Universidad Nacional de Rosario

FACULTAD DE CIENCIAS EXACTAS, INGENIERIA Y AGRIMENSURA

Informe

Trabajo práctico 1

Materia: Procesamiento del lenguaje natural

Alumnos: Pablo Pistarelli, Max Eder

Web Scraping

El principal desafío en la etapa de web scraping fue identificar la ubicación exacta de cada dato dentro de las celdas HTML de la página web. Una vez resuelto este problema, el siguiente paso fue implementar las condiciones requeridas por el enunciado: recolectar información de al menos 100 libros y asegurarse de que hubiera un mínimo de 10 libros por género.

Los datos extraídos se volcaron en un dataset que posteriormente fue depurado. En esta fase de limpieza, se eliminaron las filas que no contenían una descripción del libro y se descartaron los registros duplicados.

Embeddings

Una vez conformado el dataset, generamos los embeddings de las reseñas utilizando la librería sentence-transformers y el modelo distiluse-base-multilingual-cased. Para cada reseña, creamos una nueva columna en el dataset donde almacenamos el vector generado por el modelo. Estos embeddings nos permitieron posteriormente comparar las reseñas con las sentencias ingresadas en el menú, facilitando así la recomendación de libros basada en similitud semántica.

Menú

El menú de recomendaciones interactúa con el usuario para ofrecer tres opciones principales de búsqueda de libros: recomendación directa, elección por autor y elección por género literario. Estas opciones permiten al usuario encontrar libros según sus preferencias, ya sea explorando nuevas temáticas, autores conocidos o géneros literarios específicos.

En la elección por autor o género literario, se realiza una normalización de las entradas del usuario y de los datos del dataset para asegurar una comparación precisa y evitar posibles discrepancias debido a diferencias en mayúsculas, minúsculas o caracteres especiales. Esto garantiza que la búsqueda sea efectiva y que se obtengan resultados relevantes.

La elección de utilizar la similitud del coseno y la distancia de Levenshtein en este contexto se basa en la naturaleza de los datos de texto. La similitud del coseno es efectiva para comparar la similitud semántica entre dos textos, lo que permite encontrar libros cuyas sinopsis se relacionen más estrechamente con las preferencias del usuario. Por otro lado, la distancia de Levenshtein es útil para encontrar coincidencias parciales en cadenas de texto, como nombres de autores, lo que amplía las posibilidades de búsqueda y facilita la inclusión de variaciones en los nombres.