Artículo **Revista de Ingeniería Innovativa**
Diciembre 2020 Vol.4 No.15 37-44

Evaluación de un clasificador de textos digitales basado en el contenido semántico a través de ontologías

Evaluation of a digital text classifier based on semantic content through ontologies

HERNÁNDEZ-GARCÍA, Héctor Daniel†*, NAVARRETE-ARIAS, Dulce J., PÉREZ-BAUTISTA, Mario y PAREDES-REYES, Eliud

Instituto Tecnológico Superior del Occidente del Estado de Hidalgo, División de Ingeniería en Sistemas Computacionales, México.

ID 1er Autor: *Héctor Daniel, Hernández-García* / **ORC ID:** 0000-0001-5261-8353, **Researcher ID Thomson:** P-4823-2018, **CVU CONACYT ID:** 208146

ID 1er Coautor: *Dulce J., Navarrete-Arias* / **ORC ID:** 0000-0002-7915-068X, **CVU CONACYT ID:** 366071

ID 2do Coautor: *Mario, Pérez-Bautista* / **ORC ID:** 0000-0002-3260-906X, **CVU CONACYT ID:** 638669

ID 3er Coautor: *Eliud, Paredes-Reyes* / **ORC ID:** 0000-0003-4621-2589, **CVU CONACYT ID:** 638197

DOI: 10.35429/JOIE.2020.15.4.37.44 Recibido Julio 25, 2020; Aceptado Diciembre 30, 2020

Resumen

En la actualidad la generación de información a través de documentos de texto digitales ha incrementado exponencialmente, por lo que se tiene la necesidad de almacenar estos documentos en dispositivos de almacenamiento masivos como discos duros de alta capacidad, servidores de almacenamiento, la nube, entre otros. Sin embargo, el almacenamiento que se realiza carece de una organización temática, por lo que, realizar una búsqueda de información se vuelve complejo. Ante esta problemática, la presente publicación describe el desarrollo de un sistema que tiene como propósito clasificar un documento de texto digital basado en el contenido temático. Este sistema implementa ontologías para lograr una mejor clasificación al aprovechar sus características. El sistema se divide en cinco tareas: la primera, es implementar un autómata que realiza el conteo de palabras para crear un vector de frecuencias; la segunda tarea realiza una refinación en el vector de frecuencias para eliminar los conectores de oraciones y las preposiciones; la tercera tarea ordena el vector de la frecuencia más alta hasta las más baja; la cuarta tarea toma el conjunto más significativo del vector de frecuencias, al cual se le aplica la ontología de un dominio y busca la relación que tienen las palabras para determinar la temática del documento; y la quinta tarea consiste en organizar los documentos en una estructura de carpetas basado en los dominios identificados. El sistema se desarrolló con la metodología de desarrollo incremental. Para validar el funcionamiento del sistema se realizaron un conjunto de pruebas en un escenario controlado a fin de verificar la correcta clasificación de los documentos.

Ontologías, Clasificador de textos, Autómatas

Abstract

Nowadays, the generation of information through digital text documents has increased exponentially, so there is a need to store documents in mass storage devices such as high capacity hard discs, storage servers, the cloud and others. However, the storage that is carried out lacks a thematic organization, therefore, a search for information becomes complex. Given this problem, this publication describes the development of a system that has the purpose of classifying a digital text document based on the thematic content. This system implements ontologies to achieve a better classification by taking advantage of its characteristics. The system is divided into five tasks: the first is the implementation of a word count to create a frequency vector; The second task performs a refinement on the frequency vector to eliminate the sentence connectors and prepositions; the third task orders the vector from the highest to the lowest frequency; the fourth task takes the most significant set of frequencies vector, in which the ontology of a domain is applied and the relation that the words have to determine the thematic of the document is sought; and the fifth task is to organize the documents in a folder structure based on the identified domains. The system was developed with the incremental development methodology. To validate the operation of the system, a set of tests was carried out in a controlled scenario in order to verify the correct classification of the documents.

Ontologies, Text classifier, Automata

Citación: HERNÁNDEZ-GARCÍA, Héctor Daniel, NAVARRETE-ARIAS, Dulce J., PÉREZ-BAUTISTA, Mario y PAREDES-REYES, Eliud. Evaluación de un clasificador de textos digitales basado en el contenido semántico a través de ontologías. Revista de Ingeniería Innovativa. 2020. 4-15:37-44.

*Correspondencia al Autor (Correo Electrónico: hhernandez@itsoeh.edu.mx)
† Investigador contribuyendo como primer autor.

Introducción

La actual era de la información se caracteriza por una expansión extraordinaria de datos que son generados y almacenados de manera dispersa en distintos puntos conectados en una red tan amplia como Internet. En este mar de documentos se oculta información que puede o no ser relevante para los diversos dominios del conocimiento humano.

Una situación como ésta ha dado origen a uno de los más grandes retos que se enfrenta la sociedad actual, el cual consiste, principalmente, en la definición de nuevos modelos, estrategias y herramientas que le permitan de manera eficiente describir, clasificar, almacenar y encontrar información relevante dentro de este mar de documentos. Se han hecho estudios que demuestran que la mayoría de información que se encuentra almacenada de manera dispersa en distintos sitios conectados a través de Internet es del tipo "no estructurada" (Wilks & Catizone, 2000), es decir de tipo textual.

Este hecho enfatizó la relevancia que tuvo la aparición de la Web (Berners-Lee, Fielding, & Frystyk, 1996) con el protocolo HTTP como un mecanismo para acceder a documentos cuya información se encuentra representada por páginas HTML (Specification, 1999). Sin embargo, a pesar de contar con esta tecnología tenemos el problema de la clasificación de documentos puesto que hasta el momento el almacenamiento de documento textuales se limita a ser almacenados en un dispositivo identificándolo únicamente con un nombre y una extensión.

Este mecanismo presenta un problema al momento de buscar información ya que al solo poder ser identificados con un nombre y una extensión no es posible determinar el contenido temático que contiene el documento ya que en la mayoría de veces los usuarios colocan nombres a sus archivos los cuales no tienen ninguna relación con el contenido temáticos de éstos, estas acciones conllevan a que al momento de buscar cierta información es necesario tener que abrir el documento para determinar el contenido temático de éste.

En este trabajo se presenta una herramienta que implementa vectores de frecuencias, sobre las palabras encontradas en el documento, y ontologías que determinan la relación existente entre estas palabras para determinar el dominio o tema. Las ontologías de acuerdo con Tom Gruber (Gruber, 1995) "es una descripción, como una especificación formal de un programa, de los conceptos y relaciones que pueden formalmente existir para un agente o comunidad de agentes". En otras palabras, es una estructura jerárquica que describe instancias, conceptos, atributos y relaciones sobre un dominio o tema en específico.

Actualmente, existen diferentes trabajos que realizan la clasificación de documentos de texto digitales como StringTagger que implementa un clasificador bayesiano simple para la clasificación de cadenas de texto (Python, 2018), Método de Clasificación Automática de Textos basado en Palabras Claves utilizando Información Semántica, es una aplicación a Historias Clínicas que aprovecha la información semántica existente extrayendo 3 o 5 palabras claves de cada tipo de enfermedad para clasificar (Lopez Condori, 2014), y Regulón DB que clasifica texto mediante atributos probabilísticos de coocurrencia de palabras (Sanchez Vega, 2012). La diferencia entre los trabajos mencionados y lo que se presenta es la implementación de las características de una ontología.

Metodología

Para el desarrollo de esta herramienta se implementó el lenguaje Java y la metodología de desarrollo incremental (Pressman, 2010) que divide el desarrollo de software en incrementos funcionales como resultados de la ejecución de cinco etapas (comunicación, planeación, modelado, construcción y despliegue).

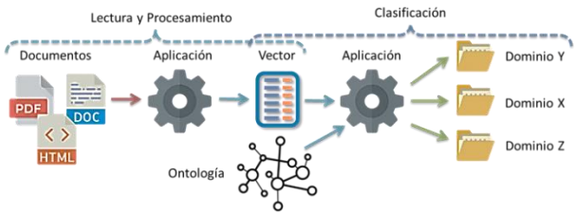

Figura 1 Funcionalidad de la herramienta

ISSN 2523-6873

Como se puede observar en la Figura 1 el funcionamiento general de la herramienta se divide en dos tareas: lectura y procesamiento de los documentos, y la clasificación de éstos. Sin embargo, la herramienta fue planeada en tres incrementos: el primero realiza la lectura y procesamiento del documento al seleccionar el archivo a clasificar, contabilizar las palabras contenidas, crear el vector de frecuencias, depurar éste último a fin de eliminar las palabras menos significativas (conectores, preposiciones, pronombres, etc.) y devolver como salida un porcentaje del vector con las palabras que obtuvieron el mayor índice de frecuencia; el segundo incremento consiste en implementar, a través de librerías, ontologías en el lenguaje Java para ser consultadas, como entrada recibe una ontología en formato OWL (Web Ontology Language) y dos palabras para devolver la descripción y relación existente entre éstas; el tercer y último incremento hace la clasificación al recibir el porcentaje de palabras significativas del vector de frecuencias, consultar la ontología, determinar el dominio o tema que representa el documento a través de identificar, en la ontología, las relaciones existentes entre las palabras contenidas en el vector y ubicar el documento en un conjunto de carpetas representativas a diferentes dominios.

Para validar el funcionamiento de la herramienta en la clasificación de documentos, se estableció un conjunto de pruebas bajo un escenario controlado que consistió en hacer un banco de 120 documentos digitalizados en inglés, de éstos 30 son del dominio de animales, 30 del dominio de enfermedades, 30 del dominio de plantas y 30 de cualquier otro dominio.

Posteriormente, conseguir 3 ontologías en formato OWL que pertenezcan a los dominios antes mencionados para clasificar los documentos. Por último, Establecer la variable de pruebas que controló el porcentaje de palabras que se tomaron del vector de frecuencias, para ésta establecieron arbitrariamente los valores de 5% y 10%. Una vez realizado lo anterior, el conjunto de pruebas se planeó con un total de 11 ejecuciones por dominio y porcentaje de palabras tomadas del vector de frecuencias, es decir, 22 ejecuciones por dominio o 66 ejecuciones en total.

Resultados y Discusión

En la Figura 2 y Figura 3 se pueden observar resultados del primer incremento que abarca la selección del documento a clasificar y la obtención de un porcentaje del vector de frecuencias, donde éste permitirá sintonizar la cantidad de palabras a tomar del vector completo para lograr la clasificación. La Figura 2 muestra el vector de frecuencias de un documento digital en el dominio de animales, mientras que la Figura 3 muestra el vector de un documento digital en el dominio de enfermedades. Como resultado del segundo y tercer incremento se tiene la lectura de ontologías en formato OWL a través de librerías y la consulta de éstas para obtener la relación entre dos palabras.

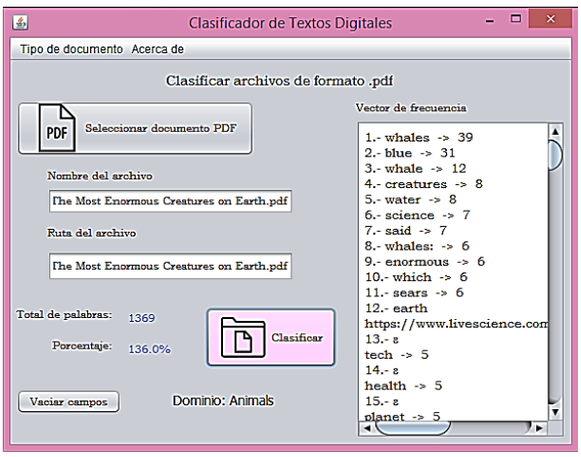

Figura 2 Interfaz del primer incremento con el dominio de animales

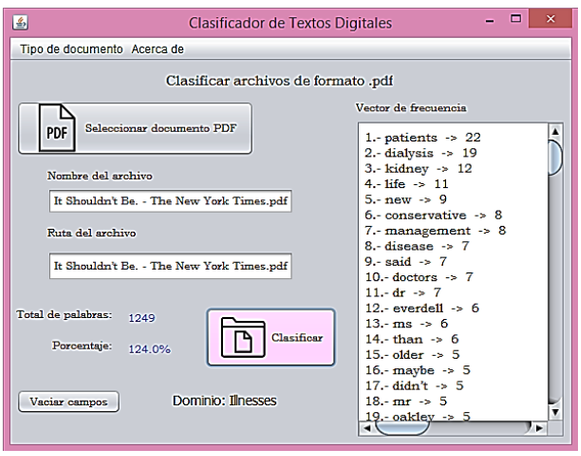

Figura 3 Interfaz del primer incremento con el dominio de enfermedades

ISSN 2523-6873

Como resultado de la tercera etapa se tiene la clasificación de documentos, por lo tanto, para validar el funcionamiento de la herramienta se implemento el escenario de pruebas descrito en la sección de metodología y los resultados obtenidos son los siguientes.

En el dominio de animales con un porcentaje del 10% en el vector de frecuencia, los resultados se pueden observar en la Tabla 1, y en el Gráfico 1, donde se tiene que de 120 documentos se tuvo una mediana de 32 documentos clasificados de los cuales 29 pertenecen al dominio y 3 fueron falsos positivos, es decir, documentos clasificados en el dominio pero que no pertenecen. Por lo tanto, ante este escenario de pruebas se tiene una eficiencia del 96.66% en la clasificación ya que de los 30 documentos que pertenecen al dominio 29 fueron clasificados, pero con una probabilidad del 10% de ser falso positivo al clasificar 3 documentos erróneamente.

Prueba	Documentos analizados	Documentos identificados	Falsos positivos	Total de documentos identificados
1	120	29	3	32
2	120	29	3	32
3	120	29	3	32
4	120	29	3	32
5	120	29	3	32
6	120	29	3	32
7	120	29	3	32
8	120	29	3	32
9	120	29	3	32
10	120	29	3	32
11	120	29	3	32

Tabla 1 Resultados con el dominio Animales con el 10% del vector de frecuencias

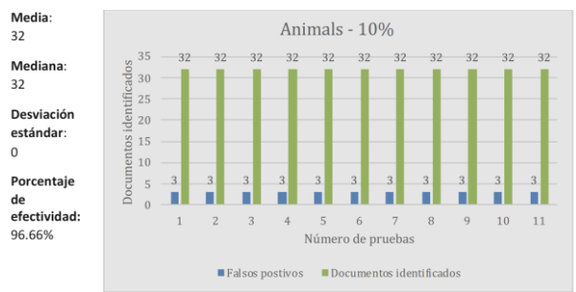

Gráfico 1 Estadísticas de los resultados con el dominio Animales con el 10% del vector de frecuencias

En la Tabla 2 se pueden observar los documentos que fueron falsos positivos en la clasificación, los dominios a los que pertenecen y la palabra contenida que provoco la clasificación falsa.

Nombre del documento	Dominio al que SI pertenece	Palabra dentro de la ontología Animals
America's Cup_ Wipeouts and wizardry mark covert battle – CNN.	Deportes	- Mule
LeBron James is first player in the NBA's top 10 all-time scoring and assists lists - BBC Sport	Deportes	- Horse
Are these bizarre music myths too good to be true_ - BBC Music.	Música	- Animal

Tabla 2 Descripción de documentos falsos positivos en el dominio de Animales con el 10% del vector de frecuencias

En el dominio de enfermedades con un porcentaje del 10% en el vector de frecuencia, los resultados se pueden apreciar en la Tabla 3, y en el Gráfico 2, donde se tiene que de 120 documentos se tuvo una mediana de 37 documentos clasificados de los cuales 30 pertenecen al dominio y 7 fueron falsos positivos. Por lo tanto, en este escenario de pruebas se tiene una eficiencia del 100% en la clasificación ya que a pesar de tener falsos positivos los 30 documentos para este dominio fueron clasificados, con una probabilidad del 23% de tener un falso positivo al clasificar 7 documentos erróneamente.

Prueba	Documentos analizados	Documentos identificados	Falsos positivos	Total de documentos identificados
1	120	30	6	36
2	120	30	7	37
3	120	30	7	37
4	120	30	7	37
5	120	30	7	37
6	120	30	7	37
7	120	30	7	37
8	120	30	7	37
9	120	30	7	37
10	120	30	7	37
11	120	30	7	37

Tabla 3 Resultados con el dominio Enfermedades con el 10% del vector de frecuencias

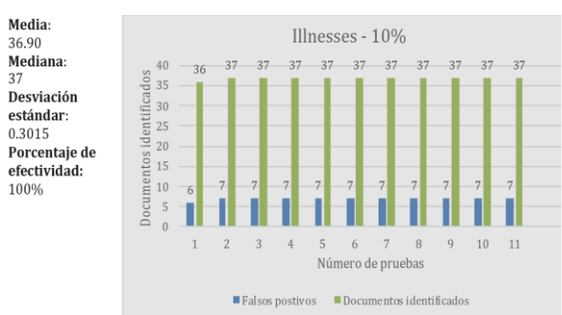

Gráfico 2 Estadísticas de los resultados con el dominio Enfermedades con el 10% del vector de frecuencias

ISSN 2523-6873

En la Tabla 4 se pueden observar los documentos que fueron falsos positivos en la clasificación, los dominios a los que pertenecen y la palabra contenida que provoco la clasificación falsa.

Nombre del documento	Dominio al que SI pertenece	Palabra dentro de la ontología Animals
An Economy in Need of Holistic Medicine - The New York Times.	Política	Health Symptoms Disease
Is There Room in 2020 for a Centrist Democrat_ Maybe One or Two - The New York Times.	Política	Health
Kamala Harris and Michael Bloomberg Clash on Medicare for All - The New York Times.	Política	Health
Clues to How Ancient Plants Handled Fungal Pests _ The Scientist Magazine.	Plantas	Infection
Deforestation Tied to Changes in Disease Dynamics _ The Scientist Magazine.	Plantas	Disease
Denmark Is Building a $12 Million Border Wall for Pigs	Animales	Health
USDA Unveils New Gene-Stacking Tool to Prevent Plant Diseases _ The Scientist Magazine	Plantas	Fungal Disease

Tabla 4 Descripción de documentos falsos positivos en el dominio de Enfermedades con el 10% del vector de frecuencias

En el dominio de plantas con un porcentaje del 10% en el vector de frecuencia, los resultados se pueden apreciar en la Tabla 5 y en el Gráfico 3, donde se tiene que de 120 documentos se tuvo una mediana de 25 documentos clasificados de los cuales 24 pertenecen al dominio y 1 fue falso positivo. Por lo que, en este escenario de pruebas se tiene una eficiencia del 80% en la clasificación ya que de los 30 documentos que pertenecen al dominio 24 fueron clasificados, pero con una probabilidad del 3.33% de ser falso positivo al clasificar 1 documento erróneamente.

Prueba	Documentos analizados	Documentos identificados	Falsos positivos	Total de documentos identificados
1	120	25	1	26
2	120	24	1	25
3	120	24	1	25
4	120	24	1	25
5	120	24	1	25
6	120	24	1	25
7	120	24	1	25
8	120	24	1	25
9	120	24	1	25
10	120	24	1	25
11	120	24	1	25

Tabla 5 Resultados con el dominio Plantas con el 10% del vector de frecuencias

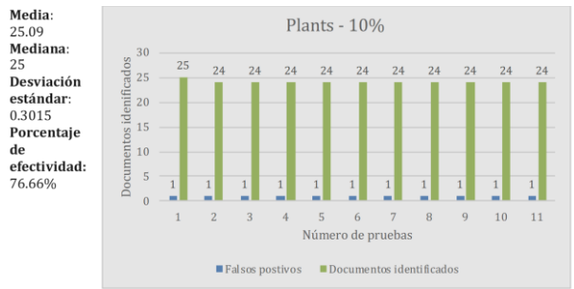

Gráfico 3 Estadísticas de los resultados con el dominio Plantas con el 10% del vector de frecuencias

En la Tabla 6 se pueden observar los documentos que fueron falsos positivos en la clasificación, los dominios a los que pertenecen y la palabra contenida que provoco la clasificación falsa.

Nombre del documento	Dominio al que SI pertenece	Palabra dentro de la ontología Plantas
Foxconn Is Reconsidering Plan for Wisconsin Factory - The New York Times.	Política	Plant

Tabla 6 Descripción de documentos falsos positivos en el dominio de Plantas con el 10% del vector de frecuencias

Los resultados anteriores fueron para los tres dominios con un porcentaje del 10% de representación del vector de frecuencias. Los resultados de estos tres dominios, pero con un porcentaje del 5% de representación se describen a continuación. En el dominio de animales con un porcentaje del 5% en el vector de frecuencia, los resultados se pueden observar en la Tabla 7 y en el Gráfico 4, donde se tiene que de 120 documentos se tuvo una mediana de 33 documentos clasificados de los cuales 30 pertenecen al dominio y 3 fueron falsos positivos.

ISSN 2523-6873

Por lo tanto, ante este escenario de pruebas se tiene una eficiencia del 100% en la clasificación ya que los 30 documentos que pertenecen al dominio fueron clasificados, pero con una probabilidad del 10% de ser falso positivo al clasificar 3 documentos erróneamente.

Prueba	Documentos analizados	Documentos identificados	Falsos positivos	Total de documentos identificados
1	120	30	3	33
2	120	30	3	33
3	120	30	3	33
4	120	30	3	33
5	120	30	3	33
6	120	30	3	33
7	120	30	3	33
8	120	30	3	33
9	120	30	3	33
10	120	30	3	33
11	120	30	3	33

Tabla 7 Resultados con el dominio Animales con el 5% del vector de frecuencias

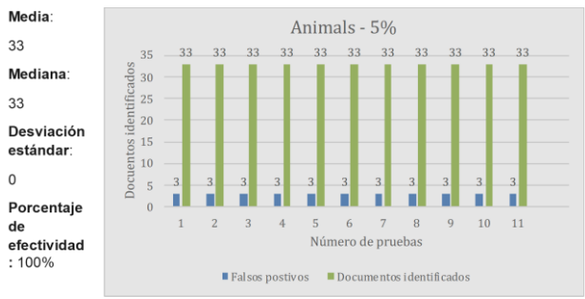

Gráfico 4 Estadísticas de los resultados con el dominio Animales con el 5% del vector de frecuencias

En la Tabla 8 se pueden observar los documentos que fueron falsos positivos en la clasificación, los dominios a los que pertenecen y la palabra contenida que provoco la clasificación falsa.

Nombre del documento	Dominio al que SI pertenece	Palabra dentro de la ontología Animales
Are these bizarre music myths too good to be true_ - BBC Music	Música	Animal
Flowers Sweeten Up When They Sense Bees Buzzing _ Smart News _ Smithsonian	Plantas	Bees
Researchers Stabbed Slabs of Meat With Cacti Spines to Learn About Puncture Strength _ Smart News _ Smithsonian	Plantas	Animal

Tabla 8 Descripción de documentos falsos positivos en el dominio de Animales con el 5% del vector de frecuencias

En el dominio de enfermedades con un porcentaje del 5% en el vector de frecuencia, los resultados se pueden observar en la Tabla 9 y en el Gráfico 5, donde se tiene que de 120 documentos se tuvo una mediana de 31 documentos clasificados de los cuales 30 pertenecen al dominio y 1 fue falso positivo. Por lo que, en este escenario de pruebas se tiene una eficiencia del 100% al clasificar los 30 documentos que pertenecen al dominio, sin embargo, se tiene una probabilidad del 3.33% de ser falso positivo al clasificar 1 documento erróneamente.

Prueba	Documentos analizados	Documentos identificados	Falsos positivos	Total de documentos identificados
1	120	30	1	31
2	120	30	1	31
3	120	30	1	31
4	120	30	1	31
5	120	30	1	31
6	120	30	1	31
7	120	30	1	31
8	120	30	1	31
9	120	30	1	31
10	120	30	1	31
11	120	30	1	31

Tabla 9 Resultados con el dominio Enfermedades con el 5% del vector de frecuencias

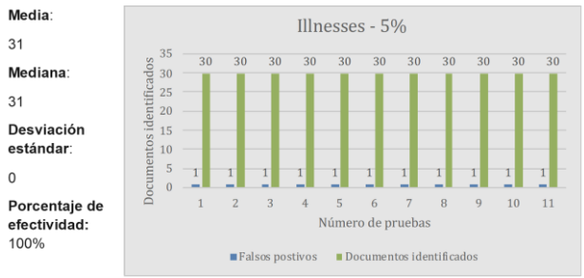

Gráfico 5 Estadísticas de los resultados con el dominio Enfermedades con el 5% del vector de frecuencias

En la Tabla 10 se pueden observar los documentos que fueron falsos positivos en la clasificación, los dominios a los que pertenecen y la palabra contenida que provoco la clasificación falsa.

Nombre del documento	Dominio al que SI pertenece	Palabra dentro de la ontología Enfermedades
An Economy in Need of Holistic Medicine - The New York Times.	Política	Symptoms Disease

Tabla 10 Descripción de documentos falsos positivos en el dominio de Enfermedades con el 5% del vector de frecuencias

ISSN 2523-6873

En el dominio de plantas con un porcentaje del 5% en el vector de frecuencia, los resultados se pueden observar en la Tabla 11 y en el Gráfio 6, donde se tiene que de 120 documentos se tuvo una mediana de 26 documentos clasificados de los cuales 25 pertenecen al dominio y 1 fue falso positivo. Por lo tanto, ante este escenario de pruebas se tiene una eficiencia del 83.33% en la clasificación ya que de los 30 documentos que pertenecen al dominio 25 fueron clasificados, con una probabilidad del 3.33% de ser falso positivo al clasificar 1 documento erróneamente.

Prueba	Documentos analizados	Documentos identificados	Falsos positivos	Total de documentos identificados
1	120	25	1	26
2	120	25	1	26
3	120	25	1	26
4	120	25	1	26
5	120	25	1	26
6	120	25	1	26
7	120	25	1	26
8	120	25	1	26
9	120	25	1	26
10	120	25	1	26
11	120	25	1	26

Tabla 11 Resultados con el dominio Plantas con el 5% del vector de frecuencias

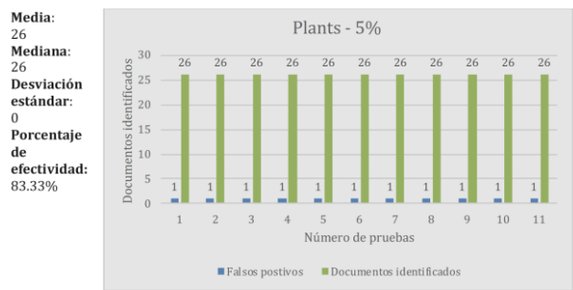

Gráfico 6 Estadísticas de los resultados con el dominio Plantas con el 5% del vector de frecuencias

En la Tabla 12 se pueden observar los documentos que fueron falsos positivos en la clasificación, los dominios a los que pertenecen y la palabra contenida que provoco la clasificación falsa.

Nombre del documento	Dominio al que SI pertenece	Palabra dentro de la ontología Plantas
Foxconn Is Reconsidering Plan for Wisconsin Factory - The New York Times.	Política	Plant

Tabla 12 Descripción de documentos falsos positivos en el dominio de Plantas con el 5% del vector de frecuencias

Al analizar los resultados de las ejecuciones en cada dominio se puede observar que existieron variaciones gracias a la variable del porcentaje en el vector de frecuencias y se puede concluir que el porcentaje del 5% es mejor ya que el vector de frecuencias contiene menos palabras que ayudan a tener una mejor clasificación al contener las palabras más significativas del documento y evitar aquellas que podrían provocar falsos positivos.

Para tener una mejor perspectiva sobre los resultados obtenidos, el Gráfico 7 muestra la comparación entre el porcentaje del 10% y 5% con respecto a los dominios utilizados.

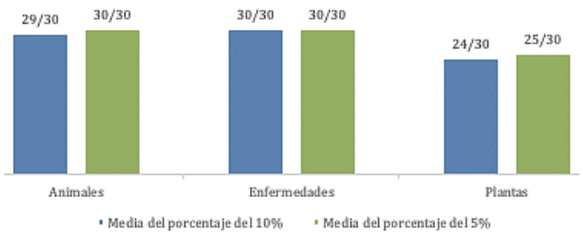

Gráfico 7 Resultados de los documentos clasificados por dominio y por porcentaje del vector de frecuencias

El Gráfico 8 también permite observar que el porcentaje del 5% del vector de frecuencias permite tener menos falsos positivos como se ve en el dominio Enfermedades, esto se debe a que al tener menos palabras tomadas del vector de frecuencias evita clasificaciones erróneas en el proceso.

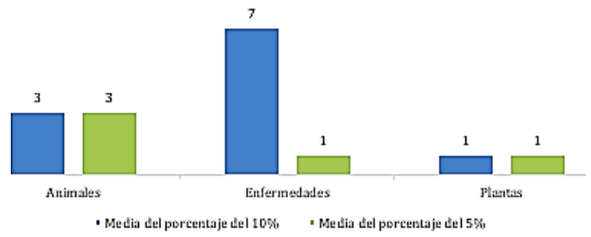

Gráfico 8 Resultados de los falsos positivos por dominio y por porcentaje del vector de frecuencias

Si se analiza el correcto funcionamiento demostrado por nuestra herramienta ante los trabajos mencionados anteriormente, se puede decir que la principal diferencia es el uso de ontologías de diferentes dominios para la clasificación lo que conlleva a que se puede clasificar documentos de texto en cualquier dominio del que trate la ontología y no solamente a uno como lo reportado en los trabajos citados.

ISSN 2523-6873

Conclusiones

El desarrollo de una herramienta que permita clasificar, en un dominio o tema, documentos de texto digitales en base a su contenido permite tener una mejor organización sobre éstos y de esta manera realizar búsquedas de información más rápidas al no tener que realizarlas en documentos que no tengan relación con el dominio o tema de la búsqueda. Otra ventaja que se tiene es que la clasificación se puede realizar bajo cualquier dominio o tema ya que solamente se requiere conseguir la ontología del dominio o tema en cuestión en formato OWL y agregarla a la herramienta. Por último, cabe mencionar que esta herramienta trabaja solamente en el idioma inglés ya que actualmente no se tiene ontologías en formato OWL en el idioma español.

Trabajo a futuro

El trabajo que se tiene planeado realizar es el de mejorar la interfaz de usuario (UI) para hacer más intuitiva y agradable la herramienta.

Referencias

Lopez Condori, R. (2014). *Método de Clasificación Atomática en Textos basado en Palabras Claves utilizando Información Semántica: Aplicación a Histórias Clínicas.* Arequipa: Universidad Nacional de San Agustín.

Berners-Lee, T., Fielding, R., & Frystyk, H. (1996). *Hypertext Transfer Protocol -- HTTP/1.0.* United States: RFC Editor.

Gruber, T. (1995). Toward Principles of the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies, 43*, 907-928.

Pressman, R. (2010). *Ingeniería del Software: un enfoque práctico.* México: McGraw Hill Education.

Python, S. C. (7 de septiembre de 2018). *StringTagger: Clasificador de Texto con Python.* Obtenido de Mi diario Python: http://www.pythondiario.com/2018/02/stringtagger-clasificacion-de-texto-con.html?m=1

Sanchez Vega, J. (2012). *Clasificación de texto mediante atributos probabilísticos de coocurrencia de palabras.* Sta. Ma. Tonantzintla: INAOE.

Specification, H. 4. (Diciembre de 1999). *HTML 4.01 Specification.* Obtenido de W3C: https://www.w3.org/TR/html401/

Wilks, Y., & Catizone, R. (2000). Can we make Information Extraction more adaptive? *Research and Development in Intelligent Systems XVI*.

ISSN 2523-6873