

Formato de inscripción proyecto de investigación

**CONSENTIMIENTO EXPRESO** Con el envío y de acuerdo con la Ley Estatutaria 1581 de 2012 de Protección de Datos y con el Decreto 1377 de 2013, los responsables del proyecto, autorizamos como Titulares de los datos que plasmamos, que éstos sean incorporados en una base de datos de responsabilidad de la Red Colombiana de semilleros de Investigación RedCOLSI, siendo tratados con la finalidad de gestión administrativa, formativa, evaluación y de información institucional, de semilleristas y de proyectos entre  otros, de conformidad con el aviso de privacidad publicado en [www.fundacionredcolsi.org](http://www.fundacionredcolsi.org/). De igual modo, los autores declaran haber sido informados que pueden ejercitar los derechos de acceso, corrección, supresión, revocación o reclamo por infracción sobre datos, mediante escrito dirigido a la dirección de correo electrónico coordinacion@fundacionredcolsi.org, indicando en el asunto el derecho que desea ejercitar, o mediante correo ordinario remitido a carrera 12 Número 53 – 42 Edificio Barcelona la Castellana Montería.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **INFORMACION GENERAL** | | | | |
| **País** | Colombia | | | |
| **Nodo** | Atlántico | | | |
| **Universidad** | Corporación Universitaria Americana | | | |
| **Nombre del Semillero** | Analítica de datos | | | |
| **Nivel de Formación** | Pregrado | | | |
| **Programa Académico** | Ingeniería De Sistemas | | | |
| **Título del Proyecto** | Algoritmo automatizado para determinar la dificultad de comprensión y legibilidad textual de Artículos científicos en inglés. | | | |
| **Autor (es)** | José Javier Torres Vanstrahlen | | | |
| **Identificación** | 1002129783 | | | |
| **Ponente (s)** | José Javier Torres Vanstrahlen | | | |
| **E-mail de Contacto** | torr  torresvanstrahlen@gmail.com | | | |
| **Teléfonos de Contacto** | 3013965056 | | | |
| **Categoría**  (seleccionar una) | Propuesta de  Investigación | Investigación  en Curso | | Investigación  Terminada |
| **Área de la investigación**  (Marque solo una opción) | Ciencias Agrarias  Biológicas y del mar  Ciencias de la Salud  Ciencias exactas y de la tierra  Ciencias humanas | | Ciencias sociales  Navales y de seguridad  Ingeniería  Lingüística artes y letras  Otra: (Mencione cuál) | |



Formato de inscripción proyecto de investigación

|  |
| --- |
| **CONTENIDO DE LA PROPUESTA DE INVESTIGACION** |
| 1. **TITULO:**   Algoritmo automatizado para determinar la dificultad de comprensión y legibilidad textual de Artículos científicos en inglés. |
| **2. INTRODUCCIÓN:**  La complejidad de un Documento o articulo científico determina el nivel de dificultad que tiene este, mediante ese nivel se puede dar a entender al lector si dicho artículo es difícil o fácil de leer y entender. En la presente investigación se establecieron unas métricas para la determinar si un documento presenta una comprensión (Muy fácil, fácil, bastante difícil, difícil, muy difícil). Utilizando diversas herramientas para el procesamiento de lenguaje natural en Python. |
| **3. PLANTEAMIENTO DEL PROBLEMA Y JUSTIFICACIÓN:**  De acuerdo con los resultados estadísticos de OECD para las pruebas PISA (2018), Colombia se encuentra entre los últimos lugares para la categoría de comprensión de lectura. Es decir que solo el 50 % de los evaluados alcanzó el nivel 2 de competencia lectora. En la comprensión de los textos escolares intervienen muchos factores. Las dificultades provienen no solo del lector, sino también del texto. Ya que el tipo de tipografía, vocabulario y estructura sintáctica, que contienen puede ser difícil de interpretar para el alumno.  La comprensión y legibilidad textual permiten que el lector pueda entender claramente el contenido del documento científico. Si un documento es difícil de leer y comprender, puede llevar a malentendidos, interpretaciones incorrectas y dificultades para aplicar los hallazgos en la práctica. Además, un documento científico que es fácil de leer y entender es más accesible a una audiencia más amplia, lo que puede aumentar su impacto y relevancia. Por la anterior mencionado, la idea de presentar una categorización de la dificultad para los diferentes artículos o textos que los estudiantes frecuentan para su formación ayuda a una mejor compresión de estos. El estudiante puede nivelarse y entender con mejor facilidad los textos que estén en un nivel que a este le favorezca. |
| **4. OBJETIVOS:**  **Objetivo General:** Determinar la dificultad de comprensión y legibilidad que poseen artículos científicos en idioma inglés.  **Objetivos Específicos:**   * Definir las métricas que para evaluar la dificultad de un artículo científico * Implementar bibliotecas de procesamiento de texto para clasificación, tokenización, derivación, etiquetado, análisis y razonamiento semántico. * Implementar métodos de procesamiento de lenguaje natural, para extraer el contenido del documento. |
| **5. REFERENTE TEORICO:**  **1.** Legibilidad del texto, métricas de complejidad y la importancia de las palabras. Este trabajo expone un estudio sobre la determinación de la edad recomendada de lectura sobre un conjunto de textos infantiles. Evaluando asi 12 medidas de complejidad propuestas por distintos autores.  **2.** Estudio de la complejidad del español para la simplificación textual. En este trabajo los autores se centran en estudiar y automatizar las métricas existentes para medir complejidad léxica para el español. Y la propuesta de una nueva métrica para medir la complejidad de manera multifacética. Estudio de la complejidad del español para la simplificación textual.  **3.** Natural Language Processing With Python's NLTK Package. Presenta un método de tokenizar e iniciar a desarrollar PLN con Python en un nivel básico. |
| **6. METODOLOGIA:**  Se implemento el lenguaje de programación abierto Python para el desarrollo del algoritmo, Python cuenta con herramientas de PLN como lo son NLTK, spaCy. Las cuales son excelentes herramientas para trabajar y enseñar lingüística computacional.  **Fase de Planeación.**  Se definieron las métricas para evaluar la compresión de los artículos. Se define que la densidad léxica equivale a que mayor número de palabras diferentes por texto mayor dificultad para la comprensión. Medir el número de palabras por oración, obteniendo así el índice de longitud oracional, y el número de frases complejas que hay por oración, a partir de un índice de frases complejas para obtener la complejidad de oración. El numero promedio de signos de puntuación se utiliza como uno de los indicadores de complejidad de este. El índice SRR, Se centra en medir el vocabulario y la estructura de oraciones para predecir la dificultad relativa de legibilidad de un texto. Un alto grado de cumplimiento de un documento determina que es un documento mas denso y con más contenido.  **Fase de Análisis.**  Las herramientas para utilizar para el análisis de texto y minería de texto ejercen una carga de tiempo y recursos a la hora de ejecutarse, mediante el análisis presente se automatizaron las funciones, ciclos y procedimientos del algoritmo para que resultara de manera óptima y sin ninguna Inter versión manual por parte del usuario. En esto se dedujo la utilización de funciones con el fin de generar recursividad y evitar los ciclos o secciones del código que manejan la parte mas repetitiva y pesada del algoritmo.  **Fase de Desarrollo.**  Desarrollo de técnicas de procesamiento de lenguaje natural para la extracción y minería de datos en el lenguaje de programación Python. La herramienta NLTK se utiliza el texto de el articulo y con ello podemos definir mediante distintos métodos; enumeración de palabras, longitud, signos, modificadores (adjetivos, adverbios), letras, oraciones, frases complejas. Se utilizo el paquete de recursos para idioma inglés de Spacy, para poder extraer y analizar el texto de los artículos. Esta es una canalización en inglés entrenada en texto web escrito (blogs, noticias, comentarios), que incluye vocabulario, sintaxis y entidades. |
| **7. RESULTADOS:**  Para un artículo (pdf), se utilizaron las librerías y herramientas ya mencionadas para extraer el texto y evaluarlo mediante diferentes técnicas de compresión. Para la Comprensibilidad; Densidad léxica, Complejidad de la oración, Complejidad sintáctica, Marcas de puntuación. Se establecieron 3 diferentes tipos de Métricas para evaluar La comprensión y legibilidad textual las cuales son; Métrica de Comprensibilidad, Legibilidad y Estructura de contenido. En estas se utilizaron técnicas de tokenizacion, extracción y separación de texto que proporciona NLTK. Entre otras técnicas y fórmulas de comprensión se encuentra la fórmula de Flesch-Kincaid.Para las Métrica Se evaluaron las siguientes Técnicas y se obtuvieron los siguientes resultados:  **Titulo Articulo:** A Review on Machine Learning and Deep Learning Perspectives of IDS.  **Abstract:** Internet of Things (IoT) is widely accepted technology in both industrial as well as academic field. The objective of IoT is  to combine the physical environment with the cyber world and create one big intelligent network. This technology has been  applied to various application domains such as developing smart home, smart cities, healthcare applications, wireless sensor  networks, cloud environment, enterprise network, web applications, and smart grid technologies. These wide emerging  applications in variety of domains raise many security issues such as protecting devices and network, attacks in IoT networks,  and managing resource-constrained IoT networks. To address the scalability and resource-constrained security issues, many  security solutions have been proposed for IoT such as web application firewalls and intrusion detection systems. In this paper,  a comprehensive survey on Intrusion Detection System (IDS) for IoT is presented for years 2015–2019. We have discussed  various IDS placement strategies and IDS analysis strategies in IoT architecture. The paper discusses various intrusions  in IoT, along with Machine Learning (ML) and Deep Learning (DL) techniques for detecting attacks in IoT networks. The  paper also discusses security issues and challenges in IoT.  **Comprensibilidad.**   * Densidad Léxica: **Nivel: Muy simplificado** * Complejidad Oración:  **Nivel: Muy simplificado** * Complejidad Sintáctica:  **Nivel: Muy simplificado** * Marcas puntuación:  **Nivel: Muy simplificado**   **Legibilidad.**   * índice SSR: **Nivel: Fácil** * Facilidad de Lectura: **Nivel: Difícil** * Árbol de Sintaxis Abstracta:  **Nivel: Muy simplificado** |
| **8. CONCLUSIONES:**  Los niveles de dificultad de comprensión de un documento se pueden obtener de diferentes técnicas lingüísticas.  En este caso se implemento más de una técnica utilizando herramientas de Procesamiento de Lenguaje Natural.  Se demuestra que un nivel de dificulta adecuado y categorizado es esencial para la ayuda y formación de los estudiantes, esto les ayuda a mantener la motivación a la hora de la comprensión lectora. |
| **9. BIBLIOGRAFIA:**   1. Rocío López-Anguita, Arturo Montejo-Raez Fernando J. Martínez-Santiago, Manuel Carlos Díaz-Galiano (2018). Legibilidad del texto, métricas de complejidad y la importancia de las palabras. <https://core.ac.uk/download/pdf/162130151.pdf> 2. Daniela Campos, Paula Contreras, Bernardo Riffo, Mónica Veliz (2014). Complejidad textual, lecturabilidad y rendimiento lector en una prueba de comprensión en escolares adolescentes.   <http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1657-92672014000300027>   1. Seobility Wiki. Frecuencia inversa de documento. <https://www.seobility.net/es/wiki/Frecuencia_inversa_de_documento> 2. Randall Araya-Camposa1, Paula Estrella, José Arguedas-Castillo, Walter Alvarez-Grijalba (2020). Estudio de la complejidad del español para la simplificación textual. <https://revistas.tec.ac.cr/index.php/tec_marcha/article/view/5478/5199> 3. Antonio Rodríguez Rosado (2021). Identificadores De Los Artículos Científicos. <https://isdfundacion.org/2021/12/03/identificadores-de-los-articulos-cientificos/> |