Marco de trabajo para evaluar la relevancia de artículos de dominio específico

Trabajo fin de Máster

Máster en Ingeniería en Sistemas de Decisión

Adrián Alonso Barriuso

9 Jul 2019

Index

Introducción

La comunidad investigadora se enfrenta cada vez a un mayor número de publicaciones y ten- dencias que deben atender a la hora hacer sus propias publicaciones. Estos tópicos y tendencias pueden ser estado del arte en el momento de su publicación en revistas o presentación en conferencias, no obstante, pueden perder relevancia a lo largo del tiempo. Por tanto, la posibilidad de obtener una medida de relevancia de un artículo puede ser de gran utilidad para la comunidad científica.

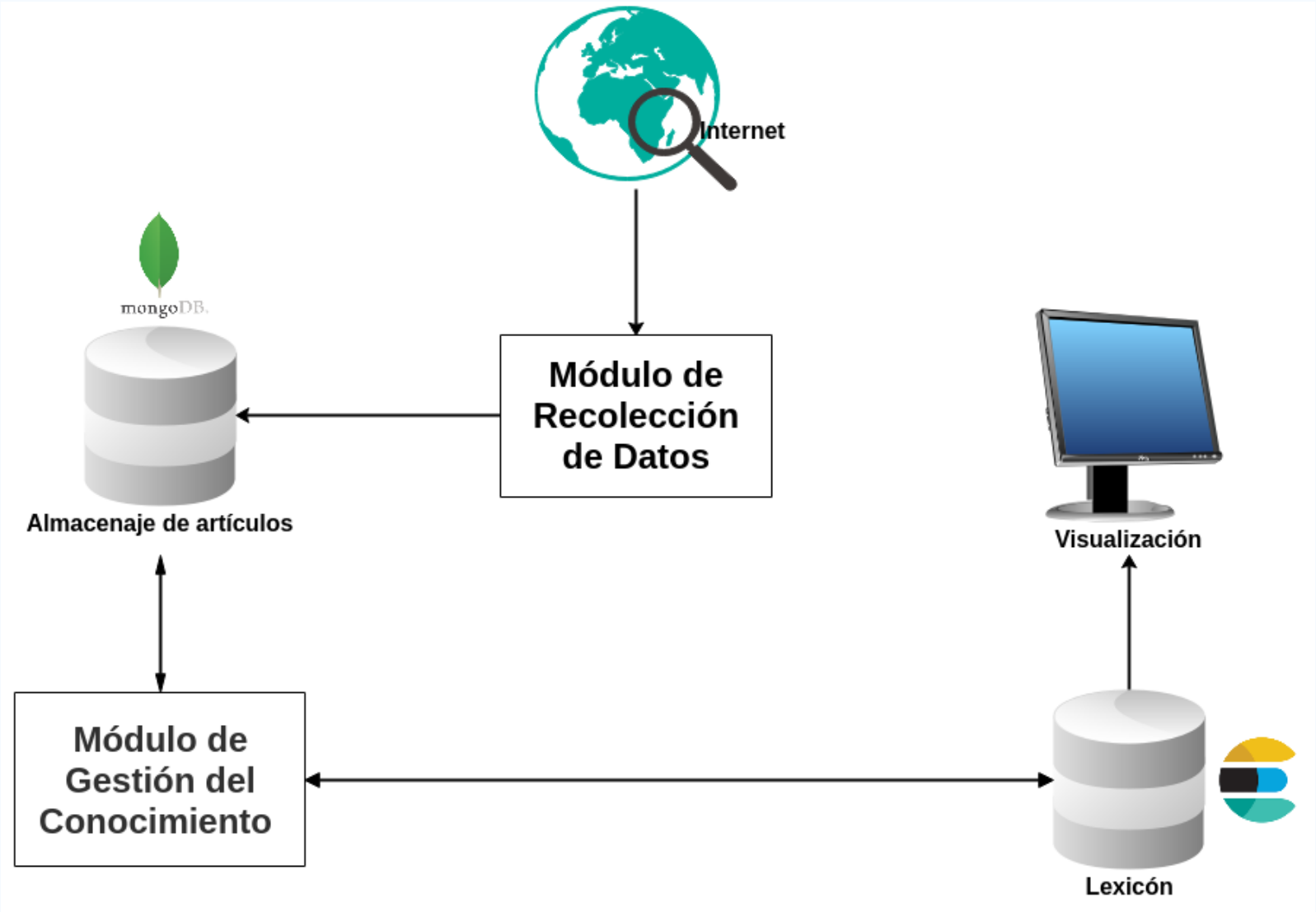
Objetivo general

El principal objetivo del presente trabajo es la creación de un sistema completo de evaluación de relevancias, lo que comprende un marco de trabajo que incluye la interfaz para la introduc- ción de documentos y a la salida devuelva la relevancia de los mismos.

Objetivos específicos

- Obtención del corpus de documentos científicos.
- Limpieza y almacenaje de los documentos.
- Construcción del lexicón de relevancias.
- Construcción de la red neuronal.
- Creación del flujo de evaluación de relevancia.

Arquitectura básica



Descarga y limpieza de datos

- Descarga, limpieza y parseo de artículos.
- Cálculo de resúmenes automáticos.
- Cálculo de reputaciones. La reputación de un artículo viene dada por la siguiente fórmula:

$$rep_p = \alpha * rep\_authors_p + (1 - \alpha) * citations_p \,, \tag{1}$$

$$rep\_authors_p = \sum_{i=1}^{n} rep_i / n \,. \tag{2}$$

$$rep_i = \omega_1 * inf\_citation\_count + \omega_2 * citation\_velocity + \omega_3 * seniority + \omega_4 * papers \,, \tag{3}$$

Parámetro	Saturación
Número de artículos	196
citationVelocity	105
influentialCitationCount	208
Seniority	34
Citas del artículo	10

Table 1: Valores de saturación de parámetros de reputación

- Obtención de matriz de términos por documentos.
- Creación del lexicón:

$$relevance(t) = \log(\frac{1}{N}\sum_{i=1}^{N} \alpha \cdot tfidf(t_i) + (1-\alpha) \cdot reputation_i), \forall t \in D \quad (4)$$

$$relevance'(t) = \beta \cdot relevance(t) + (1-\beta) \cdot occurence(t) \quad (5)$$

- Visualización
- Evaluación de nuevos artículos:

$$relevance_d = \frac{1}{N}\sum_{i=1}^{N} relevance'(t_i) \quad (6)$$

Trabajos futuros

- Optimización del cálculo de relevancias para poder considerar un mayor corpus.
- Implementar, probar y comparar otras métricas de relevancia.
- Emplear relaciones jerarquicas de ontologías médicas.
- Extracción de entidades con Scipacy
- Construir lexicones por cada año.
- Crear una red de neuronas para predecir las relevancias de los términos que no se encuentren en el lexicón.