Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de la Matanza
ISSN: 2525-1333
Vol.: 7 - Nro. 2 (DICIEMBRE 2022)

Artículo original

Determinación del umbral inferior de coincidencia aplicando medidas de edición a términos jurídicos

Determination of the lower similarity threshold applying measures of edit distance to legal terms

Lorena Matteo [(1)], *Viviana Ledesma* [(2)], *Osvaldo Spósitto* [(3)]

[(1)] Universidad Nacional de La Matanza
lmatteo@unlam.edu.ar

[(2)] Universidad Nacional de La Matanza
vledesma@unlam.edu.ar

[(3)] Universidad Nacional de La Matanza
spositto@unlam.edu.ar

Resumen:

Aplicar técnicas que ayuden a reducir el espacio de búsqueda en tareas de consultas a corpus jurídicos documentales es sumamente importante debido al volumen y diversidad de datos involucrados. Utilizando medidas de similitud léxica, en particular, aquellas basadas en cadenas de caracteres, es posible encontrar el umbral que determine el límite inferior aceptable del porcentaje de coincidencia de los términos que representan el mismo concepto. De este modo se minimiza la tarea manual de los expertos de dominio, ayudándolos a focalizarse en la revisión/validación de la similitud de aquellos términos que estén dentro de ese umbral de coincidencia. Seleccionando el término más representativo de cada concepto es posible reducir la matriz término-documento, punto de entrada para la búsqueda de información dentro del corpus.

Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de la Matanza
ISSN: 2525-1333
Vol.: 7 - Nro. 2 (DICIEMBRE 2022)

En este artículo se explica el procedimiento para encontrar el umbral de coincidencia que surge al aplicar medidas de similitud léxica a ciertos grupos de términos que representan distintos escenarios jurídicos. Estas medidas son las distancias de edición de Hamming y de Levenshtein.

Los resultados muestran que el umbral puede variar según cada escenario o medida, ayudando a los expertos a centrarse en el análisis de aquellos términos cuyo porcentaje de similitud esté dentro del umbral propuesto.

Abstract:

Applying techniques that help reduce search time in query tasks to documentary legal corpus is of great importance due to the volume and diversity of data involved. Using measures of lexical similarity, based on character strings, it is possible to find the threshold that determines the acceptable lower limit of the coincidence percentage of terms that represent the same concept. In this way, the manual task of domain experts is minimized, helping to focus on the review/validation of the similarity of those terms that are within that matching threshold. By selecting the most representative term for each concept in question, it is possible to reduce the term-document matrix, the entry point for searching for information within the corpus.

This article explains the procedure to find the coincidence threshold that arises when applying lexical similarity measures to certain groups of terms that represent different legal scenarios. These measures are the Hamming and Levenshtein edit distances.

The results show that the threshold can vary according to each scenario/measurement, helping experts to focus on the analysis of terms whose percentage of similarity is within the proposed threshold.

Palabras Clave: *Medidas de Similitud Léxica; Umbral de Similitud; Sistema de Recuperación de Información; Hamming; Levenshtein*

Key Words: *Lexical Similarity Measures; Similarity Threshold; Information Retrieval System; Hamming; Levenshtein*

Colaboradores: *Julio Bossero, Edgardo Moreno*

ISSN: 2525-1333. Vol.: 7 - Nro. 2 (DICIEMBRE-2022)

I. CONTEXTO

Este artículo se enmarca en una línea de investigación, relacionada al estudio de los Sistemas de Recuperación de Información (SRI) realizada por investigadores del Departamento de Ingeniería e Investigaciones Tecnológicas y del Departamento de Derecho y Ciencia Política de la Universidad Nacional de La Matanza. Particularmente se asocia al proyecto PROINCE, código C241, "*Implementación de un Sistema Web de Recuperación de la Información Orientado a Documentación Jurídica con el Proceso de Indexación Semántica Latente Paralelizado"*, con vigencia 2021-2022.

II. INTRODUCCIÓN

En el dominio judicial, la jurisprudencia es un factor importante como fuente de derecho; porque sus conclusiones crean una pauta para la aplicación de la ley ante situaciones jurídicas similares. Cada año el poder judicial argentino produce una gran cantidad de decisiones que se guardan en diversas formas, como ser dictámenes o expedientes, haciendo que esta fuente documental sea cada vez más voluminosa, lo que impulsa a los profesionales de la justicia a dedicar más tiempo a la búsqueda de documentos relevantes. Esto conduce a la aplicación de técnicas sofisticadas para reducir el tiempo de búsqueda y mejorar la pertinencia de los documentos recuperados.

En tal contexto, como se mencionó previamente, este grupo de investigación se encuentra trabajando en la especialización de un SRI para su utilización en un contexto jurídico. El principal objetivo es que dicho sistema permita, a partir de una consulta, recuperar documentos con características similares y útiles para la resolución de un problema legal. A su vez, se pretende diseñar y crear un corpus de referencia jurídica.

Como es sabido, es de suma importancia contar con la participación de los expertos de dominio para ir validando los términos que forman parte del corpus, muchas veces efectuando controles manuales, lo cual implica un esfuerzo considerable.

A fin de reducir ese costo de intervención manual y, para mejorar la performance en la búsqueda exhaustiva de patrones realizada inicialmente mediante el uso de expresiones regulares [1], es que surge la necesidad de aplicar técnicas complementarias que ayuden a reducir la dimensionalidad o cantidad de términos de la matriz término-documento. Dicha matriz es el punto de entrada en la búsqueda de información dentro un corpus, según se explicará más adelante.

Por tanto, si bien el procedimiento y acciones descriptos en el presente artículo surgen en principio con el objetivo de reducir los términos de la matriz de búsqueda de documentos jurídicos, esto puede aplicarse para encontrar un umbral de similitud optimo entre cualquier conjunto de términos, sea de la índole que fuera. Esto ayudará a los expertos de dominio a centrarse en el análisis manual de aquellos términos cuyo porcentaje de coincidencia se encuentre dentro del umbral sugerido.

III. SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

Un SRI, puede describirse como un conjunto de ítems de información o corpus de documentos, un conjunto de peticiones y un mecanismo que determine qué ítems satisfacen las peticiones de los usuarios. En otras palabras,

ISSN: 2525-1333. Vol.: 7 - Nro. 2 (DICIEMBRE-2022)

devuelve una lista ordenada o rankeada de documentos supuestamente relevantes para la consulta [2].

Se han ideado diferentes modelos basados en distintos paradigmas para la representación de un SRI, así como para calcular el grado de similitud entre los elementos de información para responder determinada consulta [3] [4] [5].

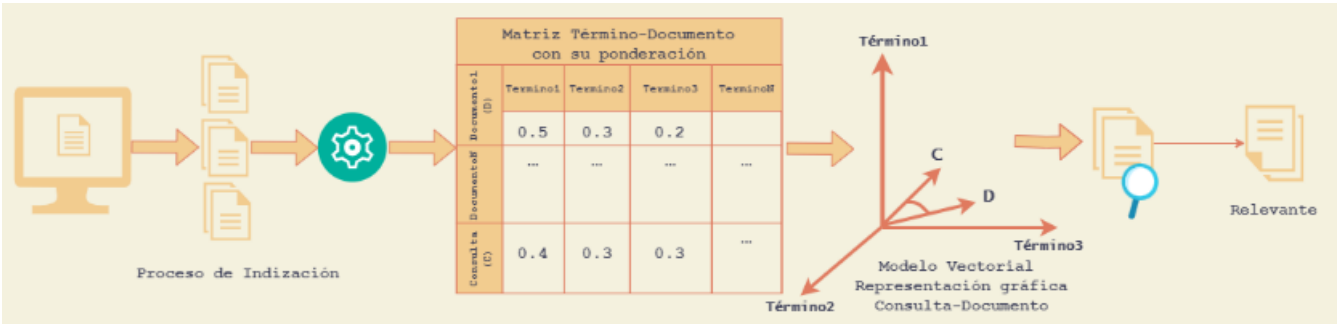

Figura 1. Proceso de búsqueda en un SRI. Fuente: Elaboración propia.

Hay tres modelos, que se consideran clásicos y son los más utilizados:

- *Booleano*: basado en la teoría de conjuntos y en el algebra de Boole. Se crea un conjunto con los elementos de la consulta y otro con los documentos, posteriormente se mide la correspondencia.
- *Vectorial*: se apoya en la idea de la importancia de un término con respecto a un documento, así como que los documentos y las consultas se pueden representar como un vector en un espacio de alta dimensionalidad. De esta manera, la consulta y los términos del documento se representan mediante dos vectores, midiéndose el grado de similitud entre ambos.
- *Probabilístico:* se calcula la probabilidad en que el documento responde a la consulta. Frecuentemente se usa retroalimentación, mediante interacción con el usuario para que indique qué documentos son más relevantes, para así reformular la consulta.

El trabajo de investigación en curso se enmarca en un modelo vectorial. Como se grafica en la Figura 1, la colección o corpus documental se representa en una matriz de término-documento. En la intersección de un término y un documento se almacena un valor numérico para denotar la importancia de tal término en el documento. Así, cada documento puede ser visto como un vector que pertenece a un espacio *n-dimensional*, donde *n* es la cantidad de términos que componen el vocabulario del corpus. En teoría, los documentos que contengan términos similares estarán cercanos entre sí sobre tal espacio. Una consulta se considera un documento más y se la mapea sobre el espacio de documentos. Entonces, a partir de una consulta dada es posible devolver una lista de documentos ordenados por distancia con los más relevantes primero.

En cualquier dominio de conocimiento, aquellos términos con significado pueden servir como descriptores para una representación lógica del contenido de documentos, así

como para las consultas en el proceso de recuperación de información [6]. Por lo tanto, una fase muy importante en un SRI es la de preparación de los documentos, esto implica que, en la fase de entrada, se deba realizar la indización y organización de la información. Según Tolosa y Bordignon, en [7] afirman que dicho proceso se puede dividir en las siguientes etapas:

- Análisis lexicográfico, se extraen las palabras y se normalizan.
- Reducción (Tokenización) de palabras vacías o de alta frecuencia.
- Lematización, se reducen palabras morfológicamente parecidas a una forma base o raíz, con la finalidad de aumentar la eficiencia de un SRI.
- Selección de los términos a indexar. Se extraen aquellas palabras simples o compuestas que mejor representan el contenido de los documentos.
- Asignación de pesos o ponderación de los términos que componen los índices de cada documento.

El trabajo que se está presentando en el presente artículo se circunscribe al proceso antes mencionado. El SRI desarrollado por este grupo de investigación, adopta inicialmente una de las representaciones más extendidas, sobre todo por su simpleza, la matriz término-documento, también llamada 'bolsa de palabras'. Es decir, se forma un vector con la frecuencia de los términos del texto, con lo cual, los documentos se caracterizan por las palabras que contienen [8].

IV. REDUCCIÓN DE DIMENSIONALIDAD DE LA MATRIZ TÉRMINO-DOCUMENTO

Las matrices conseguidas con la bolsa de palabras tienen una gran cantidad de variables o dimensiones, por no estar normalizadas, lo cual es poco útil para trabajar. Por ello, se busca una reducción de dimensionalidad, esto es llevar a una mínima cantidad posible, el número total de dimensiones que existen en el modelo del espacio vectorial.

A partir de dicha representación lógica del corpus, mediante el proceso de indización se lleva a cabo la construcción de estructuras de datos o índices a fin de brindar posteriormente soporte para la recuperación de los documentos.

Con lo anterior presente, este equipo ha propuesto, como parte de esta investigación, un algoritmo para la búsqueda y reemplazo de Entidades Nombradas (EN) utilizando Expresiones Regulares (ER). Una ER es una secuencia de caracteres que forma un patrón de búsqueda. Una EN, según [9], "*es una palabra o secuencias de palabras que se identifican como nombre de persona, organización, lugar, fecha, tiempo, porcentaje o cantidad...*". Mayor detalle de esta propuesta se incluyó en [1], donde se analizó implementar en el proceso de indización de términos de un corpus jurídico, la identificación de fechas y de referencias a EN, tales como Expediente Nº, Resolución Nº, Artículo Nº de la Ley XXX, que remiten a la norma jurídica vigente y son ampliamente utilizadas en distintos documentos judiciales.

En dicho trabajo se concluye que la aplicación de ER para encontrar EN tiene la como ventaja que:

- una vez hallada la expresión correcta, las entidades que durante la búsqueda exhaustiva

coincidan exactamente con dicho patrón serán todas las existentes en el corpus. Esto tiene mayor importancia dado que los textos legales son muy estructurados y las entidades aparecen con cierta regularidad, por otra parte,

- son fáciles de implementar ya que no necesitan más que codificar la expresión del patrón en sí, y no requieren, por ejemplo, del entrenamiento de un modelo para su reconocimiento.

Como desventaja, se sabe que estas se limitan a encontrar los patrones predefinidos, por lo cual, no es posible encontrar otra EN que no coincida con alguna de las ER existentes.

Por esta razón y considerando que el SRI debe procesar lenguaje natural, es que para ir un paso más adelante, se encaró la tarea de comparar los términos entre sí, resultando de interés detectar no sólo las coincidencias exactas entre dos términos, sino también disponer de una medida de aproximación o similitud entre estos para los casos en que la coincidencia no sea exacta. Se puso en foco la detección de términos jurídicos similares, los cuales surgen como resultado del proceso de indización y organización de la información del SRI, de este modo se pretende reducir el esfuerzo de seleccionar manualmente los términos a indexar.

Las palabras pueden ser similares léxica o semánticamente. La similitud léxica toma en cuenta si las palabras tienen secuencias de caracteres semejantes [10]. Por otro lado, las palabras tienen similitud semántica si significan lo mismo en un contexto dado, aunque léxicamente sean distintas.

Las funciones de similitud léxica han sido investigadas por décadas, existen diversos métodos o propuestas para la resolución del cálculo de la similitud de este tipo, cada una tiene sus peculiaridades según la aplicación que se le deba dar [3] [11]. Según Elmagarmid y otros [12], las distintas propuestas podrían dividirse en dos grupos: las basadas en cadenas de caracteres (distancia de edición, Brecha Afín, Smith-Waterman, Hamming, Levenshtein y Jaro, entre otras) y las basadas en tokens o secuencias de palabras (por ejemplo, Similitud de Monge-Elkan y Similitud coseno TF-IDF).

El análisis que se está presentando en este artículo se enfoca en la similitud léxica basada en cadenas de caracteres, dentro de este grupo, en la distancia de edición. Esta se define como la cantidad mínima de cambios requeridos para transformar la cadena origen en la cadena destino, en donde las operaciones permitidas se eligen de un conjunto fijo como ser la eliminación, inserción y sustitución. Como se adelantó, en este trabajo se presentan parcialmente los resultados obtenidos al comparar dos de las métricas más utilizadas en esta categoría, la distancia de Hamming (HAM) y la distancia de Levenshtein (LEV), con el objetivo de reducir la dimensionalidad de la matriz término-documento respecto de aquellos términos coincidentes. Para lograrlo, se busca encontrar un umbral de similitud aceptable que permita asumir que dos términos son representaciones de la misma EN. Ese umbral surge de comparar el porcentaje de similitud, basado en dos métricas de Precisión y Recall, ampliamente utilizadas en este tipo de ensayos como se efectuó en [16]. Estas ayudan a determinar la efectividad de las técnicas de detección de similitud de cadenas. De este modo, los expertos de dominio pueden centrarse en el análisis de aquellos términos cuyo porcentaje de similitud

esté dentro del umbral propuesto, a mayor sea su valor, mayor similitud entre los términos

IV. DISEÑO DEL EXPERIMENTO PARA LA DETERMINACIÓN DEL UMBRAL

En esta sección se explica el método aplicado para encontrar el umbral de similitud de cada grupo de términos jurídicos, creados para tal fin, en base a la efectividad resultante de las medidas de edición de caracteres.

- *Paso 1: creación de grupos de términos para representar distintos escenarios jurídicos.*

Para llevar adelante este trabajo se ha partido de una lista de 11.155 términos, es decir EN, resultantes del proceso de indización y organización de la información del SRI. La lista original contiene 3 campos: clave, término y ocurrencia en el corpus.

Con el objetivo de reducir esa lista de términos se ha recurrido a las técnicas de detección de similitud de cadenas. A modo de ensayo, se utilizó un procedimiento basado en experimentos, para ello, tal como se refleja en la Tabla 1, se armaron 5 grupos de EN significativas.

Como se mencionó anteriormente, las funciones de similitud elegidas para abordar este trabajo fueron HAM y LEV. La comparación se realizó mediante las métricas de evaluación de efectividad: precisión y recall, buscando determinar la eficacia de las funciones ante cada escenario, representado por cada grupo de entidades nombradas, Grupo EN_x, donde x identifica el caso de estudio.

Tabla 1.
Composición de experimentos por grupos de EN

Caso Estudio	Grupo EN	Cantidad Términos	Concepto
1	Grupo EN1	12	Relación con EN “Legal”
2	Grupo EN2	8	Relación con EN “Oficial”
3	Grupo EN3	16	Relación con EN “Expediente”
4	Grupo EN4	9	Relación con EN “Mediación”
5	Grupo EN5	15	Relación con EN “Fechas y Otras”
	Total EN	60	

- *Paso 2: clasificación manual de similitud entre los términos de cada Grupo de EN_x*

Para aplicar las métricas de evaluación de efectividad mencionadas es necesario que los expertos de dominio clasifiquen previamente las coincidencias reales entre cada par de términos incluidos en cada uno de los grupos de EN. Para ello se armaron matrices, donde las EN de cada grupo se colocaron en las filas y se repitieron en las columnas. Un ejemplo de ello se puede visualizar en la Figura 2, donde se muestra la clasificación de similitud para el grupo EN_2. En la intersección de cada término el experto debió realizar una clasificación manual de las coincidencias en reales/verdaderas y falsas, asignando el valor 1, cuando los considere similares o 0 en caso contrario.

ISSN: 2525-1333. Vol.: 7 - Nro. 2 (DICIEMBRE-2022)

EXPEREMIENTO CLASIFICACION MANUAL

Terminos GrupoEN2	boletinoficial	filosoficoreligi	ofici	oficial	oficializ	oficialy	oficin	suboficial
boletinoficial	1	0	0	0	0	0	0	0
filosoficoreligi	0	1	0	0	0	0	0	0
ofici	0	0	1	1	1	1	1	0
oficial	0	0	1	1	1	1	0	0
oficializ	0	0	1	1	1	1	0	0
oficialy	0	0	1	1	1	1	0	0
oficin	0	0	1	0	0	0	1	0
suboficial	0	0	0	0	0	0	0	1

Figura 2. Clasificación Manual de Similitud entre los términos del Grupo EN_2. Fuente: Elaboración propia.

Esta clasificación fue útil para comparar el resultado conseguido más tarde con la aplicación de las funciones de HAM y LEV a cada grupo, siendo dicho resultado el porcentaje de coincidencia de cada término de la matriz de similitud. De esta forma es posible evaluar la efectividad de los porcentajes de similitud, encontrando el límite inferior del umbral de coincidencia. El beneficio de esto radica en que cuando los expertos deban analizar el corpus completo, puedan enfocarse en el análisis de términos cuyo porcentaje de similitud esté dentro del umbral propuesto, reduciendo de este modo su carga de trabajo.

- *Paso 3: Cálculo de la distancia de HAM*

Esta métrica se basa en [13], es igual a la cantidad de posiciones en las que difieren ambas cadenas, y sólo permite la sustitución. Se obtiene haciendo un conteo del número de posiciones en las que los caracteres de las cadenas comparadas difieren, siendo 0 el valor resultante cuando hay coincidencia total entre las cadenas y, distinto de 0 en caso contrario. Es útil para comparar dos cadenas de caracteres de igual longitud. Es una de las métricas más simples en la que se considera el orden de los elementos.

Con lo anterior presente, el estudio realizado requirió que en un inicio las EN a comparar se ordenaran alfabéticamente. Además, se agregaron espacios en blanco a aquellas EN de la cadena de menor longitud para equiparar la cantidad de caracteres, respetando de este modo la restricción de HAM.

Se implementó una función para el cálculo, los resultados obtenidos a partir de la misma se expresaron en porcentajes de coincidencia, a modo de ejemplo, en la Figura 3 se puede observar los valores resultantes correspondientes al grupo EN_2.

Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de la Matanza
ISSN: 2525-1333. Vol.: 7 - Nro. 2 (DICIEMBRE-2022)

Terminos GrupoEN2	boletinoficial	filosoficoreligi	ofici	oficial	oficializ	oficialy	oficin	suboficial
boletinoficial	100,0	6,0	0,0	0,0	0,0	0,0	0,0	7,0
filosoficoreligi	6,0	100,0	0,0	0,0	6,0	0,0	0,0	12,0
ofici	0,0	0,0	100,0	71,0	56,0	62,0	83,0	0,0
oficial	0,0	0,0	71,0	100,0	78,0	88,0	71,0	0,0
oficializ	0,0	6,0	56,0	78,0	100,0	78,0	56,0	10,0
oficialy	0,0	0,0	62,0	88,0	78,0	100,0	62,0	0,0
oficin	0,0	0,0	83,0	71,0	56,0	62,0	100,0	0,0
suboficial	7,0	12,0	0,0	0,0	10,0	0,0	0,0	100,0

Figura 3. Porcentaje de Similitud entre términos del Grupo EN_2 usando HAM. Fuente: Elaboración propia.

Para facilitar la visualización durante el análisis de estos ensayos se utilizaron colores tipo semáforo para diferenciar los porcentajes de coincidencia: siendo la gama de verdes, según su intensidad, los más cercanos al 100%.

- *Paso 4: Cálculo de la distancia de LEV*

Esta distancia o índice también pertenece a las distancias de edición, siendo el resultado de este algoritmo dinámico la cantidad mínima de operaciones que se requiere para convertir un término en otro; entendiéndose por operaciones de edición a la inserción, eliminación o sustitución de caracteres dentro de esa EN, según se explica en [14]. Mientras mayor sea la distancia de LEV, mayor será la diferencia entre los dos términos; por ende, y al igual que en HAM, una distancia de valor igual a 0 indica que los dos términos son idénticos. Para el estudio, del mismo modo que se hizo con HAM, se implementó la función para el cálculo de LEV, en la Figura 4 se pueden observar los porcentajes resultantes para el grupo EN_2.

Tal lo mencionado en [15], esta técnica se destaca por su capacidad de detección de errores tipográficos típicos, en dicho artículo se encuentran categorizados como situaciones problemáticas, a saber: errores ortográficos y tipográficos, abreviaturas: truncamiento de uno o más términos, términos faltantes, eliminación de uno o más términos, prefijos/sufijos sin valor semántico, términos en desorden, espacios en blanco. Al trabajar con un corpus asociado a un contexto jurídico estas situaciones podrían ser corrientes, por lo que detectar la similitud entre términos con estas características es parte del objetivo de este experimento, ya que al detectarlos el experto puede decidir cuál de esos términos mejor representa a la EN en cuestión.

Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de la Matanza
ISSN: 2525-1333. Vol.: 7 - Nro. 2 (DICIEMBRE-2022)

Terminos GrupoEN2	boletinoficial	filosoficoreligi	ofici	oficial	oficializ	oficialy	oficin	suboficial
boletinoficial	100	25	36	50	36	43	36	50
filosoficoreligi	25	100	31	31	38	31	31	31
ofici	36	31	100	71	56	62	83	50
oficial	50	31	71	100	78	88	71	70
oficializ	36	38	56	78	100	78	56	50
oficialy	43	31	62	88	78	100	62	60
oficin	36	31	83	71	56	62	100	50
suboficial	50	31	50	70	50	60	50	100

Figura 4. Porcentaje de Similitud entre términos del Grupo EN_2 usando LEV. Fuente: Elaboración propia.

- *Paso 5: Obtención de medidas de Precisión y Recall*

A fin de comparar la eficacia en la detección de coincidencias utilizando las dos funciones comparadas, se aplicaron las métricas, Precisión y Recall, que permiten encontrar el límite inferior del umbral de coincidencia ya mencionado.

Acá entra en juego la clasificación manual efectuada por el experto de dominio mencionada en el Paso 2 de este apartado.

Precisión trata de responder la pregunta: ¿Qué proporción de los términos identificados como coincidentes son realmente correctos?

Es la relación entre el número de términos coincidentes identificados correctamente y el número total de términos coincidentes que ha identificado la función (de HAM o de LEV).

$$Precisión = \frac{coincidentes\ identificados\ correctamente}{número\ coincidentes\ identificados}$$

Recall, por su parte, trata de responder la pregunta: ¿Qué proporción de los términos coincidentes reales se identifica correctamente?

Es la relación entre el número de términos coincidentes identificados correctamente y el número de coincidentes que realmente hay en el grupo EN_x.

$$Recall = \frac{coincidentes\ identificados\ correctamente}{número\ coincidentes\ reales}$$

Estas medidas se calculan para cada rango de porcentajes, etiquetándolos como se ve en las Figuras 5 y 6.

PrecisiónUmbralMen60	PrecisiónUmbralMay60	PrecisiónUmbralMay70	PrecisiónUmbralMay80	PrecisiónUmbralMay90	PrecisiónUmbral100

Figura 5.Precisión para los Rangos de Umbrales de 0% a 100%.

Fuente: Elaboración propia.

RecallUmbralMen60	RecallUmbralMay60	RecallUmbralMay70	RecallUmbralMay80	RecallUmbralMay90	RecallUmbral100

Figura 6. Recall para los Rangos de Umbrales de 0% a 100%.
Fuente: Elaboración propia.

De acuerdo con los valores obtenidos, como se puede ver en la Figura 7, se descarta la columna del umbral menores al 60%, columnas *"Precisión y Recall UmbralMen60"* dado que se estaría generalizando demasiado los términos del corpus, asumiendo esas coincidencias como válidas cuando en realidad no lo son. Detenerse en ello no sería útil para los expertos dado que revisarían términos sin relación alguna y, por ende, términos que no podrían eliminarse del listado. Cabe recordar que el objetivo principal de este trabajo es facilitar las tareas reducción de la dimensionalidad de la matriz término-documento. Por otra parte, luego de los resultados obtenidos, los cuales

ISSN: 2525-1333. Vol.: 7 - Nro. 2 (DICIEMBRE-2022)

seguirán en estudio, también se descartan las columnas *"Precisión y Recall UmbralMay90"* dado que no se encontraron términos en ese rango de similitud. Finalmente, se descartan los resultados del umbral igual al 100%, columna "*Precisión y Recall Umbral100"*, ya que en dicha columna se ubican todos los términos que pertenecen a la diagonal principal de la matriz de coincidencias, es decir el cruce de cada término consigo mismo. Por tanto, los umbrales a analizar serán los comprendidos entre 60 y 99%, siempre observando los términos en orden alfabético.

- *Paso 6: Consolidación de Resultados*

A fin de visualizar los resultados de manera más clara y concisa, tal como se puede ver en las Figuras 8 y 9, se confeccionó un tablero que consolida los resultados de las métricas de efectividad resultantes para los umbrales seleccionados. Se incluye los promedios para cada una de las distancias de HAM y LEV por grupo EN_x. De esta forma es posible analizar a simple vista aquellos grupos EN_x cuyos porcentajes de similitud sean los mayores dentro de los umbrales seleccionados, ya que esos grupos contendrán términos relevantes para las tareas de revisión y reducción de las EN

GrupoEN	MetodoEdicion	Termino	CantSimilares Reales	PrecisiónUmbralMen60	PrecisiónUmbralMay60	PrecisiónUmbralMay70	PrecisiónUmbralMay80	PrecisiónUmbralMay90	PrecisiónUmbral100	RecallUmbralMen60	RecallUmbralMay60	RecallUmbralMay70	RecallUmbralMay80	RecallUmbralMay90	RecallUmbral00
GrupoEN4	Hamming	mediar	6	25%	100%	0%	100%	0%	100%	17%	33%	0%	33%	0%	17%
GrupoEN4	Hamming	mediat	3	0%	0%	0%	100%	0%	100%	0%	0%	0%	67%	0%	33%
GrupoEN5	Hamming	24/9/1991	1	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN5	Hamming	18/10/1993	2	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN5	Hamming	18/10/1993	2	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN5	Hamming	22/11/2001	1	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN5	Hamming	27/2/2006	2	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN5	Hamming	27/2/2006	2	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN5	Hamming	aÃ±o	2	0%	0%	0%	100%	0%	100%	0%	0%	0%	50%	0%	50%
GrupoEN5	Hamming	aÃ±os	2	0%	0%	0%	100%	0%	100%	0%	0%	0%	50%	0%	50%
GrupoEN5	Hamming	abiert	1	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN5	Hamming	abon	1	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN5	Hamming	abordaj	1	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN5	Hamming	acced	1	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN5	Hamming	acces	1	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN5	Hamming	acept	1	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN5	Hamming	acerc	1	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN1	Levenshtein	alegar	2	0%	0%	0%	50%	0%	100%	0%	0%	0%	100%	0%	100%
GrupoEN1	Levenshtein	alegat	2	0%	0%	0%	100%	0%	100%	0%	0%	0%	50%	0%	50%
GrupoEN1	Levenshtein	delegacion	2	0%	0%	0%	0%	0%	100%	0%	0%	0%	50%	0%	50%
GrupoEN1	Levenshtein	ilegal	2	0%	0%	0%	0%	0%	100%	0%	0%	0%	50%	0%	50%
GrupoEN1	Levenshtein	legacion	3	0%	100%	0%	0%	0%	100%	0%	33%	0%	33%	0%	33%
GrupoEN1	Levenshtein	legaj	3	0%	0%	100%	100%	0%	100%	0%	0%	33%	33%	0%	33%
GrupoEN1	Levenshtein	legajer	2	0%	0%	100%	0%	0%	100%	0%	0%	50%	0%	0%	50%
GrupoEN1	Levenshtein	legal	4	0%	0%	100%	200%	0%	100%	0%	0%	25%	50%	0%	25%
GrupoEN1	Levenshtein	legaliz	3	0%	100%	100%	0%	0%	100%	0%	33%	33%	0%	0%	33%
GrupoEN1	Levenshtein	legatari	1	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN1	Levenshtein	llegar	1	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN1	Levenshtein	supralegal	1	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	100%
GrupoEN2	Levenshtein	boletinoficial	1	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Figura 7. Precisión y Recall para los Rangos de Umbrales de 0% a 100%. Fuente: Elaboración propia.

Revista Digital del Departamento de
Ingeniería e Investigaciones
Tecnológicas de la Universidad
Nacional de la Matanza
ISSN: 2525-1333. Vol.: 7 - Nro. 2 (DICIEMBRE-2022)

Precisión y Recall Hamming por Umbrales (Entre 60 y 89%)

Etiquetas de fila	Promedio de PrecisiónUmbral May60	Promedio de PrecisiónUmbral May70	Promedio de PrecisiónUmbral May80	Promedio de RecallUmbralMay 60	Promedio de RecallUmbralMay 70	Promedio de RecallUmbralMay 80
GrupoEN1	16,67%	33,33%	29,17%	5,56%	11,81%	17,36%
GrupoEN2	18,75%	45,83%	50,00%	5,63%	18,13%	15,00%
GrupoEN3	39,27%	9,38%	9,38%	26,56%	3,65%	3,65%
GrupoEN4	50,00%	22,22%	55,56%	22,04%	7,41%	24,07%
GrupoEN5	0,00%	0,00%	13,33%	0,00%	0,00%	6,67%
Total general	**23,81%**	**18,61%**	**26,67%**	**12,25%**	**6,86%**	**11,72%**

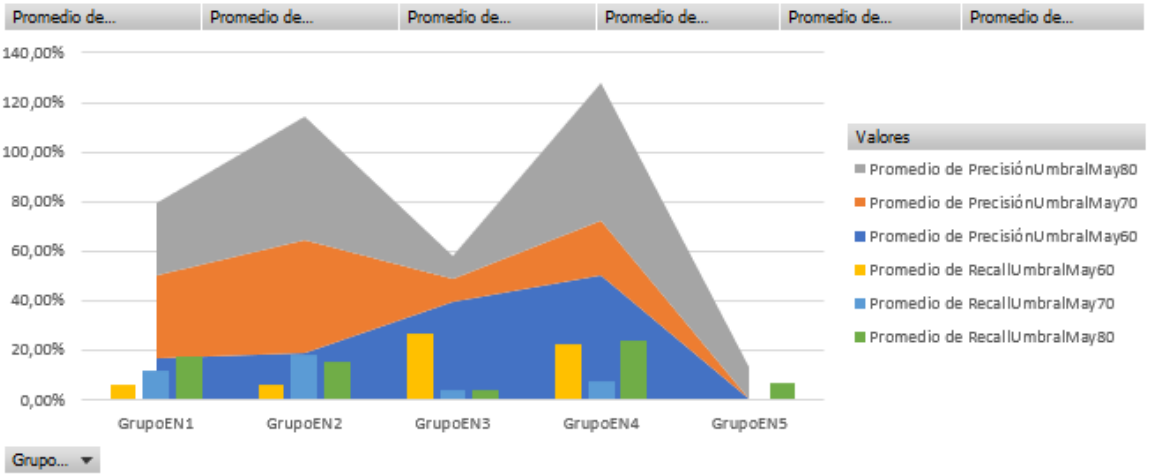

Figura 8. Tablero de control - Umbrales HAM. Fuente: Elaboración propia.

Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de la Matanza
ISSN: 2525-1333. Vol.: 7 - Nro. 2 (DICIEMBRE-2022)

Precisión y Recall Levenshtein por Umbrales (Entre 60 y 89%)

Etiquetas de fila	Promedio de PrecisiónUmbralMay60	Promedio de PrecisiónUmbralMay70	Promedio de PrecisiónUmbralMay80	Promedio de RecallUmbralMay60	Promedio de RecallUmbralMay70	Promedio de RecallUmbralMay80
GrupoEN1	16,67%	33,33%	37,50%	5,56%	11,81%	30,56%
GrupoEN2	20,83%	41,67%	0,00%	15,63%	15,63%	3,13%
GrupoEN3	18,45%	6,25%	6,25%	23,44%	4,17%	3,13%
GrupoEN4	36,11%	0,00%	11,11%	20,19%	5,56%	7,41%
GrupoEN5	6,67%	0,00%	0,00%	6,67%	0,00%	0,00%
Total general	**18,12%**	**13,89%**	**10,83%**	**14,14%**	**6,39%**	**8,47%**

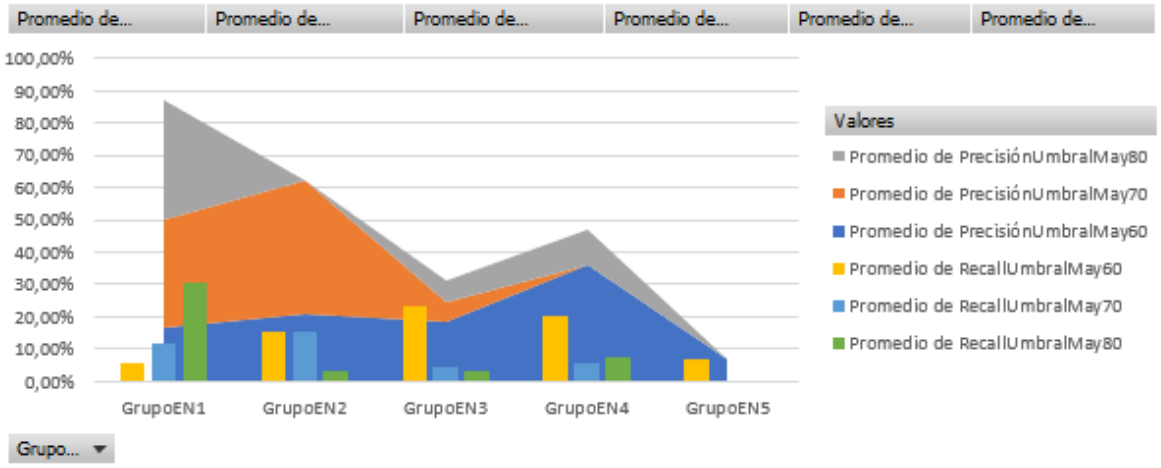

Figura 9. Tablero de control - Umbrales LEV. Fuente: Elaboración propia.

V. ANÁLISIS DE LOS RESULTADOS

Finalmente, para el análisis de los resultados del estudio realizado, se unificaron los resultados obtenidos en un tablero definitivo EN_x (ver Figura 10), lo cual permite comparar los promedios para ambas distancias de edición. Como se puede observar, la precisión de las distancias de HAM y LEV en el grupo EN_1, para el umbral mayor a 70%, es coincidente en un 33% y en aproximadamente en un 12% para el recall. En tanto, la precisión del grupo EN_2 para dicho umbral es de las mayores arrojadas por las medidas de edición. Además, el recall para ese grupo de medidas también es de los más altos. Esto denota la relevancia de que los expertos revisen la similitud de los términos que componen tal grupo. Esto es de suma importancia recordando que esa medida de efectividad indica que proporción de los términos coincidentes reales se identificó correctamente.

Gráficamente, el análisis de resultados es mucho más visible, siendo el umbral del 80%, tanto para la precisión como para recall, el porcentaje de similitud más relevante, en especial para los grupos EN_1, EN_2 y EN_4. Basándose en estos resultados, los expertos de dominio podrán

Revista Digital del Departamento de Ingeniería e Investigaciones Tecnológicas de la Universidad Nacional de la Matanza
ISSN: 2525-1333. Vol.: 7 - Nro. 2 (DICIEMBRE-2022)

focalizarse en las tareas de revisión y reducción de EN de dichos grupos en dicho umbral.

VI. CONCLUSIONES

A través del presente artículo se exhibieron los resultados obtenidos de un estudio llevado a cabo para la facilitar el proceso de detección de términos jurídicos similares.

La motivación que llevó a realizar esta actividad deriva de la importancia de acotar la matriz de término-documento, punto de entrada para el proceso de indexación, necesario para la búsqueda en el contexto de un SRI. El enfoque se puso en la búsqueda del mejor umbral de coincidencia que surge de aplicar medidas de similitud léxica a los términos resultantes del proceso de indización y organización del corpus con el que se está trabajando en la actualidad.

Precisión y Recall Hamming vs Levenshtein por Umbrales 70 y 80%

Etiquetas de fi	Promedio de PrecisiónUmbralMay70	Promedio de PrecisiónUmbralMay80	Promedio de RecallUmbralMay70	Promedio de RecallUmbralMay80
GrupoEN1	**33,33%**	**33,33%**	**11,81%**	**23,96%**
Hamming	33,33%	29,17%	11,81%	17,36%
Levenshtein	33,33%	37,50%	11,81%	30,56%
GrupoEN2	**43,75%**	**25,00%**	**16,88%**	**9,06%**
Hamming	45,83%	50,00%	18,13%	15,00%
Levenshtein	41,67%	0,00%	15,63%	3,13%
GrupoEN3	**7,81%**	**7,81%**	**3,91%**	**3,39%**
Hamming	9,38%	9,38%	3,65%	3,65%
Levenshtein	6,25%	6,25%	4,17%	3,13%
GrupoEN4	**11,11%**	**33,33%**	**6,48%**	**15,74%**
Hamming	22,22%	55,56%	7,41%	24,07%
Levenshtein	0,00%	11,11%	5,56%	7,41%
GrupoEN5	**0,00%**	**6,67%**	**0,00%**	**3,33%**
Hamming	0,00%	13,33%	0,00%	6,67%
Levenshtein	0,00%	0,00%	0,00%	0,00%
Total general	**16,25%**	**18,75%**	**6,63%**	**10,10%**

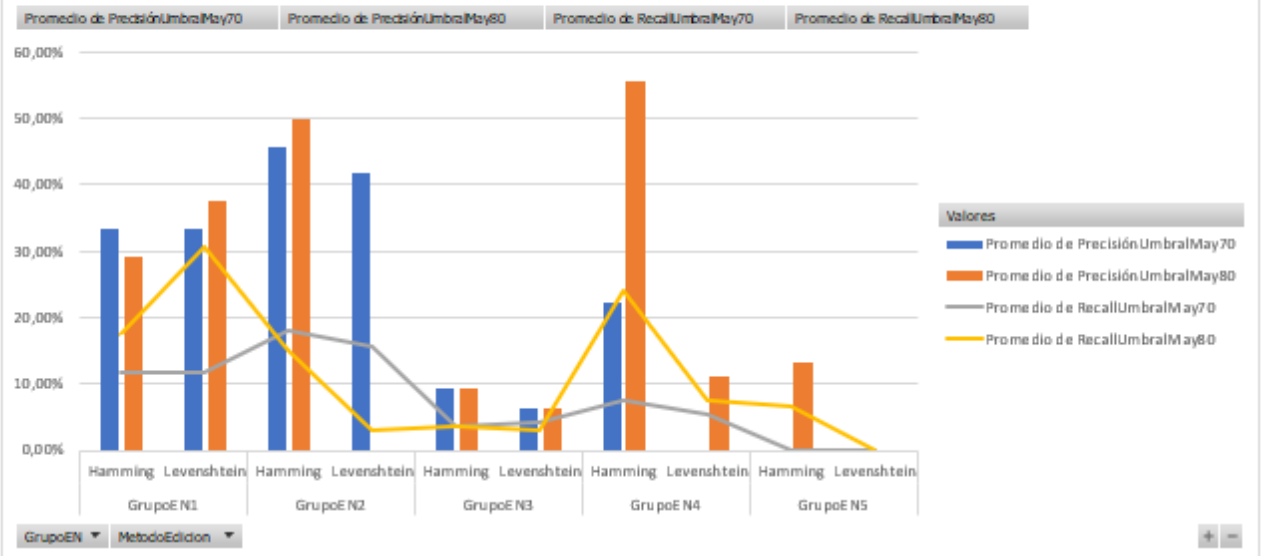

Figura 10. Tablero control - Umbrales HAM y LEV. Fuente: Elaboración propia.

Aunque este procedimiento no exime de la necesidad de contar con el experto humano, puede ser de gran ayuda para minimizar el esfuerzo implicado en su trabajo, debido a que permite acotar el volumen de términos en los que debe centrarse para elegir aquel que mejor represente a la EN.

Se ha observado que los resultados son dependientes de los términos incluidos en cada grupo de EN, y deben analizarse dentro del contexto de cada uno de los escenarios jurídicos creados.

Vale la pena mencionar que no es fácil proporcionar una solución automática, ya que se deben aplicar y adaptar varias técnicas de similitud para adecuarse a los datos concretos de que se disponen. En cuanto a este ensayo, en particular, se concluye que el límite inferior del umbral de coincidencia más relevante es el 80%. De todos modos, y como ya se ha mencionado, ha de tenerse en cuenta qué medida de edición es más fiable en cada escenario a saber: HAM debe considerarse para situaciones donde importa el orden de los caracteres de los términos en cuestión, por ejemplo, fechas, números de leyes, de expedientes, y así por el estilo. Si bien estos también pueden encontrarse fácilmente usando ER, las distancias de edición son más flexibles. En cuanto a LEV, es importante destacar que es más útil para detectar las situaciones problemáticas ya mencionadas, en dichos casos es conveniente mirar los términos del umbral con mejor precisión y recall para esa medida de edición.

VII. TRABAJOS FUTUROS

Para avanzar en esta investigación, a futuro se espera trabajar con un corpus de expedientes jurídicos de mayor volumen, así también, ampliar las medidas de similitud utilizadas.

A su vez, es necesario un análisis más exhaustivo sobre ciertos resultados llamativos dentro del tablero de control consolidado, como ser los promedios de 0% del grupo EN_5 en el umbral del 70%. Así también, será objeto de estudio el análisis de la precisión y recall para el umbralMay90, y la causa por la que no se encontraron términos en ese rango de similitud; probablemente sea necesario incluir mayor cantidad de escenarios jurídicos y por ende EN dentro de estos grupos de estudio.

Por otra parte, es necesario involucrar a mayor cantidad de expertos del dominio, para la validación de los resultados obtenidos, de cara a lograr la automatización del proceso de búsqueda del umbral de coincidencia del corpus completo.

VIII. REFERENCIAS Y BIBLIOGRAFÍA

A. Referencias bibliográficas:

[1] O. Spositto, J. Bossero, E. Moreno, V. Ledesma, & L. Matteo. “Lexical Analysis Using Regular Expressions for Information Retrieval from a Legal Corpus”, en *Computer Science – CACIC 2021*. Springer International Publishing, 2022.

[2] G. Kowalski. “*Information Retrieval Systems: Theory and Implementation*”, 1st ed. Norwell, MA, USA: Kluwer Academic Publishers, 1997.

[3] C. Lorenzetti. “Caracterización Formal y Análisis Empírico de Mecanismos Incrementales de Búsqueda basados en Contexto”. *Tesis Doctoral en Ciencias de la Computación* - Universidad Nacional del Sur. Buenos Aires, Argentina, 2011.

[4] G. Salton & M. Lesk. "Computer Evaluation of Indexing and Text Processing". J. *ACM*, 15(1): 8–36, 1968.

[5] P. Castells, M. Fernandez & D. Vallet. "An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval". *IEEE Transactions on Knowledge and Data Engineering*. 19(2): 261 – 272, 2007.

[6] J. Robredo. "Otimização dos processos de indexação dos documentos e de recuperação da informação mediante o uso de instrumentos de controle terminológico". *Ciência Da Informação*, 47(1). 2019. Disponible en: http://revista.ibict.br/ciinf/article/view/4431. Fecha de consulta: 07/02/22.

[7] G. Tolosa & F. Bordignon. "Introducción a la Recuperación de Información: Conceptos, modelos y algoritmos básicos". Universidad Nacional de Luján, Argentina, 2008. Disponible en: http://eprints.rclis.org/12243/1/Introduccion-RI-v9f.pdf. Fecha de consulta: 07/02/22.

[8] B. Harish & S. Guru & M. Shantharamu. "Representation and Classification of Text Documents: A Brief Review". *International Journal of Computer Applications, Special Issue on RTIPPR*. 1. 110 – 119, 2010.

[9] C. Sánchez Pérez. "Clasificación de Entidades Nombradas utilizando Información Global". *Tesis de Maestría*. Instituto Nacional de Astrofísica, Óptica y Electrónica. 2008. Disponible en: https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/564/1_/SanchezPCR.pdf. Fecha de consulta: 06/03/2022.

[10] W. Gomaa & A. Fahmy. "A Survey of Text Similarity Approaches". *International Journal of Computer Applications*. 68(13), 2013.

[11] I. Amón, C. Jiménez. "Funciones de Similitud sobre Cadenas de Texto: Una Comparación Basada en la Naturaleza de los Datos", Disponible en: https://repositorio.unal.edu.co/bitstream/handle/unal/69915/71644758.20104.pdf?sequence=3&isAllowed=y. Fecha de consulta: 21/09/2022.

[12] A. Elmagarmid, P. Ipeirotis, & V. Verykios. "Duplicate Record Detection: A Survey". *IEEE Transactions on Knowledge and Data Engineering*, 19 (1): 1-16, 2007.

[13] R. Hamming. "Error detecting and error correcting codes". *The Bell System Technical Journal*; Vol. XXVI, No. 2, pp. 147-160, 1950.

[14] E. Gómez Ballester, "Aportaciones a la mejora de la eficiencia de la búsqueda del vecino más cercano", pp.5,19,137, [en línea], Fecha de consulta: 7/11/2022, https://rua.ua.es/dspace/bitstream/10045/28363/1/tesis_%20evagomezballester.pdf

[15] I. Amón, C. Jiménez, "Funciones de Similitud sobre Cadenas de Texto: Una Comparación Basada en la Naturaleza de los Datos", [en línea], Fecha de consulta:21/09/2022, https://repositorio.unal.edu.co/bitstream/handle/unal/69915/71644758.20104.pdf?sequence=3&isAllowed=y

[16] I.G. Albeniz, J.R. González de Mendivil, "Estudio sobre la detección de duplicados en orígenes de datos heterogéneos", [en línea], Fecha de consulta: 22/09/2022, https://academica-e.unavarra.es/xmlui/bitstream/handle/2454/16765/TF

G_Gorostizu_Albeniz_Ion.pdf;jsessionid=6C646114AECD758F433EF12200A60A92?sequence=1

B. Bibliografía:

C. Cardellino C., M. Teruel, L. Alonso Alemany, & S, Villata. "A Low-cost, High-coverage Legal Named Entity". 2017. Disponible en: https://hal.archives-ouvertes.fr/hal-01541446/document. Fecha de consulta: 28/10/2022.

M. Cucatto. "El lenguaje jurídico y su desconexión con el lector especialista: El caso de a mayor abundamiento." Letras de Hoje, 48 (1), pp. 127-138, 2013. Disponible en: http://www.memoria.fahce.unlp.edu.ar/art_revistas/pr.9102/pr.9102.pdf. Fecha de consulta: 06/8/2022.

C. Dozier, M. Light, A. Vachher, S. Veeramachaneni & R. Wudali. "Named Entity Recognition and Resolution in Legal Text". Semantic Processing of Legal Texts, pp.27-43, 2010.

Rodríguez Inés, P. El uso de corpus electrónicos para la investigación de terminología jurídica (2008) Disponible en: https://www.tdx.cat/bitstream/handle/10803/286111/pri1de2.pdf?sequence=1. Fecha de consulta: 06/06/2022

V.I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals", Soviet Physics Doklady, pp.10:707–710, 1966. Disponible en: https://ui.adsabs.harvard.edu/abs/1966SPhD...10..707L/abstract Fecha de consulta: 08/11/2022

Recibido: 2022-11-18
Aprobado: 2022-12-23
Hipervínculo Permanente: https://doi.org/10.54789/reddi.7.2.4
Datos de edición: Vol. 7 - Nro. 2 - Art. 4
Fecha de edición: 2022-12-29