MODELO BASADO EN *NATURAL LANGUAGE PROCESSING* (NLP) PARA EL DISEÑO DE PROGRAMAS ACADÉMICOS ASISTIDO POR COMPUTADOR FACTOR 1: DENOMINACIÓN DEL PROGRAMA

MODEL BASED ON NATURAL LANGUAGE PROCESSING (NLP) FOR THE DESIGN OF ACADEMIC PROGRAMS AIDED BY COMPUTER FACTOR 1: NAME OF THE PROGRAM

Diego F. Calero[1]

Darío J. Delgado[2]

Universidad Nacional Abierta y a Distancia —UNAD—

Resumen

El proyecto se basa en un modelo en NLP (Natural Language Processing, o por sus siglas en español Procesamiento Natural de Lenguaje), el cual es una rama de Machine Learning (aprendizaje de máquinas) para el diseño de programas académicos doctorales en el marco de programas de educación superior, asistido por computador para la Escuela de Ciencias Básicas, Tecnología e Ingeniería (ECBT) de la Universidad Nacional Abierta y a Distancia. Este proyecto contempla la asistencia en la denominación del programa, es decir, su nombre. Como resultado del proyecto se tendrá un modelo asistido por computador como a manera de ideas (o "insights", en inglés) de cuáles serían los posibles nombres o títulos candidatos para nuevos programas académicos doctorales en el área de la educación superior de la Universidad.

Palabras clave: Machine Learning, NLP, doctorados, programas académicos.

Abstract

The project is based on a model in NLP (Natural Language Processing, or by its acronym in Spanish Procesamiento Natural de Lenguaje), which is a branch of Machine Learning (machine learning) for the design of doctoral academic programs within the framework of computer-assisted higher education programs for the School of Basic Sciences, Technology and Engineering (ECBT) of the National Open and Distance University. This project includes assistance in the denomination of the program, that is, its name. As a result of the project, there will be a computer-assisted model as ideas (or "insights", in English) of what would be the possible names or candidate titles for new doctoral academic programs in the area of higher education at the University.

Keywords: Machine Learning, NLP, doctorates, academic programs.

1. Introducción

El proyecto se enmarca en las mejores prácticas para el desarrollo y creación de nuevos programas académicos que la Universidad pueda ofertar en el mediano y largo plazo. Como es de conocimiento general, la educación va a avanzando y evolucionando según factores investigativos y de la industria. Y, la Universidad en su camino a la excelencia debe siempre estar a la vanguardia en educación respecto a lo que esta pasando en el mundo. Es por esto, que existe la necesidad de que la Universidad contemple y tenga un RoadMap definido sobre sus nuevos cursos de doctorados y posdoctorados que potencialmente pueda liberar en un futuro después de realizar un exhaustivo estudio y aprobaciones por los entes territoriales.

Principalmente se espera que, basado en ciertas definiciones iniciales, a través de este modelo la Universidad pueda tener una herramienta con la cual pueda identificar cuales serian los potenciales nombres para sus programas académicos que desee liberar.

2. Problemática

El planteamiento del problema deriva puntalmente del requerimiento en la propuesta de investigación (PIE) presentada por La Escuela de Ciencias Básicas, Tecnología e Ingeniería (ECBTI) de la Universidad Nacional Abierta y a Distancia, la cual tiene como título de la propuesta "Estudio de viabilidad para la creación de programas de doctorado en ingeniera bajo la modalidad virtual y a distancia en Colombia".

[1] dfcalerov@unadvirtual.edu.co
[2] dario.delgado@unad.edu.co

La propuesta de investigación objetivo de este proyecto de grado busca como fin, contribuir directamente con algunos de los puntos fundamentales de la propuesta de investigación (PIE) de la Universidad. Este se trata del desarrollo de un modelo basado en procesamiento natural de lenguaje (o NLP, por sus siglas en inglés Natural Lenguage processing) para el diseño de programas académicos asistidos por computador, el alcance de este proyecto será el factor 1: denominación del programa.

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

Por otra parte, 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" tengan la mayor cantidad de información asertiva.

3. Marco teórico

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 más comunes se tienen inglés, mandarín, español y portugués. Sin embargo, se conocen que existen mas de 6.500 diferentes tipos de lenguaje en todo el mundo (Díaz-Verdejo, 2013; RAE, s.f.).

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 (Goldberg, 2017).

Con el fin de que una computadora entienda el lenguaje humano, se debe referir a NLP (Natural Lenguage Processing). Éste está compuesto de dos conjuntos principales, el NLU (Natural Language Understanding) y el NLG (Natural Language Generation), en otras palabras, es decir, el que envía y el que recibe. NLG puede decirse que es el más 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 (Sisense, s.f.).

El uso del NLP es muy variado, y sus aplicaciones son muy amplias, entre estas se puede tener (NLP, s.f.):

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

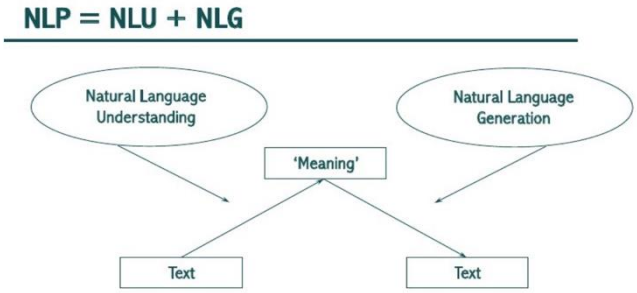


Fig. 1. Explicación NLP

Similaridad 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

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

Similaridad de Coseno

una fórmula un poco diferente y esta va relacionada con la graficación de la información en n-dimensiones si se quiere

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

4. Metodología

La metodología para utilizar es la llamada "investigación acción", o como se denomina en inglés "action-research". 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 cuenta 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 Re planificación: donde se ajusta las variables, y se vuelve a planear.

4.1 Pasos

1. Definir las primeras 20 universidades nacionales que hayan ocupado los primeros puntos del ranking nacional, durante los últimos 10 años.
a. Que tengan programas de doctorado orientado hacia la ingeniería de sistemas.
b. Nombres que aparecen ya aprobados en el SNIES en las universidades ya aprobadas.
c. Porcentaje adicional a estas.

2. Definir las primeras 20 universidades internacionales que hayan ocupado los primeros puestos del ranking latinoamericano, durante los últimos 10 años.
a. Que tengan programas de doctorado orientado hacia la ingeniería de sistemas.
b. Ranking de Shangai.

3. Datos para obtener del programa doctoral de cada universidad nacional e internacional.
a. Nombre completo del programa.
b. Descripción completa del programa.
c. Líneas de investigación del programa.

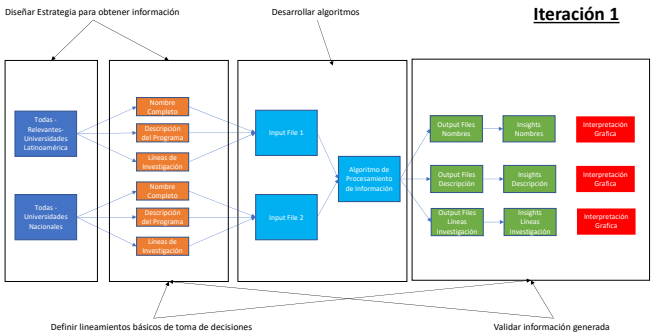
4. Correr el algoritmo para cada tipo de dato obtenido comparando universidades nacionales versus universidades internacionales.
a. Se obtendrían 3 archivos.

5. Una guía de como leer las gráficas. Y como interpretar, además de sugerencias.

6. Posibles salidas.
a. 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.
b. 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.
c. 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.

Lo anterior se ve representado por las iteraciones siguientes:

Iteración 1:



Iteración 2:

Fig. 2. Detalle del resultado.

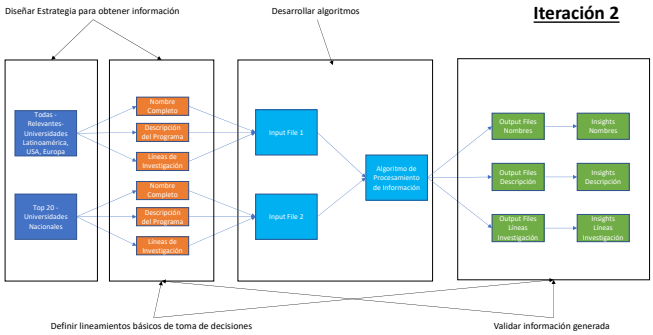
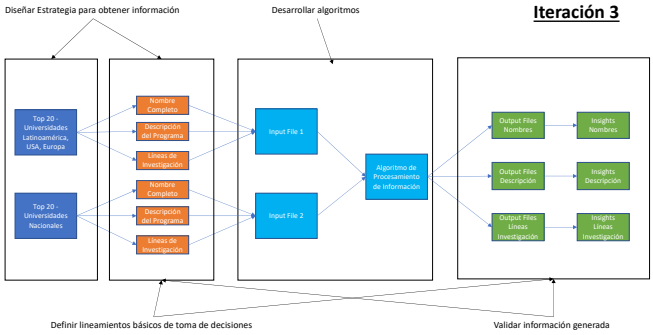


Fig. 3. Detalle del resultado.

Iteración 3:

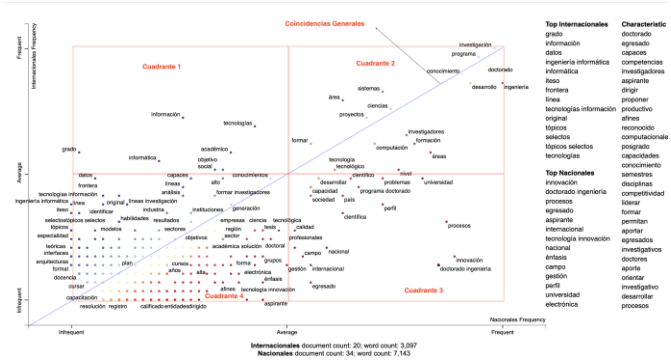


5. Resultado

Resultado de la iteración 2, se vería algo como el siguiente gráfico:

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 más infrecuente cuando X se acerca a 0 y más 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 más infrecuente cuando Y se acerca a 0 y mas frecuente cuando Y se aleja de 0.

Fig. 5. Detalle del resultado.



5.1 Coincidencias generales:

Sobre la línea central denominada “Coincidencias generales”, están todas aquellas palabras que son más frecuentes o que se utilizan de manera similar entre los dos puntos de comparación. Trayéndolo al contexto del proyecto, esto quiere decir que estas

Fig. 6. Detalle del resultado.

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 más específico, 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 (Kessler, 2017).

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.

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

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.

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.

6. Conclusiones

El seguimiento de las normas indicadas permitirá que su trabajo resulte visualmente atractivo. Esta misma plantilla se puede encontrar en formato LaTeX, en la dirección *web* oficial de las jornadas (http://www.jitel.org).

Referencias

Díaz-Verdejo, J. (2013). Ejemplo de bibliografía. *Actas de las XI Jornadas de Ingeniería Telemática, 1*(1), 1-5.

Goldberg, Y. (2017). *Neural Network Methods in Natural Language Processing (*Synthesis Lectures on Human Language Technologies).

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

Natural Language Processing (NLP) (s. f.). *Simplified: A Step-by-step Guide.* 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

RAE (s.f.). Lenguaje. *Diccionario de la lengua española.* https://dle.rae.es/lenguaje

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