ARTÍCULO DE INVESTIGACIÓN / RESEARCH ARTICLE
https://doi.org/10.14482/inde.42.02.528.748

Obtención de *Insights* a partir de un modelo de Procesamiento de Lenguaje Natural para la denominación de programas académicos en educación superior

Obtaining Insights from a Natural Language Processing model for naming academic programs in higher education

Diego F. Calero Velasco*
Darío Delgado-Quintero**
Diana M. Cardona-Román***
Albeiro Cuesta-Mesa****
Sixto Enrique Campaña Bastidas*****

* Estudiante egresado del programa de Maestría en Gestión de Tecnología de Información. Universidad Nacional Abierta y a Distancia.
Orcid ID: https://orcid.org/0000-0002-6782-4958. diegofer@ieee.org.

** Profesor, PhD, Universidad Nacional Abierta y a Distancia.
Orcid ID: https://orcid.org/0000-0001-6549-5065. dario.delgado@unad.edu.co.

*** Profesora, PhD, Universidad de los Llanos.
Orcid ID: https://orcid.org/0000-0003-0953-5178. dcardona@unillanos.edu.co.

**** Profesor, PhD, Universidad Nacional Abierta y a Distancia.
Orcid ID: https://orcid.org/0000-0001-9938-5366. albeiro.cuestas@unad.edu.co.

***** Profesor asociado, PhD, Universidad Nacional Abierta y a Distancia.
Orcid ID: https://orcid.org/0000-0001-9937-2784. sixto.campana@unad.edu.co.

Correspondencia: Darío José Delgado Quintero: Universidad Nacional Abierta y a Distancia, Escuela de Ciencias Básicas, Tecnología e Ingeniería, calle 14 Sur nº. 14 - 23, Bogotá.
Tel: 601–3443700, Ext. 1333. dario.delgado@unad.edu.co.

Resumen

El Procesamiento del Lenguaje Natural (NLP) es esencial en la Inteligencia Artificial para la interacción entre computadoras y humanos. En este se explora el uso del NLP y técnicas de visualización en la creación de un modelo para la obtención de hallazgos (*insights*) en el diseño de nuevos programas académicos. La metodología cuantitativa utiliza técnicas como tokenización y TextRank, y con el apoyo del *Scattertext Plot*, discrimina categorías sobre programas de doctorado en tecnologías de información. Esta metodología, ampliamente usada y validada por expertos, destaca particularidades, conocimientos generales y tendencias emergentes en los programas comparados. Los resultados identifican cuatro cuadrantes esenciales para la toma de decisiones en el diseño de programas, y presentan las necesidades actuales y futuras en la denominación de programas académicos. En conclusión, este estudio resalta la importancia del NLP en el diseño académico adaptado a tendencias contemporáneas y proporciona una herramienta robusta para diseñadores de programas.

Palabras clave: diseño de programas académicos, generación de conocimiento, procesamiento de lenguaje natural, visualización de información.

Abstract

Natural Language Processing (NLP) is recognized as essential in Artificial Intelligence for the interaction between computers and humans. In this article, the use of NLP and visualization techniques in the creation of a model for insights acquisition in the design of new academic programs was explored. The quantitative methodology use techniques such as tokenization and TextRank, and with the support of the Scattertext Plot, discriminate categories on doctoral programs in information technologies. This methodology, validated by experts and widely used, highlighted particularities, general knowledge, and emerging trends in the compared programs. The results identify four essential quadrants for decision-making in program design, and shows current and future needs in program naming. In conclusion, the study underscores the importance of NLP in contemporary academic design and provides a robust tool for program designers.

Keywords: academic programs design; insights generation, natural language processing; visualization approach.

Fecha de recepción: 16 de mayo de 2023
Fecha de aceptación: 15 de enero de 2024

INTRODUCCIÓN

La creación y actualización de programas académicos en las instituciones de educación superior (IES) son vitales para enfrentar las necesidades sociales y desafíos del mercado laboral [1], [2]. Al brindar habilidades y saberes pertinentes, las IES impulsan la innovación, generan conocimiento, diversifican las opciones educativas y consolidan su posición en el mercado, favoreciendo el desarrollo socioeconómico.

En Colombia, al diseñar programas académicos, las universidades analizan el mercado para identificar necesidades y oportunidades de formación [3], [4]. Un aspecto esencial es la denominación del programa, que debe encapsular contenido, perfil de egreso, tipo y área del saber. Esta denominación no solo es crucial para la identidad del programa, sino que también refleja las tendencias y cambios en el ámbito educativo, siendo común que las instituciones, en cada renovación de registro calificado, reconsideren y modifiquen estas denominaciones para mantenerse al día, renovaciones que pueden presentarse en un rango de tres a siete años. Para asignar un nombre acertado, se investiga en programas análogos, se pondera el contexto y se realizan análisis estratégicos.

No obstante, obtener información de programas nacionales e internacionales puede acarrear sesgos por selección de la muestra o análisis del diseñador. Las técnicas de Procesamiento de Lenguaje Natural (NLP) abordan estos sesgos ofreciendo análisis profundos y generando *insights* estadísticos. En el campo educativo, NLP ha facilitado la automatización de la evaluación de respuestas abiertas y la adaptación de materiales de aprendizaje, entre otros [5]. Sin embargo, el reto con NLP en este trabajo está en filtrar e identificar *insights* y en transformar datos en información útil [6].

Este trabajo aborda estos retos mediante técnicas de visualización derivadas de NLP, con el fin de identificar *insights* en programas académicos que informen decisiones estratégicas en la creación o renovación de programas, elementos que también se constituyen como innovadores por la problemática abordada y por qué la metodología puede ser adaptada a otros contextos. Incluye descripción de NLP, construcción de un modelo para obtener *insights*, contextualización de su aplicabilidad, análisis de resultados y conclusiones.

EL PROCESAMIENTO DE LENGUAJE NATURAL Y LA OBTENCIÓN DE INSIGHTS PARA LA TOMA DE DECISIONES

El Procesamiento de Lenguaje Natural (NLP) brinda herramientas computacionales para el entendimiento, comprensión y manipulación automática y sistemática de texto o habla en lenguaje natural para llevar a cabo tareas específicas [7]. Las aplica-

ciones más comunes del NLP son la traducción automática, el procesamiento y análisis de lenguaje natural, la recuperación de información multilingüe entre idiomas (CLIE), el reconocimiento de voz, la conversión de audio a texto o de texto a audio, la anotación automática de texto, la detección de similitudes, la clasificación de documentos, entre otras.

El NLP es un campo de la inteligencia artificial que se enfoca en facilitar el entendimiento del lenguaje natural a las máquinas, de manera general, se vale de cinco pasos: i) Análisis morfológico [8], proceso en el que se descompone un texto en sus componentes básicos, como palabras, frases y oraciones; ii) Análisis sintáctico [9], que analiza la estructura de las oraciones y su gramática; iii) Análisis semántico [10], que interpreta el significado del texto; iv) Análisis pragmático [11], que busca la interpretación del significado en el contexto, y v) Generación de texto [12], que busca crear texto a partir del análisis realizado en los pasos anteriores. En general, el NLP utiliza una combinación de técnicas estadísticas, de aprendizaje automático y de reglas lingüísticas para realizar cada uno de estos pasos.

Los principales usos del NLP son la identificación de *Insights* para la toma de decisiones a partir de texto no estructurado, como el texto presente en las páginas web de los programas académicos; otros usos son:

- *Análisis de sentimientos* [13], para examinar textos grandes, identificando estados emocionales en reseñas de productos o comentarios en redes sociales. Esto ayuda a las empresas a discernir opiniones de usuarios, identificar áreas de mejora y tomar decisiones informadas para mejorar su oferta.
- *Extracción de información* [14], para obtener datos relevantes de fuentes no estructuradas, como noticias, páginas web o artículos de investigación, facilitando la identificación de patrones, tendencias del mercado y la toma de decisiones en desarrollo de productos y estrategia de marketing.
- *Análisis de temas* [15], en el que el NLP identifica los asuntos principales en un texto, ayudando a las empresas a entender qué es importante para los clientes y cómo pueden adaptarse para satisfacer esas necesidades.
- *Generación de resúmenes automáticos* [16], para condensar documentos largos y complejos, ahorrando tiempo a los profesionales y destacando puntos claves para la toma de decisiones.

En resumen, el NLP facilita a las organizaciones la extracción de información valiosa de grandes textos para decisiones informadas. Este estudio se enfoca en la extracción

de información, análisis de temas y obtención de *insights* para decisiones estratégicas en programas académicos.

Antecedentes

Encontrar palabras y frases que discriminen categorías de texto es una aplicación común del NLP estadístico. Por ejemplo, encontrar las palabras más características en los discursos de un partido político en contraste con los discursos de su partido rival, puede ayudar a los politólogos a identificar discrepancias partidistas [17]. Encontrar diferencias entre los lenguajes femeninos o masculinos para encontrar características distintivas de género en el diálogo cinematográfico [18],[19], entre otros ejemplos similares [20]. En educación [21], indica que se ha utilizado para enseñanza de la física, evaluar construcciones complejas de respuestas, explorar patrones en *datasets* académicos, o para realimentación o guía automatizada. Buena parte de la interpretación de los análisis realizados mediante el NLP se basa en diversos enfoques utilizados para visualizar y resaltar ítems importantes en los documentos analizados, por ejemplo, listas simples de frecuencias de palabras, nubes de palabras, así como diagramas de dispersión basados en palabras como el *Scattertext* [19], [22], [23].

METODOLOGÍA Y DISEÑO DE LA SOLUCIÓN

Se plantea a continuación un enfoque cuantitativo basado en la discriminación de categorías utilizando NLP para la comparación de programas académicos, especialmente su denominación y las temáticas abordadas, con relación a posibles agrupaciones, buscando ayudar a los diseñadores de programa a encontrar *insights* para una adecuada denominación de un programa académico [19].

Medidas de dispersión estadística del NLP

Para realizar la validación de los temas extraídos del análisis de texto, se utilizaron las siguientes medidas de dispersión que indican qué tan cercanas o lejanas están las palabras en un corpus; en tal sentido, [24], [25] proponen un conjunto de medidas fundamentales:

- *Rango*: número de partes que contiene un determinado vocablo (v_i) y permite comparar la dispersión de los términos en el corpus (n).
- *Desviación estándar tradicional* (SD): indicado por la ecuación 1; el vocablo en el corpus es v_i, la frecuencia total del elemento en el corpus es f, el número total de términos en el corpus es n. $SD: \sqrt{\frac{\sum_{i=1}^{n}\left(v_i - \frac{f}{n}\right)^2}{n}}$ (1)

- *Dispersión de Juilland's D*: es una métrica utilizada para cuantificar la diversidad léxica en un corpus de texto (ver ecuación 2): $Julilland's\ D: 1 - \frac{SD(p)}{media(p)} \times \frac{1}{\sqrt{(n-1)}}$ (2)
- *Rosengren's S*: esta métrica permite determinar la dispersión o variabilidad de cada parte del corpus o, mejor, de la frecuencia de los vocablos en un corpus, donde el tamaño de cada parte del corpus en porcentaje es (s_i); las frecuencias del elemento en cuestión (vocablo) en cada parte del corpus es (v_i) y la frecuencia total del elemento en cuestión en el corpus es (f); tal como se muestra en la ecuación 3. $Rosengren's\ S: \left(\sum_{i=1}^{n} \sqrt{s_i \cdot v_i}\right)^2 \times \frac{1}{f}$ (3)
- *Kullback-Leibler (KL)-divergence*: medida no simétrica utilizada para precisar la similitud o la distancia entre dos distribuciones de probabilidad de palabras en diferentes corpus o conjuntos de datos de texto (ver ecuación 4). $KL\ Divergence: \sum_{i=1}^{n} \frac{v_i}{f} \times log_2\left(\frac{v_i}{f} \times \frac{1}{s_i}\right)$ (4)
- *Dispersión Promedio (DP)*: medida de diversidad léxica promedio; se calcula dividiendo la cifra total de palabras del corpus (n) entre la frecuencia de un determinado vocablo (f), con la fórmula indicada en la ecuación 5: $DP = \frac{n}{f}$ (5).
- *Dispersión Promedio Normalizada (DP norm)*: que permite comparar corpus de diferentes tamaños; se obtiene la distancia acumulada al cuadrado, que es el cuadrado de la suma acumulada de las diferencias entre la distancia real de aparición de cada vocablo (v_i) y la distancia ideal media ($\frac{f}{n}$), así como se presenta en la ecuación 6: $DP\ Norm: \sum_{i=1}^{n} \left(v_i - \frac{f}{n}\right)^2$ (6)

Contexto de la solución

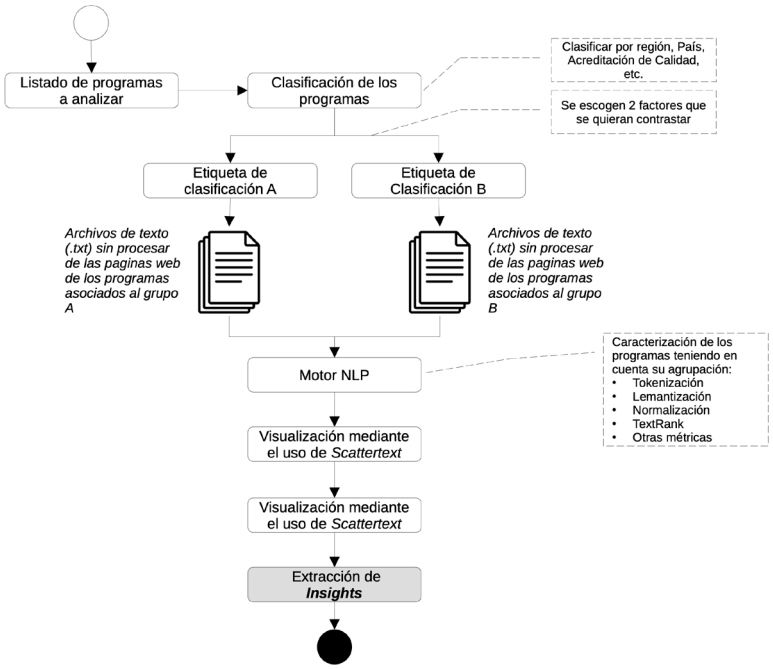
Los sitios web de los programas académicos suelen tener información relevante que describen de manera general los elementos que suelen diferenciar dicho programa de los demás programas en su área de estudio; entre otros elementos, estas páginas suelen contener: denominación del programa, información legal, nivel de formación, metodología en que se imparte, título que otorga, elementos distintivos, requisitos de ingreso y egreso, perfil de los estudiantes, perfil de sus egresados, planes de estudio. Estos sitios web condensan información relevante que puede ser utilizada para realizar inteligencia previa a la definición de denominaciones de nuevos programas o la renovación de programas existentes (justificar un cambio en su denominación); esto mediante la búsqueda de tendencias regionales, tendencias temáticas u oportunidades estratégicas. Con lo anterior se construyó una base de datos, publicada en [26], con los atributos del nombre de la universidad, un identificador para realizar una clasificación (binaria) en el caso de estudio, una clasificación entre programas internacionales y programas nacionales, y un campo con la información proporcio-

nada por la página web del programa doctoral. La información del último campo se almacenó en texto plano con los siguientes atributos: el nombre del programa de doctorado, sus objetivos, sus líneas de investigación, el perfil de los estudiantes, el perfil de los egresados [19].

Modelo planteado

En la Figura 1 se presenta el modelo propuesto para obtención de *Insights* como ayuda a la definición de la denominación de un programa académico.

A partir de una adecuada recopilación de información de programas académicos en un mismo campo de formación[1], de las páginas web institucionales, se definieron criterios de clasificación de los programas, como separación entre programas nacionales (Colombia) y programas internacionales, programas con algún tipo de acreditación de calidad, modalidades de oferta (distancia tradicional, virtual, presencial).



Fuente: elaboración propia [19].

Figura 1. Proceso metodológico para la obtención de *Insights* mediante el uso de PNL como soporte en la denominación de un programa académico

1 Inicialmente una manera de poder encontrar programas que correspondan a una misma línea temática general se puede realizar mediante la Clasificación Normalizada de Educación CINE F 2013 AC (DANE, 9 de febrero de 2023), el cual funciona como sistema de referencia para la clasificación de programas en diferentes campos de formación.

Una vez definidas las posibles clasificaciones para realizar el análisis, se escogieron los escenarios de interés; vale la pena resaltar que para este trabajo se prefirió el análisis dicotómico, el cual contempla solo dos agrupaciones, por ejemplo, comparar entre programas nacionales o internacionales[2], entre otras posibles agrupaciones del corpus y sus variables.

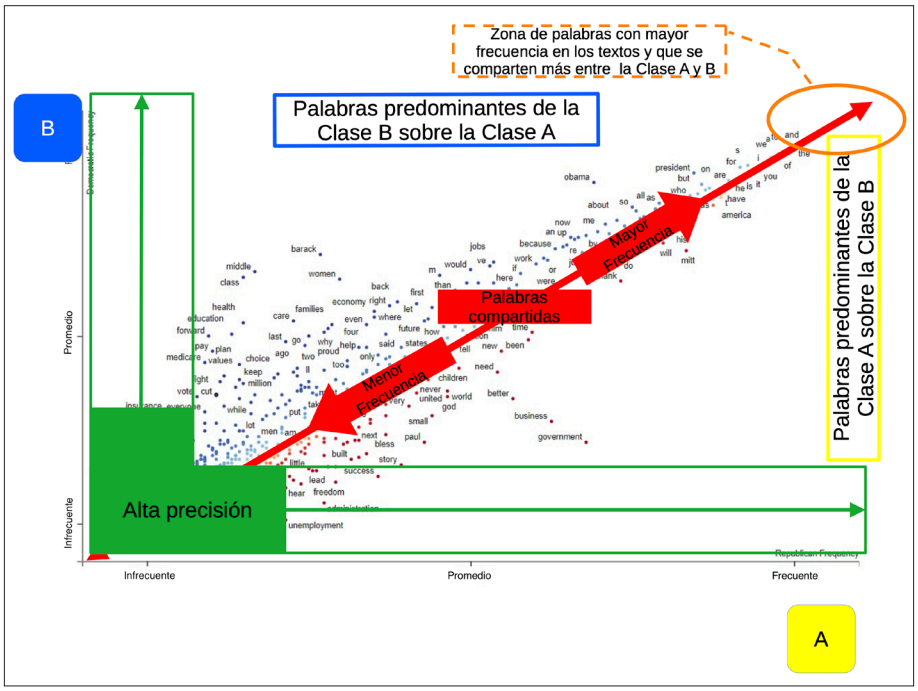
Posterior a la definición de las agrupaciones, se etiquetó la información extraída de los programas. Una vez empaquetada y etiquetada la información se realizó el análisis de los programas, utilizando un motor para el procesamiento de lenguaje natural, donde la Tokenización, lematización y normalización de la información permitieron la adecuada comparación. Luego de preprocesados los textos se aplicó el algoritmo de *TextRank* [27], el cual es un modelo de clasificación basado en grafos para el procesamiento de texto utilizado principalmente para encontrar oraciones y palabras relevantes en un texto.

Una vez identificadas las oraciones y palabras clave, y cuantificadas las frecuencias y demás métricas estadísticas, se utilizó el enfoque de visualización de información de *Scattertext* [20] para poder comparar visualmente categorías de texto. Posteriormente, al emplear Scattertext para la visualización, se hicieron indispensables medidas de dispersión como Juilland's D y KL-divergence para evaluar la uniformidad y divergencia de términos en los programas analizados. La Dispersión Promedio (DP) y la Dispersión Promedio Normalizada (DP norm) fueron cruciales para comparaciones equitativas entre corpus de distintos tamaños, mientras que la Desviación estándar (SD) y el rango de frecuencias (Range) aportaron perspectivas sobre la variabilidad y el alcance de los términos clave en diferentes categorías.

Este método presenta palabras y oraciones clave en un gráfico de dispersión (*Scatter Plot*), ver Figura 2, el gráfico presenta la información de la siguiente manera:

- Se presenta un sistema bidimensional de coordenadas cartesianas en el que cada punto representa un término y está ubicado en un punto con dos coordenadas, una en el eje-x y otra en el eje-y. Las coordenadas se derivan de las frecuencias de cada término en cada agrupación de programas académicos.

2 Debe tenerse en cuenta que ambas categorías de información deben estaren el mismo idioma.



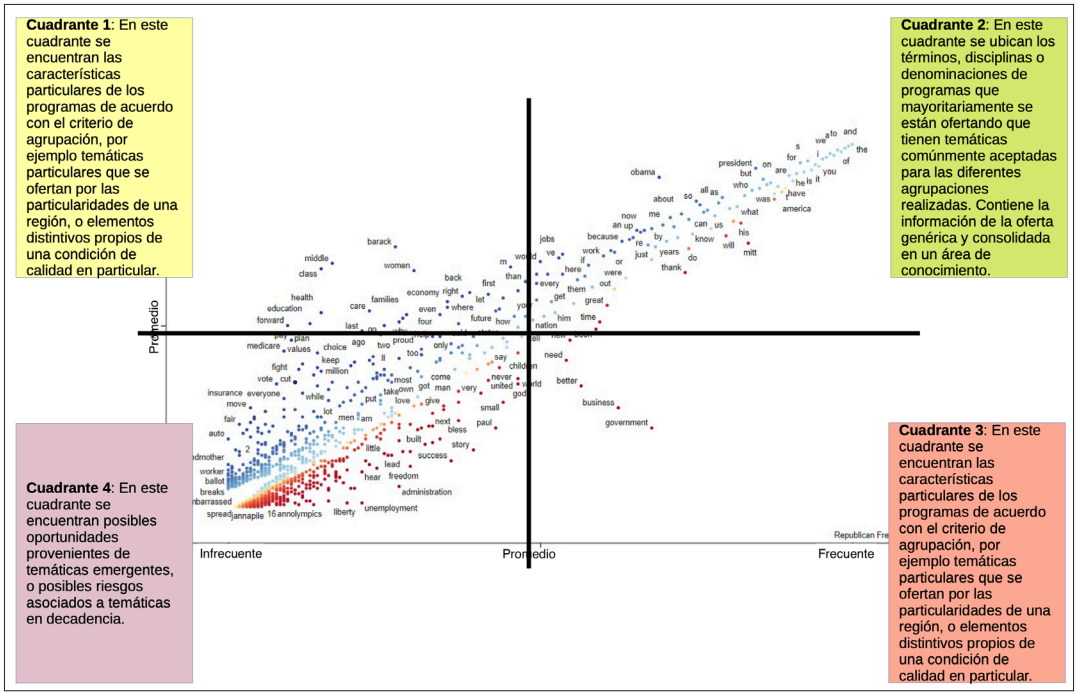
Fuente: adaptado de [28].

Figura 2. Interpretación de los resultados al comparar dos agrupaciones de texto utilizando *Scattertext Plot*

- Los ejes X y Y representan las categorías en las que se quiere comparar a los programas académicos, denotados en la imagen en amarillo para la clase A, en azul para la clase B. A su vez, los ejes representan la frecuencia de ocurrencia de palabras o frases clave en cada categoría de agrupación.
- El componente señalado en rojo representa aquellos términos clave que coinciden en menor o mayor medida para las dos categorías que se comparan.
- En verde, la franja de alta precisión, se refiere a aquellos términos clave para cada una de las categorías que tienen un poder discriminatorio alto. Es decir, son términos distintivos para cada una de las categorías.
- En naranja se presentan los términos con mayor frecuencia para ambas categorías que muestran los elementos en común que tienen.

Finalmente, una vez generado en *Scattertext Plot* se identifican los *insigths* proporcionados por el análisis. En la Figura 3 se presenta una guía para la definición de posibles *Insights* para la toma de decisiones con la información de los programas contrastados. El *Scattertext Plot* se divide en cuatro cuadrantes, adaptando el con-

cepto de cuadros semánticos abordados en la misma técnica [29], para asistir a los diseñadores de programas en la toma de decisiones.



Fuente: elaboración propia, generada en Python con scattertext.

Figura 3. Guía para la definición de Insights de acuerdo con un *Scattertext Plot*

El cuadrante 1 y el 3 permiten identificar las particularidades de los programas académicos comparados según sus agrupaciones y evidenciar temáticas de interés particular que sólo son relevantes para una región específica debido a factores políticos, estratégicos, logísticos, tecnológicos, culturales, entre otros. Esto es útil, por ejemplo, para comparar programas nacionales con internacionales en una misma área de conocimiento.

En el cuadrante 2 se encuentra la información compartida entre las dos clases comparadas, que representa el corpus de conocimiento general de la disciplina y las denominaciones mayoritariamente empleadas. Este cuadrante podría considerarse como el *status quo* o la zona de confort de los programas académicos comparados.

Finalmente, el cuadrante 4 puede ser una fuente de información estratégica, pero debe abordarse con cautela, ya que representa las temáticas emergentes de una disciplina y los cambios en la forma de denominar un programa con esas temáticas. Sin

embargo, también pueden estar ubicadas en este cuadrante temáticas o denominaciones que estén entrando en desuso [19].

RESULTADOS Y DISCUSIÓN

Se utilizó como caso de estudio la solicitud de una universidad pública que busca plantear su primer programa de doctorado en temas relacionados con tecnologías de la información, comunicaciones, ingeniería de sistemas, informática y computación. Para ello, se recopiló información de programas de doctorado [19], ver Tabla 1 y referencia [26], en el periodo mayo de 2021 a enero de 2022 que cumplieran las siguientes restricciones: i) que los programas de doctorado tengan una denominación que se relacione con los requisitos planteados por la escuela de ingenierías; ii) programas de doctorado con un registro vigente en el SNIES para el caso de programas nacionales (Colombia); iii) programas de doctorado en el contexto internacional activos y asociados al área de conocimiento definida; iv) programas internacionales extraídos de países de habla hispana o con información en idioma español.

Una vez consultada la información de los programas de doctorado, se construyó una base de datos [26] que almacenó la información a ser procesada. Vale la pena resaltar que no es información normalizada en las páginas web de los programas.

Tabla 1. Cantidad de programas académicos cuya información fue recopilada para el estudio

Origen del programa	Cantidad
Nacional	31
Internacional	26

Fuente: elaboración propia. La información ampliada puede ser consultada en [26].

Para la visualización de información (*ScatterText Plot*), ver la Figura 4 (cuyo gráfico dinámico puede ser consultado en [30]), cada punto corresponde a una palabra o frase mencionada en las páginas de presentación de los programas de doctorado tanto en un contexto nacional o internacional; los términos se colorean[3] de acuerdo con su afiliación, rojo para los programas de doctorado nacionales y azul para los programas de doctorado internacionales. Entre más cercano se encuentre un punto a la parte superior de la gráfica, más utilizado es ese término por los programas

3 Que un término aparezca en una clase u otra, no lo excluye de su aparición en una clase contraria. Simplemente toma partido por la relación entre la frecuencia de aparición en ambas clases.

en un contexto internacional, mientras más a la derecha se encuentre un punto, es más frecuentemente encontrado en los programas de doctorado en un contexto nacional. Las palabras o frases frecuentemente usadas en los programas doctorales, tanto nacionales como internacionales (tales como "Investigación", "Ingeniería", "Doctorado", "Sistemas", "Computación", "Ciencias" o "Programa"), aparecen en la esquina superior derecha del gráfico, lo cual coincide con la mayor frecuencia en las denominaciones de este tipo de programas de formación doctoral. En la contraparte izquierda aparecen aquellos términos que aparecen con menor frecuencia en las dos categorías analizadas.

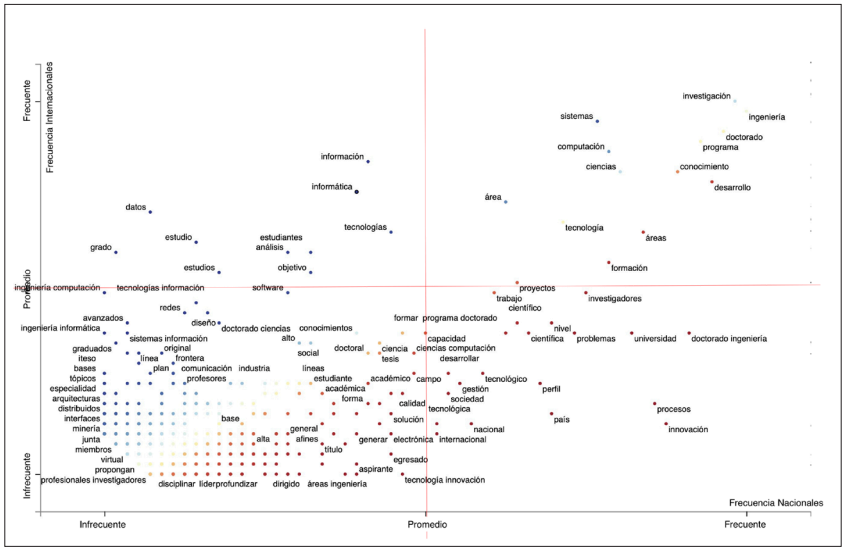
Elementos interesantes para resaltar suceden cerca de la esquina superior izquierda (cuadrante 1, ver Figura 3 y resultados en la Figura 4), en donde, por ejemplo, la palabra "datos", que representa temáticas como minería de datos, bases de datos, ciencia de datos, ingeniería de datos, análisis de datos. Son más frecuentemente encontradas en la presentación de programas extranjeros, y en contraste, en la esquina inferior derecha (cuadrante 3 en Figura 3 y Figura 4), con el término "innovación" en contextos como herramienta para la solución de problemas, innovación tecnológica, en la transferencia de tecnología, como herramienta para la productividad, como herramienta para el emprendimiento, como motor de desarrollo regional, para la sostenibilidad industrial, o el término "Doctorado en ingeniería", que de paso es la denominación más frecuentemente encontrada, son elementos propios de los programas nacionales.

Otro elemento para tener en cuenta se encuentra en el cuadrante inferior izquierdo (cuadrante 4, ver Figura 4), este cuadrante concentra la información que parece incluir la información que distingue a los programas entre sí; por ejemplo, el término "Virtual", si bien corresponde a las palabras con bajas frecuencias de aparición, hace referencia a posibles temáticas, como realidad Virtual; también hace referencia a la modalidad en la que se oferta un doctorado, o frases como "Ingeniería computación", "Tecnología información", "Ingeniería informática", "Software", que son nombres particulares que se les da a los programas de doctorado en un contexto internacional que pueden ser denominaciones emergentes o de interés.

En términos generales, es importante analizar los términos en relación con su frecuencia y dispersión en los documentos correspondientes a los programas de doctorado, sin discriminar las clases de análisis.

Se utiliza en este caso la medida de dispersión de *S de Rosengren* [22], la cual evalúa la variabilidad de frecuencias de las palabras en un corpus de texto. Esta medida permite calcular qué tan equitativamente se distribuyen las frecuencias de un término en particular en todo el corpus de textos de los programas de doctorado en las páginas web. Los términos (palabras o vocablos) tienden a aumentar en sus puntajes de dis-

persión a medida que se vuelven más frecuentes (suele existir una alta correlación)[4] (ver Figura 5).



Fuente: elaboración propia, generada en Python con scattertext.

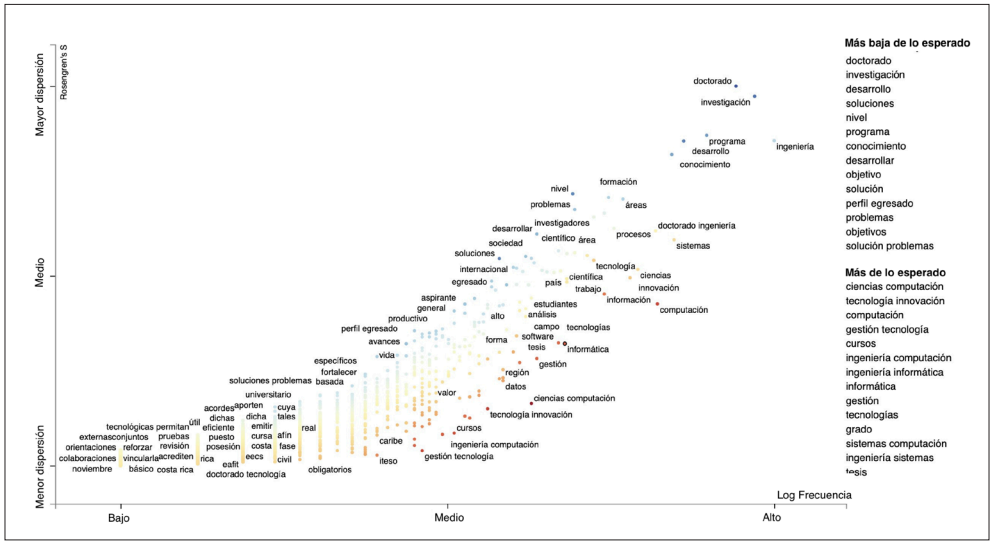
Figura 4. *ScatterText Plot* para la comparación de programas de doctorado que involucren la formación en TIC, contraste entre programas internacionales y nacionales

La Figura 5 permite ubicar los términos importantes en la presentación de los programas de doctorado: en azul, los términos con una alta frecuencia y dispersión; en rojo, aquellos con frecuencias altas pero baja dispersión. Valores altos de dispersión y frecuencia sugieren que los términos se encuentran en la mayoría de las páginas web de los programas de doctorado analizados, mientras que una baja dispersión indica que las frecuencias están concentradas en un conjunto más pequeño de palabras propias de unos pocos programas.

Para tener una idea más completa de la dispersión, en la Tabla 2 se presenta una muestra de los términos más comunes encontrados en el corpus; adicionalmente se calculan las métricas explicadas anteriormente, que permiten comparar la dispersión de los términos en el corpus, como la medida de dispersión de Juilland's D, la métrica de KL-divergence, la Dispersión Promedio (DP), la Dispersión Promedio

4 Los términos que más aparecen en las descripciones de los documentos igualmente suelen aparecer en una mayor variedad de documentos para analizar.

Normalizada (DP norm) que permite comparar corpus de diferentes tamaños, la Desviación estándar tradicional (SD) y el rango de frecuencias (Range) [19].



Fuente: elaboración propia, generada en Python con scattertext.

Figura 5. Análisis de dispersión de términos en el conjunto de textos a analizar

Tabla 2. Muestra de los primeros mayores valores de dispersión de términos en los programas de doctorado analizados ordenados de acuerdo con su frecuencia

Index	Frequency	Range	SD	Juilland's D	Rosengren's S	DP	DP norm	KL-divergence
doctorado	354	57	2,828	0,915	0,937	0,195	0,195	0,185
investigación	297	56	4,068	0,918	0,912	0,207	0,208	0,225
programa	251	50	2,518	0,897	0,816	0,295	0,295	0,42
ingeniería	193	49	5,119	0,9	0,803	0,285	0,285	0,421
desarrollo	157	44	2,867	0,881	0,802	0,275	0,276	0,444
conocimiento	144	41	2,507	0,886	0,769	0,304	0,304	0,473
formación	124	36	1,554	0,867	0,663	0,368	0,368	0,687
áreas	122	39	1,909	0,816	0,659	0,386	0,386	0,744
problemas	104	32	1,35	0,846	0,633	0,383	0,384	0,78
universidad	97	33	1,896	0,815	0,624	0,405	0,405	0,868
investigadores	91	34	1,475	0,838	0,614	0,405	0,406	0,851

Continúa...

procesos	84	27	2,145	0,81	0,586	0,452	0,452	0,918
doctorado ingeniería	80	34	2,431	0,826	0,58	0,424	0,425	0,892
sistemas	77	36	3,742	0,81	0,558	0,477	0,477	1,101

Fuente: elaboración propia.

Entre de los principales *insights* para la denominación de programas de doctorado se pueden destacar: doctorado en Ciencias de la Computación, doctorado en Tecnología e Innovación, doctorado en Ingeniería y Computación, doctorado en Innovación, doctorado en Gestión Tecnológica, doctorado en Ciencia de Datos, doctorado en Ingeniería Informática, doctorado en Gestión de la Información. Estos resultados combinan, como se indicó antes, los rasgos distintivos, con la oferta de programas y presenta un primer insumo para establecer la denominación de un programa de formación de alto nivel.

CONCLUSIONES

Utilizar el NLP en el contexto de análisis textual de información proveniente de sitios web de instituciones que ofertan programas de doctorado permitió identificar un listado de términos relevantes para la descripción de los programas. Estas palabras relevantes, principalmente, se relacionan con la denominación de los programas; razón por la cual, en una primera instancia, se puede afirmar que el enfoque de análisis permite encontrar cuál es la tendencia de denominación de los programas académicos desde la perspectiva de la información Nacional (Colombia) o Internacional.

Scattertext, como herramienta de visualización, facilita la comparación de categorías y discriminación de frecuencias de términos, ayudando a identificar elementos distintivos y comunes en programas académicos. Permite reconocer temáticas y tendencias según la categoría, así como los términos más relevantes y utilizados por las universidades en diferentes entornos.

La identificación de términos clave, tendencias y referencias a programas y universidades permite a los diseñadores enfocar sus esfuerzos en crear propuestas de valor basadas en tendencias regionales y disciplinarias diferenciadoras, gracias a herramientas como el NLP que facilitan el análisis holístico de la información.

El estudio se basó en un enfoque de clasificación de programas, pero hay múltiples categorías para contrastar, como programas acreditados, registro calificado, universidades públicas vs. privadas o modalidades virtual y presencial. Analizar diferentes agrupaciones podría proporcionar información adicional valiosa para el diseño de programas.

El NLP permitió considerar *insights para programas de doctorado* desde la extracción de información [13] y el análisis de temas [14] de un corpus no estructurado, cuyos resultados expusieron de una manera objetiva las tendencias y temas con alta frecuencia de aparición de términos relacionados con doctorado en el ámbito de las tecnologías de información y comunicación y también elementos diferenciadores con una baja frecuencia de aparición, métodos que pueden aportar en diferentes campos de aplicación.

La Figura 4 y la Figura 5 ayudan a delimitar interpretaciones de los resultados gráficos del análisis textual. Sin embargo, para identificar hallazgos estratégicos, es esencial interpretar y reconstruir términos frecuentes y significativos. Esto no solo requiere dominio del tema y comprensión de formatos de visualización, sino también la intervención de expertos disciplinares en los programas que se diseñan para validar adecuadamente los resultados.

La principal limitación que puede considerarse en el trabajo es la construcción del corpus para el tratamiento y análisis de texto, pues se tuvo como objetivo el abordaje en idioma español, dejando por fuera el análisis de denominaciones de programas de doctorado en inglés, constituyéndose en una oportunidad de mejora y extensión del modelo utilizado.

AGRADECIMIENTOS

Los autores agradecen a la Escuela de Ciencias Básicas, Tecnología e Ingeniería de la Universidad Nacional Abierta y a Distancia por la asignación de horas a través del proyecto de investigación ECBTIPIE022021 “Estudio de Factibilidad para la creación de programas de Doctorado en Ingeniería bajo la modalidad Virtual y a Distancia”. La autora Cardona-Román agradece parcialmente al proyecto C03-F02-014-2022 DGI/Unillanos.

REFERENCIAS

[1] D. A. Cueva Gaibor, “Transformación digital en la universidad actual”, *Revista Conrado*, vol. 16, nº. 77, pp. 483–489, dic. 2020.

[2] M. L. Piñero Martín, E. R. Esteban Rivera, A. R. Rojas Cotrina, C. Becerra y S. Fiorella, “Tendencias y desafíos de los programas de posgrado latinoamericanos en contextos de COVID-19”, *Revista Venezolana de Gerencia* (RVG), vol. 26, nº. 93, pp. 123–138, 2021.

[3] Ministerio de Educación [MEN], “Creación de programas académicos. Ministerio de Educación Nacional”. [En línea]. Disponible en: https://www.mineducacion.gov.co/portal/Educacion-superior/Sistema-de-Educacion-Superior/235796:Creacion-de-programas-academicos. [Accedido: 19 junio, 2023].

[4] G. Restrepo González, E. Castañeda Gómez y D. Álzate G., " ¿Tienen propuesta de valor las facultades y programas de Ingeniería en Colombia?", *Revista Ingeniería y Sociedad*, vol. 8, pp. 49–57, 2014. [En línea]. Disponible en: http://hdl.handle.net/10495/7740. [Accedido: 19 junio, 2023].

[5] T. Shaik *et al.*, "A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis", *IEEE Access*, vol. 10, pp. 56720–56739, 2022. doi: 10.1109/ACCESS.2022.3177752.

[6] K. Dhanasekaran y R. Rajeswari, "Insight into Information Extraction Method using Natural Language Processing Technique", *International Journal of Computer Science and Mobile Applications*, vol. 1, nº. 5, pp. 97–109, 2013.

[7] G. G. Chowdhury, "Natural Language Processing", *Annual Review of Information Science and Technology (ARIST)*, vol. 37, pp. 51–89, 2023.

[8] K. R. Beesley, "Morphological analysis and generation: A first step in natural language processing", en *Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation*, pp. 1–8, Julie Carson-Berndsen, Ed., SALTMIL Workshop at LREC, abr. 2004. [En línea]. Disponible en: http://lrec.elra.info/proceedings/lrec2004/ws/ws2.pdf#page=7. [Accedido: 19 junio, 2023].

[9] Pătruţ, B., Boghian, I., & Moldovan, G., "Syntactic Analysis. Lexical Disambiguation, Logical Formalisms, Discourse Theory, Bivalent Verbs", en *Natural Language Processing*, Germany, Munich: AVM–Akademische Verlagsgemeinschaft München, 2012.

[10] D. Hussen Maulud, S. R. M. Zeebaree, K. Jacksi, M. A. Mohammed Sadeeq y K. Hussein Sharif, "State of Art for Semantic Analysis of Natural Language Processing", *Qubahan Academic Journal*, vol. 1, nº. 2, pp. 21–28, mar. 2021. doi: 10.48161/qaj.v1n2a44.

[11] Y. Li, M. a Thomas, y D. Liu, "From semantics to pragmatics: where IS can lead in Natural Language Processing (NLP) research", *European Journal of Information Systems*, vol. 30, nº. 5, pp. 569–590, sep. 2021. doi: 10.1080/0960085X.2020.1816145.

[12] M. Bayer, M.-A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, y C. Reuter, "Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers", *International Journal of Machine Learning and Cybernetics*, vol. 14, nº. 1, pp. 135–150, ene. 2023. doi: 10.1007/s13042-022-01553-3.

[13] A. Rajput, "Natural Language Processing, Sentiment Analysis, and Clinical Analytics", en *Innovation in Health Informatics*, Elsevier, pp. 79–97. 2020. doi: 10.1016/B978-0-12-819043-2.00003-4.

[14] S. Singh, "Natural Language Processing for Information Extraction", jul. 2018.

[15] Y. Belinkov y J. Glass, "Analysis Methods in Neural Language Processing: A Survey", *Trans Assoc Comput Linguist*, vol. 7, pp. 49–72, abr. 2019. doi: 10.1162/tacl_a_00254.

[16] D. Khurana, A. Koli, K. Khatter y S. Singh, "Natural language processing: state of the art, current trends and challenges", *Multimed Tools Appl*, vol. 82, núm. 3, pp. 3713–3744, ene. 2023. doi: 10.1007/s11042-022-13428-4.

[17] J. R. Grimmer, "Representational Style: The Central Role of Communication in Representation", Harvard University, 2010.

[18] A. Schofield y Leo Mehr, "Gender-distinguishing features in film dialogue. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*", pp. 32–39, jun. 2016.

[19] D. F. Calero, "Modelo basado en procesamiento de lenguaje natural para el diseño de programas académicos asistido por computador factor 1: denominación del programa", [Proyecto de investigación]. Repositorio Institucional UNAD., Universidad Nacional Abierta y a Distancia, 2023. [En línea]. Disponible en: https://repository.unad.edu.co/handle/10596/56948. [Accedido 14 enero, 2024].

[20] J. S. Kessler, "Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ", mar. 2017.

[21] P. Wulff, A. Westphal, L. Mientus, A. Nowak y A. Borowski, "Enhancing writing analytics in science education research with machine learning and natural language processing -Formative assessment of science and non-science preservice teachers' written reflections", *Front Educ (Lausanne)*, vol. 7, ene. 2023. doi: 10.3389/feduc.2022.1061461.

[22] J. Richarz, S. Wegewitz, S. Henn y D. Müller, "Graph-based research field analysis by the use of natural language processing: An overview of German energy research", *Technol Forecast Soc Change*, vol. 186, p. 122139, ene. 2023. doi: 10.1016/j.techfore.2022.122139.

[23] K. Zhao, N. Shi, Z. Sa, H.-X. Wang, C.-H. Lu y X.-Y. Xu, "Text mining and analysis of treatise on febrile diseases based on natural language processing", *World J Tradit Chin Med*, vol. 6, nº. 1, p. 67, 2020. doi: 10.4103/wjtcm.wjtcm_28_19.

[24] S. Th. Gries, "Analyzing Dispersion", en *A Practical Handbook of Corpus Linguistics*, Cham: Springer International Publishing, pp. 99–118, 2020. doi: 10.1007/978-3-030-46216-1_5.

[25] S. Th. Gries, "What do (most of) our dispersion measures measure (most)? Dispersion?", *Journal of Second Language Studies*, vol. 5, núm. 2, pp. 171–205, oct. 2022. doi: 10.1075/jsls.21029.gri.

[26] D. F. Calero, "Base de datos con la información de los programas". [En línea]. Disponible en: https://docs.google.com/spreadsheets/d/1lf5j-Xss3f72rahj2Vb9dLiiIvnjQBh-/edit?usp=share_link&ouid=101796938056868559056&rtpof=true&sd=true. [Accedido: 19 junio, 2023].

[27] C. Mallick, A. K. Das, M. Dutta, A. K. Das y A. Sarkar, "Graph-Based Text Summarization Using Modified TextRank", pp. 137–146, 2019. doi: 10.1007/978-981-13-0514-6_14.

[28] James Opacich, "Interpreting Scattertext: A seductive tool for plotting text", Towards Data Science. [En línea]. Disponible en: https://towardsdatascience.com/interpreting-scattertext-a-seductive-tool-for-plotting-text-2e94e5824858. [Accedido: 19 junio, 2023].

[29] Jason Kessler, "ScaterText". https://github.com/JasonKessler/scattertext.

[30] D. F. Calero, "ScatterText Plot dinámico". [En línea]. Disponible en: https://drive.google.com/file/d/1-3br2a4uZ3Wexc2-6MsV--GNT5ect6ZK/view?usp=share_link. [Accedido: 19 junio, 2023].