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 tendencias 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.

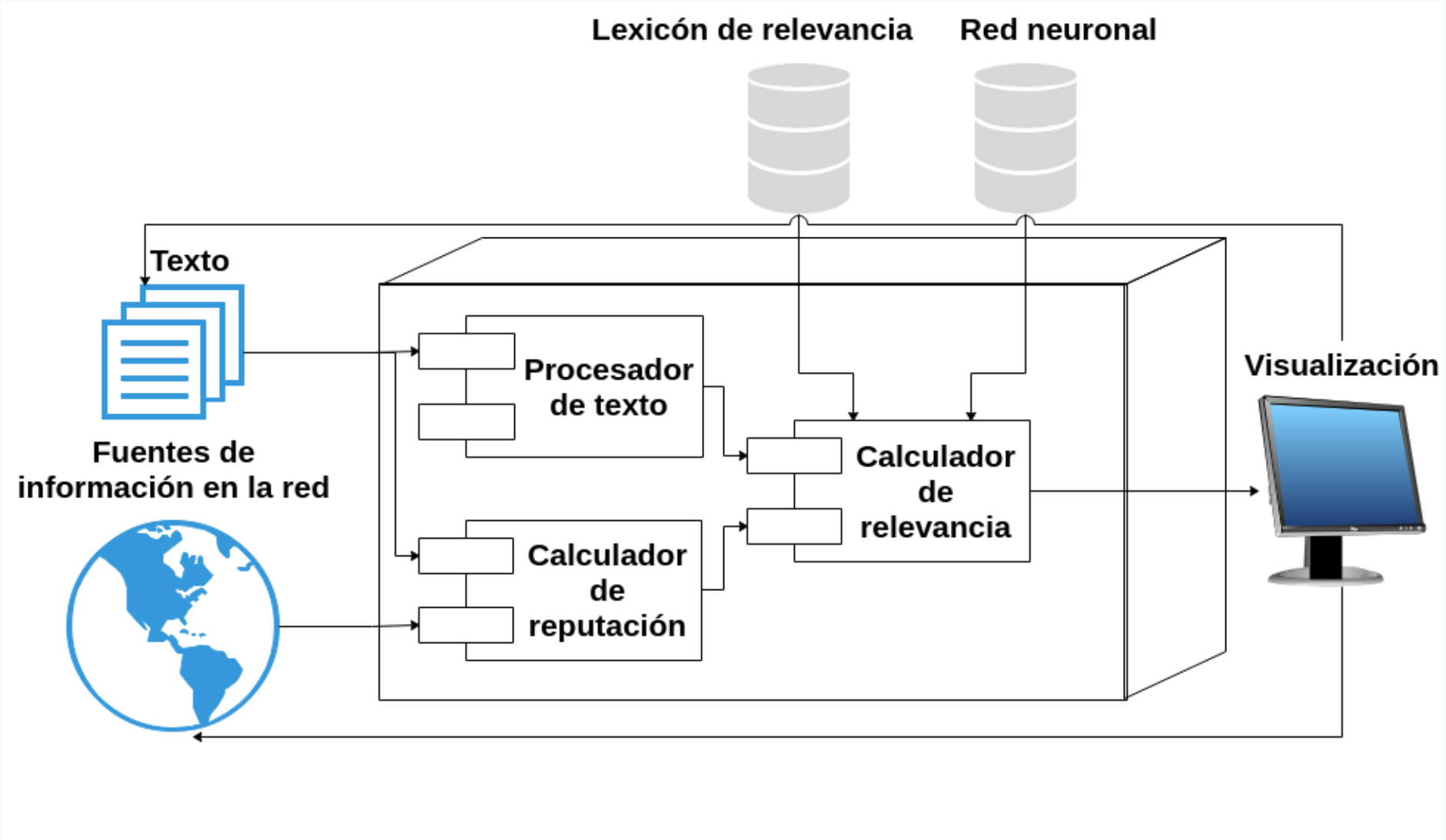
Objetivo general

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 introducción 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



Preparación de los artículos

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

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

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

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

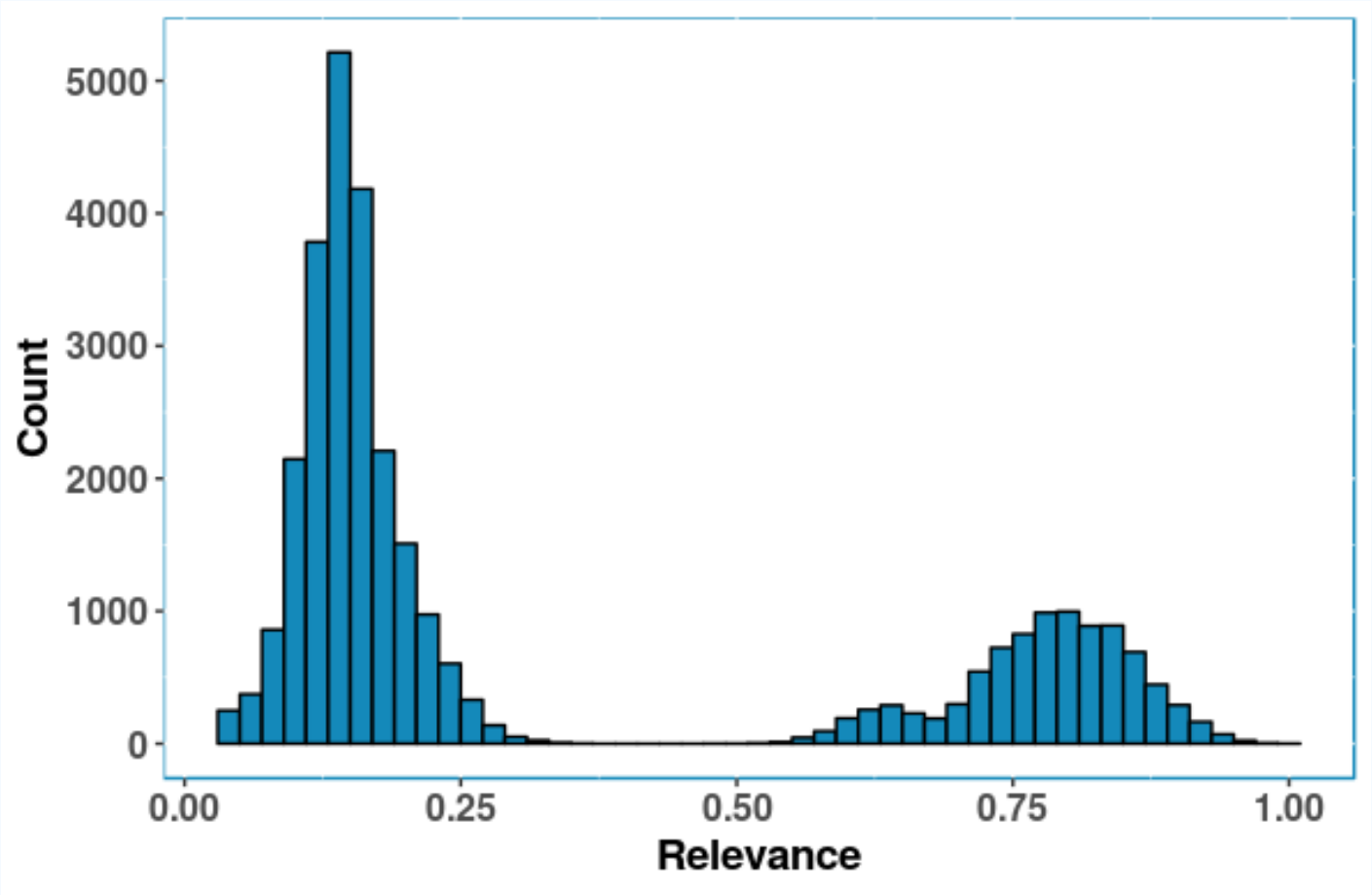
Relevancia de un término t en un corpus C:

$$rel\_lex_t = \log\left(\frac{1}{N}\cdot\sum_{p=1}^{N}\beta\cdot tfidf(t)_p + (1-\beta)\cdot rep_p\right), \forall p \in C\,, \tag{4}$$

Relevancia acumulada por año:

$$rel\_lex_t(y) = \rho\cdot rel\_lex_t + (1-\rho)\cdot rel\_lex_t(y-1)\,, \tag{5}$$

Generación del lexicón



Creación de la red neuronal

- Embedding pre-entrenado de 400.000 palabras.
- Red neuronal convolucional.
- 320.000 frases train.
- 80.000 frases test.
- Salida softmax.

Año	Acierto en conjunto de test
2015	0.964
2016	0.962
2017	0.959
2018	0.959

Table 1: Resultados de test de las redes neuronales

-Relevancia combinada entre lexicón y red neuronal:

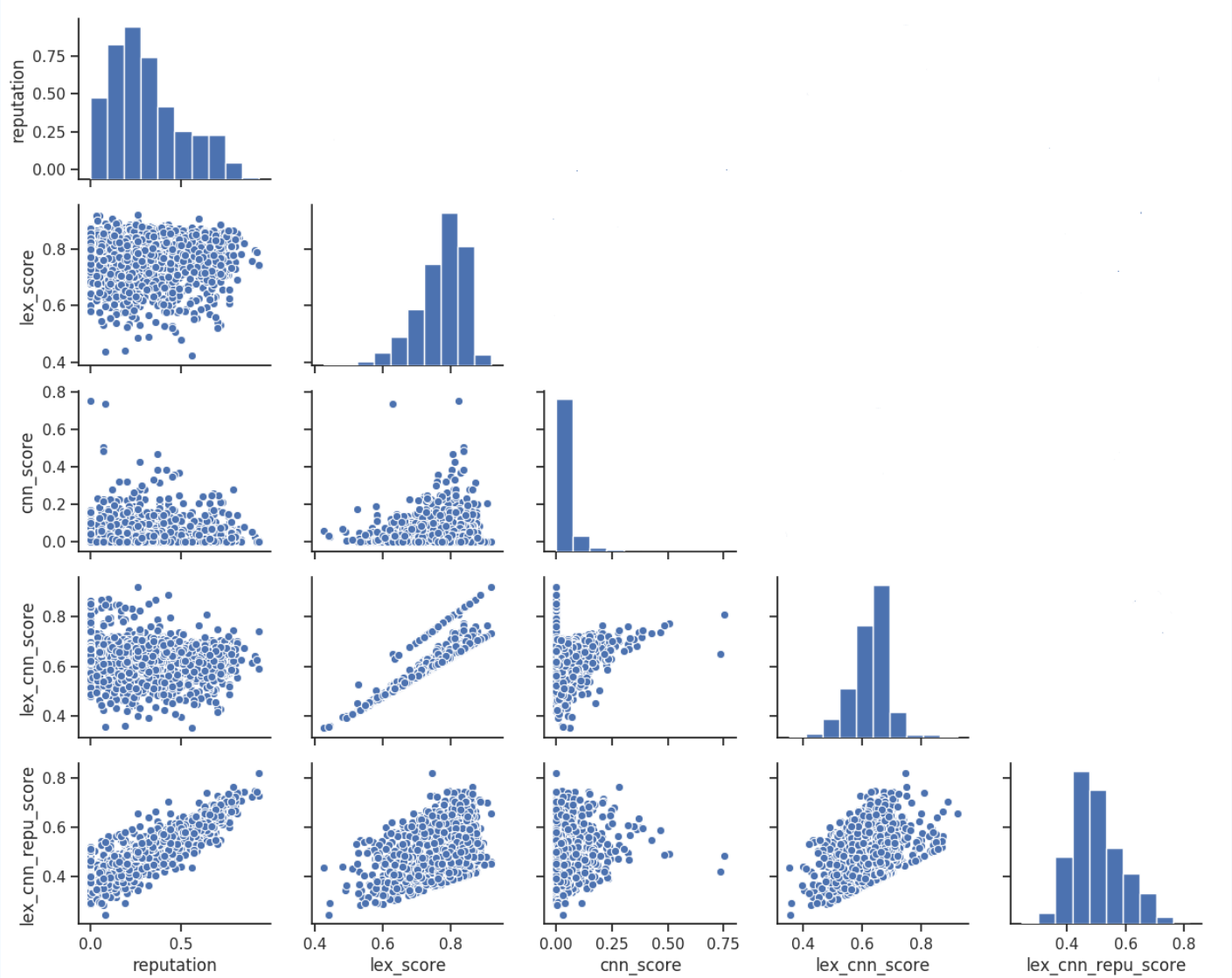
$$combined\_rel_p = \theta \cdot rel\_lex_p + (1-\theta) \cdot \frac{1}{K} \sum_{k=1}^{K} rel\_neural(s_k)\,, \tag{6}$$

-Relevancia final combinada con reputación:

$$combined\_rel\_doi_p = \gamma \cdot combined\_rel_p + (1-\gamma) \cdot rep_p\,, \tag{7}$$

Experimentos

-2000 artículos analizados por cada año:



Experimentos

reputation	lex_cnn_repu_score	lex_score	cnn_score	total_sentences	lex_cnn_score
0.93	0.819	0.744	0	83	0.744
0.79	0.764	0.863	0.282	215	0.747
0.91	0.749	0.795	0.027	104	0.642
0.92	0.746	0.787	0	92	0.629
0.81	0.745	0.841	0.144	278	0.702
0.85	0.744	0.821	0.082	218	0.673
0.64	0.741	0.809	0	114	0.809
0.81	0.74	0.854	0.052	272	0.694
0.82	0.738	0.837	0.072	181	0.684
0.76	0.737	0.864	0.15	177	0.721
0.78	0.733	0.862	0.058	101	0.701
0.7	0.73	0.75	0	91	0.75
0.93	0.728	0.739	0.008	212	0.593
0.89	0.726	0.758	0.053	134	0.617
0.8	0.723	0.833	0.03	275	0.672

Trabajos futuros

- Aumento de tamaño de los lexicones.
- Implementar, probar y comparar otras métricas de relevancia.
- Detección de correferencias de términos.
- Creación de word embeddings a partir del corpus.
- Sustituir sustantivos por conceptos de n-gramas en base a diccionarios u ontologías.