PROYECTO DE INVESTIGACION

Proyecto: Algoritmo automatizado para determinar la dificultad de comprensión y legibilidad textual de Artículos científicos en inglés.

Fecha Inscripcion De Proyecto: July 12, 2023

Palabras Claves: keywords, pln, software, investigacion, desarrollo

Facultad: Humanidades

Programa: Derecho

Grupo De Investigacion: Aglalia

Linea De Investigacion: Modelado y simulación

Semillero Investigacion: semilllero1

INVESTIGADORES PARTICIPANTES

Datos Del Investigador

Nombre: TRIANA MARTINEZ PEDRO AGUSTIN

Identificacion: 8781206

Telefono: 123

Correo: ptriana@coruniamericana.edu.co

Cvlac: https://cvlac.com

Datos del Co-Investigador

Nombre: Jose

Identificacion: 1002129783
Telefono: 3013965056
Correo: Jose@gmail.com
Formacion: Pasante
Institución en la que labora: CUA
Link CVLAC: https://cvlac.coinver.com

FECHAS DURACION PROYECTO

Fecha de inicio: July 12, 2023
Fecha de finalización: Oct. 20, 2023

DESCRIPCIÓN DEL PROYECTO

Problema y pregunta: 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.
Justificación: 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.

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.

Metodología propuesta: Fase de Planeación. Se define que la densidad léxica equivale a que mayor número de palabras diferentes por texto mayor dificultad. Medir el número de palabras por oración, obteniendo así el índice de longitud oracional, y el número de frases complejas que hay por oración para obtener la complejidad de oración. El numero promedio de signos de puntuación se utiliza como uno de los indicadores de complejidad. El índice SRR, Se centra en medir el vocabulario y la estructura de oraciones para predecir la dificultad relativa de legibilidad. 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 Interversión manual por parte del usuario. Fase de Desarrollo. Técnicas de PLN para la extracción y minería de datos en Python. La herramienta NLTK, utiliza el texto del articulo y con ello podemos definir mediante distintos métodos; enumeración de palabras, longitud, signos, modificadores, letras, oraciones, frases complejas. Se utilizo el paquete de recursos para idioma inglés de Spacy, Esta es una canalización en inglés entrenada en texto web escrito (blogs, noticias, comentarios), que incluye vocabulario, sintaxis y entidades.

Estado del arte o antecedentes: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 3. Seobility Wiki. Frecuencia inversa de documento. https://www.seobility.net/es/wiki/Frecuencia_inversa_de_documento 4. 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

Impactos esperados: Los niveles de dificultad de comprensión de un documento se pueden obtener de diferentes técnicas lingüísticas. En este caso se implementó 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.

CRONOGRAMA

Actividad	Descripción	Duración	Tipo de Duración
ACTIVIDAD 1	Definir las métricas que para evaluar la dificultad de un artículo científico	2	semanas
ACTIVIDAD 2	Implementar bibliotecas de procesamiento de texto para clasificación, tokenización, derivación, etiquetado, análisis y razonamiento semántico.	2	semanas

PRODUCTOS DERIVADOS DE LA INVESTIGACIÓN.

Generación de Nuevo Conocimiento	Desarrollo Tecnológico e Innovación	Apropiación Social del Conocimiento	Formación del Recurso Humano Ctel
Noticias Científicas	Producto Tecnológico Certificado o Validado	Participación ciudadana en CTel y creación	Dirección de Tesis de Doctorado Dirección de Trabajo de grado de Pregrado Apoyo a creación de programas o cursos de formación de Investigadores

COMITE DE ETICA

El proyecto ha sido sometido al Comité de Ética.
Descarga el acta en el Portal