Revista de investigación de sistemas e informática
17(2): 5-15 (2024)
DOI: https://doi.org/10.15381/risi.v17i2.29906

ISSN: 1815-0268 (Impreso) / ISSN: 1816-3823 (Electrónico)
Facultad de Ingeniería de Sistemas e Informática - UNMSM

Uso de herramientas de procesamiento de lenguaje natural para el análisis y desarrollo de artículos científicos de ingeniería

Use of natural language processing tools for the analysis and development of scientific engineering articles

César Aristóteles Yajure Ramírez
https://orcid.org/0000-0002-3813-7606
cyajure@gmail.com
Universidad Central de Venezuela. Caracas, Venezuela

Recibido: 15/11/2024 - Aceptado: 05/12/2024 - Publicado: 31/12/2024

RESUMEN

Es importante contar con herramientas que nos permitan extraer la información útil de textos científicos sin necesidad de leer todo su contenido. Por ejemplo, cuando se requiere determinar los tópicos que tratan en los artículos científicos, establecer la línea de investigación de un autor a través de la revisión de sus publicaciones, o diseñar el resumen de un artículo y sus palabras claves. Entonces, el objetivo de esta investigación es utilizar las técnicas de procesamiento de lenguaje natural para extraer la información útil de artículos científicos de ingeniería. Se toman los veintidós artículos publicados por un autor para utilizarlos como documento base del análisis, el cual se divide en: análisis general al conjunto de los artículos, y análisis particular por artículo publicado. Como resultado del primero se obtuvieron las palabras y bigramas claves, siendo las palabras “datos” ,“energía”, y “modelo”, las más frecuentes, y los bigramas “solar fotovoltaica”, “variables explicativas”, y “energías renovables”, los más importantes. Del análisis particular se obtuvo que los bigramas jerarquizados por cada artículo representan una buena aproximación de sus palabras claves, y además hay una alta similitud entre los resúmenes obtenidos aplicando las técnicas de lenguaje natural a los artículos publicados durante el año 2024 y sus resúmenes, siendo el obtenido con GPT2 el que presentó el mayor nivel de similitud. Con las frases claves obtenidas con SGRank se pudo determinar el tópico de los artículos respectivos.

Palabras claves: Artículos científicos, bigramas claves, frases claves, lenguaje natural, resumen de texto, similitud de texto.

ABSTRACT

It is important to have tools that allow us to extract useful information from scientific texts without having to read all their content. For example, when it is necessary to determine the topics covered in scientific articles, establish an author's line of research through the review of their publications, or design the summary of an article and its keywords. So, the objective of this research is to use natural language processing techniques to extract useful information from scientific engineering articles. The twenty-two articles published by an author are taken to use them as the base document for the analysis, which is divided into: general analysis of all the articles, and particular analysis per published article. As a result of the first, the key words and bigrams were obtained, with the words "data", "energy", and "model" being the most frequent, and the bigrams "solar photovoltaic", "explanatory variables", and "renewable energies", the most important ones. From the second analysis, it was obtained that the hierarchical bigrams for each article represent a good approximation of their keywords, and there is also a high similarity between the summaries obtained by applying natural language techniques to the articles published during the year 2024 and their summaries, being the one obtained with GPT2 presented the highest level of similarity. With the key phrases obtained with SGRank, the topic of the respective articles could be determined.

Keywords: Scientific articles, key bigrams, key phrases, natural language, text summarizing, text similarity.

I. INTRODUCCIÓN

La mayoría de las revistas indexadas predefinen la estructura que deben tener los artículos de investigación que vayan a acceder a su proceso de revisión por pares. Esta estructura es del tipo IMRDC, es decir, introducción, metodología (o materiales y métodos), análisis y discusión de resultados, y conclusiones. Una vez se tengan diseñadas estas etapas del artículo, se debe desarrollar el resumen y definir las palabras claves. El resumen es una de las partes más importantes del artículo, pues como bien lo indican Revuelta & Llorente (2024) su cometido es informar acerca del contenido del artículo, y a la vez captar la atención del lector. En cuanto a las palabras claves, su valor como metadato es: "relacionar e identificar el contenido del artículo con el área, tópico o tema dentro de una o varias disciplinas científicas" (Flores Ramírez, 2023, pág. 12). Entonces, sería de mucha utilidad contar con una herramienta que permita generar tanto las palabras claves como el resumen de una manera rápida y eficiente, o validar esos ítems, si ya están diseñados, por medio de la comparación con otros candidatos. Asimismo, cuando se está en un proceso de revisión de artículos científicos para crear el estado del arte de nuestra investigación, sería de provecho identificar los tópicos de los artículos a revisar y sus palabras y bigramas más importantes, de manera tal de no tener la necesidad de leer todo el contenido y hacer más eficiente el proceso. Por otra parte, en ocasiones hay la necesidad de identificar o contrastar el área de investigación de un autor, y una manera de hacerlo es revisar sus artículos que hayan sido publicados en revistas indexadas, y así determinar los tópicos o áreas en los que se ha desenvuelto, lo que consume una cantidad de tiempo significativa. Por lo anterior, el objetivo de esta investigación es utilizar las técnicas de Procesamiento de Lenguaje Natural (NLP) para el análisis del contenido de artículos científicos de ingeniería publicados en revistas indexadas. Como documento objeto de la investigación se tienen los veintidós artículos de investigación publicados por un autor particular, en idioma español, y durante el período 2014 - 2024. Las técnicas se aplican utilizando el lenguaje de programación Python.

Se hizo una revisión de las investigaciones previas relacionadas con el análisis de texto haciendo uso de las técnicas de NLP. Por ejemplo, Kusnetzov et al. (2024) hacen un análisis del proceso de revisión por pares de artículos científicos, evaluando las etapas en las que las herramientas de procesamiento de lenguaje natural pudieran ser de ayuda en este proceso. Consideran que estas herramientas pueden ser de beneficio antes de la revisión, en las tareas de envío y en el emparejamiento autor-artículo. Durante la revisión el aporte sería en la evaluación, revisión de escritura, y en la discusión. Después de la revisión la ayuda estaría en el desarrollo de la meta revisión, revisión del manuscrito, y en el análisis post-revisión. En (Baruni & Sathiaseelan, 2020) utilizan los algoritmos RAKE (*Rapid Automatic Keyword Extraction*) y TextRank para extraer frases claves del resumen de un artículo científico ya publicado. Concluyen que el algoritmo RAKE es el más eficiente y que entrega mejores resultados. En (Mandal & Singh, 2020) utilizan el método LSA (*Latent Semantic Analysis*) para desarrollar el resumen de textos relacionados con 51 tópicos diferentes de documentos públicos utilizados por otros autores. Concluyen que con su análisis se mejoran los resultados obtenidos por esos otros autores. En (Mohan et al., 2023) aplican las técnicas de resumen de textos GPT 2 (*Generative Pre-trained Transformer 2*) y BERT (*Bidirectional Encoder Representations of Transformers*) a un gran conjunto de datos

públicos compuesto por miles de archivos de texto de 15 líneas en promedio. Utilizan la métrica ROUGE (*Recall-Oriented Understudy of Gisting Evaluation*) para evaluar sus resultados, obteniendo que la técnica BERT arroja mejores resultados que la técnica GPT 2. En (Flayeh et al., 2022) aplican las técnicas del procesamiento de lenguaje natural utilizando el lenguaje C++ y la interfaz Visual Studio para comparar dos artículos en términos del número de verbos, preposiciones, fortaleza del artículo, entre otros. Asimismo, clasifican los verbos, sustantivos, y adjetivos y lo muestran a través de la interfaz. En (Vastrad et al., 2022) utilizan el procesamiento de lenguaje natural y el aprendizaje automático para el pronóstico del precio de las acciones considerando que si se entienden las emociones de las personas se podrían mejorar estos pronósticos. Concluyen que el modelo FinALBERT utilizado para el pronóstico del precio de las acciones, es afectado por la sintonización de los hiperparámetros, y por el tamaño del conjunto de datos. En (Priya et al., 2021) hacen un recuento de las distintas aplicaciones del NLP, incluyendo el texto escrito. Indican que la NLP es útil para el área de salud pues ayuda a mejorar la integridad y precisión de los registros médicos, identificar posibles errores en la prestación de los servicios de salud. La NLP ayuda a los autores a mejorar su escritura, corregir sus ensayos, y a crear sistemas automatizados de evaluación de escritura. En (Sandu et al., 2024) llevan a cabo un análisis bibliométrico de un conjunto de 1852 artículos científicos relacionados con el uso del NLP en la investigación de redes sociales. Desarrollan un análisis descriptivo del conjunto de artículos, una jerarquización de los artículos más citados, análisis de los n-gramas, incluyendo la jerarquización de palabras, bigramas, y trigramas. Obtienen que la fuente más relevante de artículos fue *IEEE Access* con 92 artículos publicados, el autor con más publicaciones en el área investigada fue Serker A. con 28 artículos publicados, la palabra más frecuente fue *twitter*, el bigrama más frecuente fue *social media*, y el trigrama más frecuente fue *natural language processing*. Finalmente, en (Calero Sánchez et al., 2024) utilizan el procesamiento de lenguaje natural para llevar a cabo la revisión de literatura científica de una manera más eficiente usando la base de documentos PubMed. Específicamente, aplican las técnicas NLP para la recuperación bibliográfica de la investigación del suicidio en jóvenes de la base de documentos mencionada. Concluyen que la metodología propuesta reduce los tiempos y las tareas para la selección inicial de trabajos de investigación de bases de datos bibliográficas.

El resto del artículo se distribuye de la siguiente manera. En la sección 2 se presenta la metodología utilizada en la investigación. En la sección 3 se analizan y discuten los resultados obtenidos. En la sección 4 se presentan las conclusiones derivadas de la investigación realizadas, y finalmente se muestran las referencias bibliográficas.

II. MATERIALES Y MÉTODOS

2.1. Fundamentos teóricos

Indicador TF-IDF

Previamente, debe establecerse que un token se entiende como el fragmento más pequeño al cual puede dividirse un texto, que para nuestro caso es una palabra o también pudiera ser una frase (Vajjala et al., 2020). Entonces, el TF-IDF es el conteo de la frecuencia de ocurrencia de los tokens que componen un documento, degradado por la importancia de éstos, la cual se calcula dividiendo el número de documentos por el número de documentos que contienen el token respectivo. La frecuencia del documento para un token se define como el número de documentos que lo contienen dividido por el número total de documentos, mientras que el IDF (*Inverse Document Frequency*) es exactamente lo inverso (Singh, 2023).

Si la frecuencia de ocurrencia del token en un documento relevante es TF (*Term Frequency*), en (1) se tiene la expresión para obtener el IDF (Artama et al., 2020). Luego con (2) se calcula el indicador TF-IDF para el token *i*.

$$IDF_i = \log \frac{Nro.\ Documentos}{Nro.\ de\ documentos\ que\ contienen\ el\ token\ i} \quad (1)$$

$$TF - IDF = TF \cdot IDF \quad (2)$$

Similitud del coseno

Es una métrica utilizada para medir el nivel de similitud entre dos textos, y se basa en modelos de espacio vectorial. La similitud se mide representando los textos como vectores de frecuencia de palabras y calculando el valor del coseno del ángulo entre esos vectores (Liu et al., 2023). Asimismo, en (Januzaj & Luma, 2022) establecen que, si los vectores apuntan en la misma dirección, significa que son similares. Asimismo, el valor de esta métrica varía entre -1 y 1, siendo 1 cuando el ángulo entre los vectores es de cero grados y hay máxima similitud, y valdrá -1 cuando el ángulo entre los vectores es de 180 grados y la similitud es nula.

Técnicas de preprocesamiento de texto

La tokenización es el proceso de extraer palabras a partir de una secuencia de caracteres, sabiendo que las palabras están usualmente separadas por espacios y/o signos de puntuación. Por otra parte, la remoción de palabras vacías (*stop words*) consiste en remover del texto las palabras que no aportan información significativa del tópico tratado en un texto particular, pero que son útiles para la comprensión general del texto. Ejemplos de palabras vacías son: determinadores, verbos auxiliares, pronombres, adverbios, entre otras (Albrecht et al., 2021). En algunas aplicaciones de NLP, este tipo de palabras es removida previó al desarrollo de otras tareas, pero en ciertas aplicaciones, como resumen de textos, es necesario mantenerlas en el documento. La derivación (*stemming*) es un proceso para identificar las raíces de las palabras, siendo la raíz de una palabra su parte principal después de quitarle sus prefijos y sufijos (Hagiwara, 2021). Este mismo autor indica que la lematización es el proceso de conseguir la forma base de una palabra antes de su inflexión.

Técnicas de procesamiento de lenguaje natural

La extracción de frases claves (*KeyPhrase Extraction*) es una tarea de procesamiento de información textual relacionada con la extracción automática de frases características y representativas desde un documento, que expresan todos los aspectos claves de su contenido. El resumen de texto (*Text Summarizing*) consiste en generar un resumen de un texto amplio, reteniendo la información vital del texto original. Para esta actividad se pueden aplicar métodos para realizar resumen de texto extractivo, en el cual las frases obtenidas son una réplica del texto que está siendo resumido, y también métodos para resumen de texto abstractivo, en los que las frases obtenidas son similares a la del texto original, pero no necesariamente iguales. Por otra parte, la similitud de textos (*Text Similarity*) es un procedimiento utilizado para determinar el nivel de similitud entre pares de textos a través del uso de distintas métricas (Singh, 2023).

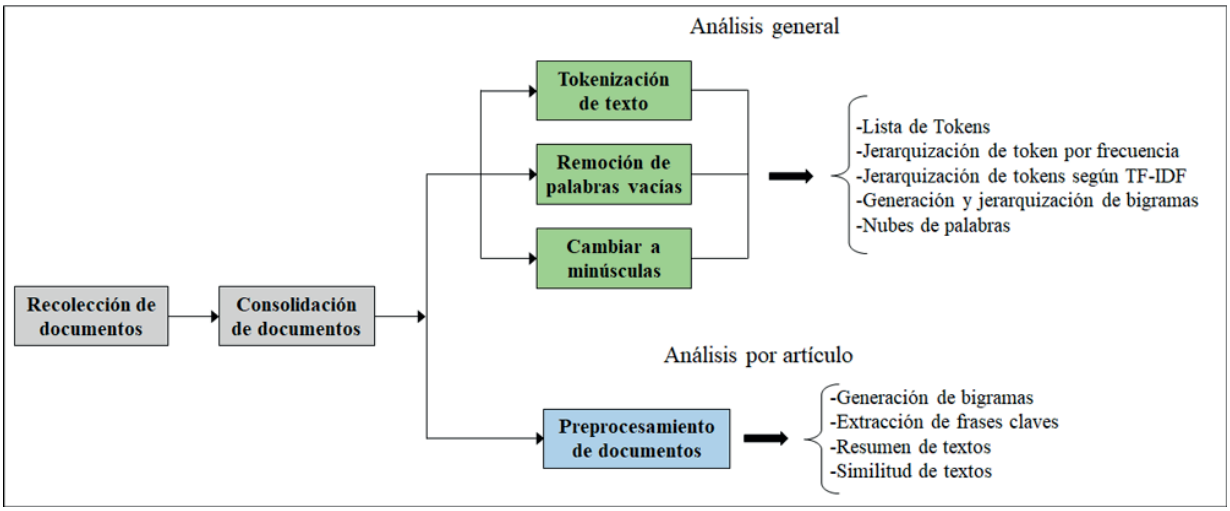
2.2. Metodología

La metodología utilizada consistió en hacer el análisis de una serie de documentos conformados por un conjunto de veintidós artículos científicos de ingeniería publicados en revistas indexadas, dividiendo dicho análisis en dos partes, tal como se muestra en la Figura 1.

El primer paso consistió en recolectar los artículos científicos publicados por el autor bajo estudio, para el período 2014-2016 (se publicaron seis artículos), y luego para el período 2022-2024 (se publicaron dieciséis artículos), puesto que tuvo una pausa de cinco años durante la etapa 2017-2021. A partir de estos artículos, se generó un archivo de hoja de cálculo que incluye una columna para la fecha de publicación, otra para el nombre de la revista donde fue publicado el artículo, la siguiente para el país en el cual se edita la revista, otra para el resumen, las palabras clave, y finalmente una para el texto correspondiente a cada artículo. El texto de cada artículo incluido en el archivo está compuesto por: el título, la introducción, la metodología y marco teórico, el análisis de resultados y las conclusiones, es decir, no se incluye: el resumen, las palabras

Figura 1

Esquema de la metodología aplicada



claves, ni las referencias bibliográficas. También se eliminan de los textos, las tablas y las figuras. Estos pasos están representados por los bloques grises de la Figura 1.

En una primera parte se hace un análisis general, obteniéndose métricas para el conjunto de los artículos, y no para artículos en específico. En la segunda parte, se aplican técnicas de procesamiento de lenguaje natural, que tiene más sentido hacerlo a los artículos en específico, es decir, de manera individual.

Para el análisis general de los artículos, se hizo un pre-procesamiento del documento que incluyó la tokenización del texto, la remoción de las palabras vacías, y cambiar a minúsculas todas las palabras restantes. Con la tokenización se obtuvo la lista de palabras, y el número de palabras, por artículo. Luego, se aplicaron técnicas que permitieron obtener una jerarquización de las palabras de acuerdo con su frecuencia de ocurrencia, y de acuerdo con la métrica TF-IDF. Asimismo, se obtuvieron sentencias compuestas por dos y tres palabras (n-gramas), y su jerarquización según su frecuencia de ocurrencia. Además, se obtuvieron nubes de palabras para el conjunto de los documentos, y para los artículos publicados por año.

En cuanto al análisis por artículo, se aplicaron varias técnicas de procesamiento de lenguaje natural: generación de bigramas, extracción de frases claves, resumen de textos, y similitud de textos. El pre-procesamiento aplicado previo a cada técnica depende del tipo de técnica correspondiente. Los bigramas obtenidos por artículo se compararon con las palabras claves de los artículos publicados, que por lo general se componen hasta por tres palabras, para determinar su nivel de coincidencia. Las frases claves que se obtienen de los artículos son útiles para determinar el tópico del que trata el artículo respectivo, y en ocasiones, cuando se obtienen frases cortas, éstas se comparan con sus palabras claves. El resumen de texto obtenido de un artículo particular aplicando NLP se compara con el resumen original del artículo, calculando su nivel de similitud.

Análisis y discusión de resultados

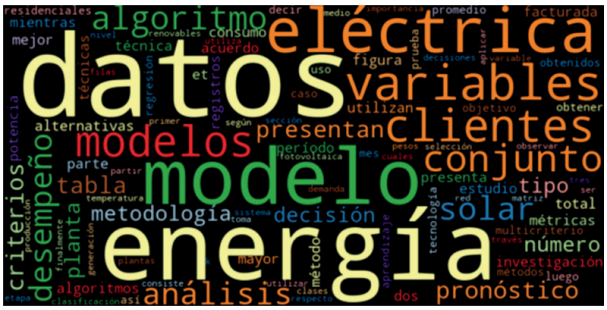
Análisis exploratorio general

El primer paso consistió en hacer la remoción de las palabras vacías. Posteriormente, se transformaron las palabras a su versión en minúscula puesto que, una misma palabra pudiera aparecer en el texto toda en minúscula, con la primera letra en mayúscula, o incluso toda en mayúscula, y para que se detecten estas tres versiones como una sola palabra, pero repetida, en lugar de tres palabras diferentes, se deben transformar todas las palabras a su versión en minúscula. Luego se generaron los tokens para cada uno de los artículos considerados, así como también los bigramas (secuencias de dos palabras).

Como primera actividad, se generó una nube de palabras considerando los tokens de todos los artículos del período considerado, la cual se presenta en la Figura 2. Se puede ver que la palabra con mayor frecuencia de ocurrencia es “datos”, seguida de la palabra “energía”.

Figura 2

Nube de palabras completo



Posteriormente, los tokens se jerarquizaron de acuerdo con su frecuencia de ocurrencia (frec), se clasificaron por año, y los resultados se presentan en la Tabla 1.

Tabla 1

Palabras jerarquizadas por frecuencia de ocurrencia

2014		2015		2016	
token	**frec**	**token**	**frec**	**token**	**frec**
carbón	43	energía	67	energía	67
alternativas	34	criterios	59	criterios	52
condiciones	25	alternativas	55	tecnología	48
promethee	24	tecnología	43	renovables	41
vapor	21	pesos	39	alternativas	40
criterios	19	eléctrica	37	solar	39
tecnología	18	multicriterio	34	multicriterio	38
pulverizado	17	tabla	32	decisión	38
ciclo	17	promethee	32	ponderación	38
energía	16	decisiones	30	tecnologías	36
2022		**2023**		**2024**	
token	**frec**	**token**	**frec**	**token**	**frec**
datos	428	datos	448	datos	160
energía	284	modelo	264	modelo	73
clientes	249	energía	211	demanda	65
eléctrica	138	modelos	173	energía	63
facturada	123	variables	165	algoritmo	62
residenciales	121	eléctrica	156	clases	60
consumo	119	desempeño	136	pronóstico	57
algoritmo	117	pronóstico	133	máxima	56
modelo	116	solar	130	eléctrica	55
variables	106	planta	113	potencia	52

Se puede observar que la mayor cantidad de palabras está a partir del año 2022, cuando también hubo mayor cantidad de artículos publicados. Para el período 2014-2016, las palabras más frecuentes fueron: energía, alternativas, criterios, y tecnología. En cuanto a los años 2022-2024, las palabras más frecuentes fueron: datos, modelo, energía, y eléctrica.

De igual forma, se jerarquizaron los tokens de acuerdo con el indicador TD-IDF, y los resultados se presentan en la Tabla 2. Se nota que la posición de las palabras cambia en la jerarquización, para cada uno de los años considerados, siendo evidente la desaparición de la palabra "energía" de los primeros lugares para el período 2014-2016, y la palabra "datos" para el período "2022-2024".

Tabla 2

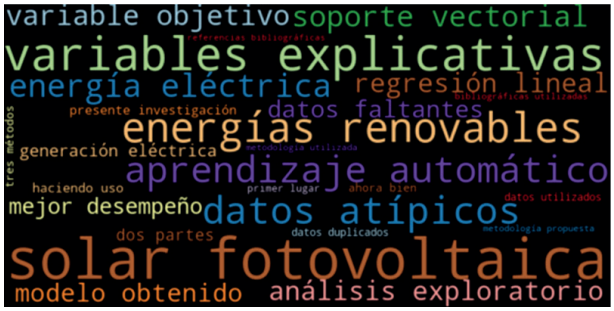
Palabras jerarquizadas según TD-IDF por año.

2014		2015		2016	
token	**tfidf**	**token**	**tfidf**	**token**	**tfidf**
carbón	78	tecnología	60	ponderación	80
pulverizado	42	promethee	51	tecnología	67
promethee	38	alternativas	49	tecnologías	50
alternativas	30	criterios	42	multicriterio	38
vapor	29	multicriterio	34	criterios	37
tecnología	25	pesos	22	alternativas	36
condiciones	25	decisiones	7	renovables	29
ciclo	17	energía	7	solar	24
criterios	13	tabla	5	decisión	16
energía	2	eléctrica	4	energía	7
2022		**2023**		**2024**	
token	**tfidf**	**token**	**tfidf**	**token**	**tfidf**
clientes	348	pronóstico	148	clases	75
facturada	172	modelo	94	máxima	70
residenciales	169	modelos	84	pronóstico	63
consumo	75	solar	81	demanda	46
algoritmo	57	planta	55	algoritmo	30
datos	43	datos	45	modelo	26
modelo	42	variables	32	potencia	19
energía	28	energía	21	datos	16
variables	21	eléctrica	16	energía	6
eléctrica	14	desempeño	14	eléctrica	6

Seguidamente, se generaron los bigramas para todo el período, y por año. En la Figura 3 se presenta la nube de bigramas incluyendo a todos los artículos de todos los años considerados, y tomando en cuenta el TF-IDF, de la que se observa que el bigrama con mayor relevancia es "solar fotovoltaica", seguida por "variables explicativas", y "energías renovables".

Figura 3

Nube de bigramas completo según TF-IDF.



Además, los bigramas se jerarquizaron por frecuencia de ocurrencia y se clasificaron por año. Los resultados obtenidos se muestran en la Tabla 3, de la cual se ver que el bigrama "energía eléctrica" es el más importante, pues aparece en los primeros tres lugares en todos los años, a excepción del 2014. De igual manera, se observa que el bigrama "aprendizaje automático" aparece sólo a partir del año 2022.

Es interesante observar que en los años 2015 y 2016 aparecen los bigramas "importancia relativa", "decisiones multicriterio", y "comparaciones pareadas", lo cual coincide con el tópico principal tratado en esos años, el cual es la toma de decisiones multicriterio aplicada a la selección de sistemas de energías renovables. De igual forma, para los años del período 2022-2024 aparecen los bigramas "energía eléctrica", "demanda máxima", "aprendizaje automático", "solar fotovoltaica", entre otros, coincidiendo con los tópicos tratados durante ese período: ciencia de datos, pronóstico de energía generada por plantas solares fotovoltaicas, entre otros.

Análisis por artículo

Generación de bigramas por artículo

Previo a la obtención de los bigramas, se aplican las tres tareas de preprocesamiento: cambio a minúsculas, tokenización, eliminación de palabras vacías. Seguidamente, obtienen los bigramas más frecuentes para cada artículo publicado por separado y se jerarquizan de acuerdo con su frecuencia de ocurrencia. Se compararon los bigramas obtenidos con las palabras claves de los artículos publicados, que por lo general están realmente compuestas de una, dos, o hasta tres palabras. Además, la cantidad de palabras claves en los artículos científicos es usualmente cinco, pero para los artículos estudiados varían entre 4 y 6 palabras claves.

Tabla 3

Bigramas jerarquizados por año

2014		2015		2016	
bigrama	**frec**	**bigrama**	**frec**	**bigrama**	**frec**
carbón pulverizado	16	energía eléctrica	34	energías renovables	27
condiciones sub-críticas	14	importancia relativa	22	comparaciones pareadas	26
ciclo combinado	11	decisiones multicriterio	16	energía eléctrica	22
carbón mineral	11	comparaciones pareadas	15	importancia relativa	20
gasificación integrada	9	energías renovables	13	decisiones multicriterio	18
lecho fluidizado	8	carbón mineral	11	tecnología solar	12
promethee ii	8	mejor tecnología	10	solar térmica	10
energía eléctrica	8	técnica ahp	10	solar fotovoltaica	10
impacto ambiental	7	conjuntos difusos	9	radiación solar	10
condiciones super-críticas	6	números difusos	9	tecnología hidráulica	9
2022		**2023**		**2024**	
bigrama	**frec**	**bigrama**	**frec**	**bigrama**	**frec**
energía eléctrica	123	energía eléctrica	133	demanda máxima	40
clientes residenciales	60	planta solar	51	energía eléctrica	24
energía facturada	60	variables explicativas	49	plantas hidroeléctricas	17
clientes regulados	43	solar fotovoltaica	49	nsga ii	17
eléctrica facturada	36	irradiancia solar	47	optimización multiobjetivo	17
componentes principales	35	temperatura ambiente	34	planta solar	15
datos atípicos	30	aprendizaje automático	33	solar fotovoltaica	15
aprendizaje automático	25	mejor desempeño	30	aprendizaje automático	15
energía total	25	red neuronal	29	análisis exploratorio	14
migrantes venezolanos	23	análisis exploratorio	28	funciones objetivo	14

Tabla 4

Bigramas en artículos con mayores coincidencias

Artículo 2014			Artículo 2015			Artículo 2016		
Palabras clave	**Bigrama**	**frec**	**Palabras clave**	**Bigrama**	**frec**	**Palabras clave**	**Bigrama**	**frec**
multicriterio	carbón pulverizado	16	multicriterio	energía eléctrica	25	ponderación de criterios	comparaciones pareadas	22
promethee	condiciones sub-críticas	14	promethee	energías renovables	12	topsis	energías renovables	15
carbón mineral	ciclo combinado	11	ahp	decisiones multicriterio	11	conjuntos difusos	importancia relativa	13
energía eléctrica	carbón mineral	11	conjuntos difusos	importancia relativa	10	energía renovable	energía eléctrica	8
impacto ambiental	gasificación integrada	9	energía renovable	promethee ii	9	energía eléctrica	decisiones multicriterio	7
	lecho fluidizado	8	energía eléctrica	promethee i	8		pareadas difusas	7
	promethee ii	8		tecnología solar	8		tecnología solar	5
	energía eléctrica	8		criterio i	8		técnica topsis	5
	impacto ambiental	7		solar fotovoltaica	8		radiación solar	5
	condiciones super-críticas	6		comparaciones pareadas	8		expresión matemática	4
Artículo 2022			**Artículo 2023**			**Artículo 2024**		
Palabras clave	**Bigrama**	**frec**	**Palabras clave**	**Bigrama**	**frec**	**Palabras clave**	**Bigrama**	**frec**
agrupamiento	energía eléctrica	26	aprendizaje automático	energía eléctrica	28	cambio climático	optimización multiobjetivo	17
componentes principales	clientes regulados	19	irradiancia solar	irradiancia solar	18	frente de Pareto	nsga ii	17
aprendizaje automático	clientes residenciales	17	red neuronal artificial	solar fotovoltaica	11	generación eléctrica	funciones objetivo	14
energía facturada	energía facturada	13	regresión lineal	temperatura ambiente	11	nsga	plantas hidroeléctricas	13
	aprendizaje automático	10	serie de tiempo	planta solar	10	racionamiento eléctrico	algoritmo nsga	13
	eléctrica facturada	7	temperatura ambiente	plantas solares	8		racionamiento eléctrico	9
	componentes principales	7		solares fotovoltaicas	8		energía eléctrica	8
	k vecinos	7		regresión lineal	8		generación eléctrica	6
	detectar patrones	5		modelo arima	8		algoritmo evolutivo	6
	algoritmo k-means	4		red neuronal	8		planta hidroeléctrica	6

En 17 de los 22 artículos publicados hubo al menos dos coincidencias entre las palabras claves y los bigramas, y en tres de los artículos no hubo coincidencia alguna. En la Tabla 4 se presenta esa comparación, considerando por cada año el artículo que tuviera mayores coincidencias entre las palabras claves y los bigramas. Se puede observar que en los años en los que hubo mayores coincidencias fueron: 2014, 2016, y 2023 con cuatro. En el caso del año 2016, se contabilizan cuatro porque "ponderación de criterios" se considera sinónimo de "comparaciones pareadas". Para el año 2023, se podrían considerar cinco coincidencias puesto que el modelo Arima se utiliza para el análisis de series de tiempo.

Los resultados presentados en la Tabla 4 son importantes ya que: permiten validar las palabras claves de un artículo científico, coadyuva en la generación de palabras claves, y da indicios sobre el tópico del que trata el artículo correspondiente.

Extracción de frases claves

Para llevar a cabo esta actividad, se consideran los cuatro artículos publicados durante el año 2024. Asimismo, se aplican cuatro métodos: TextRank, RAKE NLTK, SGRank, y KeyBERT, las cuales entregan las frases ya jerarquizadas. Es importante mencionar, que debido a las características del producto generado (frases), en este caso no se eliminaron las "palabras vacías".

Con los dos primeros métodos mencionados se obtuvieron frases con una extensión considerable en cuanto al número de palabras, lo que permite determinar el tópico del cual trata del artículo analizado. Mientras que, utilizando los últimos dos métodos, las frases obtenidas tuvieron una extensión de hasta cuatro palabras. Además, muchas de las palabras de las frases son del tipo "palabras vacías", como se puede notar en la Tabla 5, en la cual se presentan los resultados obtenidos para los métodos SGRank y KeyBERT.

De la Tabla 5 se puede observar que utilizando el método SGRank, se generan algunas frases compuestas sólo por "palabras vacías", lo cual no arroja información útil alguna. Sin embargo, las otras frases generadas entregan información asociada a los tópicos tratados en los artículos. Por ejemplo, para el artículo 1 y método SGRank se tiene el tópico de "optimización multiobjetivo aplicada a la operación de plantas hidroeléctricas", lo cual coincide con que este artículo está asociado a un problema de optimización bi-objetivo para el despacho de plantas hidroeléctricas minimizando su producción de energía eléctrica y minimizando el racionamiento eléctrico (Yajure-Ramírez C. A., 2024). Para el artículo 2 se tiene "hora de ocurrencia de la demanda máxima", y el tópico de este artículo es el pronóstico de la hora de ocurrencia de la demanda eléctrica máxima. De igual manera, para el artículo 3 y método SGRank, se tiene el tópico de "calidad de la energía eléctrica entregada por una planta solar fotovoltaica". En cuanto a los resultados asociados al método KeyBERT, estos se pueden relacionar más con las palabras claves de los artículos publicados. Por ejemplo, para el artículo 1 se tiene: plantas hidroeleçtricas, generación hidroeléctrica, para el artículo 2: demanda eléctrica máxima, para el artículo 3: planta solar fotovoltaica, y para el artículo 4: fenómenos climáticos. En (Eldallal & Barbu, 2023) utilizan los métodos SGRank, TextRank, y KeyBERT (además de otros cuatro métodos) para extraer las frases claves de una serie de artículos

Tabla 5

Frases claves generadas para los artículos del 2024

Método	Artículo 1	Artículo 2	Artículo 3	Artículo 4
SGRank	las plantas hidroeléctricas	de la demanda eléctrica máxima	la calidad de la energía	para el pronóstico
	de energía eléctrica	de ocurrencia de la demanda	calidad de la energía eléctrica	el mes de
	de las plantas	la hora de	una planta solar fotovoltaica	de los datos
	de optimización multiobjetivo	el método SMOTE	y	de entrada
	y	los datos	del factor de potencia	en el
KeyBERT	de plantas hidroeléctricas	demanda eléctrica máxima	plantas solares fotovoltaicas	fenómeno climático considerado
	plantas hidroeléctricas en	de demanda eléctrica	solar fotovoltaica seguidamente	el fenómeno climático
	plantas hidroeléctricas con	demanda eléctrica crece	solares fotovoltaicas para	climático estas condiciones
	de generación hidroeléctrica	demanda eléctrica	solar fotovoltaica que	climático considerado fue
	plantas hidroeléctricas restringiendo	la demanda eléctrica	solares fotovoltaicas de	climático considerado

de investigación, previamente clasificados en tres categorías: ciencias de la computación, historia, y probabilidades.

Resumen de texto

Para efectos de obtener el resumen, como texto base se utiliza un artículo publicado en el año 2023, asociado a la selección multicriterio de modelos de pronóstico de consumo de energía eléctrica (Yajure-Ramírez C. , 2023). Con el fin de realizar el análisis, se toma la introducción, metodología, y conclusiones del artículo. Se utilizan cuatro métodos de resumen de texto extractivo: BERT, GPT2, XLNet, y LSA. Para cada uno de ellos se obtuvo un resumen del artículo mencionado, con longitudes de 176, 177, 195, y 272 palabras respectivamente. Seguidamente, el resumen del artículo publicado, de 234 palabras, se compara con cada uno de los resúmenes obtenidos, aplicando una prueba de similitud de texto con el fin de verificar que tan confiables son estas pruebas para efecto de generar el resumen de un artículo de investigación. Se consideraron tres métodos de prueba de similitud, y los resultados encontrados se presentan en la Tabla 6.

Tabla 6

Puntajes de similitud con el resumen del artículo

Métodos	roberta	bert	distiluse
GPT2	0,770	0,959	0,701
XLNet	0,718	0,917	0,665
BERT	0,702	0,894	0,592
LSA	0,626	0,888	0,479

En la primera columna de la tabla están los cuatro métodos utilizados para encontrar el resumen del texto asociado al artículo bajo estudio. Por otra parte, en la primera fila se ubican los tres métodos utilizados para encontrar el nivel de similitud entre el resumen original del artículo publicado, y cada uno de los resúmenes generados con los cuatro métodos mencionados. Los números dentro de la tabla indican el nivel de similitud encontrado con respecto al resumen del artículo, y la métrica utilizada fue similitud coseno, cuyos valores varían entre -1 y 1. Wang & Dong (2020) indican que, dependiendo del tamaño de los documentos, es recomendable utilizar la similitud coseno por encima de la métrica de la distancia euclidiana. Entonces, en la Tabla 6 los valores más cercanos a la unidad indican que hay buena similitud entre el resumen del artículo publicado y el resumen obtenido con el método respectivo, por lo que se puede decir que el resumen obtenido con el método GPT2 es el mejor para cada uno de los métodos de nivel de similitud utilizados.

Este resultado es interesante, puesto que primero: podría utilizarse este procedimiento para validar el resumen de un artículo previo a su publicación, y segundo: se podría utilizar el procedimiento para generar el resumen durante el proceso de diseño del artículo respectivo. Tan et al. (Tan, Kieuvongngam, & Niu, 2020) utilizan los métodos BERT y GPT2 para desarrollar el resumen de artículo de investigación médica relacionada con el COVID-19.

CONCLUSIONES

Se presentó una metodología para la extracción de información útil de artículos científicos de ingeniería publicados en revistas indexadas, utilizando las técnicas de procesamiento de lenguaje natural. Se dividió el análisis en dos partes: un análisis general al conjunto de los artículos, y un análisis específico por artículo, dependiendo de la técnica de procesamiento aplicada. Esta información consistió en palabras, bigramas, y frases claves, el resumen de textos, y la determinación de la similitud de textos.

Del análisis general se pudieron jerarquizar las principales palabras utilizadas en los artículos, de acuerdo con su frecuencia de ocurrencia, y visualizarlas a través de nubes de palabras, resaltando las palabras “datos” ,“energía”, y “modelo”. De la jerarquización de los bigramas de acuerdo con la métrica TF-IDF siendo los principales “solar fotovoltaica”, “variables explicativas”, y “energías renovables”.

En cuanto al análisis por artículo, se generaron los bigramas, y los más importantes según su frecuencia de ocurrencia, se compararon con las palabras claves de los artículos, resultando que en diecisiete de los artículos hubo al menos dos coincidencias, y en tres de ellos no hubo coincidencia alguna. Aplicando el método SGRank para la extracción de frases claves a los cuatro artículos publicados durante el año 2024, se pudo determinar el tópico tratado en tres de estos artículos. Al comparar el resumen de uno de los artículos publicados en el año 2023, con los cuatro resúmenes obtenidos al aplicar las métricas GPT2, XLNet, BERT y LSA, se obtuvo que el obtenido con GPT2 presentó mayor nivel de similitud con el resumen original del artículo, para tres técnicas de nivel de similitud diferente, y considerando la métrica similitud coseno.

REFERENCIAS

[1] Albrecht, J., Ramachandran, S., & Winkler, C. (2021). *Blueprints for Text Analytics Using Python.* Sebastopol, CA: O’Reilly Media, Inc.

[2] Artama, M., Sukajaya, I., & Indrawan, G. (2020). Classification of official letters using TF-IDF method. *Journal of Physics*, 1-7. DOI 10.1088/1742-6596/1516/1/012001.

[3] Baruni, J., & Sathiaseelan, J. (2020). Keyphrase Extraction from Document Using RAKE and TextRank Algorithms. *International Journal of Computer Science and Mobile Computing*, 83-93. DOI: 10.47760/IJCSMC.2020.v09i09.009.

[4] Calero Sánchez, M., González González, J., Sánchez Berriel, I., Burillo-Putze, G., & Roda García, J. (2024). El Procesamiento de Lenguaje Natural en la revisión de literatura científica. *Revista Española de Urgencias y Emergencias*, 184-195. https://www.reue.org/wp-content/uploads/2024/07/REUE_Vol3_Num3_2024_F2.pdf.

[5] Eldallal, A., & Barbu, E. (2023). BibRank: Automatic Keyphrase Extraction Platform Using Metadata. *MDPI Information*, 1-13. doi.org/10.3390/info14100549.

[6] Flayeh, A., Hamodi, Y., & Zaki, N. (2022). Text Analysis Based on Natural Language Processing (NLP). *2022 2nd International Conference on Advances in Engineering Science and Technology (AEST)* (págs. 774-778). Babil, Iraq: doi: 10.1109/AEST55805.2022.10413039.

[7] Flores Ramírez, J. A. (2023). EL VALOR DE LAS PALABRAS CLAVE EN LOS ARTÍCULOS CIENTÍFICOS. *Gestión I+D*, 11-13. http://saber.ucv.ve/ojs/index.php/rev_GID/article/view/25193.

[8] Hagiwara, M. (2021). *Real-World Natural Language Processing.* Shelter Island, NY: Manning Publications Co.

[9] Januzaj, Y., & Luma, A. (2022). Cosine Similarity – A Computing Approach to Match Similarity Between Higher Education Programs and Job Market Demands Based on Maximum Number of Common Words. *International Journal of Emerging Technologies in Learning*, 258-268. DOI: https://doi.org/10.3991/ijet.v17i12.30375.

[10] Kuznetsov, I., Afzal, O., Dercksen, K., Dycke, N., Goldberg, A., Hope, T., . . . Wang, J. (2024). What Can Natural Language Processing Do for Peer Review? *ArXiv*, 1-43. https://arxiv.org/abs/2405.06563.

[11] Liu, Z., Zhu, J., Cheng, X., & Lu, Q. (2023). Optimized Algorithm Design for Text similarity Detection Based on Artificial Intelligence and natural Language Processing. *Procedia Computer Science*, 195-202. https://doi.org/10.1016/j.procs.2023.11.023.

[12] Mandal, S., & Singh, G. (2020). LSA Based Text Summarization. *International Journal of Recent Technology and Engineering*, 150-156. DOI:10.35940/ijrte.B3288.079220.

[13] Mohan, G., Kumar, R., Parathasarathy, S., Aravind, S., Hanish, K., & Pavithria, G. (2023). Text Summarization for Big Data Analytics: A Comprehensive Review of GPT 2 and BERT Approaches. En R. Sharma, G. Jeon, & Y. Zhang, *Data Analytics for Intenet of Things Infrastructure* (págs. 247-264). Springer Nature. https://doi.org/10.1007/978-3-031-33808-3_14.

[14] Priya, B., Nandhini, J., & Gnanasekaran, T. (2021). An Analysis of the Applications of Natural Language Processing in Various Sectors. *Smart Intelligent Computing and Communication Technology*, 598-602. doi:10.3233/APC210109.

[15] Revuelta, G., & Llorente, C. (2024). *Elaboración de abstracts o resúmenes en el entorno sanitario.* Barcelona, España: Centro de Estudios de Ciencia, Comunicación y Soiedad de la Universidad Pompeu Fabra.

[16] Sandu, A., Cotfas, L.-A., Stănescu, A., & Delcea, C. (2024). A Bibliometric Analysis of Text Mining: Exploring the Use of Natural Language Processing in Social Media Research. *MDPI Applied Sciences*, 1-34. https://doi.org/10.3390/app14083144.

[17] Singh, J. (2023). *Natural language processing in the real-world : text processing, analytics, and classification.* Boca Raton, FL: CRC Press.

[18] Tan, B., Kieuvongngam, V., & Niu, Y. (2020). Automatic Text Summarization of COVID-19 Medical Research Articles using BERT and GPT-2. *ArXiv*, 1-13. https://arxiv.org/pdf/2006.01997.

[19] Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). *Practical Natural Language Processing.* Sebastopol, CA: O'Reilly Media, Inc.

[20] Vastrad, R., Devali, A., Urs, R., & D, N. (2022). Analysis of Text Data for Stock Prediction. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 3391-3395. https://doi.org/10.22214/ijraset.2022.43132.

[21] Wang, J., & Dong, Y. (2020). Measurement of Text Similarity A Survey. *MDPI Information*, 1-17. doi:10.3390/info11090421.

[22] Yajure-Ramírez, C. (2023). Multi-criteria methodology based on data science for the selection of the optimal forecast model for residential electricity consumption. *Scientia et Technica*, 108-116. DOI: https://doi.org/10.22517/23447214.25335.

[23] Yajure-Ramírez, C. A. (2024). Resolución del problema de optimización bi-objetivo para el despacho de plantas hidroeléctricas en condiciones de bajo caudal. *Revista ESPOL*, 32-43. https://doi.org/10.37815/rte.v36n1.1146.

Financiamiento:

Propia.

Conflictos de interés:

El autor declara no tener conflictos de interés.

Contribuciones de autoría:

El autor realizó todo el proceso de la investigación.