Interdisciplinary Congress of Renewable Energies - Industrial Maintenance - Mechatronics and Informatics
Booklets

RENIECYT - LATINDEX - Research Gate - DULCINEA - CLASE - Sudoc - HISPANA - SHERPA UNIVERSIA - Google Scholar DOI - REDIB - Mendeley - DIALNET - ROAD - ORCID

Title: Evaluación de un Clasificador de Textos Digitales basado en el Contenido Semántico a través de Ontologías

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

Editorial label ECORFAN: 607-8695
BCIERMMI Control Number: 2020-04
BCIERMMI Classification (2020): 211020-0004

Pages: 12
RNA: 03-2010-032610115700-14

ECORFAN-México, S.C.
143 – 50 Itzopan Street
La Florida, Ecatepec Municipality
Mexico State, 55120 Zipcode
Phone: +52 1 55 6159 2296
Skype: ecorfan-mexico.s.c.
E-mail: contacto@ecorfan.org
Facebook: ECORFAN-México S. C.
Twitter: @EcorfanC

www.ecorfan.org

Holdings
Mexico Colombia Guatemala
Bolivia Cameroon Democratic
Spain El Salvador Republic
Ecuador Taiwan of Congo
Peru Paraguay Nicaragua

Introducción

Metodología

Resultados

Conclusiones

Trabajo a Futuro

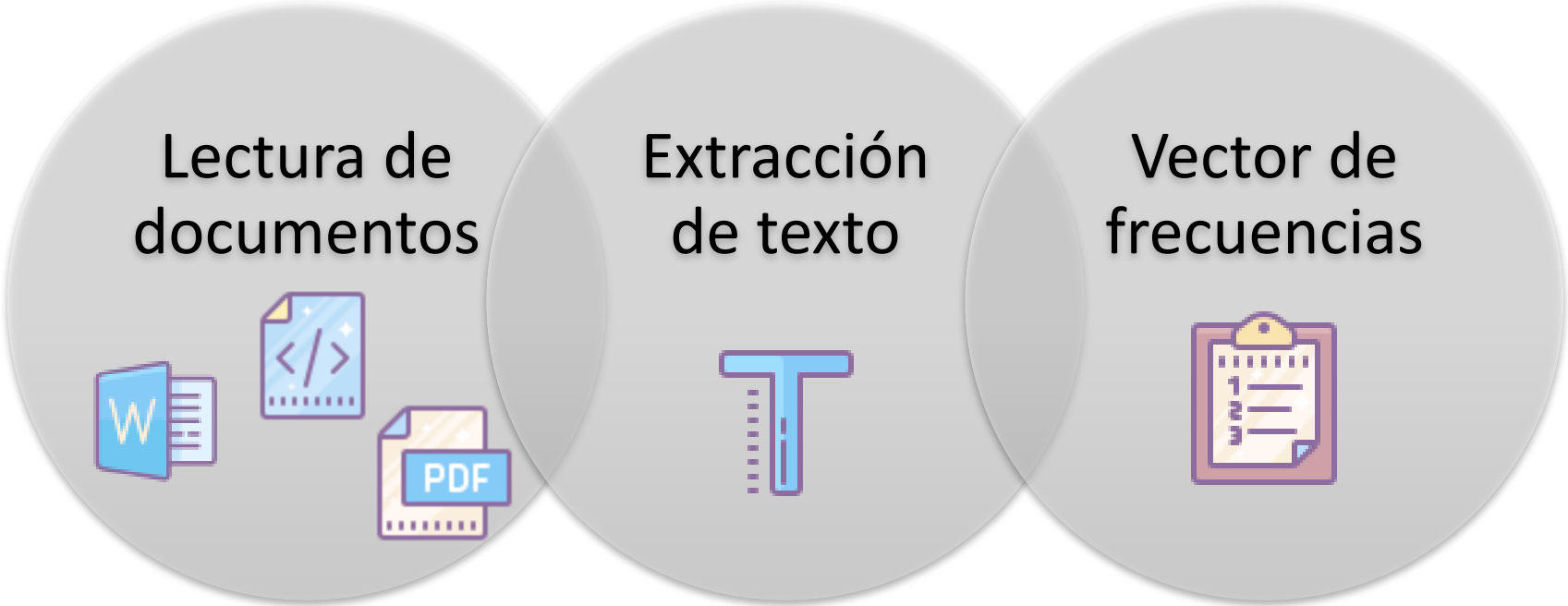
Referencias

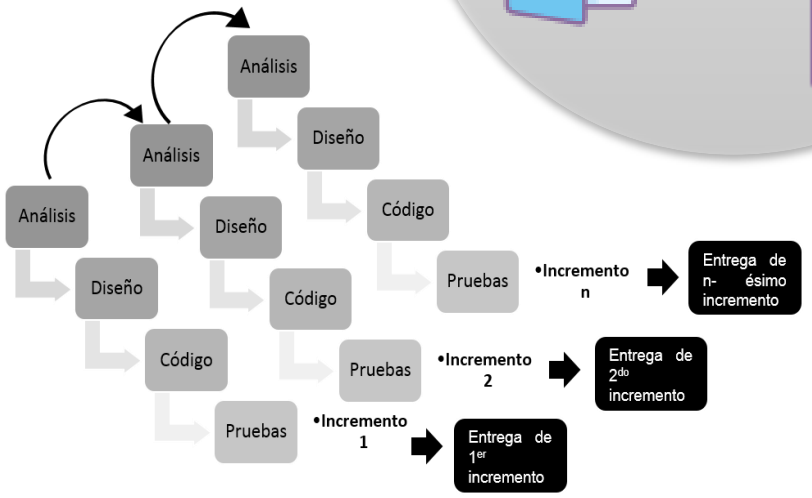
Introducción

- Continuo crecimiento e innovación de documentos digitales.
- Existen herramientas para la ayuda de clasificación de textos.
- Herramienta ofimática que realice la clasificación de textos digitales por medio de ontologías.

Metodología

Primer incremento



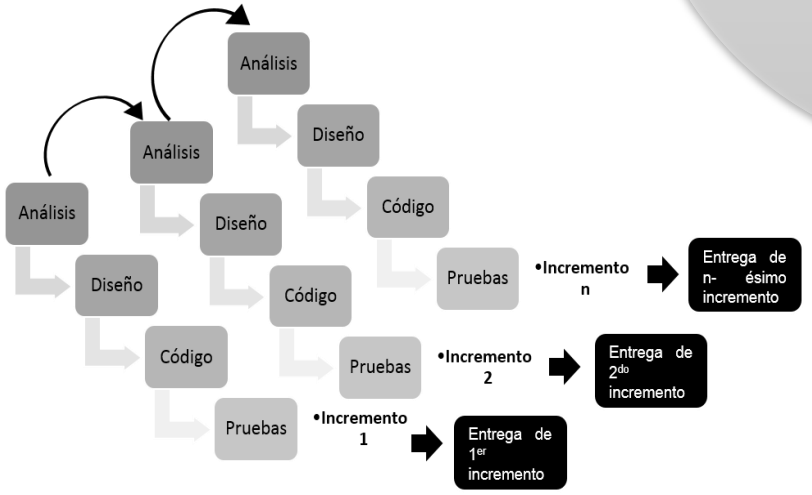


Metodología

Segundo incremento

Investigación de ontologías en formato owl y rdf

Lectura de ontología en formato owl en Java



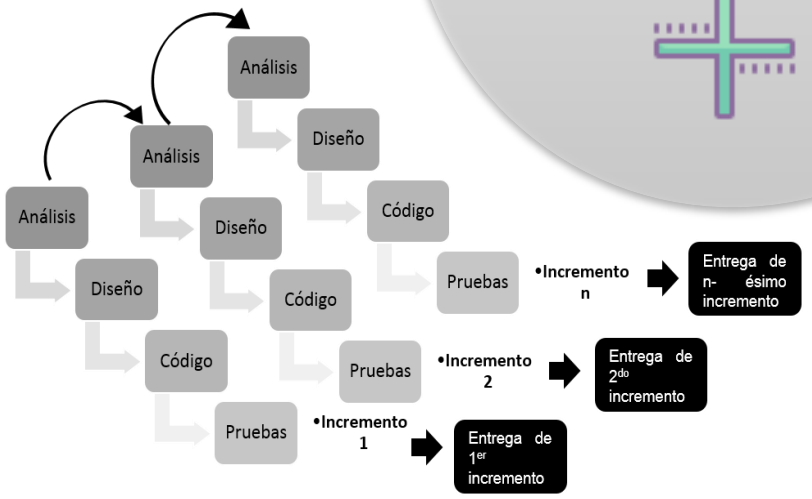
Metodología

Tercer incremento

Añadir la implementación de la ontología.

Organizar cada texto en una carpeta representativa al dominio identificado.

Métricas de evaluación con los resultados obtenidos.



Resultados

Escenario de pruebas

- 120 documentos
- 30 documentos para dominio Animales
- 30 documentos para dominio enfermedades
- 30 documentos para dominio Plantas
- 30 documentos sin dominio en particular

Ejecuciones

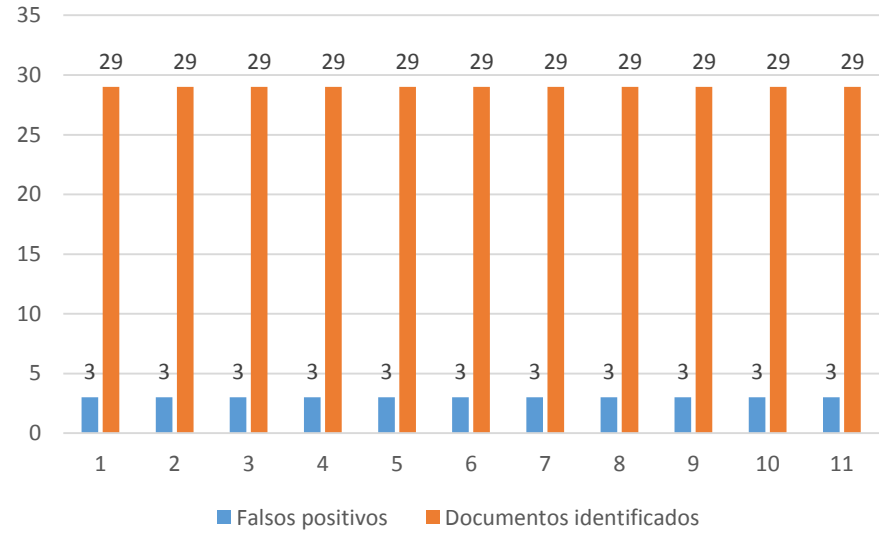
- 1 variable experimental
  - % del vector de frecuencias
- 11 ejecuciones por dominio para un valor del 10% de palabras del vector de frecuencias
- 11 ejecuciones por dominio para un valor del 5% de palabras del vector de frecuencias

Resultados
Dominio Animales

10% del vector de frecuencias

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

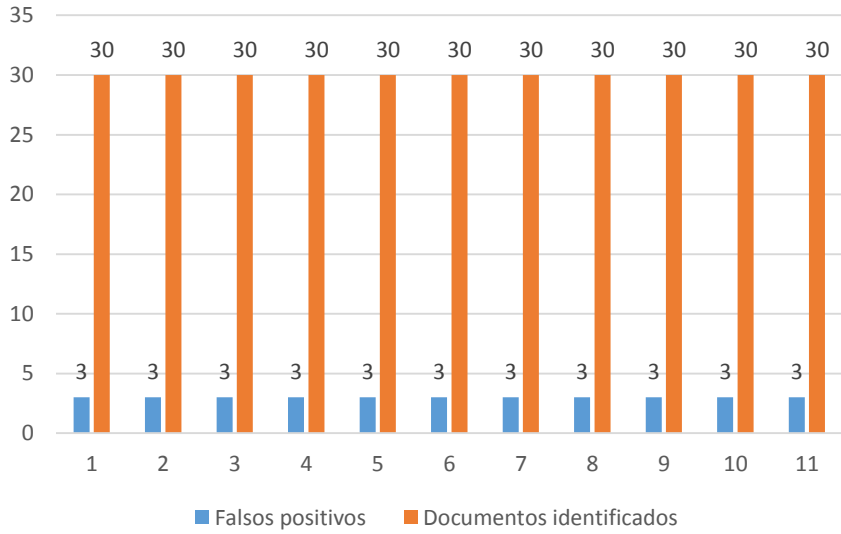
Media: 32
Mediana: 32
Desviación estándar: 0
Efectividad: 96.66%
Falso Positivo: 10%



5% del vector de frecuencias

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

Media: 33
Mediana: 33
Desviación estándar: 0
Efectividad: 100%
Falso Positivo: 10%

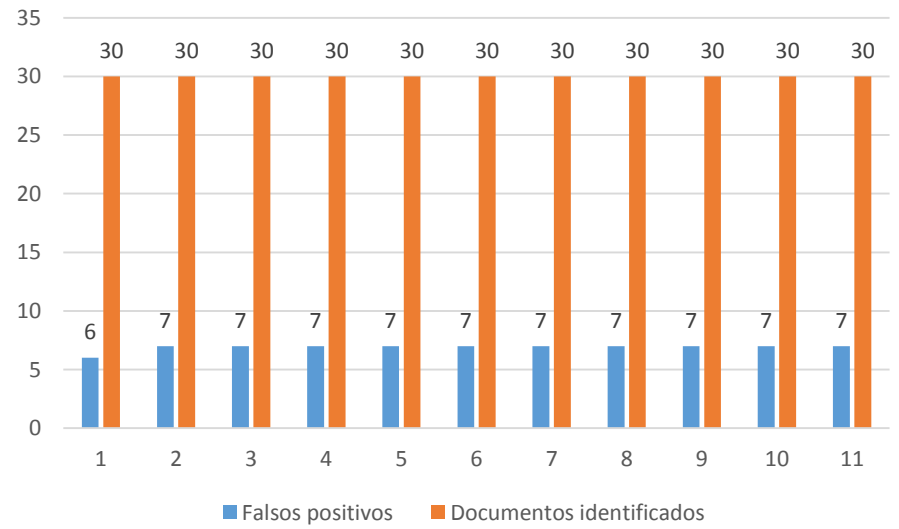


Resultados
Dominio Enfermedades

10% del vector de frecuencias

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

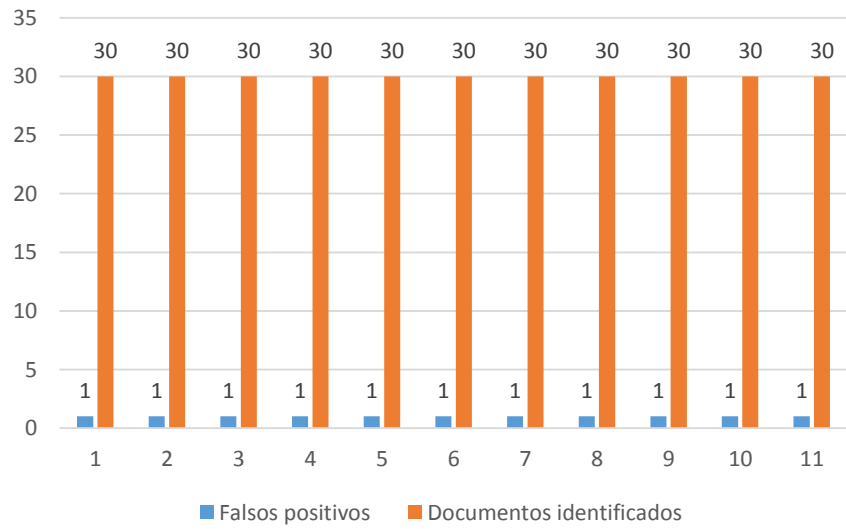
Media: 36.90
Mediana: 37
Desviación estándar: 0.3015
Efectividad: 100%
Falso positivo: 23%



5% del vector de frecuencias

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

Media: 31
Mediana: 31
Desviación estándar: 0
Efectividad: 100%
Falso positivo: 3.33%

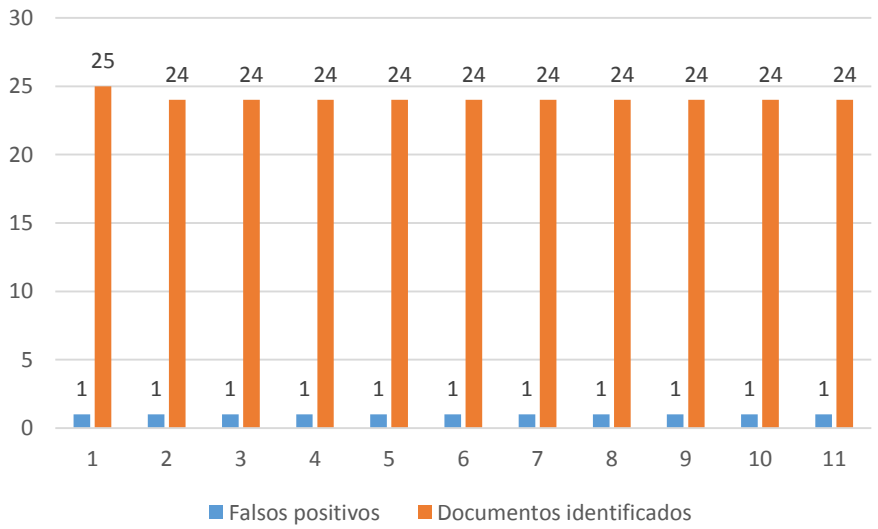


Resultados
Dominio Plantas

10% del vector de frecuencias

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

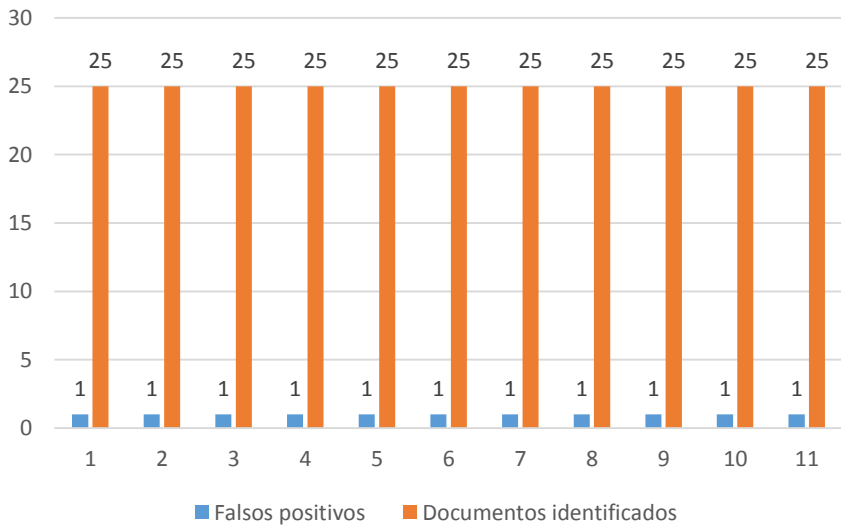
Media: 25.09
Mediana: 25
Desviación estándar: 0.3015
Efectividad: 80%
Falso positivo: 3.33%



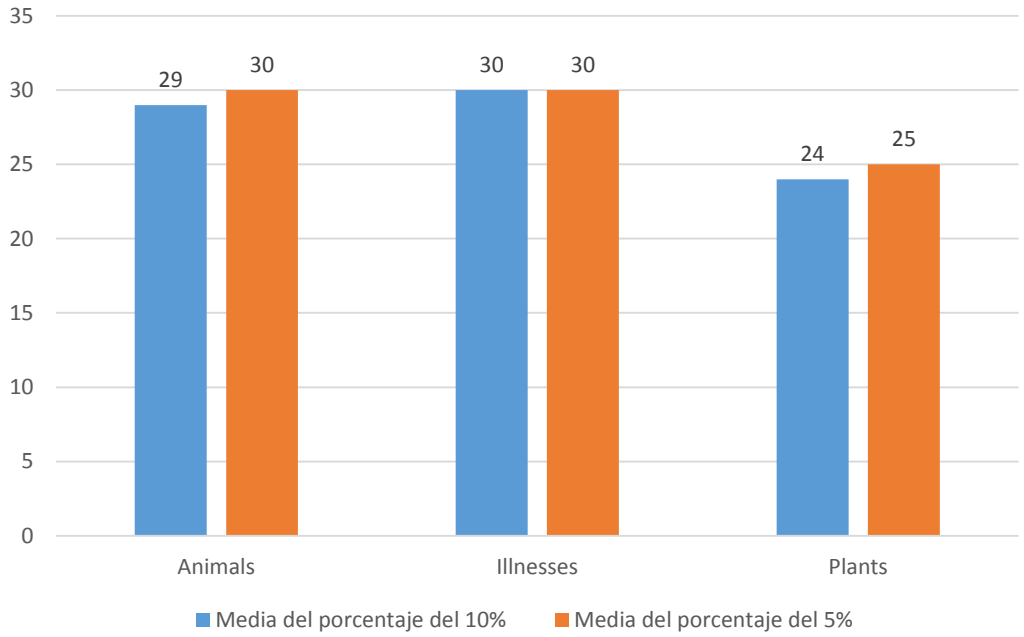
5% del vector de frecuencias

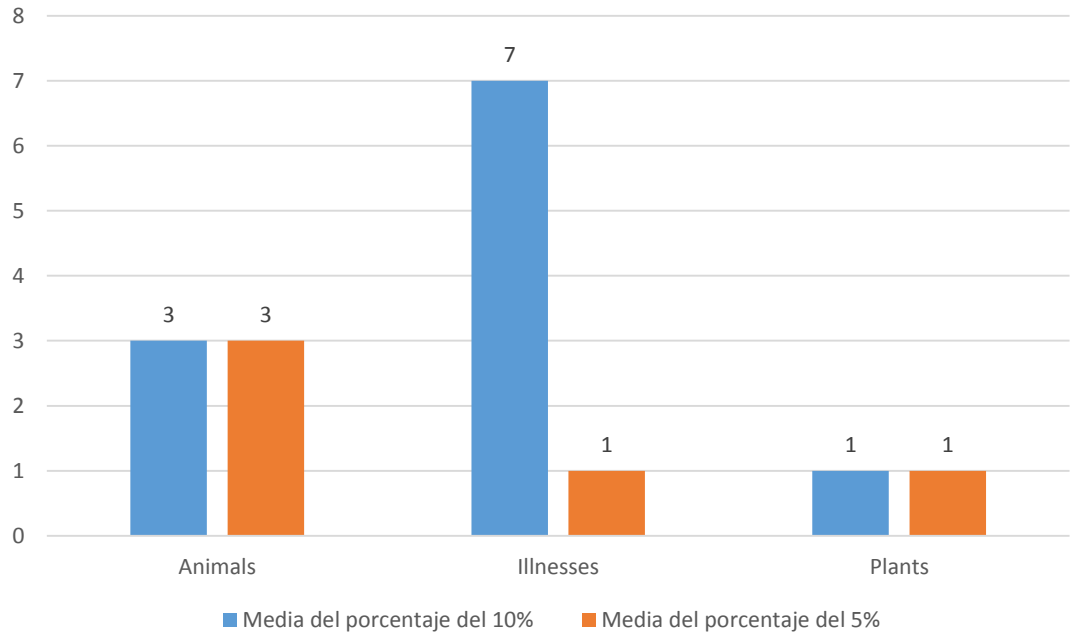
Prueba	Documentos analizados	Documentos identificados	Falsos positivos	Total de documentos identificados
1	120	25	1	26
2	120	25	1	26
3	120	25	1	26
4	120	25	1	26
5	120	25	1	26
6	120	25	1	26
7	120	25	1	26
8	120	25	1	26
9	120	25	1	26
10	120	25	1	26
11	120	25	1	26

Media: 26
Mediana: 26
Desviación estándar: 0
Efectividad: 83.33%
Falso positivo: 3.33%



Resultados





- Existieron variaciones gracias a la variable experimental del porcentaje en el vector de frecuencias
- Se concluye que el porcentaje del 5% es mejor
    - El vector de frecuencias contiene menos palabras que ayudan a tener una mejor clasificación
    - Evita falsos positivos.

Concusiones

- Permite tener una mejor organización sobre archivos.
- Permite realizar búsquedas de información más rápidas.
- La clasificación se puede realizar bajo cualquier dominio o tema.
- Esta herramienta trabaja solamente en el idioma inglés ya que actualmente no se tiene ontologías en formato OWL en el idioma español.

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*.

ECORFAN®