

Elección del dataset:

El presente documento tiene como finalidad evaluar los recursos y conjuntos de datos disponibles en el repositorio AmazonReviews2023, el conjunto de datos de Retailrocket, y un estudio adicional publicado en arXiv (referencia 2307.09688). Estos recursos serán analizados en el contexto de proyectos relacionados con sistemas de recomendación y recuperación de información en el ámbito del comercio electrónico.

1. AmazonReviews2023

El repositorio AmazonReviews2023 ofrece un amplio conjunto de herramientas y datos esenciales para investigaciones avanzadas en sistemas de recomendación y recuperación de información. Este conjunto de datos abarca reseñas de productos de Amazon desde 1996 hasta 2023, proporcionando una cobertura temporal y temática extensa. Sus componentes principales incluyen:

- **Datos de reseñas y productos:** Más de 570 millones de reseñas y 48 millones de productos distribuidos en 33 categorías distintas.
- **Procesamiento de datos:** Scripts diseñados para transformar los datos en conjuntos de entrenamiento, validación y prueba, facilitando así la evaluación de modelos. (Disponible en GitHub)
- **Modelos preentrenados BLalR:** Modelos de lenguaje enfocados en capturar la relación entre descripciones de productos y su contexto lingüístico, lo que mejora la calidad de las recomendaciones. (Disponible en GitHub - BLalR)
- **Amazon-C4:** Un conjunto de datos adicional creado para evaluar la comprensión del contexto y la relevancia en la recuperación de productos. (Disponible en GitHub - Amazon-C4)

Valoración:

AmazonReviews2023 se presenta como una opción idónea para el desarrollo de algoritmos de recomendación avanzados. Los modelos BLalR y Amazon-C4 proporcionan oportunidades significativas para mejorar el entendimiento semántico y la recuperación de productos, lo que incrementa la precisión de las recomendaciones ofrecidas.

Referencia académica:

Hou, Y., Li, J., He, Z., Yan, A., Chen, X., & McAuley, J. (2023). *Bridging Language and Items for Retrieval and Recommendation*.

2. Retailrocket dataset

El conjunto de datos de Retailrocket proporciona información detallada sobre la actividad de los usuarios en una plataforma de comercio electrónico. Los datos están organizados de la siguiente manera:

- **Eventos web:** Registro de interacciones de los usuarios, tales como visualizaciones de productos, adiciones al carrito y compras.
- **Propiedades de los artículos:** Información detallada de los productos, incluyendo descripciones y características específicas.
- **Árbol de categorías:** Estructura jerárquica que organiza los productos en diversas categorías para facilitar la navegación y búsqueda.

Valoración:

Este conjunto de datos es particularmente útil para la construcción y evaluación de sistemas de recomendación en escenarios que involucran la navegación web y el comportamiento del usuario.

Referencia académica:

Roman Zykov, Noskov Artem, and Anokhin Alexander. (2022). Retailrocket recommender system dataset [Data set]. Kaggle.
<https://doi.org/10.34740/KAGGLE/DSV/4471234>