Optimización de recomendaciones de productos en tiendas en línea a través del análisis de reseñas y aprendizaje no supervisado.

Maestría en inteligencia analítica de datos (MIAD) - Aprendizaje no supervisado
Grupo 22: Elvin Rodrigo Méndez, Juan Jose Ovalle, William Morales
Enlace repositorio GitHub

Resumen.

Este proyecto se enfoca en mejorar la clasificación de las reseñas (reviews) de productos generadas por los usuarios de tiendas en línea mediante la aplicación de técnicas de aprendizaje no supervisado. En el creciente mundo del comercio electrónico (e-commerce), la clasificación y precisión en las reviews de productos son bastante importantes para la satisfacción del cliente y el éxito del negocio. A través del aprendizaje no supervisado, buscamos optimizar la forma en que los usuarios interactúan con las reviews de los productos y toman decisiones de compra mejor informadas.

Para la ejecución del proyecto, utilizamos un conjunto de reseñas de Amazon, el cual fue adquirido durante un evento organizado por la empresa Factored. Dentro del análisis de los datos incluiremos el preprocesamiento de datos para limpiar y estructurar la información, la extracción de características relevantes de las reseñas utilizando técnicas de procesamiento de lenguaje natural (NLP) y la clusterización de las reseñas.

El objetivo final es proporcionar una mejor experiencia al momento de comprar dando especial énfasis a las reseñas generadas por los usuarios que han adquirido los productos.

Introducción.

En los últimos años, especialmente después de la pandemia, las compras en línea han dejado de ser un tabú y se han convertido en una práctica común para adquirir productos a través de estas plataformas en línea, sin embargo, la cantidad de productos disponibles y las diferentes marcas que ofrecen funcionalidades similares demuestran que ya no es únicamente el costo del producto lo que influye en la decisión de compra por parte de los usuarios, es aquí donde las reseñas y las recomendaciones generadas por quienes han comprado los productos, juegan un papel importante para mejorar la experiencia final de compra.

El objetivo de este proyecto es desarrollar un proceso de clasificación no supervisada, que apoye en la selección de productos con base en las reseñas de los usuarios y así poder tomar decisiones de compra mejor informadas.

Revisión Preliminar de Antecedentes en la Literatura.

En investigaciones anteriores se han utilizado técnicas de aprendizaje no supervisado y el análisis de sentimientos para abordar problemas similares en la industria del e-commerce, sin embargo, las recomendaciones se centran específicamente en grupos de usuarios con preferencias similares, nuestro objetivo es mejorar la forma en que se clasifican las reseñas

sobre productos específicos utilizando técnicas similares de aprendizaje no supervisado, mejorando la selección de predictores utilizando análisis de componentes principales.

En el artículo “E-commerce Recommender System Using PCA and K-Means Clustering” aplican técnicas de aprendizaje no supervisado para mejorar la recomendación de productos a usuarios basados en preferencias de grupos similares, PCA se utiliza para reducir la cantidad de características a utilizar y así mejorar el procesamiento de los modelos.

Por otro lado, en el artículo “Fake Review Detection using Principal Component Analysis and Active Learning” se basan en las reseñas de los usuarios, pero con el objetivo de detectar aquellas que son falsas a través de aprendizaje supervisado, el uso de PCA en este caso se encuentra enfocado en reducir las características de las reseñas.

Descripción Detallada de los Datos

Para abordar el proyecto se utilizarán dos sets de datos, los cuales corresponden a productos y reseñas de usuarios que han comprado en Amazon.

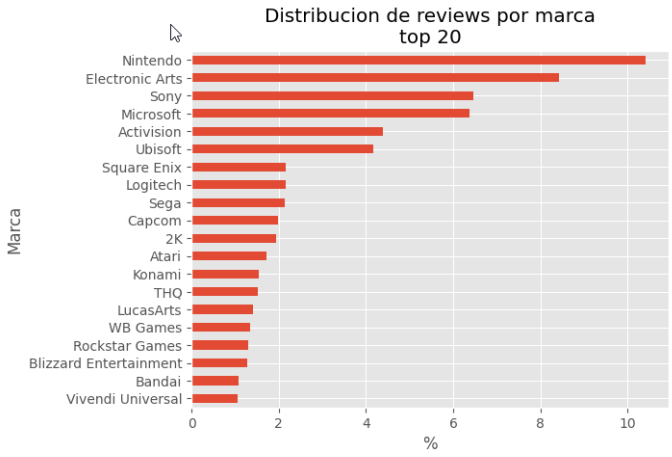
- ✓ **Reseñas de productos de Amazon:** Este es un set de datos de reseñas de productos de Amazon. Este set de datos contiene 82 millones de reseñas únicas de alrededor de 20 millones de usuarios.

Columna	Tipo de Dato	Descripción
asin	Cadena de Texto	Identificador único (ASIN) del producto asociado a la reseña.
overall	Cadena de Texto	Calificación general otorgada por el revisor al producto en forma de estrellas o puntos.
reviewText	Cadena de Texto	Texto completo de la reseña escrita por el usuario.
reviewerID	Cadena de Texto	Identificador único del revisor de productos.
reviewerName	Cadena de Texto	Nombre o seudónimo del revisor.
summary	Cadena de Texto	Resumen de la reseña escrita por el usuario.
unixReviewTime	Cadena de Texto	Fecha de la reseña en formato UNIX (marca de tiempo).
verified	Cadena de Texto	Indicador de si la reseña ha sido verificada por la tienda en línea
vote	Cadena de Texto	Número de votos que ha recibido la reseña de otros usuarios.
image	Cadena de Texto	Enlace o referencia a una imagen relacionada con el producto o la reseña.
style	Cadena de Texto	Detalles sobre el tipo de producto que se está revisando.

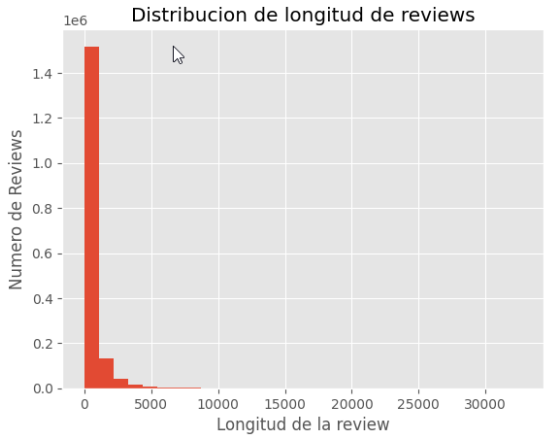
- ✓ **Metadatos de Amazon:** Contiene la descripción y metadatos de los productos incluidos en el set de datos.

Columna	Tipo de Dato	Descripción
also_buy	Secuencia	Productos que los compradores también adquirieron junto con el producto.
also_view	Secuencia	Productos que los usuarios también vieron al explorar el producto.
asin	Cadena de Texto	Identificador único (ASIN) del producto.
brand	Cadena de Texto	Marca o fabricante del producto.
category	Secuencia	Categorías a las que pertenece el producto.
date	Cadena de Texto	Fecha asociada al producto o su listado en la plataforma.
description	Secuencia	Información detallada sobre el producto, incluyendo características y especificaciones.
feature	Secuencia	Características del producto.
fit	Cadena de Texto	Compatibilidad del producto.
image	Secuencia	Enlaces o referencias a imágenes relacionadas con el producto.
main_cat	Cadena de Texto	Categoría principal del producto.
price	Cadena de Texto	Precio del producto.
rank	Cadena de Texto	Rango o posición del producto en relación con otros productos de la misma categoría.
similar_item	Cadena de Texto	Identificador de un producto similar o relacionado.
tech1	Cadena de Texto	Información técnica adicional relacionada con el producto.
tech2	Cadena de Texto	Detalles técnicos adicionales del producto.
title	Cadena de Texto	Mombre del producto.

Teniendo en cuenta el volumen de datos, para el análisis exploratorio de los datos en esta primera entrega se hará uso únicamente de la categoría de videojuegos.



- ✓ Nintendo, EA, Sony, Microsoft y Activision juntas representan el 30% de los datos, mientras que las demás marcas contribuyen con entre un 1% y un 2% cada una.



- ✓ La longitud promedio de las reseñas es de 500 caracteres, y aproximadamente el 75% de los datos tiene una longitud de alrededor de 550 caracteres o menos. Sin embargo, algunas reseñas tienen una longitud inusual, siendo la más larga de 32,721 caracteres

Propuesta Metodológica.

Para bordar el proyecto, se usará el siguiente enfoque metodológico.

- ✓ **Análisis de los datos.**
  Iniciaremos con un análisis descriptivo de los datos disponibles, este paso nos permitirá comprender la naturaleza de la información con la que trabajaremos.
- ✓ **Preprocesamiento de datos.**
  Realizaremos limpieza y preprocesamiento de los datos, esto incluirá la eliminación de valores nulos y la tokenización de las reseñas.
  Debido al tamaño del conjunto de datos, aprovecharemos herramientas en la nube, como BigQuery de Google, para llevar a cabo este proceso.
- ✓ **Extracción de características.**
  Aplicaremos técnicas de procesamiento de lenguaje natural (NLP) para extraer características relevantes de las reseñas de los usuarios, identificaremos aspectos clave que nos ayudarán a comprender mejor el contenido de las reseñas
- ✓ **Reducción de Dimensionalidad.**
  Utilizaremos técnicas de reducción de dimensionalidad para simplificar el conjunto de datos, manteniendo solo las características más relevantes, esto nos permitirá trabajar de manera más eficiente y centrarnos en los aspectos más importantes de las reseñas.
- ✓ **Clustering de Productos.**
  Aplicaremos algoritmos de clustering para agrupar reseñas similares en función de sus características, esto nos ayudará a identificar patrones y tendencias en las opiniones de los usuarios y a comprender cómo los productos se agrupan en función de la retroalimentación de los clientes.

Referencias.

Faisal, M., & Sifat , A. (2019). Fake Review Detection using Principal Component Analysis and Active Learning. Information Technology Articles, http://doi.org/10.5120/ijca2019919418

Andra, D., & Baizal, A. (2022). E-commerce Recommender System Using PCA and K-Means Clustering. Information Technology Articles, https://doi.org/10.29207/resti.v6i1.3782