computing se toman las primeras 20 palabras de forma descendente comenzando por la palabra con mayor frecuencia como se ve en la Tabla 14.

|  |  |
| --- | --- |
| **Palabra** | **Cantidad** |

computadoras 12 uso 10 proyectos 7 programas 6 región 5

|  |  |
| --- | --- |
| software | 5 |
| educación | 5 |

estudio 5 mejorar 5 más 5 educativos 5 docentes 4 américa 4 resultados 4 pueden 4 latina 4

desarrollo 4

tecnologías 3 computadora 3 especialmente 3

Tabla 14. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del texto (d). Las palabras sombreadas son las keywords que se encontraron en este texto.

A continuación en la Tabla 15 se puede observar el resumen de las keywords de Cloud Computing obtenidas en el texto (d) con sus respectivos pesos.

|  |  |
| --- | --- |
| Palabras | Peso |
| software | 4,01 |
| *W*= | 4,01 |

Tabla 15. Palabras identificadas en el texto (d) con sus respectivos pesos y peso total *W.*

Al obtener el peso del texto (d) se puede calcular la distancia que tiene respecto a la Temática de Cloud Computing con la ecuación (3), el proceso se puede ver a continuación.

𝑊 = 4,01

(3) 𝐷𝑇𝑑 = (𝑃𝑇 − 𝑊))

𝐷𝑇𝑑 = (79,14 − 4,01))

𝐷𝑇𝑑 = 5644,51

La distancia obtenida del texto (d) respecto a la temática de Cloud Computing *DTd* No está dentro del rango definido anteriormente para las distancias mínimas y máximas que debe tener un texto que pertenezca a la temática de Cloud Computing, por lo cual se puede deducir que el texto (d) no hace referencia a esta temática.

 𝐷𝑇𝑑 ≤ 3225,10

### 4.1.5 Prueba con el texto (e), título: Comida Saludable

El texto (e) fue seleccionado de una URL de la web (Ver Anexo 5) y fue procesado con el contador de frecuencias (Ver Anexo 2).

Para la medida de la distancia respecto a las keywords de Cloud Computing se toman las primeras 20 palabras en orden descendente comenzando por la palabra de mayor frecuencia como se ve en la tabla 16.

|  |  |  |  |
| --- | --- | --- | --- |
| **Palabra** | | **Cantidad** | |
| peso | 10 |
| para | 10 |
| cuerpo | 8 |

dieta 8 bajar 7 ejercicio 7 nuestro 6 solo 5 ayudara 5 calorías 5 dietas 4 perder 4 esto 3 grasa 3 sola 2 por 2 mejor 2 quemar 2 sino 2 nos 2

Tabla 16. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del texto (e).

Este texto (e) que se refiere al tema de comida saludable y no contiene ninguna de las keywords de Cloud Computing por lo tanto su peso representativo para esta temática es:

𝑊 = 0

La distancia *DTe* que se tiene respecto a la temática de Cloud Computing es la que se calcula a continuación usando la ecuación (3):

1. 𝐷𝑇𝑒 = (𝑃𝑇 − 𝑊)) 𝐷𝑇𝑒 = (79,14 − 0))

𝐷𝑇𝑒 = (79,14))

𝐷𝑇𝑒 = 6263,13

La distancia obtenida del texto (e) respecto a la temática de Cloud Computing *DTe* No está dentro del rango definido para las distancias mínimas y máximas que debe tener un texto que pertenezca a la temática de Cloud Computing, por lo cual se puede deducir que el texto (e) no hace referencia a esta temática.

1. 0 ≤ 𝐷𝑇𝑒 ≤ 3225,10

### 4.1.6 Prueba con el texto (f), título: ¿Qué el Cloud Computing y qué beneficios aporta a la empresa?

El texto (f) fue seleccionado de una URL de la web (Ver Anexo 5) y fue procesado con el contador de frecuencias (Ver Anexo 2).

Para la medida de la distancia respecto a las keywords de Cloud Computing se toman las primeras 20 palabras de forma descendente comenzando con la palabra de mayor frecuencia como se ve en la Tabla 17.

|  |  |
| --- | --- |
| **Palabra** | **Ocurrencia** |
| servicio | 15 |
| computing | 13 |
| cloud | 13 |
| web | 11 |
| internet | 10 |
| empresas | 6 |

about 6 tienda 5 sobre 5 ejemplo 5 desarrollo 5 caso 5

trabajo 4

tendencias 4 disponer 4

puede 4

plataforma 4 online 4 nube 4 empresa 4

Tabla 17. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del Texto (f) de prueba. Las palabras sombreadas son las keywords que se encontraron en este texto.

El texto (f) tiene keywords de Cloud Computing que están resumidas con su respectivo peso en la Tabla 18.

|  |  |
| --- | --- |
| Palabras | Peso |
| computing | 4,55555556 |
| cloud | 4,75925926 |
| web | 4,33333333 |
| internet | 4,55555556 |
| servicio | 4,37037037 |
| plataforma | 3,92592593 |
| nube | 4,74074074 |
| W= | 31,24 |

Tabla 18. Palabras identificadas en el texto (f) con sus respectivos pesos y peso total W.

Al obtener el peso del texto (f) se puede calcular la distancia que tiene respecto a la Temática de Cloud Computing usando la ecuación (3), el proceso se puede ver a continuación.

𝑊 = 31,24

(3) 𝐷𝑇𝑓 = (𝑃𝑇 − 𝑊))

𝐷𝑇𝑓 = (79,14 − 31,24))

𝐷𝑇𝑓 = (47,9))

𝐷𝑇𝑓 = 2294,42

Según el resultado de *DTf* obtenido anteriormente se puede observar que el texto (f) pertenece a la temática de Cloud Computing puesto que la distancia está dentro del rango definido anteriormente, como se ve a continuación:

 𝐷𝑇𝑓 ≤ 3225,10

### 4.1.7 Prueba con el texto (g), título: Calentamiento Global

El texto (g) fue seleccionado de una URL de la web (Ver Anexo 5) y se procesó con el contador de frecuencias (Ver Anexo 2).

Para la medida de la distancia respecto a las keywords de Cloud Computing se toman las primeras 20 palabras en orden descendente comenzando por la palabra de mayor frecuencia como se ve en la Tabla 19.

|  |  |
| --- | --- |
| **Palabra** | **Ocurrencias** |

calentamiento 27

contaminación 19

energía 17

automóviles 12

disminuir 11

carbono 11

año 11

causa 10

bióxido 10

vehículos 8

podemos 8

está 8

tecnologías 8

estados 8

plantas 8

emisiones 7

electricidad 7

unidos 7

generadoras 7 cada 6

Tabla 19. Primeras 15 palabras y sus ocurrencias en orden mayor a menor del texto (f).

El texto (g) hace referencia al Calentamiento Global y no contiene ninguna de las keywords de Cloud Computing por lo tanto su peso representativo para esta temática es:

𝑊 = 0

Al obtener el peso del texto (g) se puede calcular la distancia que tiene respecto a la Temática de Cloud Computing usando la ecuación (3), el proceso se puede ver a continuación.

(23) 𝐷𝑇𝑔 = (𝑃𝑇 − 𝑊))

𝐷𝑇𝑔 = (79,14 − 0))

𝐷𝑇𝑔 = 6263,13

Se puede observar por el resultado de la distancia del texto (g) que este no pertenece a la temática de Cloud Computing puesto que no está en el rango señalado en la ecuación (4):

(4) 0 ≤ 𝐷𝑇𝑔 ≤ 3225,10

### 4.1.8 Prueba con el texto (h), título: Cuentos Vargas Llosa

El texto (h) fue seleccionado de una URL de la web (Ver Anexo 5) y se procesó con el contador de frecuencias (Ver Anexo 2).

Para la medida de la distancia respecto a las keywords de Cloud Computing se toman las primeras 20 palabras en orden descendente comenzando por la palabra de mayor frecuencia como se ve en la Tabla 20.

|  |  |
| --- | --- |
| **Palabra** | **Ocurrencias** |

cuento 18

como 15

las 15

para 14

trama 14

este 13

miguel 13

sus 12

principal 11

pero 11

hay 10

flora 10

esta 9

lector 9

vargas 9

primera 8

sin 8

leonidas 8

llosa 8

quien 7

Tabla 20. Primeras 15 palabras y sus ocurrencias en orden mayor a menor del texto (h).

El texto (h) es referente a un cuento del escritor Mario Vargas Llosa y no contiene ninguna de las keywords de Cloud Computing por lo tanto su peso

representativo para esta temática es:

Este se refiera al Calentamiento Global y no tiene ninguna de las keywords de

Cloud Computing por lo tanto:

𝑊 = 0

Al obtener el peso del texto (h) se puede calcular la distancia que tiene respecto a la Temática de Cloud Computing usando la ecuación (3), el proceso se puede ver a continuación.

(3) 𝐷𝑇ℎ = (𝑃𝑇 − 𝑊))

𝐷𝑇ℎ = (79,14 − 0))

𝐷𝑇ℎ = 6263,13

La distancia obtenida del texto (h) respecto a la temática de Cloud Computing *DTh n*o está dentro del rango definido anteriormente para las distancias mínimas y máximas que debe tener un texto que pertenezca a la temática de Cloud Computing, por lo cual se puede deducir que el texto (h) no hace referencia a esta temática.



### 4.1.9 Prueba con el texto (i), título: Espantos De Agosto- Garcia Marquez

El texto (i) fue seleccionado de una URL de la web (Ver Anexo 5) y se procesó con el contador de frecuencias (Ver Anexo 2).

Para la medida de la distancia respecto a las keywords de Cloud Computing se toman las primeras 20 palabras en orden descendente comenzando por la palabra de mayor frecuencia como se ve en la Tabla 21.

|  |  |  |  |
| --- | --- | --- | --- |
| Palabra | | Ocurrencias | |
| cuento | 18 |
| como | 15 |
| para | 14 |
| trama | 14 |
| este | 13 |
| miguel | 13 |
| principal | 11 |
| pero | 11 |
| flora | 10 |
| esta | 9 |
| lector | 9 |
| vargas | 9 |
| primera | 8 |
| leonidas | 8 |
| llosa | 8 |
| quien | 7 |
| aquí | 7 |
| hasta | 7 |
| dijo | 7 |
| novela | 6 |

Tabla 21. Primeras 15 palabras y sus ocurrencias en orden mayor a menor del texto (i).

El texto (i) es referente a un cuento del escritor Gabriel García Márquez y no contiene ninguna de las keywords de Cloud Computing por lo tanto su peso

representativo para esta temática es:

𝑊 = 0

Al obtener el peso del texto (i) se puede calcular la distancia que tiene respecto a la Temática de Cloud Computing usando la ecuación (3), el proceso se puede ver a

continuación.

1. 𝐷𝑇𝑖 = (𝑃𝑇 − 𝑊))

𝐷𝑇𝑖 = (79,14 − 0))

𝐷𝑇𝑖 = 6263,13

La distancia obtenida del texto (i) respecto a la temática de Cloud Computing *DTi n*o está dentro del rango definido anteriormente para las distancias mínimas y máximas que debe tener un texto que pertenezca a la temática de Cloud Computing, por lo cual se puede deducir que el texto (i) no hace referencia a esta temática.

1. 0 ≤ 𝐷𝑇𝑖 ≤ 3225,10

### 4.1.10 Prueba con el texto (j), título: Las empresas hablan sobre La Nube

El texto (j) fue seleccionado de una URL de la web (Ver Anexo 5) y fue procesado con el contador de frecuencias (Ver Anexo 2).

Para la medida de la distancia respecto a las keywords de Cloud Computing se toman las primeras 20 palabras de forma descendente comenzando con la palabra de mayor frecuencia como se ve en la Tabla 22.

|  |  |
| --- | --- |
| Palabra | Ocurrencia |
| nube | 9 |
| servicio | 7 |
| empresas | 6 |
| aplicaciones | 6 |
| web | 5 |
| servidores | 5 |

ofrecen 5 empresa 5 argentina 5 nosotros 4 microsoft 4 lanacion 4 google 4 capacidad 4 visitas 3 viene 3 usuarios 3

|  |  |
| --- | --- |
| software | 3 |
| nueva | 3 |
| marca | 3 |

Tabla 22. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del Texto (j). Las palabras sombreadas son las keywords que se encontraron en esté texto.

El texto (j) tiene keywords de Cloud Computing que están resumidas con su respectivo peso en la Tabla 23.

|  |  |
| --- | --- |
| Palabras | Peso |
| nube | 4,74074074 |
| aplicaciones | 4,11111111 |
| web | 4,33333333 |
| servicio | 4,37037037 |
| software nueva | 4,01851852 |
| *W=* |  |

Tabla 23. Palabras identificadas en el texto (j) con sus respectivos pesos y peso total w.

Al obtener el peso del texto (j) se puede calcular la distancia que tiene respecto a la Temática de Cloud Computing usando la ecuación (3), el proceso se puede ver a continuación.

𝑊 = 24,44

*(3)* 𝐷𝑇𝑗 = (𝑃𝑇 − 𝑊))

𝐷𝑇𝑗 = (79,14 − 24,44))

𝐷𝑇𝑗 = (54,7))

##### 𝐷𝑇𝑗 = *2992,09*

Según el resultado de *DTj* obtenido anteriormente se puede observar que el texto (j) pertenece a la temática de Cloud Computing puesto que la distancia está dentro del rango definido anteriormente, como se ve a continuación:

 𝐷𝑇𝑗 ≤ 3225,10

### 4.1.11 Prueba con el texto (k), título: El Tiempo En Campana

El texto (k) fue seleccionado de una URL de la web (Ver Anexo 5) y fue procesado con el contador de frecuencias (Ver Anexo 2).

Para la medida de la distancia respecto a las keywords de Cloud Computing se toman las primeras 20 palabras de forma descendente comenzando con la palabra de mayor frecuencia como se ve en la Tabla 23.

|  |  |
| --- | --- |
| Palabra | Ocurrencia |

grados 18 cielo 12

|  |  |
| --- | --- |
| viento | 11 |
| temperatura | 10 |
| mañana | 10 |
| día | 8 |
| tarde | 8 |
| estará | 8 |
| máxima | 8 |
| pronóstico | 7 |
| mínima | 6 |
| entre | 6 |
| sol | 6 |

|  |  |
| --- | --- |
| nube | 6 |

|  |  |
| --- | --- |
| hoy | 5 |
| débil | 5 |
| soplará | 5 |
| noche | 5 |
| poco | 5 |
| frío | 5 |

Tabla 244. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del Texto (k). La palabra

que aparece sombreada es una keyword que se encontró en esté texto l w.

El texto (k) tiene una de las keywords de Cloud Computing que está detallada con su respectivo peso en la Tabla 24.

|  |  |
| --- | --- |
| Palabras | Peso |
| nube |  |
| *W=* | 4,74074074 |

Tabla 255. Palabras identificadas en el texto (k) con sus respectivos pesos y peso total w.

Al obtener el peso del texto (k) se puede calcular la distancia que tiene respecto a la Temática de Cloud Computing usando la ecuación (3), el proceso se puede ver a continuación.

𝑊 = 4,74

*(3)* 𝐷𝑇𝑘 = (𝑃𝑇 − 𝑊))

𝐷𝑇𝑘 = (79,14 − 4,74))

𝐷𝑇𝑘 = (74,26))

##### 𝐷𝑇𝑘 = *5514,54*

Según el resultado de *DTk* obtenido anteriormente se puede observar que el texto (k) no pertenece a la temática de Cloud Computing puesto que la distancia no está dentro del rango definido anteriormente, como se ve a continuación:

 𝐷𝑇𝑘 ≤ 3225,10

### 4.1.12 Prueba con el texto (L), título: Formacion de las Nubes

El texto (L) fue seleccionado de una URL de la web (Ver Anexo 5) y fue procesado con el contador de frecuencias (Ver Anexo 2).

|  |  |
| --- | --- |
| Palabra | Ocurrencia |
| nube | 53 |

|  |  |
| --- | --- |
| agua | 26 |
| aire | 21 |
| forma | 21 |
| hielo | 17 |
| cristales | 16 |
| capa | 12 |
| gotas | 12 |
| del | 12 |
| puede | 12 |
| este | 11 |
| base | 11 |
| aspecto | 11 |
| pueden | 10 |
| tienen | 10 |

Para la medida de la distancia respecto a las keywords de Cloud Computing se toman las primeras 20 palabras de forma descendente comenzando con la palabra de mayor frecuencia como se ve en la Tabla 25.

|  |  |
| --- | --- |
| constituido | 10 |
| forman | 10 |
| gris | 10 |
| parte | 9 |
| entre | 9 |

Tabla 266. Primeras 20 palabras y sus ocurrencias en orden mayor a menor del Texto (L). La palabra

que aparece sombreada es una keyword que se encontró en esté texto.

El texto (L) tiene una de las keywords de Cloud Computing que está detallada con su respectivo peso en la Tabla 26.

|  |  |
| --- | --- |
| Palabras | Peso |
| nube |  |
| *W=* | 4,74074074 |

Tabla 277. Palabras identificadas en el texto (L) con sus respectivos pesos y peso total w.

Al obtener el peso del texto (L) se puede calcular la distancia que tiene respecto a la Temática de Cloud Computing usando la ecuación (3), el proceso se puede ver a continuación.

𝑊 = 4,74

*(3)* 𝐷𝑇𝑙 = (𝑃𝑇 − 𝑊))

𝐷𝑇𝑙 = (79,14 − 4,74))

𝐷𝑇𝑙 = (74,26))

##### 𝐷𝑇𝑙 = *5514,54*

Según el resultado de *DTl* obtenido anteriormente se puede observar que el texto (L) no pertenece a la temática de Cloud Computing puesto que la distancia no está dentro del rango definido anteriormente, como se ve a continuación:

 𝐷𝑇𝑘 ≤ 3225,10

## 4.2 ANALISIS DE RESULTADOS

Luego de realizar las pruebas de clasificación de textos utilizando la técnica propuesta se puede clasificar cada texto según el peso que se obtuvo respecto a las keywords de Cloud Computing. En la Tabla 24 se observa el resumen de los resultados obtenidos con la relación de cada texto con su respectivo peso, por otro lado en la figura 10 se pueden observar gráficamente las distancias obtenidas de cada texto respecto a Cloud Computing[[8]](#footnote-8).

|  |  |  |  |
| --- | --- | --- | --- |
| Texto Analizado | Distancias Peso w | | |
| a. Datacenters de Nube Argentina de Telecom, socios estratégicos del resguardo de información de las empresas. | 2869,74 25,57 | | |
| b. Cloud Computing | 1695,79 37,96 | | |
| c. Tendencias de los servicios de Cloud Computing en el mercado Español. | 2944,14 24,88 | | |
| d. Educación y Computadoras – Enseñanzas de América Latina | | 5644,51 | 4,01 |
| e. Comida Saludable | | 6263,13 | 0 |
| f. ¿Qué el Cloud Computing y qué beneficios aporta a la empresa? | | 2294,42 | 31,24 |
| g. Calentamiento Global | | 6263,13 | 0 |
| h. Cuentos Vargas Llosa | | 6263,13 | 0 |
| i. Espantos De Agosto- Garcia Márquez | | 6263,13 | 0 |
| j. Las empresas hablan sobre La Nube | | 2992,09 | 24,44 |
| k. El Tiempo En Campana | | 5514,54 | 4,74 |

L. Formación de las Nubes 5514,54 4,74

Tabla 288. Después del análisis se obtuvo que los textos que se muestran sombreados (a, b, c,f y j)

hacen referencia a temática de Cloud Computing).

Con la técnica propuesta se lograron identificar los textos que hacen referencia a la temática de Cloud Computing (a, b, c, f y j) puesto que su distancia se encuentra

dentro del rango definido así como se puede observar en la figura 10.

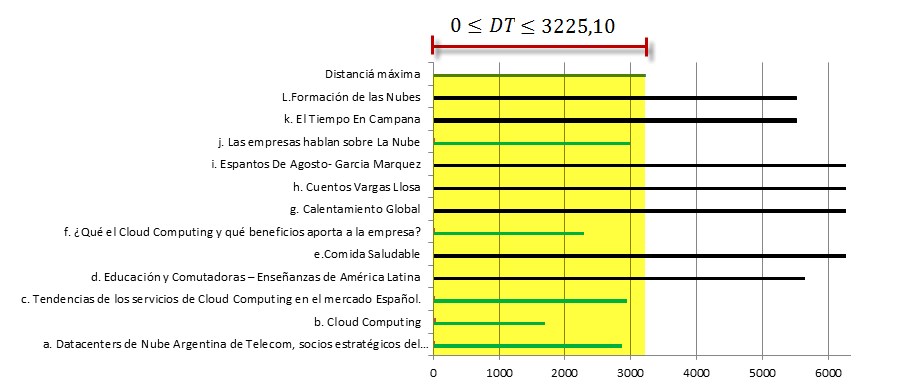


Figura 10. Proceso de prueba de distancia con 12 textos, de los cuales se obtuvo que 5 hacen

referencia la temática de Cloud Computing.

En la Figura 10 se puede observar que los textos que tienen su distancia dentro de la parte sombreada están dentro del rango de distancias que pertenecen a la temática de Cloud Computing, por otro lado los que superan este rango son los que se alejan de dicha temática.

La técnica propuesta permitió identificar de 12 textos desconocidos 5 que hacen referencia a la temática de Cloud Computing. El resumen del proceso realizado se puede observar en la figura 11, en donde se pueden identificar las etapas por las que pasaron cada uno de los textos hasta llegar a la conclusión de poder clasificar los textos que corresponden a la temática Cloud Computing.

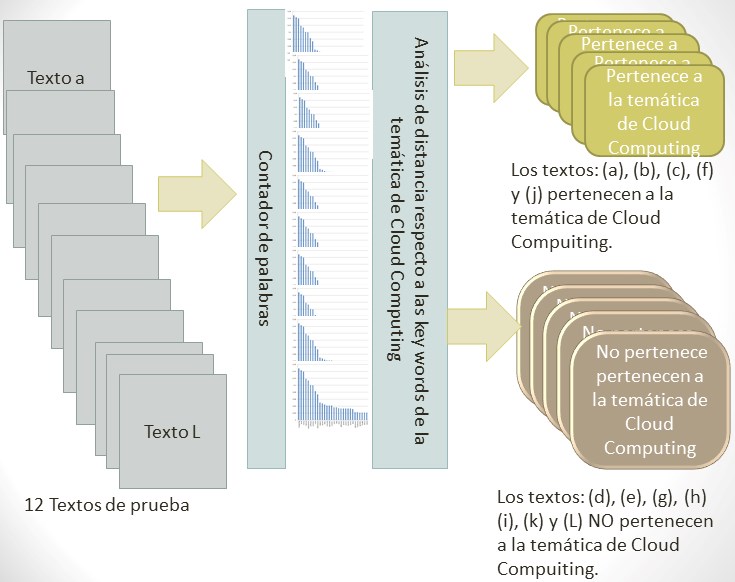


Figura 11. Proceso de prueba de distancia con 12 textos, de los cuales se determinó que 5 pertenece a la temática de Cloud Computing.

Luego de poner en práctica la técnica de clasificación propuesta en esta tesis basada en resultados estadísticos y ocurrencias de palabras relevantes respecto a una temática inicial conocida, se puede decir que la técnica es eficiente y se puede usar para la clasificación de textos de una temática en particular.

# CAPÍTULO 5

# CONCLUSIONES Y TRABAJOS FUTUROS

## 5.1 CONCLUSIONES

Se realizó el procesamiento del contenido de tres textos tomados de sitios Web de forma manual. Los tres textos hacían referencia a un mismo tema, a fin de estudiar el procedimiento automático más apropiado para lograr establecer las keywords que permitan representar dicho tema. Los resultados fueron validados con una encuesta que permitió establecer cotas superiores e inferiores de frecuencia de aparición de ciertas palabras. Con el fin de establecer la representatividad de las keywords se trabajó con ponderaciones, tomando como parámetros los pesos asignados a cada keyword.

Combinando la frecuencia y la ponderación, se obtuvo un escalafón de relevancias que puede aplicarse a cualquier texto a fin de definir si el mismo puede ser clasificado dentro de esta misma temática. Se propuso y se evaluó una distancia que tiene un texto cualquiera, respecto a la temática que hacen referencia las keywords.

Se logró clasificar de forma adecuada doce nuevos textos tomados de la web, determinando su pertenencia a la temática representada por las keywords, usando la fórmula de distancia propuesta. Como resultado, se obtuvo que cinco de los textos pertenecen a la misma temática; los otros quedaron fuera del umbral de pesos establecido, no cumpliendo con la distancia mínima que permite inferir su pertenencia de la temática.

La técnica propuesta es útil para la clasificación de textos de acuerdo a un tema específico y con una distancia propia, siendo comparable con la técnica de clustering [23-24], la cual usa la distancia euclidea para determinar cercanía entre textos de acuerdo con los caracteres que tiene cada uno y agrupando en cada cluster textos similares, ya sea por su redacción o frecuencias de palabras, dependiendo del algoritmo con el que se implemente esta técnica.

Resulta útil implementar la técnica de clustering si se desean clasificar textos similares en grupos. Entre tanto, si se quiere tomar un grupo grande de textos y clasificarlos, seleccionando solo los que corresponden a un tema en particular, se puede hacer de forma directa con la técnica propuesta en esta tesis.

Mediante la implementación de la técnica propuesta, definiendo una distancia entre palabras referentes a una temática en particular, es posible establecer cotas automáticas que permitan ser usadas como indicadores de contenidos temáticos. Esto, combinado con la extracción automática de palabras relevantes, permite concluir que esta metodología es reproducible para textos nuevos y aplicable de manera automática.

## 5.2 TRABAJOS FUTUROS

Queda pendiente para futuros trabajos la evaluación estadística de la precisión, luego de automatizar la etapa de extracción de contenidos textuales de diferentes URLs, con la posibilidad de establecer funcionalidades adicionales tales como la posibilidad de identificar de manera aleatoria los sitios en estudio o definirlos según el interés del usuario. Por otro lado, para una etapa posterior a la clasificación, queda por realizar el proceso de análisis de texto semántico, usando raíces de palabras para establecer tendencias, opiniones, o puntos de vista positivos, negativos o neutros sobre algún tema.

Con la implementación de la técnica propuesta en éste trabajo se pude dar inicio a nuevas estrategias que permitan identificar el autor de un texto por la forma y conectores que usa al escribir, o por ejemplo identificar si un texto fue tomado de una sala de chat o de un texto científico, determinando patrones que se encuentran en cada uno de estos tipos te textos. En futuros trabajos se podrá diseñar un sistema autómata en el cual la información sea tomada directamente de la web y se procese clasificando contenidos textuales por un tema de interés.

Para posteriores investigaciones, se puede realizar una comparación entre la técnica propuesta en esta tesis que usa una distancia propia entre palabras, con algoritmos que implementen distancias diferentes, como la distancia normalizada de google o la levenshtein [25-26]. Sería interesante validar los resultados de cada técnica para conocer la eficiencia de una respecto a la otra, así como también combinarlas y usarlas en diferentes instancias secuenciales de análisis, usando alguna de estas para la clasificación por temáticas y otra para la identificación de opiniones positivas, negativas o neutras de los textos.

# REFERENCIAS

1. Davenport, T.H. e L. Prusak. Working Knowledge: how Organizations Manage what they Know. Boston: Harvard Business Scholl Press, 1998
2. EMEC Corporation (2013), Revelan gran brecha en Big Data: menos del 1% de la información digital es analizada, Recuperado de http://diarioti.com/revelan-granbrecha-en-big-data-menos-del-1-de-la-informacion-digital-esanalizada/33564?lang=es.
3. Codd, E. F., A (1970) Relational Model of Data for Large Shared Data Banks,

ACM

1. Codd, E. F., Futher Normalization of the Data Base Relational Model, PrenticeHall
2. Seppo Puuronen, Mykola Pechenizkiy, Data Mining Researcher, Who is Your

Customer? Some Issues Inspired by the Information Systems Field, IEEE

10.1109/DEXA.2006.81, 2006, pp. 4-5.

1. Jiang Shian, et alii, Data Types Generalization for Data Mining Algorithm, IEEE

10.1109/ICSMC.1999.823352, 1999, p 4-6

1. Oracle Data base 11g (2012). Oracle Database 11g para Data Warehousing e

Inteligencia de Negocios, Recuperado de: http://www.oracle.com/technetwork/es/documentation/317492-esa.pdf

1. Oracle Data base 12g (2013). Oracle Data Base New Features Guide 12c Release,

Recuperado de: http://docs.oracle.com/cd/E16655\_01/server.121/e17906.pdf

1. Nestorov, Svetlozar, et alii, Ad-Hoc Association-Rule Mining within the Data

Warehouse. IEEE 01174605, Proceedings of the 36th Hawaii International

Conference on System Sciences (HICSS’03) 0-7695-1874-5/03 © 2002 IEEE].

1. Witold Pedrycz y José Valente de Oliveira, (2007). Soft Cluster Ensembles. En,

Advances in Fuzzy Clustering and its Applications, Wiley & Sons Ltd, ISBN 978-0470-02760-8.

1. Chen, Yan, et alii, (2009), General Data Mining Model System Based on Sample Data Division, IEEE 10.1109/KAM.2009.142.
2. Muthukumar, A. y Nadarajan, R, (2007), TDML: A Data Mining Language for transaction databases, Fourth International Conference on Fuzzy Systems and

Knowledge Discovery IEEE 10.1109/FSKD.2007.558

1. Khan, Danish, (2008), CAKE - Classifying, Associating & Knowledge

DiscovEry An Approach for Distributed Data Mining (DDM) Using PArallel Data

Mining Agents (PADMAs), IEEE 10.1109/WIIAT.2008.236

1. Cieslak, David A, et alii, (2006) Troubleshooting Distributed Systems via Data

Mining, IEEE 10.1109/HPDC.2006.1652163

1. Nirkhi, Smita M, (2010), Potential use of Artificial Neural Network in Data

Mining, IEEE 10.1109/ICCAE.2010.5451537

1. Eladio Blanco López, et alii, (2011), Análisis automático de emociones en la Red

Internacional e-Culturas, Revista Electrónica de Investigación y Docencia (REID),(

ISSN-e 1989-2446, Nº. 5, 2011), p 53-68

1. Aldahawi, Hanaa A. ; Allen, Stuart M. (2013),Twitter Mining in the Oil

Business: A Sentiment Analysis Approach, IEEE Digital Object Identifier:

10.1109/CGC.2013.101

1. Servia-Rodriguez, Sandra ; Fernandez, et alii (2013),Tag Clustering Algorithms for Mining Twitter Users' Interests, IEEE Digital Object Identifier:

10.1109/SocialCom.2013.102

1. AMD White Paper: Big Data (2010), Big Data It’s not just for Google Any

##### More. Recuperado de http://sites.amd.com/us/Documents/Big\_Data\_Whitepaper.pdf)

1. Hong Hu, Jiuyong Li, Ashley Plank (2006) A Comparative Study of

Classification ethods for Microarray Data Analysis, published in CRPIT, Vol. 61

1. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data

Classification (2013), International Journal Of Computer Science And Applications

V6, ISSN: 0974-1011

1. Fabrizio Sebastiani (2002), Machine Learning in Automated Text Categorization Jounal ACM Computing Surveys.
2. Haojun Sun,Shantou Univ, et alii (2008), A Document Clustering Method Based on Hierarchical Algorithm with Model Clustering, IEEE Digital Object Identifier :

10.1109/WAINA.2008.45

1. Singh, V.K., Tiwari, N.,Garg, S. (2011), Document Clustering Using K-Means,

Heuristic K-Means and Fuzzy C-Means, IEEE Digital Objectidentifier:

10.1109/CICN.2011.62

1. Balahur, Alexandra, Alicante, et alii (2008), A feature dependent method for opinion mining and classification, IEEE Digital Object Identifier :

10.1109/NLPKE.2008.4906796

1. Bin Shi, Liying fang (2009), Classification of Semantic Documents Based on WordNet, IEEE Digital Object Identifier: 10.1109/EEEE.2009.15

# ANEXOS

#### ANEXO 1

En esté anexo se encuentran las fuentes de las cuales se tomaron los tres diferentes textos que hacen referencia al tema de Cloud Computing. Estos textos fueron analizados para obtener la keywords referentes a esta temática.

Título del primer texto analizado: Definición de Cloud computing Disponible en:

http://www.e-conomic.es/programa/glosario/definicion-cloud-computing

Título del primer texto analizado: Cloud computing Disponible en:

Fuente del segundo texto analizado:

http://www.tecnologiahechapalabra.com/tecnologia/glosario\_tecnico/articulo.asp?i=3024

Título del primer texto analizado: Cloud computing Disponible en:

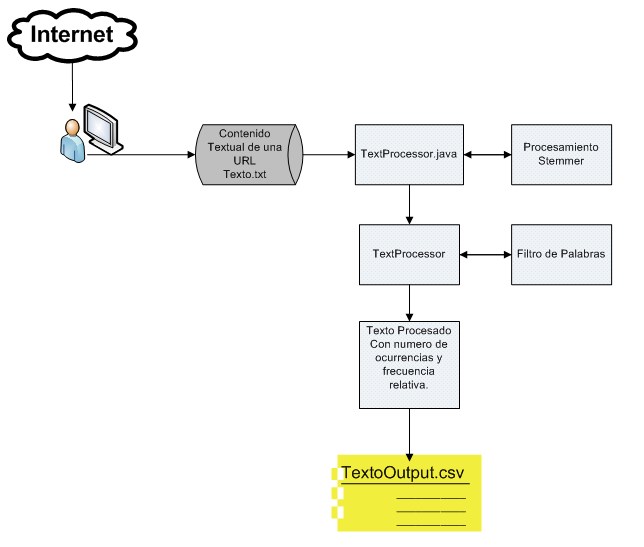
http://www.ceditec.etsit.upm.es/index.php?option=com\_content&view=article&id=21808&It emid=1439&lang=en

##### ANEXO 2

Código programa realizado en Java para contar las palabras de los textos, este programa lo desarrollamos los integrantes del proyecto jazz del grupo de investigación de la Universidad de Palermo, es de libre consulta y uso con fines investigativos.

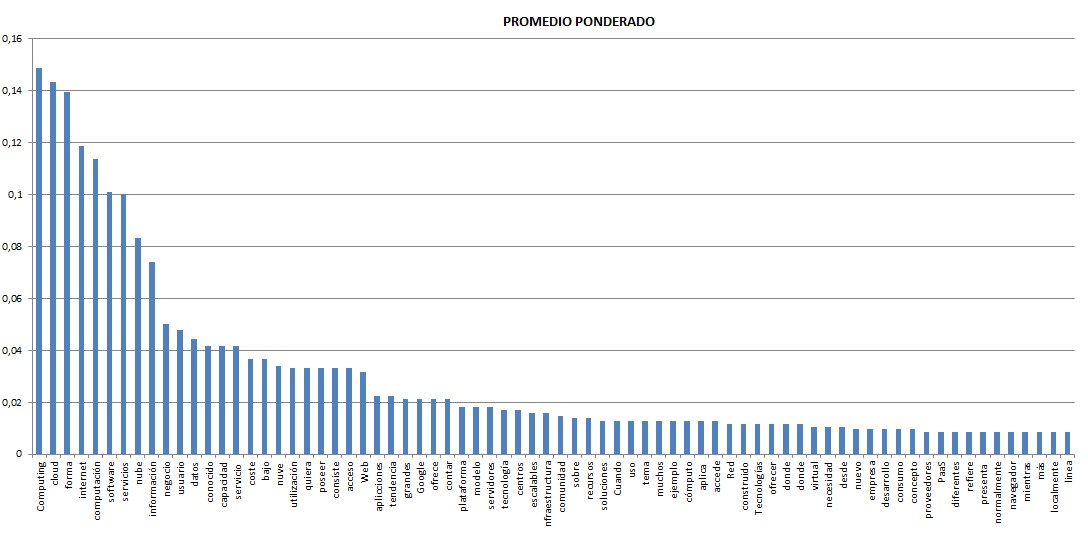
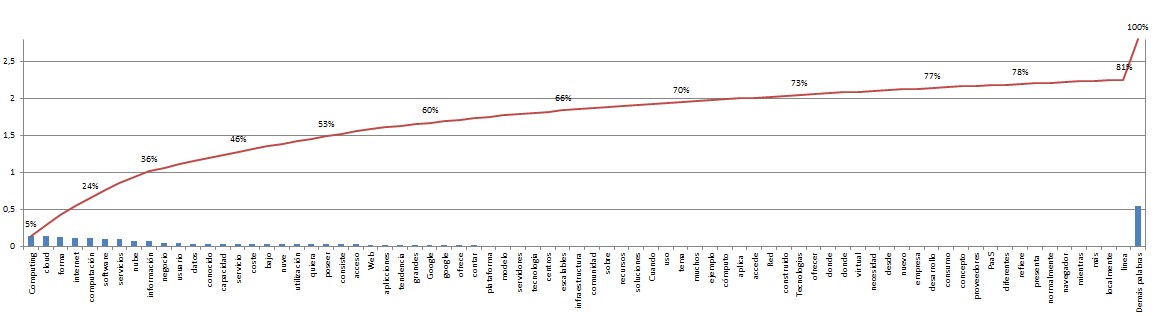
Consultar código fuente en:

https://www.dropbox.com/sh/vz3rhtccrm5d34f/N13ePiVKzI



##### ANEXO 3

Graficas de promedio ponderado de todas las palabras contendidas en los tres textos analizados y análisis de Pareto.



##### ANEXO 4

**La encuesta está disponible en el siguiente link:**

https://docs.google.com/forms/d/1CqQV\_63zRu3rwC009Oa9rGa8CCFO6i\_ZqZcU17br1Z0/ viewform

Contenido de la encuesta:

Interpretación de textos

La presente encuesta está dirigida a los miembros del Ai Group de la Universidad de Palermo y otras personas que quieran colaborar con el desarrollo de una investigación y tesis sobre de Opinion Mining que será presentada para optar por el grado de Magister en Tecnologías de La Información.

La presente encuesta la puede resolver si lee previamente los tres textos que se encuentran en el siguiente link:

https://sites.google.com/site/textos123tesis/home

Por favor léalos con atención y y luego responda a conciencia.

Mil gracias por su tiempo y colaboración.

Realizado por: Lic. Fabian Riaño Santiesteban

Principio del formulario

1. **Seleccione un titulo que más se ajuste a los tres textos que leyó previamente. \***



INFORMACIÓN DIGITAL



CLOUD COMPUTING



NUEVA ERA DEL INTERNET



SOLUCION INFORMÁTICA

1. **De la siguiente lista de palabras seleccione un valor de 0 a 5 según la importancia que cada palabra agrega al contenido delos TRES textos leídos previamente. \***

Aplicaciones

Cloud

Computing

Datos

Escalables Tendencia

Información

Internet

Modelo

Nueva

Plataforma

Red

Servicio

Servicios Software

Usuario

Web

Nube

Computación

Negocio **ANEXO 5**

En esté anexo se encuentran las fuentes de las cuales se tomaron los 12 textos con los cuales se probó

1. Datacenters de Nube Argentina de Telecom, socios estratégicos del resguardo de información de las empresas. Disponible en:

http://fortunaweb.com.ar/2012-11-08-108223-datacenters-de-nube-argentina-de- telecomsocios-estrategicos-del-resguardo-de-informacion/

1. Cloud computing , texto de prueba. Disponible en:

http://www.belt.es/noticiasmdb/home2\_noticias.asp?id=6995

1. Tendencias de los servicios de Cloud Computing en el mercado Español. Disponibl en:

http://www.revistacloudcomputing.com/2012/04/tendencias-de-los-servicios-de- cloudcomputing-en-el-mercado-espanol/

1. Educación y computadoras – Enseñanzas de América Latina. Disponible en:

http://www.iadb.org/es/noticias/articulos/2011-04-11/educacion-y-computadoras-en-americalatina,9334.html

1. Comida Saludable. Disponible en:

http://comidasaludable.net/que-es-mas-recomendable-usar-dietas-o-ejercicio-para-bajar-depeso/

1. ¿Qué es el cloud computing y qué beneficios aporta a la empresa? .Disponible en:

http://tendenciasweb.about.com/od/tendencias-web/a/Qu-E-Es-El-Cloud-Computing-Y-Qu-EBeneficios-Aporta-A-La-Empresa.htm

1. Calentamiento Global. Disponible en:

http://www.nrdc.org/laondaverde/globalwarming/f101.asp#1

1. . Cuentos Vargas Llosa. Disponible en:

http://www.roland557.com/ficcion/vargas\_llosa.htm

1. Espantos De Agosto- Garcia Marquez. Disponible en:

http://www.ciudadseva.com/textos/cuentos/esp/ggm/espantos\_de\_agosto.htm

1. Las empresas hablan sobre La Nube. Disponible en:

http://www.lanacion.com.ar/1208338-las-empresas-hablan-sobre-la-nube

1. El Tiempo En Campana

http://tiempocampana.blogspot.com.ar/2014/03/templado-con-sol-y-algunas-nubes\_17.html

L. Formación de las Nubes

http://www.alboxclima.com/meteorologia/clasificacion.htm

1. Una base de datos normalizada tiene relaciones a nivel lógico entre las tablas, por otro lado están las bases de datos no normalizadas que cuentan con una gran cantidad de datos redundantes y ocupan generalmente más espacio de almacenamiento, pero son ideales para la obtención de relaciones entre datos, siendo útiles para procesos de *Business Intelligene* BI. [↑](#footnote-ref-1)
2. Opinion Mining hace es una técnica que permite analizar las opiniones de las personas respecto a un mimo tema. Ejemplo:Análizar los Topics Trends de Twitter [↑](#footnote-ref-2)
3. El data entry es un ejemplo de inserción manual de información puesto que son hechos por personas; por otro lado, existen sistemas que permiten el almacenamiento de información digital de forma automática, como triggers, logs de sistemas, web services, entre otros. [↑](#footnote-ref-3)
4. Data mining es un concepto usado para referirse al procesamiento de alto volúmenes de información, proceso que se realiza gracias a que hoy en día es posible la recolección masiva de datos, se dispone de computadoras de alto procesamiento y algoritmos de Data Mining. 5 Web Mining es una de las técnicas de data mining usadas para extraer información de la web, determinando patrones entre los datos y transformándolos en conocimiento. [↑](#footnote-ref-4)
5. El espacio euclideo consta mínimamente de dos dimensiones (x, y) en el cual se pueden identificar puntos coordenados, por lo que es posible hallar la distancia entre estos de manera vectorial. [↑](#footnote-ref-5)
6. La medida que permite clasificar a un texto dentro de la temática de Cloud Computing o descartarlo se hace con los pesos (entre más peso tenga la palabra, más relevante es) asignados a cada Keyword y la distancia propuesta en esta tesis. [↑](#footnote-ref-6)
7. Se propone una distancia normalizada (No presenta valores negativos) como se ve en la ecuación 4 teniendo en cuenta los valores *PT* y *w* [↑](#footnote-ref-7)
8. Cuanto mayor sea al peso de un Texto quiere decir que es más relevante y va a ser más cercano a la Temática de Cloud computing. [↑](#footnote-ref-8)