Modelo basado en Procesamiento de Lenguaje Natural para el diseño de programas académicos asistido por computador

Diego F. Calero Velasco, Pregrado en Ingeniería, Darío J. Delgado, PhD en Ingeniería
Universidad Nacional Abierta y a Distancia, Colombia, dfcalerov@unadvirtual.edu.co
Universidad Nacional Abierta y a Distancia, Colombia, dario.delgado@unad.edu.co

Abstracto– En este trabajo se presenta creación de un modelo asistido por computadora basado en NLP (por sus siglas en español Procesamiento de Lenguaje Natural), el cual es una rama del Aprendizaje de Maquinas, para el diseño de perspectivas de entrada en la asistencia para la creación de programas académicos doctorales. Este trabajo contempla la asistencia por computador en la generación de perspectivas para ala toma de decisiones en la denominación del programa académico, entre estos, puede incluir nombre del programa, descripción del programa y línea de investigación del programa. Este modelo asistido por computador contempla la comparación, a través de librerías de la tecnología NLP, y la obtención información de múltiples fuentes de información principalmente académicos del ámbito nacional y del internacional sobre programas doctorales existentes. Posterior a la comparación de las fuentes de información, las librerías NLP arrojan datos y graficas para la realización de análisis comparativos consistentes a la información suministrada. Como resultado del trabajo propuesto, se obtuvo un modelo computacional basado en NLP que esta generando perspectivas para la asistencia en la creación de programas académicos doctorales basados en programas doctorales actualmente existentes del ámbito nacional e internacional. El propósito principal es dar diferente perspectiva a los interesados de denominaciones de programas con el objetivo de que los tomadores de decisiones tengan suficiente evidencia documental nacional e internacional para que puedan tomar una decisión sobre la denominación del programa académico.

Palabras Clave—Aprendizaje de Máquinas, Librería NLP, Programas Doctorales, Asistencia Computacional.

I. Introducción

El proyecto se enmarca en el desarrollo de un modelo para la denominación de nuevos programas doctorales, para que las instituciones interesadas puedan eventualmente crear y ofertar en un mediano o largo plazo dicho programa académico. Como es de conocimiento general, la educación avanza y evoluciona según factores investigativos y de la industria. Y las instituciones educativas en su camino a la excelencia, deben siempre estar a la vanguardia y la innovación en educación respecto a lo que esta pasando tanto en lo académico como en la industria a nivel nacional e internacional. Es por esto, que existe la necesidad de que las instituciones contemplen y tengan un mapa de ruta definido sobre sus nuevos programas doctorales en ingeniería que potencialmente pueda liberar en un futuro después de realizar un exhaustivo estudio y aprobaciones por los entes regulatorios. Como resultado del trabajo principalmente se espera que, basado en ciertas definiciones iniciales, a través de este modelo asistido por computador y tecnología NLP, las instituciones interesadas puedan tener herramientas que les proporcionen diferentes perspectivas que les ayuden a identificar como podría ser el potencial nombre que podrían tener para un programa académico doctoral en ingeniera basado en programas académicos doctorales existentes tanto a nivel nacional e internacional.

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

II. Problemática

La propuesta de investigación objetivo de este trabajo busca como fin, contribuir directamente con algunas instituciones educativas que están en búsqueda de creación de programas académicos doctorales en ingeniería para que a través de herramientas computacionales asistidas y con el uso de tecnologías modernas, puedan tener perspectivas de como podrían llamar a su doctorado en ingeniería. Cabe resaltar que, si bien el alcance de este proyecto es únicamente el nombre o denominación del programa, también puede extenderse a la descripción del mismo o incluso a sus líneas de investigación.

La idea en términos generales no es solo de ingresar cierta cantidad de información y que un algoritmo orientado a objetivos y NLP predetermine un nombre. Sino que el objetivo principal, es que La ECBTI pueda llegar a tener suficientes fuentes referenciales tanto académicas (IEEE, SFIA, etc.), referencias regionales (Plan Desarrollo UNAD, COMPES 3975, etc.) como referencias laborales (GARTNER, US Ocupational Outlook Handbook, etc.) con el fin de que la o las personas designadas como tomadores de decisiones por la ECBTI, tengan material suficiente para tomar decisiones, para la denominación del programa doctoral.

Para el ejercicio de este proyecto, se tomará únicamente referentes académicos nacionales e internacionales, este modelo procesará la información de los marcos referenciales mencionados, de manera que pueda mostrar en un espacio dimensional, la información recopilada. Sirviendo así, como asistente para que los "tomadores de decisiones", teniendo la mayor cantidad de información asertiva.

III. Marco Teórico

Según el diccionario de la Real Academia Española define lenguaje como "facultad del ser humano de expresarse y comunicarse con los demás a través del sonido articulado o de otros sistemas de signos". La mayor parte de la población mundial habla por lo menos un idioma, entre los mas comunes se tienen inglés, mandarín, español y portugués. Sin embargo, se conocen que existen mas de 6500 diferentes tipos de lenguaje en todo el mundo [1].

En cuanto a las computadoras y sistemas sucede exactamente lo mismo. El lenguaje que habla el computador A debe ser el mismo que habla el computador B, para que la información trasmitida sea entendible por el receptor. Para que una computadora o sistema, pueda entender el lenguaje humano, debe conocer las reglas del idioma que la persona habla. Este es el primer problema que el computador enfrenta, porque dentro del mismo idioma dependiendo de la región, una palabra puede significar algo, que en otra región que hablen el mismo idioma, esto se le denomina ambigüedades lingüísticas [2].

Con el fin de que una computadora entienda el lenguaje humano, se debe referir a NLP (por sus siglas en ingles, Procesamiento de Lenguaje Natural). Éste está compuesto de dos conjuntos principales, el NLU (por sus siglas en ingles, Entendimiento del Lenguaje Natural) y el NLG (por sus siglas en ingles, Generación del Lenguaje Natural), en otras palabras, es decir, el que envía y el que recibe. NLG puede decirse que es el mas sencillo, este genera la información en texto o voz, realizando una oración coherente y comprensible. A diferencia del NLU, debe ser capaz de mapear la data recibida entendiendo aun las ambigüedades léxicas, semánticas y referenciales [3].

Fig. 1 Explicación grafica NLP [6]

El uso del NLP es muy variado, y sus aplicaciones son muy amplias, entre estas se puede tener [4]:

1) Análisis sentimental: los famosos "me gusta", "me divierte" de las redes sociales.
2) Reconocimiento de voz: asistentes de voz como Siri de Apple® o Alexa de Amazon®.
3) Chatbots: respuestas predeterminadas a preguntas frecuentes.
4) Traducción en vivo de conversaciones.
5) Corrección de ortografía: como lo realiza Microsoft® Word o ciertas aplicaciones de mensajería instantánea.
6) Buscadores de palabras: en los principales motores de búsqueda.
7) Anuncios Publicitarios: usualmente cuando se busca algo, puede salir minutos después como sugerido en redes sociales o por correo.

Existen ciertas formulas que son muy utilizadas por NLP para la comparación de textos y palabras una vez que son procedas al nivel donde puede ser discriminadas en diferentes IDs o tokens. Entre varias de ellas, se tienen las siguientes mas importantes:

A. *Similadirdad de Jaccard o Intersección de Jaccard*

Es la intersección de dos textos o documentos, dividido sobre la unión de estos. Como resultado, se tiene un valor entre 0 a 1, donde 1 son dos textos idénticos y 0 son dos textos completamente diferentes. La formula (1) que describe esta intercesión es la siguiente:

$$J(doc1, doc2) = \frac{doc1 \cap doc2}{doc1 \cup doc2} \qquad (1)$$

B. *Similaridad de Coseno*

Una formula (2) un poco diferente y esta va relacionada con la graficación de la información en n-dimensiones si se quiere

$$Similarity = \cos(\emptyset) = \frac{A * B}{\|A\| * \|B\|} = \frac{\sum_{i=1}^{n} AiBi}{\sqrt{\sum_{i=1}^{n} Ai^2} * \sqrt{\sum_{i=1}^{n} Bi^2}} \qquad (2)$$

IV. Método

La metodología para utilizar en el proyecto es la bien conocida y llamada "Investigación Acción". Esto con el objetivo de tener investigación continua, al tiempo de que la investigación pueda ser ejecutado y/o probada al mismo tiempo.

En términos generales, esta metodología cuanta con un ciclo de investigación la cual es justo lo que se requiere para el objetivo del proyecto. Dentro de esta metodología se tiene:

1) Componente de Planificación: donde se planifica una ruta ejecución
2) Componente de Acción: donde se ejecuta y prueba la planificación
3) Componente de Evaluación: donde se revisa los resultados objetivos, versus los resultados esperados

4) Componente de Replanificación: donde se ajusta las variables, y se vuelve a planear.

Con el fin de desarrollar esta metodología, para el método propuesto del programa asistido por computador se tendrán los siguientes pasos a desarrollar.

A. Paso 1: Definición Ámbito Nacional

Definir las primeras 20 universidades nacionales que hayan ocupado los primeros puntos del ranking nacional, durante los últimos 10 años en Colombia. Los criterios de selección serán:

1) Que tengan programas de doctorado orientado hacia la ingeniería de sistemas.

2) Nombres que aparecen ya aprobados en el SNIES (base de datos del Ministerio Nacional de Educación de Colombia) en las universidades ya aprobadas.

3) Relevancia nacional.

B. Paso 2: Definición Ámbito Internacional

Definir las primeras 20 universidades internacionales que hayan ocupado los primeros puestos del ranking latinoamericano, durante los últimos 10 años. Los criterios de selección serán:

1) Que tengan programas de doctorado orientado hacia la ingeniería de sistemas.

2) Ranking de Shangai.

C. Paso 3: Obtener Datos Nacionales e Internacionales

Datos para obtener del programa doctoral de cada universidad nacional e internacional

1) Nombre completo del programa

2) Descripción completa del programa

3) Líneas de investigación del programa

D. Paso 4: Preparación de Información

Una vez la información haya sido recolectada, la información debe presentarse en el formato correcto para que pueda ser leída por el algoritmo. Como resultado deberían obtenerse tres tipos de archivos en Excel, uno para nombre del programa, otro para descripción del programa y otro para líneas de investigación.

1) ID: Id dentro del archivo, debe ser único

2) Universidad: Nombre o abreviación del nombre de la Universidad

3) Tipo: Si la universidad es nacional o internacional

4) Texto: Aquí debe contener la información mas relevante por cada archivo, y debe contener en primer lugar el nombre del programa, descripción del programa y finalmente las líneas de investigación del programa.

Es importante recordar los nombres de los archivos en tipo Excel. Estos deben mencionarse en la ruta dentro del algoritmo para que sean leídos.

E. Paso 5: Revisión algoritmo NLP

Revisar y correr el algoritmo para cada tipo de los archivos obtenidos, el algoritmo comparara los datos y generara graficas.

1) Revisar que cada vez que se corra el algoritmo, la ubicación del archivo y el nombre del archivo este correctamente referenciado en el algoritmo

2) Se obtendrán tres archivos diferentes en formato HTML, para lectura desde su lector de HTML preferido o desde su navegador de preferencia.

F. Paso 6: Interpretación de graficas obtenidas

Una guía de como leer las graficas. Y como interpretar, además de sugerencias. Las posibles salidas que se tendrán son las siguientes:

1) Si el enfoque del nombre que se le quiere denominar a un programa doctoral quiere mostrar una completa relación entre el área nacional e internacional, se sugiere utilizar las palabras del costado superior derecho de cada archivo.

2) Si el enfoque del nombre que se le quiere denominar a un programa doctoral quiere mostrar un profundo objetivo de competencia internacional, se sugiere usar las palabras del costado superior izquierdo

3) Si el enfoque del nombre que se le quiere denominar a un programa doctoral quiere mostrar un profundo objetivo de competencia nacional, se sugiere utilizar el costado inferior derecho.

La siguiente imagen evidencia el proceso que se describió para el desarrollo del método asistido por computador:

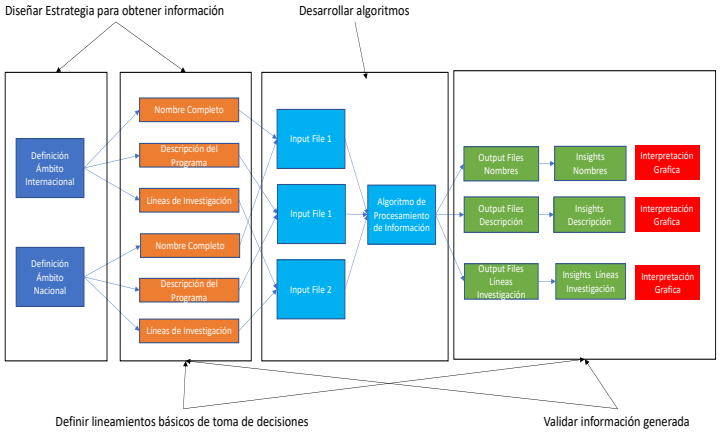

Fig. 2 Flujo del desarrollo

V. RESULTADO

Tal como se mencionó, se deben obtener tres graficas diferentes de los tres archivos procesados por el algoritmo que ya se ejecuto a este punto. Con el fin de ejemplificar los resultados se describe el resultado de solo uno de los archivos, en este caso, para continuar con el ejercicio se utiliza el archivo generado sobre los nombres de los programas académicos doctorales nacionales e internacionales.

Fig. 3 Resultado Ejemplo Nombres de Programas doctorales de Universidades Nacionales e Internacionales

Las siguientes consideraciones se toman, teniendo en cuenta que el eje X representa que tan frecuente o infrecuente se utilizan palabras en el aspecto nacional, siendo mas infrecuente cuando X se acerca a 0 y mas frecuente cuando X se aleja de 0. Para el eje Y, se representa que tan frecuente o infrecuente se utilizan palabras en el aspecto internacional, siendo mas infrecuente cuando Y se acerca a 0 y mas frecuente cuando Y se aleja de 0.

Para la interpretación de la grafica, se divide esta grafica en cuatro cuadrantes a lo largo del eje X y el eje Y. Adicionalmente, se traza una línea central x-distante entre el eje X y el eje Y, que se denominara coincidencias generales. La división se realizará de la siguiente manera, según se observa en la siguiente figura:

En el escenario completo, debería generarse tres graficas como estas cada una conteniendo nombre del programa,

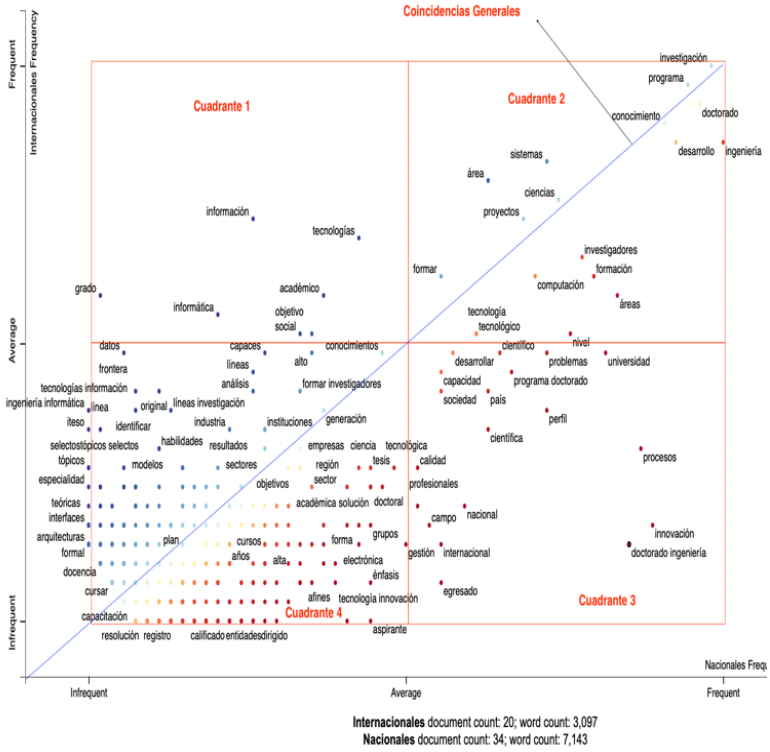

Fig. 4 Resultado divido en cuadrantes principales

A. *Coincidencias Generales*

Sobre la línea central denominada "Coincidencias Generales", están todas aquellas palabras que son mas frecuentes o que se utilizan de manera similar entre los dos puntos de comparación. Trayéndolo al contexto del proyecto, esto quiere decir que estas palabras que se encuentran en esta franja son palabras que se usan de manera simular tanto en el aspecto nacional como en el internacional, a razón de la misma frecuencia.

Siendo un poco mas especifico, las palabras que interceptan en el cuadrante 2 sobre la línea coincidencias generales, son palabras que tanto en el aspecto internacional como en el nacional se utilizan mucho para describir programas doctorales. Y las palabras que interceptan el cuadrante 4 sobre la línea coincidencias generales, son palabras que ni en el aspecto internacional ni en el aspecto nacional son muy usadas [5].

B. *Cuadrante 1*

El cuadrante 1 está compuesto en el eje X entre "infrecuente" y "promedio", y en el eje Y entre "promedio" y "frecuente". Las palabras que caigan sobre este cuadrante quieren decir que son palabras muy usadas en la descripción

de programas doctorales a nivel internacional, pero en el aspecto nacional son palabras que poco se usan.

C. *Cuadrante 2*

El cuadrante 2 está compuesto en el eje X entre "promedio" y "frecuente", y en el eje Y entre "promedio" y "frecuente". Las palabras que estén sobre este cuadrante quieren decir que son palabras muy usadas, tanto en el aspecto nacional como en el internacional, a la hora de describir programas doctorales

D. *Cuadrante 3*

Este cuadrante está compuesto en el eje X entre "promedio" y "frecuente, y en su eje Y entre "infrecuente" y promedio. Las palabras que estén sobre este cuadrante quieren decir que son palabras muy usadas en el aspecto nacional pero no muy usadas en el aspecto internacional, a la hora de describir programas doctorales.

E. *Cuadrante 4*

Este cuadrante está compuesto en el eje X entre "infrecuente" y "promedio, y en el eje Y entre "infrecuente" y "promedio". Aquellas palabras que estén sobre este cuadrante quieren decir que son palabras muy poco usadas tanto en el aspecto internacional como en el nacional, al momento de describir programas académicos doctorales.

VI. Conclusiones

El modelo que se presenta en este trabajo es un modelo asistido por computadora a través algoritmos de la tecnología de las librerías de NLP, el cual claramente puede ser tomado únicamente como una perspectiva o indicio de como potencialmente podría llamarse un programa académico doctoral en ingeniería, basado en la información suministrada. Sin embargo, se resalta que la calidad de la información obtenida es directamente proporcional a la calidad de la información suministrada como base de la información de otros programas académicos doctorales en ingeniería existentes actualmente en el ámbito nacional e internacional.

Se destaca también, que este modelo debe ser tomado en cuenta únicamente como una perspectiva mas para los tomadores de decisiones de las instituciones al momento de considerar la creación de un programa académico doctoral. Si bien la información puede ser muy útil, la denominación final del programa debe ser responsabilidad total de las personas encargadas por las instituciones para dicha labor.

El modelo puede proporcionales datos importantes como los visto en las imágenes anteriores, estos datos comparativos pueden dar ideas del tipo de programa que una institución quiera crear. Por ejemplo, si el programa que desea ser creado vaya mas orientado sobre el enfoque que las instituciones internacionales están otorgando, o si desea ser mas sobre el rasgo nacional. O incluso si desea ser muy innovador y usar nombres que las instituciones nacional e internacional poco usan.

Finalmente, el modelo puede ser utilizado no únicamente con fuentes académicas, sino como se menciono al comienzo de este articulo también pueden tenerse referentes laborales e industriales, si lo que se busca es tener un programa académico doctoral con características que la industria realiza hoy en día.

Agradecimientos

Agradecimientos a la Universidad Nacional Abierta y a Distancia, especialmente a la Escuela de Ciencias Básicas, Tecnología e Ingeniería, y al programa de Maestría en Gestión de Tecnología de la Información.

Referencias

[1] lenguaje | Diccionario de la lengua española. (s. f.-b). «Diccionario de la lengua española» - Edición del Tricentenario. https://dle.rae.es/lenguaje

[2] Goldberg, Yoav. (2017). Neural Network Methods in Natural Language Processing (Synthesis Lectures on Human Language Technologies). Page 1

[3] What is Natural Language Understanding? | Sisense. (s. f.). Sisense. https://www.sisense.com/glossary/natural-language-understanding/

[4] Natural Language Processing (NLP) Simplified : A Step-by-step Guide. (s. f.). Data Science Articles and Whitepapers | Data Science Awards | Data Science Consultancy. https://datascience.foundation/sciencewhitepaper/natural-language-processing-nlp-simplified-a-step-by-step-guide

[5] Kessler, J. S. (2017). Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. Proceedings of ACL 2017, System Demonstrations.

[6] C. K. GN. "NLP vs NLU vs NLG (Know what you are trying to achieve) NLP engine (Part-1)". Medium. https://towardsdatascience.com/nlp-vs-nlu-vs-nlg-know-what-you-are-trying-to-achieve-nlp-engine-part-1-1487a2c8b696 (accedido el 12 de marzo de 2022).