UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

FACULTAD 3

PCA: Procedimiento para la extracción automatizada de conceptos en corpus documentales legales

Trabajo de Diploma para optar por el título de
Ingeniero en Ciencias Informáticas

Autor:
Lindamelia Sánchez Mondeja

Tutor:
MSc Julio Cesar Díaz Vera
Ing. Guillermo Manuel Negrín Ortiz

La Habana, junio de 2019
"Año 61 de la Revolución"

DECLARACIÓN DE AUTORÍA

Declaro ser la autora de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

_____ _____

Lindamelia Sánchez Mondeja — Julio Cesar Díaz Vera

Autora — Tutor

_____

Guillermo Manuel Negrín Ortiz

Tutor

DATOS DE CONTACTO

Síntesis del Tutor

Ing. Guillermo Manuel Negrín Ortiz:

Ingeniero en Ciencias Informáticas. Graduado de la Universidad de Ciencias Informáticas en el año 2014. Profesor Asistente del Departamento de Ingeniería y Gestión de Software, Facultad 3 en las asignaturas de Sistemas de Bases de Datos I y II.

MSc. Julio César Díaz Vera:

Ingeniero en Telecomunicaciones y Electrónica. Graduado de la Universidad Central de las Villas Martha Abreu en el año 2003. Máster en Gestión de Proyectos Informáticos. Profesor Auxiliar del Departamento de Ingeniería y Gestión de Software, Facultad 3 en las asignaturas de Sistemas de Bases de Datos I y II.

DEDICATORIA

A mis padres y hermano,

Por su ejemplo de perseverancia y dedicación,

por su estimulante optimismo y confianza,

por su amor sin límites

A mis padres por todo su amor y apoyo incondicional durante toda mi vida.

A mi hermano por hacerme ver que no existen imposibles.

A mis tutores Julio y Guille por dedicarme más tiempo del que podían y debían.

A mi profesora Dariela por su apoyo en estos años.

A todos los profesores que de una forma u otra contribuyeron con mi formación profesional.

RESUMEN

La extracción de conceptos relevantes en un corpus documental es un área de la minería de texto que ha alcanzado mucho interés en la última década en tareas de análisis de sentimiento y recuperación de información, entre otras. En el campo del gobierno electrónico puede ser utilizado para contribuir a la comprensión de las leyes por parte de los ciudadanos. Para el lenguaje español no existen técnicas estándares como si lo hay para el inglés, que ayuden a llevar a cabo el proceso automatizado de extracción de conceptos. En este trabajo se desarrolla una propuesta con base estadística y por tanto independiente del idioma para extraer los conceptos relevantes en corpus documentales legales escritos en el idioma español. Los resultados alcanzados mejoran el desempeño de los extractores estadísticos en las métricas de $Precision$ y $Recall$.

Palabras claves: <minería de texto, extracción de conceptos, extractores estadísticos >

ABSTRACT

Relevant concept extraction from documental corpuses is an area of text mining which has gained much interest over the last decade, primarily in tasks regarding sentiment analysis and information retrieval. In the field of e-government could contribute to achieve a better understanding of the law. There are no techniques for Spanish like those for the English that help to carry out the automatized concept extraction process. This research develops a statistical based proposal hence language non-dependent to extract relevant concepts from legal documents written in Spanish. The results for the measures $Precision$ *and* $Recall$ *obtained offer better results comparing to the other approaches.*

Keywords: < **text mining, concept extraction, statistical extractors** >

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

El uso de las Tecnologías de la Información y las Comunicaciones (TIC) ha generado el importante concepto de gobierno electrónico, relacionado directamente con la gestión pública y participación ciudadana (Abarca, 2013). El gobierno electrónico es una de las áreas de mayor crecimiento en los últimos años, el cual se ha venido constituyendo como un elemento posible y factible para la modernización del Estado y el reforzamiento de la democracia.

Desde el punto de vista social, desde la última década del pasado siglo, los individuos han optado por comunicarse a través de medios electrónicos. Por lo tanto, una administración al servicio de los ciudadanos debe permitir interacciones de la misma forma en que ellos lo hacen con sus pares, es decir usando medios electrónicos.

Dentro de los esfuerzos para implementar sistemas del gobierno electrónico se encuentra, en la India, el sistema público de redirección de quejas en línea (OPGRS, por sus siglas en inglés). Este sistema permitió el descubrimiento de relaciones significativas antes no conocidas (Dwivedi et al. 2011). En los Estados Unidos se ha creado un portal de servicios orientado al usuario, basado en las tecnologías de la web 2.0, que son accesibles desde varios canales (Sun, Ku, Shih 2015). Por otro lado en España, existe la iniciativa para la implementación de gobiernos electrónicos municipales (E. Calver, S. de Juana Espinosa 2019). En Cuba se ha implementado varios sistemas informáticos en esta área, por ejemplo, el Sistema de Gestión Fiscal, el Sistema de Informatización de los Tribunales Populares Cubanos, el Observatorio de Cámara de Comercio, entre otros proyectos de los que la Universidad de Ciencias Informáticas ha sido participe.

Los ejemplos citados anteriormente abordan la gestión de gobierno, mediante el uso de plataformas en líneas, que tienen disponibles las leyes vigentes para facilitar su acceso. Sin embargo, este acceso no incide directamente en la comprensión de la ley como tal y este es un elemento clave para que el gobierno electrónico se desarrolle de manera fluida.

En ese sentido es fundamental presentar los conceptos legales de una manera en que puedan ser interpretados por los ciudadanos. La construcción de glosarios de términos manuales es una tarea que conlleva gran cantidad de recursos y un gran número de expertos, además, no hay garantías de que la explicación en el glosario sea suficiente. Por lo que, la construcción de modelos automatizados resulta una tarea importante en esta investigación.

La extracción de conceptos que son relevantes en un texto, es un área de investigación ampliamente abordada en la minería de texto. Ha sido aplicada en diferentes áreas del conocimiento: recuperación de información y clasificación de documentos (Alksher et al. 2016), asociada a determinar relevancia

en los textos, una de las tareas más difíciles, dada a su naturaleza difusa. Sin embargo, la mayoría de las investigaciones realizadas tienen como idioma objetivo el inglés y existen muchas menos investigaciones para el idioma español.

En este trabajo se propone un procedimiento para extraer conceptos relevantes de una palabra a partir de textos legales que sean independiente del idioma, teniendo en cuenta que la relevancia del concepto va a depender del área en particular en el que vaya a ser utilizado.

Las variantes de extracción de conceptos pueden clasificarse en: lingüísticas, estadísticas o híbridas (Okumura, Miura 2015). Las variantes lingüísticas e hibridas usan filtros sintácticos los cuales son muy difíciles de implementar en el lenguaje español puesto que requerirían la intervención de un lingüista, con el que no se cuenta para el desarrollo de la investigación.

Una vez determinado que no se pueden aplicar variantes lingüísticas a la extracción de conceptos en el idioma español, se hace necesario buscar una variante para este idioma. Las variantes estadísticas logran una independencia del lenguaje, por lo tanto son una aproximación válida para abordar este problema en español. A partir de este momento se hará referencia indistintamente a independencia del lenguaje como extracción de conceptos en el idioma español.

Problema a resolver: ¿Cómo extraer de manera automática los conceptos en un corpus documental legal?

Objeto de estudio: Procesamiento de lenguaje natural.

Objetivo general: Definir un procedimiento para la extracción automatizada de conceptos relevantes en corpus documental legal.

Campo de acción: Técnicas de extracción de conceptos.

Objetivos específicos:

1. Establecer el marco conceptual con las principales definiciones y términos para realizar el proceso de extracción de conceptos.
2. Construir el procedimiento para la extracción de los conceptos en un caso de estudio.
3. Validar la propuesta.

Se define como tareas a cumplir para el desarrollo de la investigación:

- Búsqueda bibliográfica.
- Selección de la bibliografía relevante.
- Análisis de la bibliografía relevante.

- Definición de las etapas del procedimiento.
- Selección de algoritmos para calcular la frecuencia y varianza relativa de las palabras.
- Implementación de los algoritmos seleccionados.
- Chequeo de la efectividad de la implementación.
- Selección de los datasets para la experimentación.
- Selección de las métricas de calidad.
- Cálculo de las métricas de calidad.
- Comparación con los extractores $Tf - Idf,\ Zhou,\ Syllables$.
- Análisis de los resultados.

Posibles resultados: PAC: Procedimiento para la extracción automatizada de conceptos en corpus documental legal.

Estructura de la tesis

La tesis se estructura en tres capítulos.

Capítulo 1. Fundamentación Teórica. En este capítulo se aborda los conceptos asociados a la investigación para el desarrollo de la propuesta de solución, centrándose en el estudio del marco teórico conceptual asociado a la investigación y se construye el estado de arte asociado a la misma. Para el desarrollo del estado de arte se utiliza el método de revisión sistemática el cual permite llevar a cabo una evaluación minuciosa y confiable de las investigaciones realizadas dentro de la temática abordada en la tesis, donde las referencias bibliográficas toman el lugar de los datos. Lo cual implica que sean las referencias las que son analizadas para aportar evidencias a cada una de las preguntas de la investigación.

Capítulo 2. Propuesta de solución. En este capítulo se describe el método para la realización de las fases del procedimiento de extracción automatizada de conceptos (PAC). De cada fase del proceso se mostrará los fundamentos teóricos, herramientas y técnicas utilizadas en su implementación mostrando la aplicabilidad a partir de un caso de estudio, en el que se considera como corpus legal el texto de la Constitución de la República de Cuba.

Capítulo 3. Validación de la solución. En este capítulo se realiza la validación de la solución mediante la metodología de experimentación y los resultados alcanzados. Se detalla la forma en que se selecciona los umbrales de varianza relativa para cada uno de los pares de palabras a partir de los cuales se obtienen valores de precisión y recuperación satisfactorios. Al finalizar se muestran un grupo de los conceptos encontrados.

CAPÍTULO 1. FUNDAMENTACIÓN TEÓRICA

Introducción

En este capítulo se establece el marco teórico conceptual asociado a la investigación y se construye el estado de arte asociado a la misma. Para el desarrollo del estado de arte se utiliza el método de revisión sistemática el cual permite llevar a cabo una evaluación minuciosa y confiable de las investigaciones realizadas dentro de la temática abordada en la tesis, donde las referencias bibliográficas toman el lugar de los datos. Lo cual implica que sean las referencias las que son analizadas para aportar evidencias a cada una de las preguntas de la investigación.

La sección dedicada al estado de arte será dividida en tres etapas: planificación, conducción y análisis de los resultados obtenidos, con el fin de poder registrar los pasos seguidos y ofrecer la información correspondiente o ponerla a disposición de otros investigadores interesados.

1.1. Marco Conceptual

Uno de los elementos fundamentales para la futura comprensión de los elementos detallados en esta investigación, está asociada a determinar una noción compartida del término concepto. No existe una definición formal, dentro de la rama de la inteligencia artificial para este término, pero de manera intuitiva se asume por los investigadores que el término concepto puede definirse como, una palabra o una secuencia de palabras que poseen algún valor semántico. Por ejemplo, mientras que palabras como ‘presidente’ y ‘república’ pueden ser consideradas conceptos, las palabras como ‘y’, ‘de’, ‘o’ no poseen gran valor semántico. Las primeras palabras poseen valor semántico intrínseco, tienen un significado, mientras que las últimas pertenecen a la clase de palabras funcionales y no tienen ningún significado relevante.

Los conceptos pueden estar formados por más de una palabra, por ejemplo la unión de ‘presidente’ y ‘república’ forman un nuevo concepto compuesto ‘presidente de la república’. Este es más específico que los conceptos simples que lo forman. De hecho, no se hace referencia a cualquier presidente, sino específicamente al presidente de la república. Desde el punto de vista de república, no se refiere a cualquier institución de la república, sino a su presidente. En esta investigación no se tendrán en cuenta los conceptos multi-palabras, por temas asociados al tiempo disponible para el desarrollo.

El proceso de extracción de conceptos consiste en identificar aquellas frases (conjunto de palabras) o palabras simples que pueden tener un sentido conceptual (Torre 2017).

Para poder llevar a cabo el proceso de extracción de conceptos, (Zhang 2016) los subdivide en tres categorías:

- Enfoque lingüístico: se analizan los patrones lingüísticos, formulados a partir de un conjunto de categorías gramaticales.
- Enfoque estadístico: se analizan las probabilidades de ocurrencia de un término y su relevancia en el texto, con independencia total del lenguaje.
- Enfoque híbrido: se unen las categorías previamente mencionadas para una mejor exactitud de la respuesta deseada.

Las técnicas necesarias para implementar los métodos antes referidos están contenidas dentro del procesamiento del lenguaje natural (Zhang 2016), un área interdisciplinaria de las ciencias de la computación y la lingüística computacional. Desde el punto de vista computacional, pueden utilizarse dos enfoques para desarrollar las tareas de extracción de conceptos; automatizados cuando no requieren la intervención de un agente humano y semi-automatizado cuando de una manera u otra se implica al mismo (Zubrinic, Kalpic, Milicevic 2012).

El proceso de extracción de conceptos comúnmente ocurre sobre textos que son agrupados en diferentes ficheros, a esta agrupación de texto se le denomina corpus documental y son clasificados normalmente al área temática a la que pertenecen. De esta forma un corpus documental que esté compuesto por un conjunto de leyes y/o normas jurídicas es denominado corpus documental legal.

1.2. Estado del arte

1.2.1 Planificación

En esta etapa se definen las preguntas por las cuales estará encausada la investigación, las definiciones de interés, palabras claves, estrategias de búsquedas y criterios de inclusión y exclusión. Los siguientes parámetros definen las preguntas en las que estará enfocada la investigación de esta tesis.

- PI (1): ¿Cuáles son los artículos más citados?
- PI (2): ¿Qué técnicas se utilizan para la extracción automatizada de conceptos?
- PI (3): ¿Cuáles son los algoritmos utilizados en las investigaciones actuales?
- PI (4): ¿Qué tipos de validación se usan en la extracción automatizada de conceptos?
- PI (5): ¿Qué tecnologías son utilizadas para la extracción automatizada de conceptos?
- PI (6): ¿Cuáles son los *dataset*[1] utilizados en los artículos seleccionados?

A continuación, se presenta los criterios de inclusión y exclusión (ver en la tabla I), los que servirán como filtros en el proceso de selección bibliográfica, con el objetivo de lograr calidad en el proceso.

[1] conjunto de datos

Tabla I Criterios de inclusión, exclusión y calidad

Inclusión	Estudios en inglés.
	Artículos científicos, Conferencias.
	Estudios relevantes sobre la extracción automatizada de conceptos.
Exclusión	Investigaciones que no tengan la calidad deseada.
	Estudios escritos en idiomas que no sean inglés.
	Publicaciones antes del 2014.
Calidad	Estudios con diferentes propuestas de solución.
	Estudios con resultados completos.

En esta fase se decide hacer una selección teniendo en cuenta el resumen, las palaras claves, las técnicas, algoritmos propuestos y los resultados, para garantizar una mayor calidad de la revisión. Se deciden emplear las siguientes fuentes bibliográficas, para el proceso de búsqueda de los documentos.

- *ACM Digital Library: dl.acm.org*
- *IEEE Xplore Digital Library: ieeexplore.ieee.org*
- *Science Direct: www.sciencedirect.com*
- *Springer : www.springer.com*

Luego de haber seleccionado los métodos de búsqueda en las fuentes bibliográficas anteriormente citadas, se genera la siguiente cadena: (*Concept Extraction OR Automatized Concept Extraction) AND (Text Mining)*, para identificar los documentos válidos a consultar.

1.2.2. Ejecución del Plan.

Esta fase cuenta de cinco pasos claves (1) realizar la búsqueda en las bases de datos seleccionadas; (2) comparar los resultados de las búsquedas para excluir artículos repetidos; (3) aplicar criterios de inclusión, exclusión y calidad; (4) evaluar todos los estudios que superaron la revisión inicial; y (5) síntesis de datos (Bonidia et al. 2018).

En la siguiente figura se muestra un diagrama de flujo donde la primera fase consistió en ejecutar las cadenas de búsqueda en todas las bases de datos, que encontraron un gran conjunto de 1120 documentos. Por lo tanto, para lograr una mejor precisión y confiabilidad, se utilizó la herramienta **StArt** (Estado del arte a través de revisiones sistemáticas). Esta herramienta garantiza una mayor calidad de resultados, optimizando las fases de ejecución de la revisión sistemática facilitando el proceso de investigación.

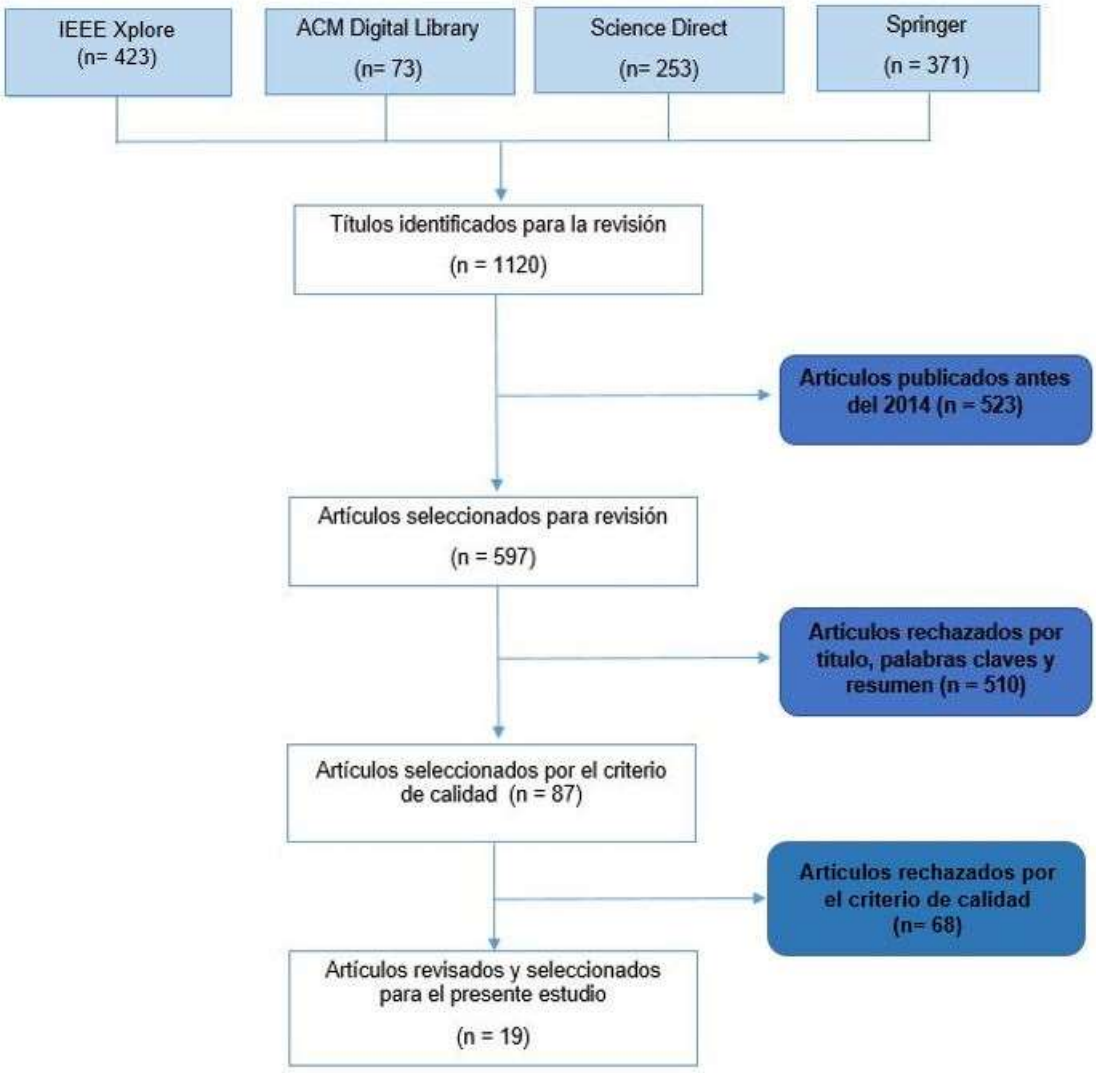


Figura 1: Proceso de selección de artículos

Finalmente, después de la preselección de los trabajos, se realizó una síntesis de los datos, con el objetivo de aplicar una evaluación basada en los criterios de calidad establecidos. De ese modo, de los 87artículos, 68 fueron eliminados (no han demostrado los métodos o las técnicas aplicadas), lo que dio lugar a un conjunto final de 19 artículos con información relevante.

1.2.3. Resultados.

En esta sección, se presentan los resultados de la búsqueda. Por lo tanto, a continuación, se reflejarán cada uno de los pasos realizados que responderá a las cuestiones planteadas al comienzo de la investigación.

RPI (1): ¿Cuáles son los artículos más citados?

Los trabajos seleccionados serán presentados en la Tabla II:

Tabla II : IDs de los artículos seleccionados y cantidad de referencias.

ID	Referencias	ID	Referencias
1	13	16	15
2	7	17	18
3	36	18	25
4	21	19	15
5	11		
6	34		
7	30		
8	41		
9	32		
10	11		
11	30		
12	9		
13	21		
14	32		
15	17		

Su estudio podrá proveer un reporte sobre el panorama actual de las múltiples variantes de solución sobre la extracción automatizada de conceptos. Este reporte refleja que dentro de los más citados se encuentran el artículo 8, 3 y 6 respectivamente.

Para proporcionar una visión general de las temáticas que se han propuesto en los artículos, fueron categorizados en cuatro clases: Comprensión de documentos, Artículos Científicos, Medicina y Educación y Literatura. Para un mejor entendimiento de las mismas en la tabla III quedan reflejados.

Tabla III Artículos agrupados por temáticas

Temática	Artículos
Compresión de documentos	[1] [4] [17] [5]
Educación y Literatura	[2] [6] [12] [8] [13]
Artículos Científicos	[3] [10] [14] [16] [18] [19]
Medicina	[7] [15] [11] [9]

- Comprensión de documentos: tratan de la capacidad de dar sentido al lenguaje libre de formato en documentos no estructurados.
- Artículos Científicos: se evidencian variantes al problema de la extracción automática de frases clave de documentos científicos.
- Medicina: se propone una fuente textual en conceptos, temas, relaciones y resúmenes para examinarla y clasificarla rápida y fácilmente asociados a la detección de patrones para la detección de síntomas e irregularidades biológicas.
- Educación y Literatura: se ha desarrollado y evaluado en el dominio de la educación científica, incorpora conceptos clave importantes para desarrollar una comprensión sólida del dominio, demuestra una alta precisión en la identificación de la concordancia de la oración cuando existe un acuerdo entre los expertos humanos en la calificación de la concisión.

En la distribución de los textos por temáticas antes mencionadas (ver Tabla III), se puede observar que, la temática de Educación y Literatura, representa la mayor cantidad de trabajos asociados, con un total de 5, seguido por Artículos Científicos, Comprensión de textos, y los referentes a medicina.

RPI (3): ¿Qué técnicas se utilizan para la extracción automatizada de conceptos?

En la siguiente tabla se muestra un resumen de las técnicas utilizadas para el proceso de extracción de conceptos.

Tabla IV: Técnicas utilizadas por artículos.

Artículos	Técnica utilizada
1	Extracción y abstracción
2	TF-IDF
3	Generar un árbol de análisis
4	Árbol de dependencia, Clúster, Relevancia marginal máxima
5	Análisis de redes
6	TF-IDF
7	NN- basado en redes neuronales
8	Análisis de redes
9	PMI y TF-IDF
10	TF-IDF
11	Análisis de redes
12	Patrones sintácticos
13	Entropía, TF-IDF

14	Análisis de redes
15	Patrones sintácticos
16	Patrones sintácticos
17	Programación lineal y regresión lineal
18	Zhou y Slater
19	Syllables

- **Extracción y abstracción:**

El resumen abstractivo interpreta el contenido del texto original, entiende la relación semántica entre las oraciones y produce un resumen. Considerando que, el resumen extractivo recupera sólo las oraciones más relevantes del documento fuente, manteniendo así una baja redundancia. El resumen automático de texto encuentra su aplicación en el área de medios masivos, motores de búsqueda, área de noticias, área de bolsa (Naik 2017).

- **Tf-Idf:**

$Tf - Idf$ (Frecuencia de término - Frecuencia inversa de término) calcula la relevancia de las palabras en los documentos. Esencialmente, esta técnica mide la importancia de una determinada palabra en un documento con respecto a otros documentos de la misma colección (Zhang 2016). Básicamente, una palabra es más importante en un documento determinado cuanto más aparece en ese documento, pero si esa palabra aparece en otros documentos, su importancia disminuye. Las palabras que son muy frecuentes en un solo documento tienden a ser más valiosas que las palabras comunes que aparecen en más documentos, como artículos o preposiciones. Formalmente, siendo W una palabra, la importancia de W para un documento d_j en un corpus D, se define de la siguiente manera:

$$Tf - Idf(W, d_j) = Tf(W, d_j) \times Idf(W, d_j) = \frac{f(W,d_j)}{size(d_j)} \times log \frac{||D||}{||\{d:W \in d\}||} \quad (1)$$

donde $||D||$ significa el número de documentos en el corpus D; $||\{d: W \in d\}||$ es el número de documentos que contienen el término W y $\text{size}(\text{d}_\text{j})$ el número de palabras en el documento d_j. Para evitar descentramientos hacia documentos más largos, la probabilidad $f(W, d_j)$ / tamaño (d_j) del término W en el documento d_j se usa comúnmente en lugar de la frecuencia absoluta $f(W, d_j)$(Wu et al. 2017).

- **Zhou y Slater:**

Es una métrica propuesta por (Zhou, Slater 2003) para calcular la relevancia de las palabras individuales en un texto. Como suma, se pueden encontrar palabras relevantes en ciertas áreas de los

textos, ya sea formando parte de temas locales o relacionados con contextos locales, formando grupos en esas áreas. Por otro lado, las palabras comunes y menos relevantes deben aparecer al azar en todo el texto, sin formar agrupaciones significativas. Esta técnica mide la relevancia de una palabra según la posición de ocurrencia de cada texto del corpus. Para una palabra w, los autores comienzan con una lista $L_w = \{-1, t_1, t_2, \ldots, t_m, n\}$, donde t_m representa la posición de la $m - ésima$ ocurrencia de la palabra w en el texto y n representa el número total de palabras en el mismo texto. Luego, obtienen u, que es básicamente la separación promedio entre apariciones consecutivas de la palabra w para el caso de una distribución uniforme de las ocurrencias.

$$u = (n + 1)/(m + 1) \quad (2)$$

El siguiente paso consiste en el cálculo de la separación promedio entre apariciones consecutivas reales de la palabra w en el texto. Se utilizan 3 ocurrencias consecutivas para cada cálculo.

$$d(t_i) = \frac{t_{i+1} - t_{i-1}}{2} \quad i = 1,2, \ldots, m \quad (3)$$

Finalmente, la puntuación de la palabra w se mide con la ecuación cuatro. Siendo la información sobre si t_i pertenece o no a un clúster en $\delta\ (t_i)$ y en $v\ (t_i)$ la separación normalizada a la distancia promedio, Γ (w) tiene el valor de v (t_i) cuando t_i pertenece a un clúster y cero en caso contrario (Zhou, Slater 2003).

$$\Gamma(w) = \frac{1}{m} \sum_{i=1}^{m} \delta\ (t_i) \times\ v\ (t_i) \quad (4)$$

- **Entropía:**

Según (Ismail 2018), para cada clúster se asume la entropía como:

$$entropía(D_i) = -\sum_{j-1}^{k} Pr_i(C_j) \times\ log_2\ Pr_i(C_j) \quad (5)$$

Dónde Pr_i es la proporción de puntos de la clase (C_j) ubicados en el clúster i o D_i. La entropía total de todo el agrupamiento (que considera todos los clústeres) es:

$$entropía_{total}\ (D) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} \times\ entropía(D_i) \quad (6)$$

- **Generar un árbol de análisis:**

La técnica propuesta se basa en la generación de un árbol de análisis para cada oración en el texto de entrada. Los árboles de análisis de oraciones se cortan en subárboles para extraer las ramas de las frases candidatas (es decir, nombre, verbo, etc.). Los subárboles se combinan luego utilizando etiquetas de partes del discurso para generar una lista plana de frases candidatas.

Finalmente, el filtrado se realiza utilizando reglas heurísticas y las frases redundantes se eliminan para generar una lista final de características candidatas.

La técnica propuesta se ajusta para determinar el valor óptimo para el tamaño de la ventana de contexto del parámetro y luego se compara con las técnicas convencionales existentes basadas en n-gramas y frases nominales(Aman et al. 2018).

- **Patrón sintáctico:**

Se utiliza para extraer conceptos de un determinado texto, utilizando el método supervisado con enfoque de patrón sintético léxico para definir las propiedades de los conceptos e identificar la semántica adicional relacionada con los atributos de los conceptos (Guo 2018).

- **Syllables:**

Según (Neves, Santos, Machado 2007), propone dos métricas para calcular el núcleo de una palabra w según las relaciones con sus palabras de éxito (las palabras que aparecen justo después de w-ecuación 7) y con sus predecesores (todas las palabras aparecen justo antes de w-ecuación 8).

$$Sc_{suc}(w) = \sqrt{\frac{1}{||Y||-1} \times \sum_{y_i \in Y} \left(\frac{p(w,y_i)-p(w,.)}{p(w,.)}\right)^2} \quad (7)$$

$$Sc_{pre}(w) = \sqrt{\frac{1}{||Y||-1} \times \sum_{y_i \in Y} \left(\frac{p(y_{i,w})-p(.,w)}{p(.,w)}\right)^2} \quad (8)$$

$$p(w,.) = \frac{1}{||Y||} \times \sum_{y_i \in Y} p(w,y_i) \qquad p(w,.) = \frac{1}{||Y||} \times \sum_{y_i \in Y} p(y_i,w) \quad p(a,b) = \frac{f(a,b)}{N} \quad (9)$$

Donde Y es el conjunto de palabras en el corpus, $||Y||$ representa su tamaño y N es el número de ocurrencias de palabras que aparecen en el corpus. $f(a,b)$, es la frecuencia de aparición de las dos gramas (a,b) en el mismo corpus. La puntuación final es dada por $Sc(w)$:

$$Sc(w) = \frac{Sc_{pre}(w) + Sc_{suc}(w)}{2} \quad (10)$$

- **NN- basado en redes neuronales:**

La red neural permite realzar las acciones aplicadas en la extracción de cualquier palabra de los textos. Se buscan "palabras destacadas", que pueden describir temas de documentos y luego encontrar documentos que coincidan con las consultas de los usuarios. El modelo de red neuronal consta de varios nodos. A cada nodo se le asigna una palabra de una consulta de búsqueda definida por el

usuario. La salida de la red es una lista de palabras individuales obtenidas del texto (Toti, Rinelli 2016). Esa lista incluye solo las primeras 200 palabras de cada documento en el corpus.

- **Árbol de dependencia, Clúster, Relevancia marginal máxima:**

Árbol de dependencia y Relevancia marginal máxima: Un árbol de dependencias organiza las estructuras para que haya relaciones explícitas entre las estructuras, donde cada rama contiene el peso máximo que puede presentar, es decir, la información mutua entre x_i y x_j es un máximo. La distribución de probabilidad de la distribución de árbol para una dependencia de árbol $P_t(x)$ es una aproximación óptima a $P(x)$ si y solo si su árbol de dependencia es de máximo peso (Biyabangard 2015).

Clúster: identifica grupos de documentos relacionados o extraer temas del conjunto de contenido. Se utiliza para investigar los tipos de registros en un conjunto de contenido o identificar conjuntos de documentos similares (Beil et al. 2014).

- **Programación lineal y regresión lineal:**

Aborda el resumen del texto como un problema de mochila, es decir, maximiza la cobertura de la información relevante para tomar en cuenta la configuración de restricciones, especialmente el umbral para la longitud del resumen de salida, mientras que la regresión lineal se utiliza para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes y un término aleatorio (Malin, Thronesbery, Throop 2016).

RPI (4): ¿Cuáles son los métodos utilizados en las investigaciones actuales?

En la tabla V se resumen los métodos utilizados para cada artículo seleccionado para la investigación.

Tabla V: Relación de artículos por métodos.

Artículos	Método utilizado
1	Resumen basado en Reglas
2	Asociación ponderada de minería de reglas
3	Extracción de frases clave
4	Resumen de documentos múltiples utilizando la descomposición del tensor, HOSVD por sus siglas en inglés.
5	Resumen automático de documentos múltiples que emplea la estructura de oraciones del sujeto, predicado, objeto y complemento, SPOC por sus siglas en inglés.
6	Resumen de documentos múltiples (MDS)
7	-
8	-

9	Método para evaluar y visualizar información relevante para guiar el proceso de elaboración
10	-
11	Creación de vocabulario de palabras visuales e indexación textual de informes médicos.
12	Manera jerárquica para generar ontología automática.
13	Patrón Regex
14	*Automatic Key Extraction (AKE)*
15	-
16	-
17	Programación lineal integral resumen de un solo documento
18	Zhou y Slater
19	Isla

- **Resumen basado en Reglas**:

Durante el procesamiento, el documento se pasará a través de pasos de pre procesamiento. Las palabras clave se extraerán del documento y, según el umbral calculado, las palabras clave se eliminarán. Además, el documento estará sujeto a varios métodos de extracción de características en los que las oraciones se representan como un vector de características. Se escribirá una regla y todas las oraciones se pasarán a través de esta regla. Finalmente, las oraciones serán ordenadas y solo se seleccionarán las oraciones superiores en función de la extensión del resumen definido para producir el resumen (Naik 2017).

- **Asociación ponderada de minería de reglas:**

La relación entre los conceptos se expresa directamente como la probabilidad de ocurrencia común. Las reglas de asociación no solo pueden extraer los conceptos que tienen una fuerte correlación con el concepto objetivo, sino que también pueden extraer la relación entre conceptos (Zhang 2016).

En primer lugar, se obtiene el conjunto de texto relacionado con el dominio de la ontología extendida desde la web; en segundo lugar, se obtiene el conjunto de conceptos y el conjunto de transacciones a través del pre-procesamiento de documentos, y obtenemos conceptos que tienen la relación semántica a través del análisis de asociación ponderada (Zhang 2016).

- **Resumen de documentos múltiples utilizando la descomposición del tensor**:

El método propuesto es un resumen de varios documentos. En la tarea de documentos múltiples, hay varios documentos por tema, cada uno de ellos habla sobre el tema desde diferentes vistas. Por lo tanto, en varios documentos, existe la probabilidad de referirse a un importante concepto inseparable con oraciones de apariencias variadas. Sin embargo, un resumen bien estructurado comprende las oraciones importantes con los conceptos menos repetitivos(Biyabangard 2015). En cuanto al tema, las similitudes semánticas entre oraciones se miden según *WordNet*[2] y eliminan la redundancia, de modo que el resumen resultante carece de conceptos y redundancias repetidas.

- **SPOC:**

La estructura de la oración se extrae de un árbol de dependencias utilizando unas reglas definidas. Las reglas extraerán elementos del árbol que coincidieron con relaciones determinantes para cada tipo de estructura de oración correspondiente. El conjunto de relaciones determinantes se decide observando los árboles de dependencia a partir del conjunto de oraciones (Reztaputra 2017).

- **MDS:**

Extrae las oraciones más centrales de varias fuentes textuales, los textos se representan en redes, donde los nodos representan oraciones y las aristas se establecen en función del número de palabras compartidas. A diferencia de trabajos anteriores, la identificación de términos relevantes se guía por la caracterización de nodos a través de mediciones dinámicas de redes complejas, incluida la simetría, la accesibilidad y el tiempo de absorción (Tohalino, Amancio 2017).

- **Método para evaluar y visualizar información relevante para guiar el proceso de elaboración:**

El algoritmo se ha desarrollado y evaluado en el dominio de la educación, en el que la concisión se refiere al grado en que una oración incorpora conceptos clave importantes para desarrollar una comprensión sólida del dominio. El método funciona computando y aprovechando automáticamente el grado de similitud semántica entre oraciones de recursos y conceptos de dominio estándar diseñados por expertos humanos para varios dominios STEM (Trovati 2017).

- **Creación de vocabulario de palabras visuales e indexación textual de informes médicos:**

Propone un nuevo enfoque para la anotación semántica automática de imágenes médicas. Se utiliza el modelo de bolsa de palabras para caracterizar el contenido visual de la imagen médica combinada con descriptores de texto basados en la técnica de frecuencia de documentos de frecuencia inversa y

[2] Base de datos léxica del Idioma inglés

reducida por semántica latente para extraer la coocurrencia entre el texto y los términos visuales(Zhang et al. 2017).

- **Zhou y Slater**

El método utiliza las fluctuaciones de densidad de una palabra para calcular un índice que mide su grado de agrupamiento. Las palabras altamente significativas tienden a formar grupos, mientras que las palabras comunes se reparten esencialmente de manera uniforme en un texto. Si una palabra no es rara, la métrica es estable cuando movemos cualquier aparición individual de esta palabra en el texto (Zhou, Slater 2003).

- **Patrón Regex:**

El patrón se basa en el procesamiento del lenguaje natural que utiliza el etiquetado en sus reglas. Se propone un método que usa patrones para extraer conceptos para el desarrollo de la ontología del *Hajj*[3]. Comparándose con un sistema de aprendizaje ontológico prominente,*Text2Onto* y los patrones con el patrón *Qterm* que está diseñado específicamente para el dominio Solah en el Corán (Ismail 2018).

- **AKE**:

Proponen un conjunto de sistemas de extracción de palabras clave automáticas Algoritmo para la extracción de palabras claves (KEA) y Búsqueda sin supervisión de conocimiento para conceptos de creación de instancias en ontologías ligeras (KUSCO) y Campos Aleatorios Condicionales (CRF). A diferencia de KEA y KUSCO, que son herramientas bien conocidas para la extracción automática de palabras clave, CRF necesita un pre-procesamiento adicional (Geadas, Alves, Ribeiro 2014).

- **Isla:**

Según (Neves, Santos and Machado 2007), se extraen las palabras relevantes locales. La idea del método de las Islas es que una palabra w es relevante si su puntaje es consistentemente más alto que sus vecinos inmediatos. Si $r\ (w)$ es el puntaje de una palabra dada por $Sc\ (w)$ con el análisis de sílabas), la relevancia de w viene dada por la ecuación siguiente:

$$Relevancia\ (w) = \begin{cases} 1 & r(w) > 0.9 \times max(Avg_{pre}(w), Avg_{suc}(w)) \\ 0 & en\ otro\ caso \end{cases} \quad (11)$$

RPI (5): ¿Qué tipos de validación se usan en la extracción automatizada de conceptos?

[3] http://corpus.quran.com/ontology.jsp

Para validar cada una de las propuestas de solución de los trabajos previamente seleccionados, se muestra una relación por trabajos de las métricas utilizadas para dichas validaciones. En la Figura 2 quedan evidenciados.

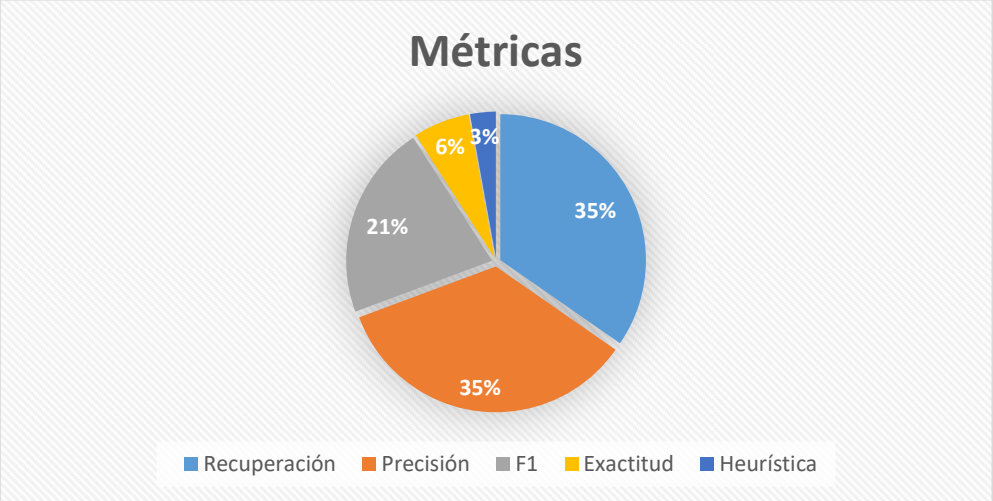


Figura 2: Métricas más utilizadas para validar los resultados obtenidos.

Exactitud: es la proporción de oraciones a las que se asigna el nivel de concisión correcto (Bethard 2017).

Precisión: es una suma ponderada de precisiones sobre todas las clases, donde el peso de la precisión para una clase es proporcional al número de oraciones en esa clase (Ismail 2018).

Recuperación o *Recall*: mide la probabilidad de que si se selecciona un concepto como frase clave, el método lo identifica correctamente. Da la proporción de los conceptos clave extraídos correctamente entre todos los conceptos claves (Aman et al. 2018).

F_1: es la media armónica de precisión y recuperación (Zhou, Slater 2003). Normalmente, la precisión disminuye a medida que aumenta la recuperación y viceversa.

Heurística: se emplea para mejorar los valores resultantes de la métrica exactitud.

De esta forma, por lo antes expuesto, se puede decir que, dentro de las métricas más utilizadas en la validación de métodos para la extracción de conceptos, se encuentran la precisión y recuperación con un 35 por ciento de ocurrencia en el total de los 18 artículos seleccionados.

RPI (7): ¿Qué tecnologías son utilizadas para la extracción automatizada de conceptos?

1. NLTK librería de Python para el procesamiento del lenguaje natural.

El kit de herramientas de lenguaje natural, o más comúnmente **NLTK,** es un conjunto de bibliotecas y programas para el procesamiento del lenguaje natural (PLN) simbólico y estadísticos en el lenguaje de programación Python. Proporciona interfaces fáciles de usar para recursos corporales y léxicos, junto con un conjunto de bibliotecas de procesamiento de texto para clasificación, tokenización, derivación, etiquetado, análisis y razonamiento semántico, para bibliotecas PLN de gran solidez industrial(P et al. 2018)(Bird 2017).

NLTK está disponible para Windows, Mac OS X y Linux. La ventaja de esta es un proyecto gratuito, de código abierto y dirigido por la comunidad.

2. Etiquetado de voz o POS por sus siglas en inglés con MXPost Tagger.

Etiquetador de voz máxima para entropía en el lenguaje de programación Java, este a partir de un modelo de tuplas encuentra la etiqueta POS más probable para una palabra (Mejri, Akaichi 2014).

3. ROGUE.

Sub-estudio orientado a recordar para la evaluación de *gisting*[4] (ROGUE por sus siglas en inglés) ayuda a medir la calidad de un resumen contando las superposiciones de unidades entre el resumen del candidato y un conjunto de resúmenes de referencia escritos. Puede generar tres tipos de puntajes: memoria, $precisio \quad y \; F_1$ (Merchant 2018).

RPI (7): ¿Cuáles son los dataset utilizados en los artículos seleccionados?

Se identifican los dataset utilizados por cada uno de los artículos seleccionados. (Ver Tabla VI)

Tabla VI: Relación de artículos y dataset utilizados

Artículo	Dataset
1	Compresión de documentos
2	Ontología de envíos
3	Compresión de documentos
4	Colección de artículos de noticias
5	Documentos en Portugués
6	Documentos de literatura
7	Reportes de Foster
8	Reportes Policiales en Inglés
9	Publicaciones Científicas acerca de Ciencias de las Computación y Física.

[4] busca la idea principal o el punto más importante en un texto escrito

10	Imágenes médicas
11	Artículos científicos
12	Fragmentos del Corán
13	Artículos acerca de Ciencias de las Computaciones
14	Reportes médicos
15	Colección de datos de cámaras digitales
16	Artículos Científicos
17	Artículos de las CNN
18	Compresión de documentos
19	Documentos en Portugués

1.3. Procedimiento para la extracción automatizada de conceptos (PAC).

Aunque una metodología con enfoque lingüístico, pudiese lanzar resultados más satisfactorios para la investigación, la misma depende de que los patrones lingüísticos y reglas sintácticas, se definan correctamente (Grishman 2015), puesto que pueden originar pérdida de conceptos relevantes en el corpus a analizar.

Por lo que se propone una metodología con enfoque estadístico con independencia total del lenguaje, la cual sigue la estructura siguiente:

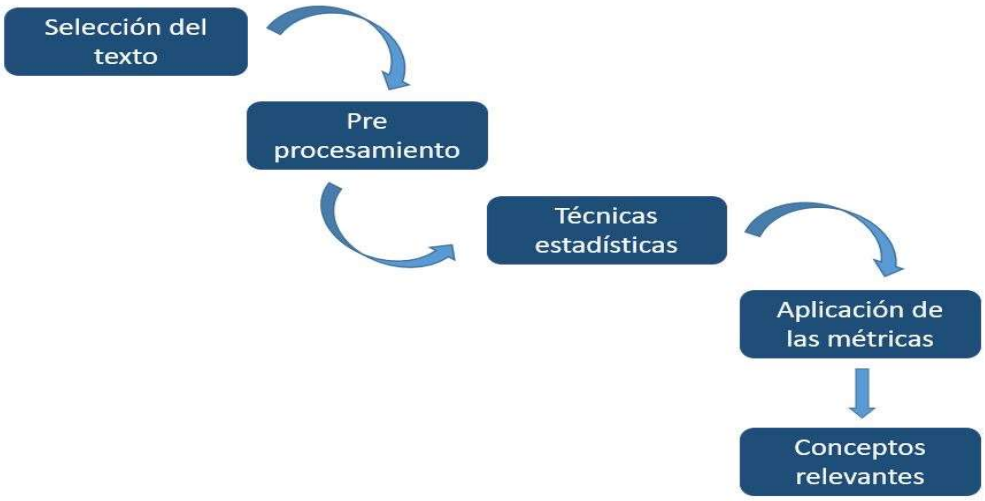


Figura 3: Procedimiento para la extracción automatizada de conceptos (PAC)

El procedimiento propuesto cuenta con 5 fases expuestas en la Figura 3:

1.3.1 Selección del texto:

Esta fase se enfoca a determinar el tipo de texto de entrada que sea válido para la extracción de conceptos. En la revisión sistemática de la bibliografía se determina que los textos válidos para este tipo de trabajo son corpus documentales que están compuestos por múltiples textos (Bonidia et al. 2018). Este enfoque coincide con la necesidad para este trabajo que implica la utilización de un corpus documental en el dominio legal, que va a estar compuestos por varias leyes.

1.3.2. Pre procesamiento:

En esta fase para iniciar la tarea de pre procesamiento del texto seleccionado, se llevarán a cabo algunas de las técnicas de la minería de texto que permitirán obtener la información deseada de nuestro texto; la **tokenización**, es el proceso de dividir un documento mediante la identificación de espacios en blanco o los signos de puntuación (P et al. 2018), una vez se ha finalizado la tokenización. Además, se eliminarán las **palabras vacías** ya que estas no aportan significado relevante al documento. La lista de palabras vacías será tomada de las definidas en la biblioteca NLTK, en su versión 2.4.

1.3.3. Técnicas estadísticas:

Establecer la frecuencia de ocurrencia de los términos:

Se utilizarán técnicas estadísticas, para la implementación de esta técnica se utiliza la biblioteca NLTK en el lenguaje de programación Python en su versión 2.7.

1.3.4. Extracción de los conceptos:

Esta fase se dedica a establecer un ranking de los conceptos detectados, teniendo en cuenta las métricas de clasificación estándar; recuperación y precisión implementadas en Python en su versión 2.7.

- $\boldsymbol{Precision}$**:** es una suma ponderada de precisiones sobre todas las clases, donde el peso de la precisión para una clase es proporcional al número de oraciones en esa clase (Kocbek et al. 2016).
- $\boldsymbol{Recall}$: mide la probabilidad de que si se selecciona un concepto como frase clave, el método lo identifica correctamente. Da la proporción de los conceptos clave extraídos correctamente entre todos los conceptos claves (Kocbek et al. 2016).

1.3.5. Presentación de los conceptos relevantes:

Para la presentación de conceptos al usuario se decide aplicar las técnicas de tabulación y nube de palabras.

1.4. Conclusiones parciales.

El marco conceptual de referencia para soportar los fundamentos teóricos de la investigación permite definir el proceso de extracción de conceptos, así como las técnicas más empleadas y los principales dominios sobre el cual se aplica. Se introduce también, la minería de textos como proceso para la conversión de la información, del lenguaje natural al computacional y así procesar la información necesaria para el estudio. Además, se propone un procedimiento que permite establecer un orden y análisis de los artefactos para alcanzar el resultado deseado. Todo lo anterior sustenta la propuesta de creación de un extractor de conceptos con el empleo de técnicas estadísticas para brindarle independencia total del lenguaje, que es la propuesta de esta investigación.

CAPÍTULO 2. DESCRIPCIÓN DE LA PROPUESTA

Introducción

En el presente capítulo se detallan las fases de PAC. De cada fase del proceso se mostrarán los fundamentos teóricos, herramientas y técnicas utilizadas en su implementación mostrando la aplicabilidad a partir de un caso de estudio, en el que se considera como corpus legal el texto de la Constitución de la República de Cuba.

2.1. Selección del texto

La selección del texto se encarga de la creación del corpus documental a ser utilizado en el proceso de minería de texto. Un corpus documental agrupa a los documentos en un formato tal que pueden ser utilizados como entrada de los algoritmos de minería de texto, la estructura de datos en la que son almacenados de manera típica es bolsas de palabras.

El formato a utilizar en este trabajo para los documentos que conformarán el corpus documental es el formato de texto plano con codificación estándar UTF-8. Los ficheros que contengan el texto deben encontrarse almacenados bajo un único directorio raíz.

Para la creación del corpus documental a partir de ficheros de texto plano, utilizando las herramientas disponibles en NLTK, se utiliza la clase PlaintextCorpusReader que extiende de CorpusReader.

En el constructor de esta clase se define los atributos root (una cadena de texto que contiene la dirección donde estarán almacenados los ficheros de texto) y fileids (un listado de ficheros o una expresión regular a partir de la cual se puede obtener el listado).

Los ficheros leídos son separados en oraciones o palabras de acuerdo al interés del usuario, conformando de esta manera la bolsa de palabras con la que se va a trabajar. Las funciones que se encargan de esta operación están implementadas en la clase PlaintextCorpusReader se denominan *sents()* y *words().* La definición de las mismas es presentada a continuación:

Nombre de la función: *sents()*

- Devuelve el corpus dado como una lista de oraciones, cada una codificada como una lista de palabras.

Nombre de la función: *words()*

- Devuelve el corpus dado como una lista de palabras.

Siendo la primera la encargada de crear la bolsa de oraciones y la segunda la bolsa de palabras[5].

[5] El modelo de bolsa de palabras es una representación simplificada utilizada en el procesamiento del lenguaje natural y la recuperación de información. En este modelo, un texto (como una oración o un documento) se representa como la bolsa (conjunto múltiple) de sus palabras, sin tener en cuenta la gramática e incluso el orden de las palabras, pero manteniendo la multiplicidad. Se utiliza comúnmente en los métodos de clasificación de documentos donde la (frecuencia de) aparición de cada palabra se usa como una característica para entrenar a un clasificador.

En este trabajo a modo de prueba de concepto se construye un corpus documental a partir del texto de la Constitución de la República de Cuba. Cada uno de los artículos de la constitución se almacenan en un fichero texto con la siguiente denominación art***.txt, donde las *** representan el número del artículo. En este caso no son necesarias grandes tareas de limpieza, por tanto la fase de pre procesamiento se resume en la eliminación de los signos de puntuación, aun cuando es válido notar que podría agregarse la tarea de lematización, para contribuir a obtener mayor precisión en los resultados.

La bolsa con la que se decide trabajar fue de palabras y por tanto se utilizó la función correspondiente a la creación de la bolsa de palabras.

De la bolsa de palabras se eliminaron los signos de puntuación utilizando las siguientes líneas de código:

```
quotations = ['.','¿':'¿',',','¿';']
words = [x for x in words if x not in quotations]
```

Figura 4 Fragmento de código para la eliminación de los signos de puntuación de la bolsa de palabras

2.2. Establecer la frecuencia de ocurrencia de los términos.

Comúnmente las palabras que no tienen significación semántica, por ejemplo: la, el, y, con, entre otras; tienden a estar asociadas con una mayor cantidad de palabras que aquellas que tienen mayor significación semántica como es el caso de los conceptos.

Para capturar la noción de distancia fija se utiliza una ventana a partir de la cual se miden las frecuencias de coocurrencias. En este trabajo se decide utilizar una ventana de tamaño ocho, que va a estar compuesta por cuatro posiciones antes de la palabra estudiada y cuatro posiciones después.

En la tabla VII se muestran un grupo de relaciones entre palabras utilizando el esquema de distancia fija.

Tabla VII: Frecuencia de co-ocurrencia de pares de palabras para diferentes posiciones relativas

Pares	Frecuencia relativa por posición
(decretos_estado)	[22, 0, 22, 0, 'decretos', 0, 0, 0, 0]
(decretos_las)	[66, 110, 132, 0, 'decretos', 0, 0, 0, 0]
(ausencia_de)	[0, 0, 0, 3, 'ausencia', 0, 0, 0, 0]
(demás_funciones)	[15, 0, 0, 0, 'demás', 45, 0, 15, 0]
('el_constitución')	[400, 0, 160, 80, 'el', 0, 0, 0, 480]

Como se evidencia en la tabla VII, en los pares de palabras, una sola palabra tiende a co-ocurrir en distancias fijas relativamente entre sí. Por ejemplo, la palabra ‘estado’, ocurre 22 veces antes de ‘decretos’, dos posiciones antes de ‘estado’ y cuatro posiciones antes, mientras que ‘las’ ocurre 66, 110,132 veces cuatro, tres y dos posiciones antes respectivamente. De manera similar ocurre con la

palabra ‘de’, que ocurre tres veces una posición antes de la palabra ‘ausencia’. Esta propiedad permite definir un grupo de métricas que pueden ser utilizadas para evaluar la carga semántica de una palabra y decidir si es un concepto o no.

2.2.1 Especificidad.

Un mecanismo para medir la especificidad de los conceptos está asociado a la frecuencia de co-ocurrencia del concepto con el resto de las palabras de la bolsa. Comúnmente se limita a una distancia fija la evaluación de esta métrica en aras de reducir la carga computacional y aumentar la exactitud de la misma. Esto aprovecha el hecho de que las palabras fuertemente correlacionadas con carga semántica aparecen juntas en el texto.

Para una palabra individual w de un corpus documental, $B_w = [b_1, b_2, b_3, \dots, b_m]$ es la lista de todas las palabras vecinas únicas de w . Cada vecino b_i se produce en diferentes posiciones en relación con w, dentro de una ventana con tamaño s. La posición de b_i puede ser positiva o negativa, se determina considerando que w está en el centro de la ventana. Para cada par (w, b_i), se obtiene una lista de $X\ (w, b_i)$, contando las frecuencias de co-ocurrencia por distancia relativa entre w y b_i.

Para una X, dada una (w, b_i), la siguiente métrica calcula la varianza relativa de la distribución de frecuencias en $X\ (w, b_i)$:

$$Re\ _{var}(X_{(w,b_i)}) = \frac{1}{s(s-1)} \times \sum_{j=1}^{s} \left(\frac{x_j - \bar{x}}{\bar{x}}\right) \quad (12)$$

donde x_j es el valor del elemento j-th de la lista $X\ (w, b_i)$: y s es la longitud de la lista (el tamaño de la ventana); representa el valor promedio de las frecuencias en $X\ (w, b_i)$:

$$\bar{x} = \frac{1}{s} \times \sum_{j=1}^{s} x_j \quad (13)$$

Debe notarse que, aunque $X\ (w, b_i)$ representa una ventana que va de $-s/2$ a $s/2$, $Rel_{var(.)}$ calcula la varianza relativa independientemente del orden de sus elementos.

Para entender mejor el mecanismo de $Rel_{var(.)}$, en las Figuras 4 y 5, se muestra la distribución de frecuencias para dos pares de palabras que aparecen en el corpus (personas, libre) y (personas, de)

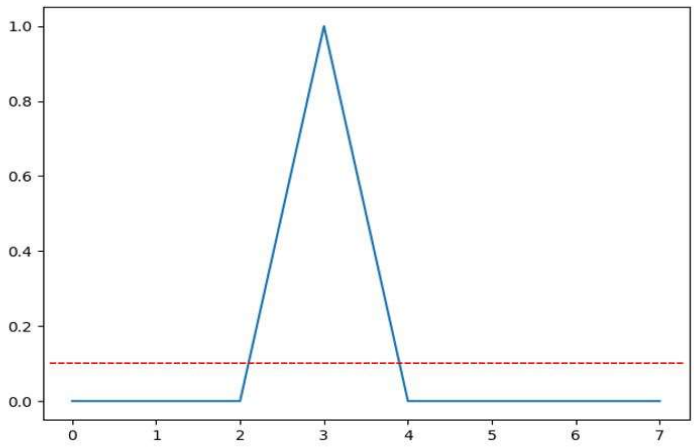


Figura 5: Representación de la frecuencia de co-ocurrencia para el par (personas, libre), $X(personas, libre) = [0, 0, 0, 1, 'personas', 0, 0, 0, 0]$, $Rel_{var(.)} = 1.0, \bar{x} = 0.125$

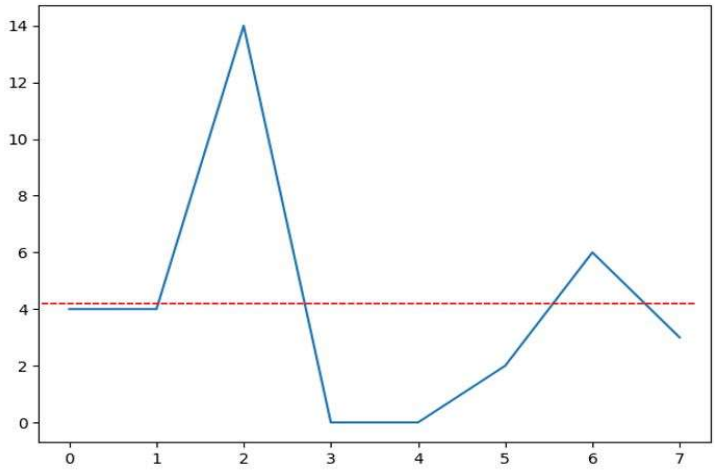


Figura 6: Representación de la frecuencia de co-ocurrencia para el par (personas, de), $X(personas, de) = [4, 4, 14, 0, 'personas', 0, 2, 6, 3]$, $Rel_{var(.)} = 0.147, \bar{x} = 4.125$

$Rel_{var(.)}$ mide, esencialmente, las distancias normalizadas desde los puntos (en este caso, frecuencias por posición) hasta un valor promedio (la frecuencia promedio). Estas distancias normalizadas se elevan al cuadrado para que los valores con signos diferentes no se cancelen entre sí. El valor máximo de 1.0 se da a las listas donde todas las frecuencias, excepto una, son 0, como en el par (personas_libre), que existe un pico claro en la posición que precede a la palabra 'persona'. Para el par (personas_de) en la Figura 6, ya que todas las frecuencias están alrededor del valor promedio, no hay una preferencia obvia para que el par co-ocurra en una posición fija, teniendo, por lo tanto, un valor $Rel_{var(.)}$ más bajo.

Por lo tanto, los pares (w, bi) muestran que es preferible que ocurran en posiciones fijas, ya que son más valorados que los pares que generalmente aparecen dispersos. En la siguiente tabla se muestra algunos valores de $Rel_{var(.)}$, para los pares de palabras extraídos del corpus.

Tabla VIII Algunos valores de Rel_{var} para el par ($personas, b_1$)

Pares	Frecuencia relativa por posición	$Rel_{var}(.)$
personas_libre	[0, 0, 0, 1, 'personas', 0, 0, 0,0]	1.000
personas_estado	[0, 0, 0, 0, 'personas', 0, 1, 0, 1]	0.428
personas_el	[2, 0, 0, 0, 'personas', 2, 1, 0, 0]	0.268
personas_la	[7, 1, 0, 0, 'personas', 3, 1, 0, 5]	0.193
personas_de	[4, 4, 14, 0,'personas', 0, 2, 6, 3]	0.147

Para medir la especificidad de una palabra w en un corpus documental, siendo $B = [b_1 , \dots, b_m]$, la lista de todas las palabras m en el corpus. Se utiliza la siguiente ecuación:

$$RDist_w = [Rel_{var(X_{(w,b_1)})}, Rel_{var(X_{(w,b_2)})}, \dots, Rel_{var(X_{(w,b_m)})}] \quad (14)$$

Donde $X\ (w, b_i)$ es la lista de las frecuencias de co-ocurrencia de la palabra b_i cerca de la palabra w (considerando una ventana de tamaño fijo), y $Rel_{var\left(X_{(w,b_i)}\right)}$ es el valor $Rel_{var(.)}$ para un par (w, b_i).

Finalmente en la ecuación 16 se muestra la medida para la especificidad de w

$$Spec(w) = Rel_{var}(RDist_w) \quad (15)$$

La idea subyacente sobre $Spec(w)$ es que, si una palabra simple w está fuertemente asociada (tiene valores $Rel_{var(.)}$ más altos) con unas pocas palabras en el corpus, y está débilmente asociada con el resto de ellas, entonces w es un concepto bastante específico. Este mecanismo se puede entender al observar la siguientes figuras, que muestran la distribución de $RDist_w$ para las palabras 'tribunal' y 'a'.

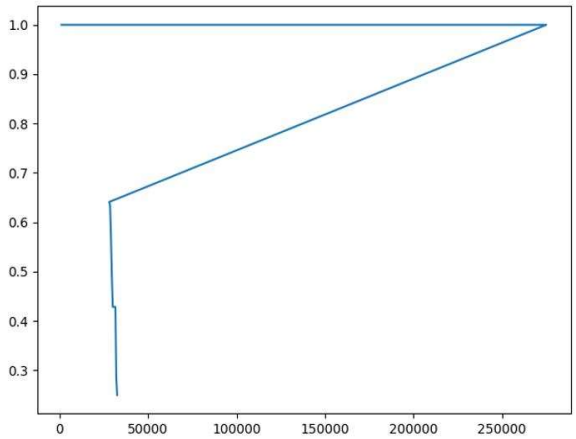


Figura 7: Distribución ordenada para los valores de $Re\ _{var}$ para los pares $(tribunal, b_i)$

En la Figura 7, se muestra que la palabra tribunal tiene altos valores de Rel_{var} con unas pocas palabras del corpus. Luego decrece el valor Rel_{var} hasta que alcanza rápidamente el valor cero. En otras palabras, se muestra que la palabra 'tribunal' se relaciona fuertemente (en términos de posiciones fijas) con unas pocas palabras del corpus, y luego se relaciona cada vez menos y menos con todas las demás palabras del corpus hasta que llega a cero: estas son palabras con menor influencia sobre la palabra 'tribunal', y la mayoría no aparecen en absoluto cerca de la 'tribunal'.

Por otro lado, en la Figura 7, los valores de Rel_{var} para la palabra 'a' disminuyen muy lentamente. Básicamente, la palabra 'a' mantiene la tendencia a tener relaciones de distancia fija con muchas más palabras que 'tribunal'. Como Rel_{var} (Ecuación 13) mide la tendencia a la aparición de "picos" en las listas de valores numéricos, el valor $Re\ _{var}$, para la distribución en la Figura 7 ('tribunal') es mayor que el valor $Re\ _{var}$, para la distribución en la Figura 8 ('a').

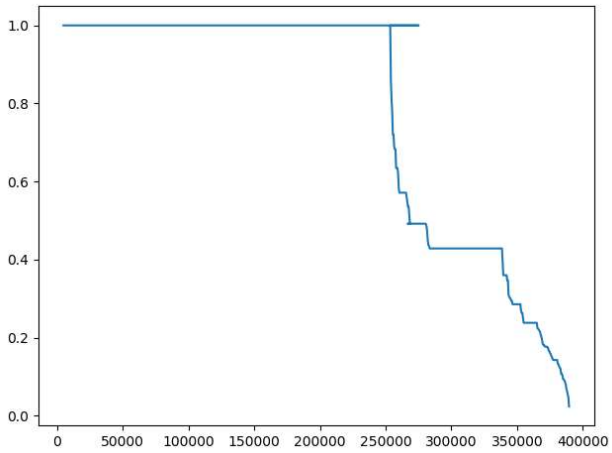


Figura 8: Distribución ordenada para los valores de $Re\ _{var}$ para los pares (a, b_1)

La siguiente tabla muestra algunos ejemplos de valores de $Spec(w)$. Como referencia, la columna de los números de pares de (w, b_i), mide aquellos pares (w, b_i) cuyos valores $Rel_{var(X_{(w,b_i)})} > 0$.

Tabla IX: Especificidad de algunas palabras para el corpus de la Constitución de la República de Cuba

w	Cantidad de pares (w, b_i)	$Spec(w)$
persona	63	0.001
base	25	0.003
secretario	37	0.004
en	893	0.0001
del	581	0.0002
o	488	0.0003

2.3. Herramienta

PyCharm:

Es un entorno de desarrollo integrado (IDE) utilizado específicamente para el lenguaje de programación Python. Proporciona análisis de código, un depurador gráfico, un comprobador de unidades integrado, integración con sistemas de control de versiones y es compatible con el desarrollo web con Django y *Data Science* con Anaconda[6]. PyCharm es un IDE multi-plataforma, con versiones de Windows, macOS y Linux.

2.4. Bibliotecas

NLTK:

El **kit de herramientas de lenguaje natural,** o más comúnmente **NLTK,** es un conjunto de bibliotecas y programas para el procesamiento del lenguaje natural (PLN) simbólico y estadístico para el lenguaje de programación Python. NLTK incluye demostraciones gráficas y datos de muestra. NLTK está destinado a apoyar la investigación y la enseñanza en PLN o áreas muy relacionadas, que incluyen la lingüística empírica, las ciencias cognitivas, la inteligencia artificial, la recuperación de información, y el aprendizaje automatizado.

Matplotlib:

Es un paquete de gráficos en 2D y 3D del lenguaje de programación Python. Está diseñado para poder crear gráficos simples y complejos con solo unos pocos comandos genera gráficos, histogramas, espectros de potencia, gráficos de barras, gráficos de errores, diagramas de dispersión (Ari, Ustazhanov 2014).

Algunas de las ventajas del uso de matplotlib sobre MATLAB[7] son:

- Gran control de cada elemento en una figura, incluyendo el tamaño de la figura.
- Salida de alta calidad en muchos formatos, incluidos PNG, PDF, SVG, EPS y PGF.
- GUI (***Graphical user interface***) para explorar figuras de forma interactiva y soporte para la generación sin cabeza de archivos de figuras (útil para trabajos por lotes). Matplotlib, muy adecuado para generar figuras para publicaciones científicas, ya que todos los aspectos de la figura se pueden controlar mediante programación. Esto es importante para la reproducibilidad y es conveniente cuando se necesita regenerar la figura con datos actualizados o cambiar su apariencia.
- Basado en Python, un moderno lenguaje de programación orientado a objetos, adecuado para el desarrollo de software a gran escala.
- Libre, de código abierto, sin servidores de licencias.

[6] https://blog.jetbrains.com/pycharm/2019/04/collaboration-with-anaconda-inc/

[7] es un sistema de cómputo numérico que ofrece un entorno de desarrollo integrado con un lenguaje de programación propio.

- Soporte de gráficos vectoriales escalables nativos (SVG).
- Se puede incrustar en una interfaz gráfica de usuario para el desarrollo de aplicaciones.
- La codificación es tan fácil que el usuario puede entenderla y ampliarla.

2.5. Estándar de codificación

PEP 8 (*Python Enhancement Proposals*)[8] está dedicada a la recopilación de los estándares seguidos por los desarrolladores de Python.

Entre las convenciones principales se tienen:

- Usar cuatro espacios para indentar.
- Es posible usar solo tabulaciones u ocho espacios para código antiguo que haya sido escrito así. Por ningún motivo se han de mezclar espacios y tabulaciones.
- Se deben separar las funciones de nivel superior y las clases con dos líneas en blanco, mientras que los métodos dentro de clases los podemos separar con una sola línea. También se pueden usar líneas en blanco dentro de las funciones para separar bloques que guardan cierta correlación lógica.
- las sentencias *import* deben de estar siempre en la parte superior del archivo agrupadas de la siguiente manera: librería estándar, librerías de terceros e *import's* de la aplicación local.
- Usar espacios alrededor de los operadores aritméticos.
- Limitar los tamaños de línea a 72 caracteres como máximo, si bien se puede continuar líneas largas con el símbolo '\', es recomendable el uso de paréntesis.

2.6. Lenguaje de Programación:

Python (versión 2.7)[9] lenguaje simple por su sintaxis por lo que es más fácil de leer para el programador. Presenta un tipado dinámico siendo esto una posibilidad que se le brinda al usuario para cambiar el tipo de variable sin tener que declararla. Actualmente domina el mundo de la Inteligencia Artificial, la ciencia de datos, entre otros. Python dispone de potentes librerías y gran cantidad de módulos. Destacado por se multi-paradigma, esto significa que más que forzar a los programadores a adoptar un estilo particular de programación, permite varios estilos: programación orientada a objetos**,** programación imperativa y programación funcional. Para el caso específico de Python 2.7 introduce una nueva clase OrderedDict en el módulo de colecciones que itera sobre las claves y los valores en el orden en el que insertaron.

[8] https://www.python.org/dev/peps/pep-0008/

[9] https://docs.python.org/2.7/

2.7. Conclusiones parciales

Se realizó la implementación del procedimiento del extractor de conceptos y a partir de una prueba de conceptos, utilizando como corpus documental el texto de la Constitución de la Republica, se logró demostrar que la propuesta realizada constituye una alternativa viable para la extracción de conceptos relevantes en un corpus documental legal.

CAPÍTULO 3. VALIDACIÓN DE LA PROPUESTA

Introducción

En este capítulo se presenta la metodología de experimentación y los resultados alcanzados. Se detalla la forma en que se seleccionan los umbrales de varianza relativa para cada uno de los pares de palabras a partir de los cuales se obtienen valores de precisión y recuperación satisfactorios. Al finalizar se muestran un grupo de los conceptos encontrados.

3.1. Metodología de experimentación.

Para determinar si una palabra se considera concepto a partir de la definición de $Spec$, dada en la ecuación 16, se requiere un valor de umbral. Las palabras que tengan un valor de $Spec$ superior al umbral se consideraran conceptos y el resto no.

Podría utilizarse un umbral definido por el usuario pero sería muy difícil que este se ajustara a la realidad, por ello se decidió utilizar como umbral el punto de corte de las métricas $Recall$, $Precision$ y F_1, considerando todos los posibles umbrales y mostrando los valores de las tres funciones en una gráfica, se permite identificar visualmente el punto de corte.

Se calcularán las siguientes métricas:

$$Precision = \frac{\#(true\_concepts \cap considered\_concepts)}{\#considered\_concepts} \qquad (16)$$

$Precision,$ algunas veces llamado valor predictivo positivo, mide la proporción de falsos positivos provocadas por la aplicación del método, es decir cuantas palabras consideradas conceptos por el método, son realmente conceptos ($true\_concepts$). El valor de $Precision$ se encontrará en el intervalo [0,1] y su valor será más efectivo cuanto más cerca del uno se encuentre, lo que remarca el hecho de que todas las palabras en $true\_concepts$ fueron clasificadas como conceptos por el algoritmo. Para que este comportamiento sea el esperado se requiere que el total de conceptos del *dataset* de entrenamiento aparezca en $true\_concepts$.

$$Recall = \frac{\#(true\_concepts \cap considered\_concepts)}{\#true\_concepts} \qquad (17)$$

$Recall$, mide la cantidad de falsos negativos generados por el método, o sea cuantos $true\_concepts$ del total de $true\_concepts$, fueron correctamente clasificados como conceptos por el método.

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Reca} \qquad (18)$$

F_1, es la media armónica entre la $Precision$ y $Recall$, que tiende esencialmente al valor más bajo.
Para el cálculo de las métricas se requiere contar con dos listados de palabras:

- $true\_concepts$, que contiene el listado de palabras clasificadas manualmente como conceptos.
- $considered\_concepts$, que contiene el listado de palabras clasificas como conceptos por el algoritmo.

Todos los experimentos se ejecutaron en una computadora con las siguientes características:

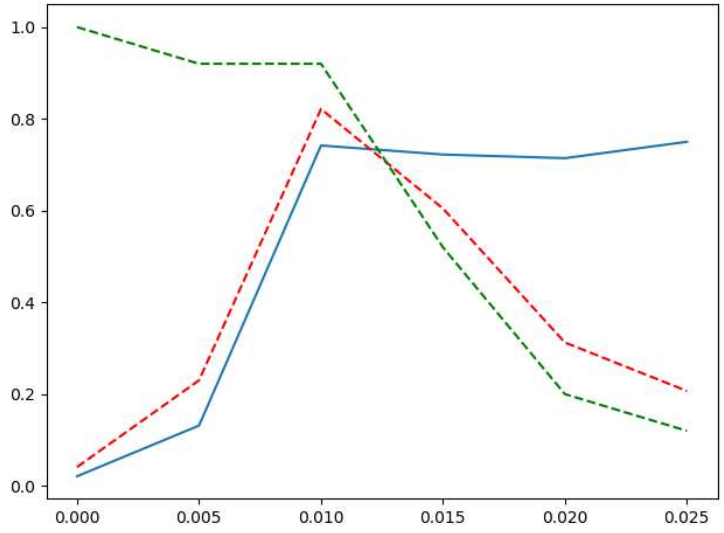
- Procesador: Intel (R) Core (TM) i7-6500U CPU @ 2.50GHz @ 2.59GHz
- Memoria: RAM 4,00 GB

Se van a realizar dos experimentos siguiendo un enfoque de aprendizaje semi-supervisado. En el primero se utilizarán como corpus documental el total de artículos de la Constitución de la República y se realizará la clasificación manual de una pequeña muestra de los conceptos presentes, aproximadamente el dos por ciento del total del corpus, aproximadamente 25 conceptos. Este experimento tiene como objetivo determinar el umbral de $spec$ que se utilizará para considerar un concepto como relevante. Para ello se propone utilizar como umbral el valor de $Spec$ para el cual se cortan las gráficas de las métricas $Precision$ y $Recall$ y F_1. Para graficar las métricas se variarán los $Spec$ desde 0.0 hasta 0.025.

Se evaluará el comportamiento del punto de corte en la extracción de conceptos cuando se añaden documentos al corpus documental. Para ello se comprobarán 4 casos que corresponden a incorporar al corpus: caso 1 Gaceta Oficial de la República de Cuba # 30, caso 2 Gaceta Oficial de la República de Cuba # 31, caso 3 Gaceta Oficial de la República de Cuba # 32, caso 4 Gaceta Oficial de la República de Cuba # 33. Se extraerán los conceptos relevantes para cada caso y se calcularán las métricas de $Precision$, $Recall$ y F_1 para comprobar la viabilidad de la propuesta.

El segundo experimento se realiza para evaluar el desempeño del clasificador con respecto a algunos de los clasificadores en el estado del arte, $Zhou,\ Tf-Idf,\ Syllables$.

3.2. Resultados.



$Precision$ ___	0.81
$Recall$ ----	0.88
F_1 ----	0.84

Figura 9 Representación de las métricas $Precision$ y $Recall$ y F_1 para el corpus de la Constitución de la República de Cuba.

En la matriz de confusión de dos clases, para este caso, cada columna de la matriz representa el número de predicciones realizadas por el modelo, y cada fila los valores reales en la observación. (Provost 2014).

		Extractor	
		+	-
Manual	+	TP	FN
	-	FP	TN

Con lo cual los conteos quedan divididos en 4 clases, TP, FN, FP y TN, que significan lo siguiente:

- **TP – *True Positives:*** Son el número verdaderos positivos, es decir, de predicciones correctas para la clase +.
- **FN – *False Negatives:*** Son el número de falsos negativos, es decir, la predicción es negativa cuando realmente el valor tendría que ser positivo. A estos casos también se les denomina errores de tipo II.
- **FP – *False Positives*:** Son el número de falsos positivos, es decir, la predicción es positiva cuando realmente el valor tendría que ser negativo. A estos casos también se les denomina errores de tipo I.
- **TN – *True Negatives:*** Son el número de verdaderos negativos, es decir, de predicciones correctas para la clase -.

Las métricas de $Precision$ y $Recall$ se pueden calcular a partir de la matriz de confusión de la siguiente manera:

$$Recall = \frac{tp}{(tp + fn)} \quad (19)$$

$$Presición = \frac{tp}{(tp + fp)} \quad (20)$$

Tabla X Matriz de confusión para el caso 1

		Extractor	
		+	-
Manual	+	25	3
	-	6	TN

$Precision$	0.80
$Recall$	0.89

Tabla XI Matriz de confusión para el caso 2

		Extractor	
		+	-
Manual	+	23	2
	-	5	TN

$Precision$	0.82
$Recall$	0.92

Tabla XII Matriz de confusión para el caso 3

		Extractor	
		+	-
Manual	+	23	1
	-	5	TN

$Precision$	0.82
$Recall$	0.95

Tabla XIII Matriz de confusión para el caso 4

		Extractor	
		+	-
Manual	+	31	7
	-	6	TN

$Precision$	0.83
$Recall$	0.81

Se realiza un segundo experimento para comparar la propuesta con los clasificadores en el estado del arte, $Zhou$, $Tf-Idf$, $Syllables$. Se puede apreciar que la propuesta supera la precisión de todos los clasificadores. Para la métrica de $Recall$ también es mejor en la mayoría de los corpus salvo para el caso dos, donde el método $Tf-Idf$ es ligeramente superior al procedimiento desarrollado. Esto muestra que la propuesta realizada es una variante a tener en cuenta para la obtención de conceptos relevantes en un corpus documental legal.

En la tabla siguiente se muestra una comparación del PAC con algunos de los enfoques mencionados en el capítulo uno de esta tesis. Las bases para esta comparación para palabras simples fueron las siguientes: ya que cada método proporciona su propia métrica de puntuación, la comparación con los umbrales del PAC no tiene sentido, por lo que se realiza utilizando los resultados que maximizan el valor de F_1 de cada método.

Tabla XIV Valores de $Precision$ y $Recall$ para diferentes enfoques para los artículos de la Constitución de la República de Cuba.

Enfoque	$Precision$	$Recall$
PAC	0.81	0.88
$Tf-Idf$	0.58	0.85
$Zhou$	0.65	0.73
$Syllables$	0.66	0.78

Tabla XV Valores de $Precision$ y $Recall$ para diferentes enfoques para los artículos de la Constitución de la República de Cuba y la Gaceta Oficial de la República de Cuba #30 (Caso 1).

Enfoque	$Precision$	$Recall$
PAC	0.80	0.89
$Tf-Idf$	0.60	0.70
$Zhou$	0.62	0.75
$Syllables$	0.68	0.76

Tabla XVI Valores de $Precision$ y $Recall$ para diferentes enfoques para los artículos de la Constitución de la República de Cuba y la Gaceta Oficial de la República de Cuba # 31 (Caso 2).

Enfoque	$Precision$	$Recall$
PAC	0.82	0.92
$Tf-Idf$	0.63	0.83
$Zhou$	0.68	0.77
$Syllables$	0.69	0.80

Tabla XVII Valores de $Precision$ y $Recall$ para diferentes enfoques para los artículos de la Constitución de la República de Cuba y la Gaceta Oficial de la República de Cuba # 32 (Caso 3).

Enfoque	$Precision$	$Recall$
PAC	0.82	0.95
$Tf - Idf$	0.67	0.88
$Zhou$	0.66	0.76
$Syllables$	0.68	0.80

Tabla XVIII Valores de $Precision$ y $Recall$ para diferentes enfoques para los artículos de la Constitución de la República de Cuba y la Gaceta Oficial de la República de Cuba # 33 (Caso 4).

Enfoque	$Precision$	$Recall$
PAC	0.83	0.81
$Tf - Idf$	0.67	0.80
$Zhou$	0.68	0.75
$Syllables$	0.70	0.81

Tabla XIX Comparación de la métrica $Precision$ para cada uno de los extractores

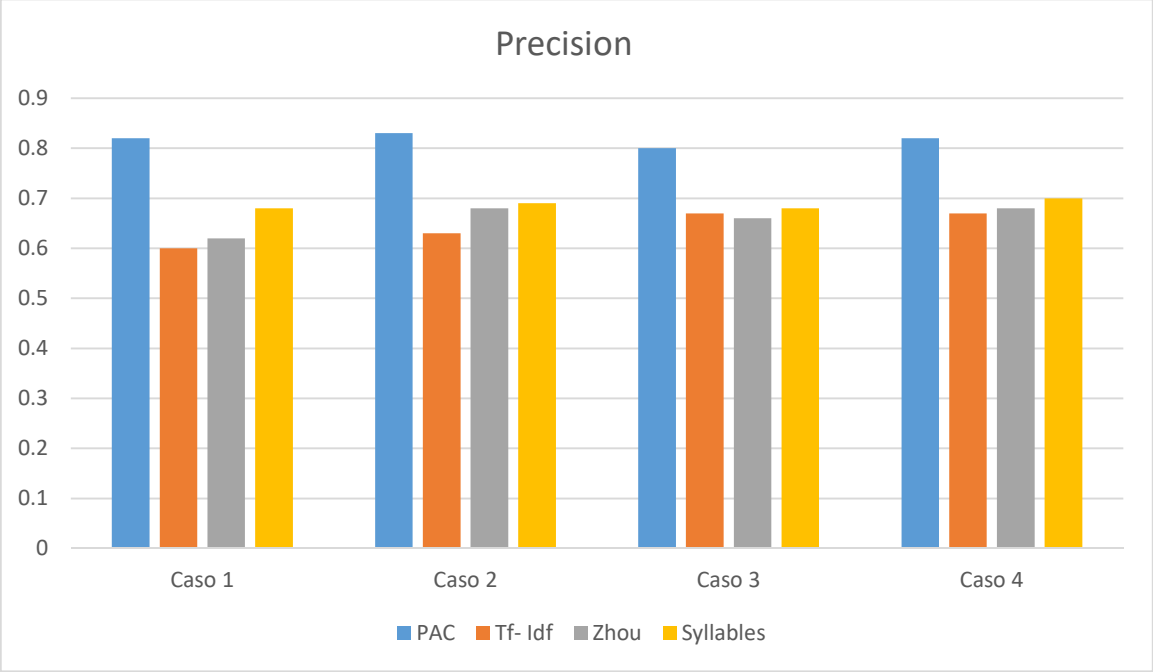
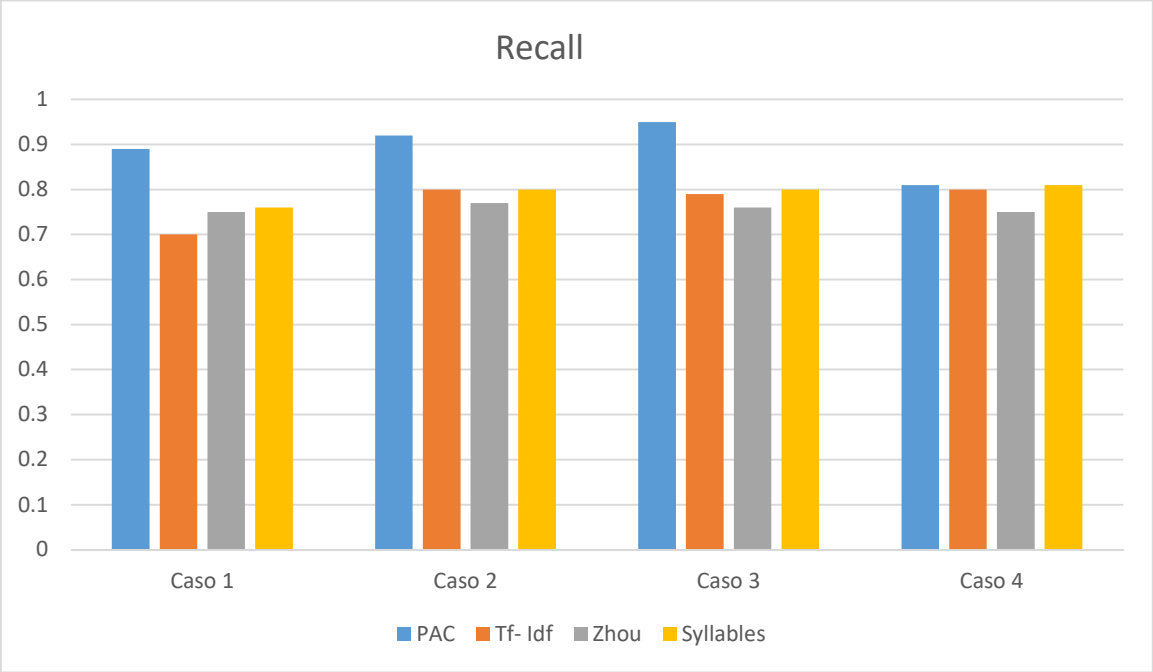


Tabla XX Comparación de la métrica $Recall$ para cada uno de los extractores



3.3. Análisis resultados.

En el primer experimento se logró determinar a partir del punto de corte de las métricas un valor para el umbral de $Spec$ de 0.011. Al utilizar este valor para determinar los conceptos relevantes en los casos (uno, dos, tres y cuatro), se alcanzaron valores de precisión superiores al 80 por ciento lo que se consideran valores adecuados en comparación con las propuestas en el estado del arte.

A pesar de que estos resultados son prometedores, todavía existe capacidad para mejorar, debido a que conceptos como ‘plan’ y ‘planes’ son detectados como semánticamente diferentes y podría trabajarse en la lematización de los tokens para evitar estos casos.

PAC, muestra resultados más altos que los otros métodos en la extracción de conceptos de una sola palabra. Con respecto a los extractores de una sola palabra, aunque $Tf - Idf$ está destinado a trabajar solo en documentos, se adaptó de tal manera que la puntuación de una palabra se otorgó por su puntuación máxima de $Tf - Idf$ obtenida para algún documento, considerando todos los documentos del corpus. Aunque el $Recall$ es bastante bueno en promedio, la precisión baja proviene del hecho de que algunos conceptos son relativamente frecuentes en el corpus, alcanzando valores de Idf más bajos. En cuanto al enfoque de $Zhou$, sus núcleos se expresan al medir sus capacidades para formar agrupaciones locales en un corpus. Sin embargo, en las pruebas se observó que conceptos raros se ven perjudicados por esta métrica, ya que su tendencia a formar agrupaciones se ve muy disminuida

por su falta de ocurrencias. Finalmente, aunque el enfoque de $Syllables$ tiene un puntaje, en promedio, más alto que los otros métodos, tiende a dañar conceptos más pequeños.

3.4. Conclusiones parciales

Se realizó la validación de la propuesta comparando la misma con los clasificadores estadísticos del estado del arte ($Zhou, Tf - Idf, Syllables$) a través de las métricas $Recall$ y $Precision$ en cinco corpus documentales diferentes. Los resultados alcanzados permiten asegurar que el método propuesto tiene mejor desempeño que las variantes en el estado del arte.

CONCLUSIONES GENERALES

- Del análisis del estado del arte se puede concluir que la técnica $Tf - Idf$ es la más utilizada en la implementación de extractores de conceptos con enfoque estadístico. Las métricas más utilizadas para evaluar la efectividad de los extractores son $Precision, Recall, F_1$.
- La aplicación del extractor de conceptos al corpus documental a la Constitución de la República de Cuba arrojó métricas de $Precision$ y $Recall$ superiores a 0.80. Estos valores son consistentes a los alcanzados en el estado del arte lo que permite demostrar la validez del método.
- Una vez realizada la experimentación se puede validar que el método propuesto sobrepasa en rendimiento a los métodos en estado del arte en casi todos los dataset analizados y siendo comparados en el único dataset en el que no lo sobrepasa, esto permite asegurar que el método constituye una alternativa viable para la extracción de conceptos relevantes.

RECOMENDACIONES

- Incluir nuevas tareas de pre-procesamiento asociados a la eliminación de palabras con el mismo significado pero distinta gramática, como el caso de ‘plan’ y ‘planes’.
- Incluir un método para la extracción de conceptos multi-palabras.

REFERENCIAS BIBLIOGRÁFICAS

ALKSHER, M.A., AZMAN, A., YAAKOB, R., KADIR, R.A., MOHAMED, A. and EISSA, M., 2016. A review of methods for mining idea from text.

AMAN, M., SAID, A.B.I.N., JADID, S. and KADIR, A., 2018. Key Concept Identification : A Sentence Parse Tree-Based Technique for Candidate Feature Extraction from Unstructured Texts. , vol. 3536, no. c, pp. 1–11. DOI 10.1109/ACCESS.2018.2875135.

ARI, N. and USTAZHANOV, M., 2014. Matplotlib In Python.

BEIL, F., ESTER, M., BC, B. and VA, C., 2014. Frequent Term-Based Text Clustering.

BETHARD, S., 2017. Towards Automatic Identification of Core Concepts in Educational Resources.

BIRD, S., 2017. NLTK Documentation.

BIYABANGARD, A., 2015. Word Concept Extraction Using HOSVD for Automatic Text Summarization.

BONIDIA, R.P., RODRIGUES, L.A.L., AVILA-SANTOS, A.P., SANCHES, D.S. and BRANCHER, J.D., 2018. Computational Intelligence in Sports : A Systematic Literature Review. , vol. 2018.

DWIVEDI, Y.K., RANA, N.P., CHEN, H. and WILLIAMS, M.D., 2011. A Meta-analysis of the Unified Theory of Acceptance and Use of Technology A Meta-analysis of the Unified Theory of Acceptance and Use of Technology (UTAUT). , no. September. DOI 10.1007/978-3-642-24148-2.

E. CALVER, S. DE JUANA ESPINOSA, J.V., 2019. E-Government Implementation, Work Process Changes and Competency Training in Spanish Town Councils. , vol. 5, pp. 5–18. DOI 10.17951/ijsr.2016.5.5.

GEADAS, P., ALVES, A. and RIBEIRO, B., 2014. Ensemble Learning for Keyword Extraction from Event Descriptions.

GRISHMAN, R., 2015. Information Extraction.

GUO, H., 2018. Characterizing treatment strategies of intrahepatic bile duct cancer by literature mining. *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*, pp. 43–48. DOI 10.1109/ICHI-W.2018.00013.

ISMAIL, R., 2018. Concepts Extraction in Ontology Learning Using Language Patterns for Better Accuracy. *2018 4th International Conference on Computer and Technology Applications (ICCTA)*, pp. 122–126.

KOCBEK, S., CAVEDON, L., MARTINEZ, D., BAIN, C., MAC, C., HAFFARI, G., ZUKERMAN, I. and VERSPOOR, K., 2016. Text mining electronic hospital records to automatically classify admissions against disease : Measuring the impact of linking data sources. *Journal of Biomedical Informatics* [en línea], vol. 64, pp. 158–167. ISSN 1532-0464. DOI 10.1016/j.jbi.2016.10.008. Disponible en: http://dx.doi.org/10.1016/j.jbi.2016.10.008.

MALIN, J.T., THRONESBERY, C. and THROOP, D.R., 2016. Content Coding and Search for Risk Assessment : Problems and Solutions.

MEJRI, M. and AKAICHI, J., 2014. A Survey of Textual Event Extraction from Social Networks. ,

MERCHANT, K., 2018. NLP Based Latent Semantic Analysis for Legal Text Summarization. *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1803–1807.

NAIK, S.S., 2017. EXTRACTIVE TEXT SUMMARIZATION BY FEATURE-BASED SENTENCE EXTRACTION USING RULE-BASED. , pp. 2–6.

NEVES, J.M., SANTOS, M.F. and MACHADO, J.M., 2007. *Progress in Artificial Intelligence: 13th Portuguese Conference on Artificial Intelligence, EPIA 2007, Workshops: GAIW, AIASTS, ALEA, AMITA, BAOSW, BI, CMBSB, IROBOT, MASTA, STCS, and TEMA, Guimarães, Portugal, December 3-7, 2007, Proceedings*: Springer.

OKUMURA, N. and MIURA, T., 2015. Automatic Labelling of Documents Based on Ontology. , pp. 34–39.

P, S.C., SOUNDARYA, J., PRIYADHARSINI, R. and BHARATHI, B., 2018. Data Analysıs of Natural Language Queryıng Usıng NLP Interface. , vol. 13, no. 8, pp. 5792–5795.

PROVOST, K. and, 2014. Confusion Matrix. , pp. 8–10.
REZTAPUTRA, R., 2017. Sentence Structure-based Summarization for Indonesian News Articles. , pp. 0–5.
SUN, P., KU, C. and SHIH, D., 2015. Telematics and Informatics An implementation framework for E-Government 2 . 0. *TELEMATICS AND INFORMATICS* [en línea], vol. 32, no. 3, pp. 504–520. ISSN 0736-5853. DOI 10.1016/j.tele.2014.12.003. Disponible en: http://dx.doi.org/10.1016/j.tele.2014.12.003.
TOHALINO, J.V. and AMANCIO, D.R., 2017. Extractive Multi Document Summarization using Dynamical Measurements of Complex Networks. , DOI 10.1109/BRACIS.2017.41.
TORRE, M.C.J. de la, 2017. *Nuevas técnicas de minería de textos: aplicaciones*. ISBN 9788491632436.
TOTI, D. and RINELLI, M., 2016. On the road to speed-reading and fast learning with CONCEPTUM. , DOI 10.1109/INCoS.2016.30.
TROVATI, M., 2017. Dependency Networks Extractions from Textual Sources - A Case Study in Criminology. , no. September.
WU, J., CHOUDHURY, S.R., CHIA, A., LIANG, C. and GILES, C.L., 2017. HESDK : A Hybrid Approach to Extracting Scientific Domain Knowledge Entities. , vol. 2017, no. June, pp. 1–4.
ZHANG, D., SONG, W., LIU, L., DU, C. and ZHAO, X., 2017. Investigations in Automatic Humor Recognition. , DOI 10.1109/ISCID.2017.160.
ZHANG, L., 2016. The Research of Concept Extraction in Ontology Extension Based on Extended Association Rules. , pp. 111–114.
ZHOU, H. and SLATER, G.W., 2003. A metric to search for relevant words. , vol. 329, pp. 309–327. DOI 10.1016/S0378-4371(03)00625-3.
ZUBRINIC, K., KALPIC, D. and MILICEVIC, M., 2012. The Automatic Creation of Concept Maps from Documents Written Using Morphologically Rich Languages. , vol. 39, pp. 12709–12718.