REVISTA INCAING
ISSN 2448 9131

Propuesta de aplicativo web para el desarrollo de perfiles ocupacionales en el entorno educativo superior, utilizando técnicas para el procesamiento de lenguaje natural (Abril de 2021)

Laura C. Riaño, Juan S. Castellón, Cesar Y. Barahona, miembros Universidad de Cundinamarca

Resumen: Tomando en conjunto el ámbito educativo y ocupacional, en un país como Colombia, en el cual el 22,1% de los asalariados desean cambiar de empleo, según un informe de observatorio laboral de la Universidad del Rosario, es válido afirmar que en su mayoría se debe a la inconformidad de los trabajadores de acuerdo a su trabajo, ya que no se desempeñan en un ámbito que se adecue a sus conocimientos; esto porque en gran parte su desarrollo educativo suele ser diferente a la oferta laboral. Es por esto que se pretende desarrollar un aplicativo web, con el fin de poder realizar una búsqueda dependiendo de las necesidades de diferentes empresas en cuanto a su oferta laboral, que tenga un mayor porcentaje de coincidencia con los perfiles ocupacionales de distintas universidades. Incorporando de igual forma un control sobre los perfiles que se inscribirán y los usuarios que tendrán acceso a dicha plataforma. Todo esto haciendo uso de técnicas de procesamiento de lenguaje natural, mediante las cuales se realizará un proceso de tokenización y stemming como metodologías principales, de igual forma se utilizarán técnicas NLP (procesamiento de lenguaje natural) que permitan un mejor procesamiento, tales como, lematización, identificación de entidades y anotaciones morfosintácticas, todo esto bajo el análisis básico de Kinosearch, que nos permitirá realizar unas serie de acciones para identificar y unificar sinonimias y eliminar stopwords.

Índice de Términos – Perfil ocupacional, procesamiento de lenguaje natural, sinonimia semántica.

I. INTRODUCCIÓN

La capacidad de ejercer profesionalmente en un oficio que se adapte a nuestros conocimientos y nuestro perfil ocupacional resulta ser más baja de lo pensado, teniendo en cuenta que la oferta laboral resulta ser diferente a la demanda; por ende resulta primordial entender y analizar dicha brecha, pues actualmente la tasa de desempleo es Colombia es de 16,8%, en esta cifra también se abarca los profesionales que no han podido acceder a un trabajo formal y decente y esto también es motivo del estancamiento para el crecimiento y desarrollo del país , teniendo en cuenta que Colombia tiene la segunda menor productividad laboral de la región según lo demuestra el Consejo Privado de Competitividad en su más reciente informe, Colombia ocupa el puesto 11 de 13 países evaluados en América Latina. Además de esto se le suma la dificultad que tienen los empleados para ocupar vacantes, donde el país ocupa el quinto puesto, entre

Documento recibido el 9 de octubre de 2001. (Anote la fecha en que usted presentó su documento para su revisión.) Este trabajo fue apoyado en parte por los U.S. Department of Commerce under Grant S123456 (reconocimiento al patrocinador y apoyo financiero va aquí). los títulos del Documento deben ser escritos en letras mayúsculas y minúsculas, no todas las mayúsculas. Evite escribir fórmulas extensas con subíndices en el título; Utilice Fórmulas cortas que identifiquen los elementos (por ejemplo, "Nd-Fe-B"). No escriba "(invitados)" en el título. Escriba los Nombres completos de los autores en el campo autor, pero no es necesario. Ponga un espacio entre los autores.

F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (corresponding author to provide phone: 303-555-5555; fax: 303-555-5555; e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar. colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

ocho países de la región. Sin embargo, esto es el resultado de las bajas capacidades tecnológicas y profesionales del mercado laboral colombiano. Para el rector de la Universidad del Rosario, José Manuel Restrepo, esta situación limita el desempeño económico del país, a diferencia de otros países de la región, donde el crecimiento se da de la mano con la formación laboral. Teniendo en cuenta dicho compartimiento, en el que se maneja indicadores cualitativos y cuantitativos, se logra deducir que a pesar de contar con estos resultados no se maneja ningún propósito u objetivo que ayude a mejorar dichas dificultades, para esto es necesario contar con la ayuda del desarrollo ingenieril para explorar más a fondo y tener resultados concisos para el mejoramiento y desarrollo del trabajo decente y digno.

De acuerdo con lo anterior, se ha decidido proponer la idea principal del proyecto, la cual se encamina en generar un análisis para las diferentes universidades, que sirva de estudio en el momento de renovar su pensum estudiantil, ya que éste debe basarse en las necesidades actuales del mercado, por lo tanto, lo que se busca en dicho proyecto es obtener estadísticas que sirvan para estas actividades basándose en las diferentes búsquedas profesionales realizadas por los usuarios.

II. MÉTODO

A. Metodología

Para el desarrollo del proyecto se implementa parte de la metodología ágil SCRUM, junto con el campo de estudio de procesamiento de lenguaje natural, en éste campo se realizan diferentes técnicas.
Para ahondar más en esto, se plantea organizar la explicación de las diferentes técnicas y tecnologías de forma que el lector pueda captar y agrupar una idea de los diferentes procesos que se proyectan para el desarrollo de dicha aplicación.

Profundizando en el proceso que se llevará a cabo, se debe entender qué es el procesamiento del lenguaje natural, éste es un campo de estudio que se enfoca en la compresión mediante un computador del lenguaje humano, acá se abarca parte de la ciencia de datos, inteligencia artificial y la lingüística.
En el procesamiento del lenguaje natural, los computadores analizan el lenguaje humano, y lo interpretan para que de esta forma sea más fácil utilizarlo.
También se puede resaltar que el proyecto se basará en la sinonimia semántica, que resulta ser la mayor cantidad de atributos compartidos entre dos o más conceptos al que hace referencia una palabra. [10]

Ahora, en el procesamiento de lenguaje natural, se puede utilizar diferentes técnicas y esto se logra separando varios procesos tales como:

- **Análisis y normalización:** Cuando se habla de un proceso como el análisis y la normalización de textos, se hace referencia a la elección de los términos que mejor pueden representar el contenido de las consultas y la transformación de los términos seleccionados con el objetivo de reducirlos a formas canónicas que faciliten su posterior uso.[5]
- **Búsqueda:** Éste proceso se encuentra encargado de buscar las semejanzas entre la consulta y los datos almacenados en la base de datos.

Los anteriores términos se encuentran como procesos principales puesto que, se necesitan como enfoque para desglosar los subprocesos que se desarrollaran dentro de estos para poder llevar a cabo un procesamiento de lenguaje de manera correcta y aproximada.
No obstante, los procesos subyacentes de los anteriormente nombrados son de vital importancia, porque son ellos los que realizarán el desarrollo del procesamiento. Para esto se procederá a exponer las diferentes técnicas que se han investigado para su posterior manejo en el desarrollo del proyecto. Se encuentran técnicas tales como:

- **Stemming**: éste método consiste en la reducción de una palabra a su *stem* o supuesta raíz mediante la eliminación de sus terminaciones, obteniendo de esta forma la posible partícula que contiene la semántica básica del concepto, así los términos a comparar se verían reducidos a una sola cadena permitiendo correspondencia entre los mismos. Si bien el objetivo principal del stemming es reducir las diferentes formas lingüísticas de una palabra a una forma común o raíz y de esta forma facilitar el acceso a la información durante el posterior proceso de búsqueda, paralelamente se está reduciendo el número de términos diferentes del sistema, entonces, se da como consecuencia una segunda reducción de los recursos de almacenamiento requeridos.
  Stemming se ha caracterizado por utilización de algoritmos clásicos entre los cuales se destacan el algoritmo de Porter, que es el algoritmo más popular, y el algoritmo de Lovins. En estos algoritmos se pueden encontrar dos fases: una fase de eliminación de sufijos en base a una lista prefijada de los mismos, y una fase de

recodificación de la cadena resultante de acuerdo a una serie de reglas. [9]

- **Tokenización**: En éste proceso se realiza el análisis léxico, o tokenización, el cual consiste en la conversión de una secuencia de caracteres en una secuencia de palabras candidatas a ser adoptadas como términos índice.
  Esto sirve para comprender el contexto o desarrollar el modelo para el procesamiento de lenguaje natural. La tokenización ayuda a interpretar el significado del texto al analizar la secuencia de las palabras. Para esto existen diferentes métodos y bibliotecas disponibles, tales como, NLTK, Gensim, Keras entre otras.
  La tokenización se puede realizar para separar palabras u oraciones. [8]

- **Lematización:** Esta consiste en relacionar una palabra flexionada o derivada con su forma canónica o lema, y un lema es básicamente la palabra en su expresión principal; se entiende que la lematización es un proceso lingüístico que consiste en tomar una palabra flexionada, es decir, una palabra en forma plural, conjugada o demás; de esta palabra hallar el lema correspondiente.
  El lema termina siendo una palabra que se encuentra como entrada en un diccionario.
  Resulta primordial la utilización de la lematización en el desarrollo del proyecto, puesto que, lematizar implica estandarizar, desambiguar, segmentar e incluso etiquetar, teniendo en cuenta que dicho proceso ayudará para desarrollar y sacar los lemas de ciertas palabras para su posterior desarrollo. Cabe resaltar que, la lematización también resulta ser una alternativa para el stemming ya que esta permite abordar los complejos fenómenos flexivos del lenguaje, pero se ha tomado la decisión de dejar ambos procesos teniendo en cuenta que ninguno asegurar un 100% de asertividad y lo que se busca en esto es acercarse lo más posible a la sinonimia. [6]

- **Tagging parto f speech o etiquetado gramatical:** El etiquetado gramatical (part-ofspeech tagging, POS tagging o POST) es el proceso que recibe como entrada un texto y devuelve como salida un conjunto de pares de la forma palabra-etiqueta gramatical, basado en su definición y su contexto.
  Resulta primordial su uso puesto que, las categorías gramaticales dan como consecuencia una gran utilidad por la amplia cantidad de información que dan acerca de una palabra.
  Entiéndase que, al saber si una palabra es un sustantivo o un verbo permite de manera más acertada la forma de interpretarla y sirve para encontrar entidades nombradas, que permite encontrarlas en texto y en otras tareas de extracción de información.
  El etiquetado gramatical resulta difícil de procesar porque una misma palabra puede representar dos o más categorías diferentes, por esto se debe realizar correctamente el etiquetado, que implicar marcar la palabra como un verbo en la primera oración y como sustantivo en la segunda.
  En la siguiente imagen, tomada de internet, se puede observar la ejemplificación del proceso de etiquetado de la frase “Enfermo grave de rabia”

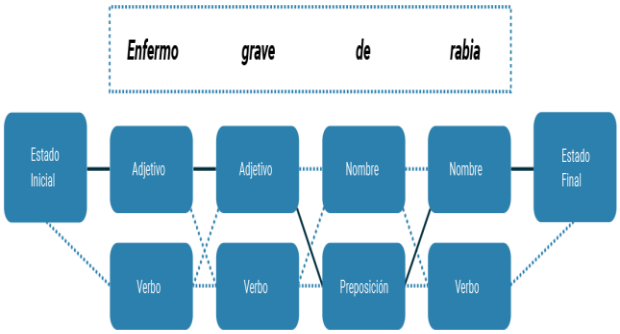

Fig. 1. Ilustración de etiquetado gramatical (part-of-speech taggin, POS tagging o POST) Recuperado de https://medium.com/soldai/etiquetado-gramatical-a418278e115c

El estado inicial indica donde comenzará el proceso de etiquetado. Al llegar al estado final, un etiquetado correcto habrá determinado que “enfermo” y “grave” son adjetivos y no formas conjugadas de los verbos “enfermar” y “gravar”, que “de” es una preposición y no el nombre de la letra D, y que “rabia” es un nombre y no el verbo “rabiar”.

Al finalizar esto, se continuará con otro proceso de procesamiento del lenguaje natural.

- **LShallow parsing / Chunks o análisis sintáctico superficial:** Consiste en recuperar la estructura sintáctica o árbol sintáctico asociado a cada oración. Los algoritmos que llevan a cabo el análisis global proporcionan la estructura asociada a la oración, cuando esta pertenece al lenguaje definido por una gramática. En caso contrario, cuando la oración no pertenece al lenguaje definido, el análisis falla.
  De igual forma se debe implementar la unificación, que es un proceso que combina la información obtenida. En el estudio de lenguaje natural, además del análisis sintáctico, se debe analizar el sentido, interpretación y significado de las palabras. La semántica de una palabra o frase puede capturarse mediante estructuras formales que cumplan una serie de características: verificables, sin ambigüedad, precisas y expresivas.[7]

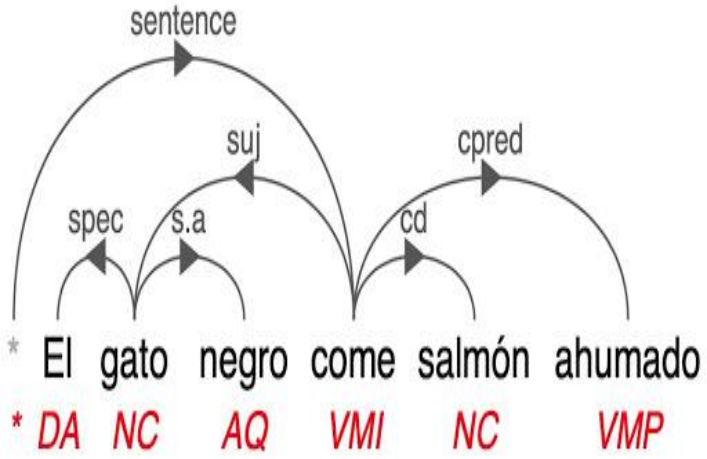

Fig. 2. Análisis sintáctico superficial / fragmentos (Shallow parsing / chunks). Recuperado de http://liceu.uab.cat/~joaquim/language_technology/NLP/PLN_analisis.html

- **Bag of words o bolsa de palabras:** es un método que se utiliza en el procesamiento de lenguaje natural para representar documentos ignorando el orden de las palabras. En éste modelo, cada documento parece una bolsa que contiene algunas palabras. En palabras más específicas, dicho método contiene algunas palabras del diccionario que se utilizarán posteriormente; y que presenta ventajas como su facilidad de uso y la eficiencia computacional.[2]

B. Propuesta de elaboración

Para el desarrollo de esta propuesta se ha decidido tomar en cuenta los diferentes procesos que se explicaron anteriormente, ya que cada uno de ellos aporta un valor fundamental para el desarrollo teniendo en cuenta que ninguno asegura por si solo un acercamiento total o parcial a la sinonimia semántica, por ende, el proceso que se ha estudiado y elegido para el análisis ha sido el siguiente:

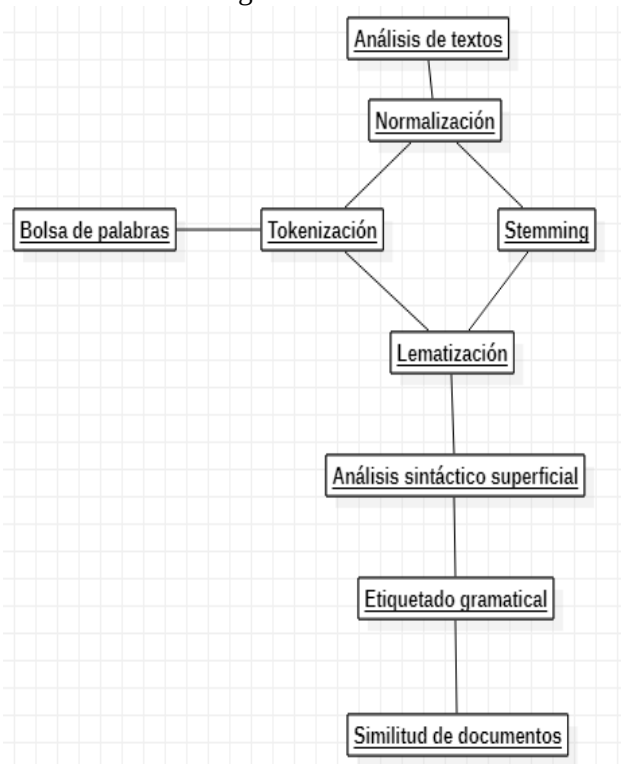

Fig 3. Desarrollo de procesamiento de lenguaje natural por pasos

Esto se concibe desarrollar gracias al uso de diferentes herramientas dadas por el lenguaje Python, en las cuales se ha decidido seguir con el funcionamiento de las técnicas de básicas de procesamiento de lenguaje natural, pero sumando también técnicas más refinadas de NLP.

Las herramientas a utilizar son:

- **NLTK**: Esta herramienta es un conjunto de bibliotecas y programas para el procesamiento de lenguaje natural simbólico y estadístico para el lenguaje de programación Python. NLTK incluye demostraciones gráficas y datos de muestra. NLTK está destinado a apoyar la investigación y la enseñanza en

procesamiento de lenguaje natural o áreas muy relacionadas.

De igual forma, esta librería también soporta Inteligencia Artificial y ya fue probada en diferentes proyectos entre ellos se encontró uno publicado en la revista de computo aplicado. [4]

- **Gensim:** Es un robusto kit de herramientas de modelado de vectores de código abierto y tópicos implementado en Python. Utiliza NumPu, SciPy y opcionalmente Cython para el rendimiento. Gensim está específicamente diseñado para manejar grandes colecciones de texto, utilizando el flujo de datos y algoritmos incrementales eficientes, que lo diferencian de la mayoría de los otros paquetes de software científicos que solo se enfocan en el procesamiento por lotes y en memoria.

  En la investigación realizada para esta herramienta se logró encontrar diferentes artículos en los cuales explican la forma en la que utilizaron e implementaron, entre ellos se puede nombrar un artículo de análisis de similitud en documentos de texto. [1]
- **Freeling**: Es una biblioteca que proporciona funcionalidades de análisis de idiomas (análisis morfológico, detección de entidades con nombre, etiquetado de PoS, análisis, desambiguación de sentido de palabra. Etiquetado de roles semánticos, etc) para una variedad de idiomas.

  Freeling también proporciona una interfaz de línea de comandos que se pueden utilizar para analizar textos y obtener la salida en el formato deseado, ya sea XML, JSON, CoNLL.

  Esta librería también está concebida como una librería sobre la cual se puedan desarrollar potentes aplicaciones de procesamiento de lenguaje natural, y orientado a facilitar la integración con las aplicaciones de niveles superiores de los servicios lingüísticos que ofrece.[3]
- **SpaCy**: Es una biblioteca de procesamiento de lenguaje natural Python diseñada específicamente con el objetivo de ser una biblioteca útil para implementar sistemas listos para producción. Es particularmente rápido e intuitivo, por lo que es un competidor superior para las tareas de procesamiento de lenguaje natural.

  SpaCy también devuelve un etiquetado POS completo, no solo indica la función de la palabra en la oración, sino otros datos tales como tiempo verbal, persona, número, modo, género, entre otros.[8]

- **Textacy**: Es una biblioteca de Python para realizar una variedad de tareas de procesamiento de lenguaje natural, construida sobre la biblioteca spaCy de alto rendimiento. Con los fundamentos de tokenización, etiquetado de parte del discurso, análisis de dependencia, etc. Principalmente se ha decidido su utilización para el uso de texto descargado directamente desde la web.[3]

La estructura de la propuesta anteriormente mencionada se logra evidenciar en la siguiente tabla:

TÉCNICA	HERRAMIENTA	APORTE
Tokenización	NLTK – Gensim	En la tokenización, se toma como primer paso el análisis de un texto utilizando ciertos módulos de la biblioteca, teniendo ya dicho texto se puede utilizar el módulo BeautifulSoup para limpiar el texto capturado. Posteriormente se separa en tokens el texto y así se calcula la distribución de frecuencia de esos tokens para su uso.
Normalización	NLTK	En la normalización se puede poner en igualdad de condiciones todo el texto, eliminando la puntuación, los números equivalentes de palabras, etc.
Stemming	NLTK – Snowball	Para esto se utiliza

		Snowball que es un pequeño lenguaje de procesamiento implementado en ANSI C para la creación y uso de algoritmos de stemming, también se dispone de PyStemmer que es un wrapper de Snowball para Python.
Lematización	Spacy – Freeling	En la lematización, después de realizar un barrido utilizando Freeling, teniendo una lista de lemas completa, el proceso de lematización toma en consideración la probable clase de palabra (adjetivo, verbo, sustantivo, etc).
Tagging parto f speech o etiquetado gramatical	Spacy	Se realiza mediante funciones de NLTK para etiquetar morfológicamente una oración o texto y especificando un token, de esta forma calcula frecuencias y generaliza una categoría gramatical para dicho token.
Shallow parsing / chunks o análisis sintáctico superficial	Spacy	Éste análisis sintáctico busca localizar un segmento o grupo de palabras para identificar su clase.
Bag of words o bolsa de palabras	Spacy	Para realizar éste proceso se toman dos mensajes, de los cuales se extraen todas las palabras presentes en ambos mensajes, se eliminan los símbolos, y se genera un conjunto de palabras presentes en cada mensaje, luego cada mensaje puede ser representado utilizando dicha bolsa de palabras.
Similitud de documentos	Gensim - Tensorflow	Estas dos librerías ofrecen funciones fundamentales para la comparación en similitud de documentos.

Tabla I. Relación entre librerías, procesos y el aporte dado por cada proceso utilizando la librería dada.

CONCLUSIONES

Se puede concluir que, teniendo en cuenta todo lo que conlleva el procesamiento de lenguaje natural, es posible incorporar sus diferentes técnicas y tecnologías para llevar a cabo su funcionamiento optimo y eficaz, ya que cada una de ellas ofrece un proceso, solución y resultado distintos, que si se lograse unificar podría dar una aproximación a la sinonimia semántica.

RECONOCIMIENTO

Agradecimientos a la Universidad de Cundinamarca.

REFERENCES

[1] Belén, A., & Oscar, R. (2018). *Mediante Técnicas De Ciencia De Datos Basadas En Aprendizaje Profundo (Deep Learning)*. 246–250.

[2] Ernesto Sarmiento Torres, C., Diaz, N., & Vargas Cañas, R. (2020). Clasificación de noticias criminales basada en procesamiento del lenguaje natural y algoritmos de aprendizaje automático. *Revista Ibérica de Sistemas e Tecnologías de Información*, 117–129.

[3] Inform, S., & Polit, T. U. (n.d.). *Analizadores Multiling ¨ues en FreeLing*.

[4] Luis, V. J. (2020). *Cómputo Aplicado*. *4*, 44. https://www.ecorfan.org/spain/researchjournals/

Computo_Aplicado/vol4num13/Revista_de_Computo_Aplicado_V4_N13.pdf#page=32

[5] Matemática, F. De, Computación, F., & Bracco, A. (n.d.). *Normalización de Texto en Español de Argentina*. 0–67.

[6] Miranda, C. H., Guzmán, J., & Salcedo, D. (2016). Minería de opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hotels. *Procesamiento de Lenguaje Natural*, *56*, 25–32.

[7] Pla, F. (2000). *Etiquetado Léxico y Análisis Sintáctico Superficial basado en Modelos Estadísticos.*

[8] Talamé, L., Cardoso, A., & Amor, M. (2019). Comparación de herramientas de procesamiento de textos en español extraídos de una red social para Python. *XX Simposio Argentino de Inteligencia Artificial (ASAI 2019)-JAIIO 48 (Salta)*, *1*, 53–67. http://sedici.unlp.edu.ar/handle/10915/87854

[9] Watzlawik, M., & Valsiner, J. (2012). The Making of Magic: Cultural Constructions of the Mundane Supernatural. *The Oxford Handbook of Culture and Psychology*, *2*(6), 1930–1938. https://doi.org/10.1093/oxfordhb/9780195396430.013.0038

[10] Zapico, M., & Vivas, J. (2015). La sinonimia desde una perspectiva linguístico-cognitiva. Medición de la distancia semántica. *Onomazein*, *32*(2), 198–211. https://doi.org/10.7764/onomazein.32.11

Biografía Autor(es) Profesor Cesar Yesid Barahona Rodríguez, Facultad Ingeniería de Sistemas, Universidad de Cundinamarca, Magister en Sistemas Computacion, Colombia.
Estudiante Laura Camila Riaño Gamboa, Facultad Ingeniería de Sistemas, Universidad de Cundinamarca.
Estudiante Juan Sebastián Castellón Ramos, Facultad de Ingeniería de Sistemas, Universidad de Cundinamarca.