Recomendador de Revistas Científicas mediante Ciencia de Datos

En la era actual, la explosión de publicaciones científicas presenta un desafío creciente para los investigadores. **Elegir la revista más adecuada para sus trabajos.**

Nuestro objetivo principal es diseñar un sistema que, a partir del texto de un artículo, recomiende automáticamente la revista científica más idónea.

Presentado por: Eduardo Ortega Zerpa

Datos utilizados: La base de nuestro sistema

Para entrenar y evaluar nuestro recomendador, hemos compilado un robusto conjunto de datos compuesto por **artículos científicos publicados entre 2020 y 2024** en destacadas revistas de Elsevier.

Consideramos cuatro revistas clave, cubriendo diversas áreas:

- Applied Ergonomics
- Expert Systems with Applications
- Journal of Visual Communication and Image Representation
- Neural Networks

Cada artículo procesado incluye los siguientes campos esenciales:

- **Título**
- **Resumen**
- **Palabras clave**
- **Año**
- **DOI**
- **Revista** (la etiqueta de clase para nuestro modelo)

El texto de entrada utilizado por los modelos se construye concatenando el título, el resumen y las palabras clave, creando un documento unificado por artículo.

Extracción automática y preparación de datos

Configuración Global

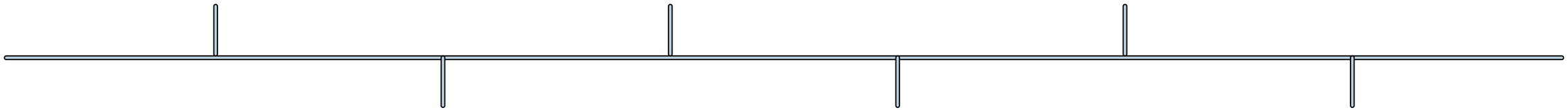
Establecimiento de API key, rangos de años y revistas objetivo.

Filtrado por DOI

Aseguramiento de que solo se procesan artículos de Elsevier.

Enriquecimiento con OpenAlex

Adición de palabras clave y conceptos para mayor contexto semántico.



Búsqueda en Scopus

Identificación de artículos relevantes por revista y año.

Recuperación de ScienceDirect

Obtención de detalles completos (resumen, título, etc.).

Normalización y Almacenamiento

Organización final de los datos en formato JSON, estructurados por revista y año.

Metodología de Clasificación: Dos Enfoques

Implementamos dos enfoques complementarios para la clasificación de artículos, combinando la eficiencia de los modelos clásicos con la capacidad semántica del Deep Learning.

1. Modelo Clásico: TF-IDF + SVM

- **Representación:** TF-IDF (Term Frequency-Inverse Document Frequency), para ponderar la importancia de las palabras.
- **Clasificador:** Linear SVM (Support Vector Machine).
- **Implementación:** Desarrollado con scikit-learn.
- **Ventajas:** Eficiente computacionalmente y muy robusto para vocabulario específico.

2. Modelo de Deep Learning: BERT

- **Modelo:** BERT (bert-base-uncased), un modelo de Transformers pre-entrenado.
- **Entrenamiento:** Fine-tuning con la librería Transformers de Hugging Face.
- **Características:** Tokenización subword y comprensión del contexto global del texto.
- **Capacidad:** Mayor riqueza semántica y mejor manejo de la ambigüedad.

El conjunto de datos se dividió de forma estratificada: **80% para entrenamiento** y **20% para test**, asegurando una distribución equitativa de las clases.

Resultados Principales: Comparativa y Rendimiento

La evaluación de ambos modelos reveló diferencias significativas en su capacidad para recomendar revistas científicas, destacando las fortalezas de cada enfoque.

Modelo TF-IDF + SVM

- **Rendimiento:** Muy buen desempeño en revistas con vocabulario altamente específico, logrando un **F1-score superior a 0.9** en áreas como "Applied Ergonomics" y "Expert Systems with Applications".
- **Desafíos:** Presentó dificultades en revistas con un solapamiento temático más pronunciado, donde la distinción semántica es más complicada.

Modelo BERT

- **Rendimiento:** Mostró una **mejora clara en todas las revistas**, con un incremento notable en aquellas con temática más generalista o interconectada, como "Journal of Visual Communication and Image Representation".
- **Ventaja:** Su mayor capacidad de generalización, gracias a la comprensión del contexto semántico.

Métricas de Evaluación

F1-score por revista

Métricas de Evaluación

Accuracy por revista

Métricas de Evaluación

Matriz de confusión normalizada

Conclusiones Finales

Sistema Automático Desarrollado

Hemos logrado construir un sistema efectivo de recomendación de revistas científicas.

Sistemas Híbridos

La combinación de enfoques es clave para la escalabilidad.

Eficiencia del Modelo Clásico

Rápido, eficiente y una excelente línea base para el problema.

Superioridad de BERT

Ofrece un mejor rendimiento global y generaliza en dominios complejos.

Desafío Persistente

El solapamiento semántico entre revistas sigue siendo una complejidad.