CADS - ReBaF

Proyecto Final - Henry - Data Science

https://github.com/UMazzucco/CADS-PF

¿Quiénes somos?

Colombia Argentina Data Solutions es una empresa internacional de tecnología que brinda soluciones de procesamiento de datos para mantener a tu empresa en la cima del mercado.

“Scientia potentia est", ‘El conocimiento es poder’, El Leviatán, Thomas Hobbs, 1668.

CADS
Data Solutions

Nuestro Equipo

* Álvarez Mateo, Software Developer, Especialista en Machine Learning.

* Mazzucco Uriel, Data Engineer, Especialista en Big Data.

* Pilla María del Pilar, Data Analyst, Especialista en Business Intelligence.

* Rojas Martín, Software Developer, Especialista en Natural Language Processing.

Objetivos

Amazon busca mejorar la conexión entre usuarios y vendedores, utilizando la perspectiva de los compradores.

Review Based Features ataca el problema desde dos frentes:

Un sistema de recomendación basado en las puntuaciones que los usuarios dan a los productos.

Un procesamiento de texto que nos brinda información agregada sobre los patrones de compra de los usuarios.

Tecnologías

Puesto que tenemos 42 Gb de datos, las capacidades de cómputo locales no son suficientes.

Por ello trabajaremos con diversas herramientas de Google Cloud Services.

Optamos por abrir un clúster Dataproc , permitiéndonos trabajar en un sistema Hadoop a través de Spark.

Para la visualización utilizamos Looker,almacenando los datos procesados en Google Storage y conectándolos a través de Big Query.

Pipeline

Los datos se obtienen a través de una url aportada por el cliente:

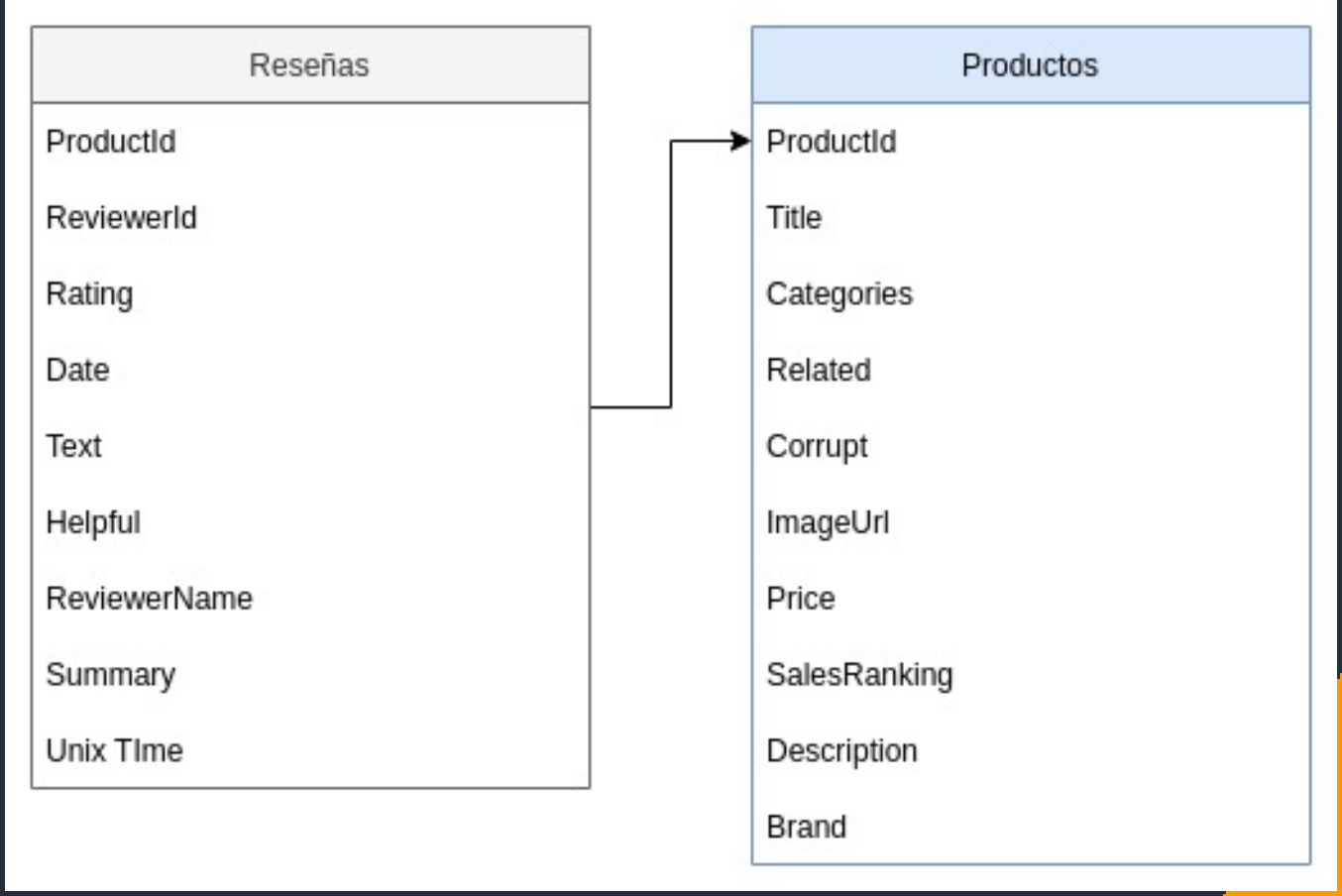
http://jmcauley.ucsd.edu/data/amazon/links.html

Através de comandos bash se obtienen los archivos, se los descomprime y se los transfiere al sistema hadoop.

El resultado son dos archivos .json: uno con reseñas y el otro con metadata de los productos.

Dichos datasets ya poseían una limpieza previa, quedando sin duplicados y sólo las reseñas de productos con al menos 5 reseñas.

DER


Reseñas
ProductId
ReviewerId
Rating
Date
Text
Helpful
ReviewerName
Summary
Unix TIme
Productos
ProductId
Title
Categories
Related
Corrupt
ImageUrl
Price
SalesRanking
Description
Brand

ETL

Productos

Comenzamos por eliminar las columnas que no nos brindaban información útil.

Normalizamos los nombres de las columnas restantes.

Eliminamos los productos que no poseían nombre o id.

Quitamos los registros cuyos id no se encontraba entre las reseñas.

Normalizamos las categorías.

ETL

Reseñas

Comenzamos por eliminar las columnas que no nos brindaban información útil.

Normalizamos los nombres de las columnas restantes.

Rellenamos con valor '' los falores faltantes en la columna texto.

Quitamos los registros cuyos id no se encontraba entre los productos.

Formateamos la fecha.

EDA

Del análisis exploratorio obtenemos:

* La calificación promedio de productos es 4.23.

* Obtenemos aproximadamente 30 millones de reseñas, acerca de 1 millón de productos, realizadas por 2 millones de usuarios.

* Cada producto fue reseñado por 30 usuarios en promedio.

* Cada usuario reseñó 15 productos en promedio.

* A tener en cuenta: se estima que entre el 3% y el 10% de los compradores reseña el producto.

Procesamiento de texto

Sobre el texto de las reseñas aplicamos dos procesos:

Un análisis de sentimiento de la librería NLTK, para utilizar en el modelo de recomendación.

Una búsqueda de palabras clave para determinar si se habla sobre calidad, la facilidad de uso o el precio del producto, y si se hace de manera positiva o negativa.

Como resultado obtuvimos tres columnas nuevas con valores -1, 0 ó 1.

Finalmente unimos las tablas para realizar las predicciones y la visualización.

Tabla Final

Fila	categories	title	calidad	related	precio	facilidadUso	sentiment	reviewTime	rating	reviewerId	productId
1	Books	Red Adam's Lady	0	, 0812823354, 0062273574, 0843933585, 0843931140, 0553583557, 0312956029, 0440614155, 0380871556, 0671737627, 0425259269,,	0	0	5	2009-06-18	5	AMVV8VYDTLA78	0002216973
2	Books	Red Adam's Lady	0	, 0812823354, 0062273574, 0843933585, 0843931140, 0553583557, 0312956029, 0440614155, 0380871556, 0671737627, 0425259269,,	0	0	5	2011-12-31	5	AHCOCJHM388I7	0002216973
3	Books	Red Adam's Lady	0	, 0812823354, 0062273574, 0843933585, 0843931140, 0553583557, 0312956029, 0440614155, 0380871556, 0671737627. 0425259269..	1	0	5	2013-08-26	5	ACUJMLOJEVYTB	0002216973

Delta

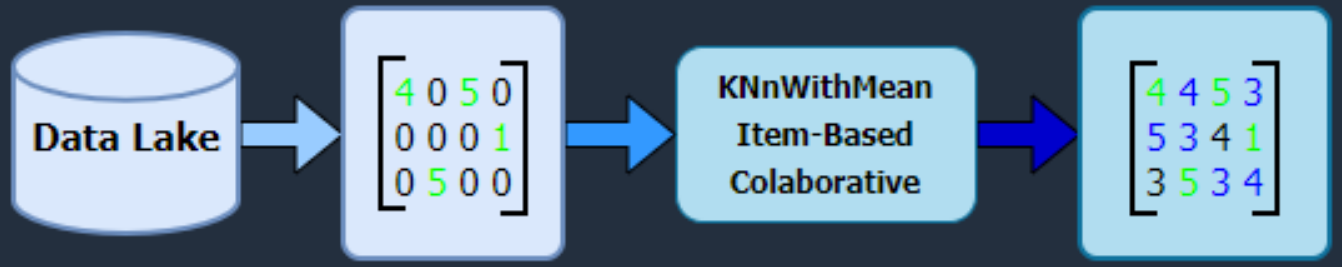
Implementar un sistema de etiquetas review-based para guiar al usuario.

Automatizar el proceso a la ingesta dinámica de información.

Trazar perfiles de compra relacionados con cada usuario.

Modelo de Recomendación

Aplicamos un modelo de filtro colaborativo, que trabaja sobre una matriz usuarios-productos, tomando los valores de las reseñas y aprendiendo de ellos para predecir los campos vacíos.

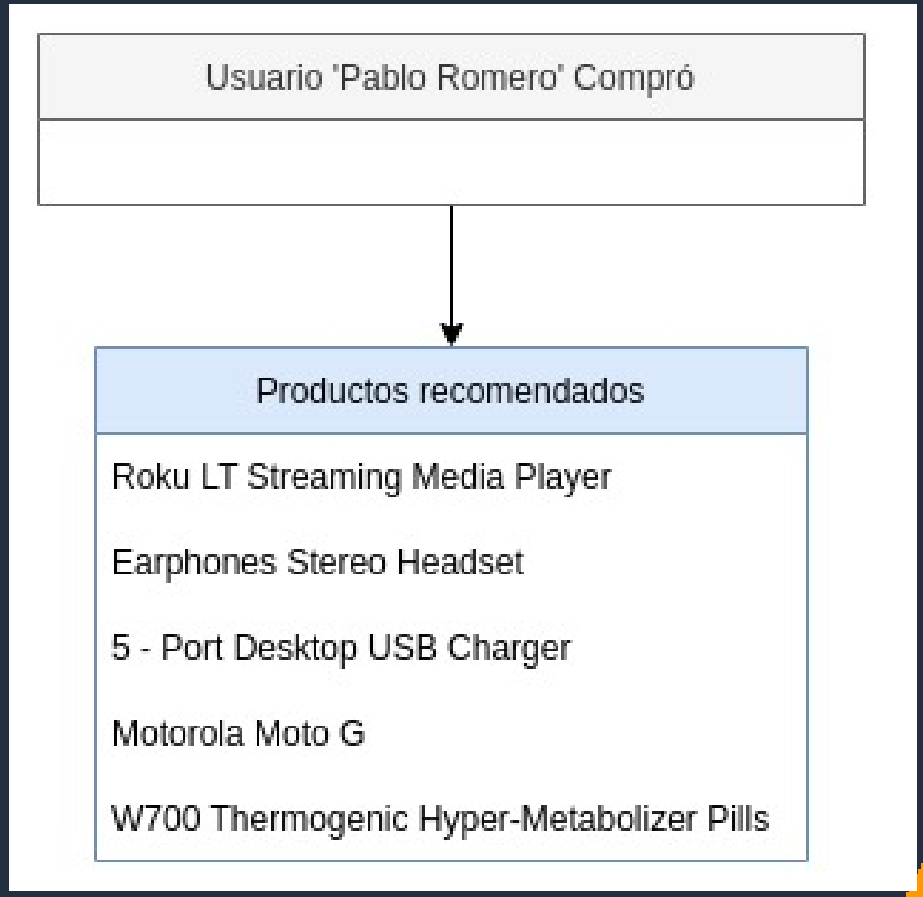


Algunas Recomendaciones:

Usuario 'A1MBZNBQCYCTBD' Compró
Coffee Decaf 24 K-Cups for Keurig Brewers
1001 Things to Spot in Fairyland
Bungee Jumper
Coffee Italian Roast, 50 K-Cup for Keurig Brewers
Grasshoppers Women's Highview Slip-On Loafer
Nite Ize Spokelit Bicycle Light
Dressy Doll Clothing

Productos recomendados
Put Me in the Zoo
Are You My Mother?
How the Grinch Stole Christmas!
The Mitten
LEGO Ultimate Building Set - 405 Pieces
Panasonic DMP-BD35K Blu-ray Player
Fisher-Price I Can Play Basketball
Polar Express
Iced Tea 16 K-Cups for Keuring Brewers
Grasshoppers Women's Highview Slip-On Loafer

Más Recomendaciones:



amazon

Elegí la categoría: Electronics (1)

Escribí acá el nombre del producto:: Goo... (1)

Elegí un año

Elegí un trimestre

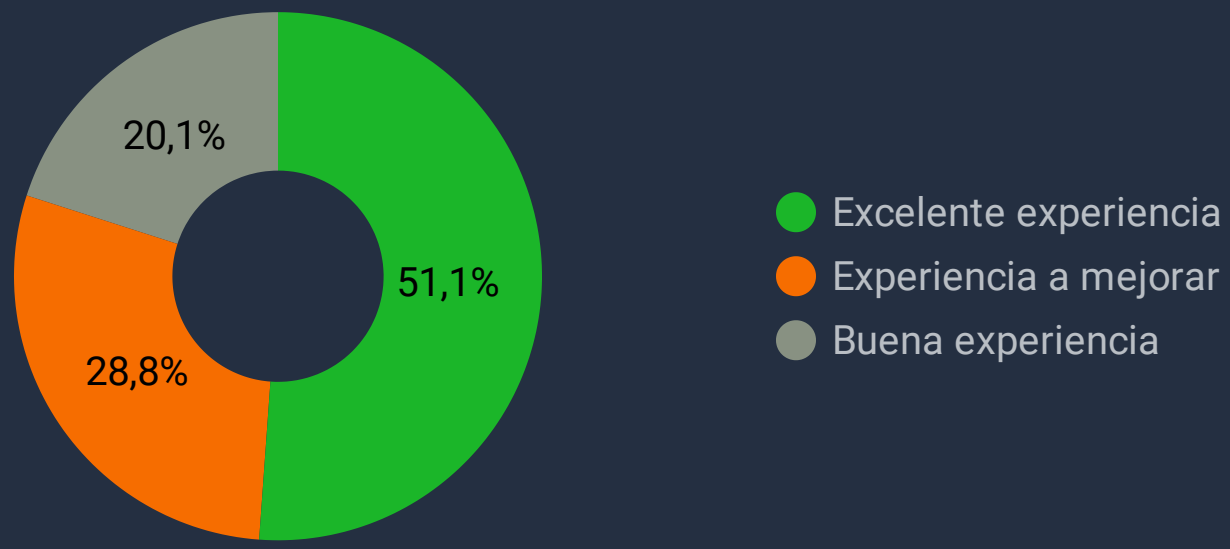
Elegí un mes

CSAT
67,2 %

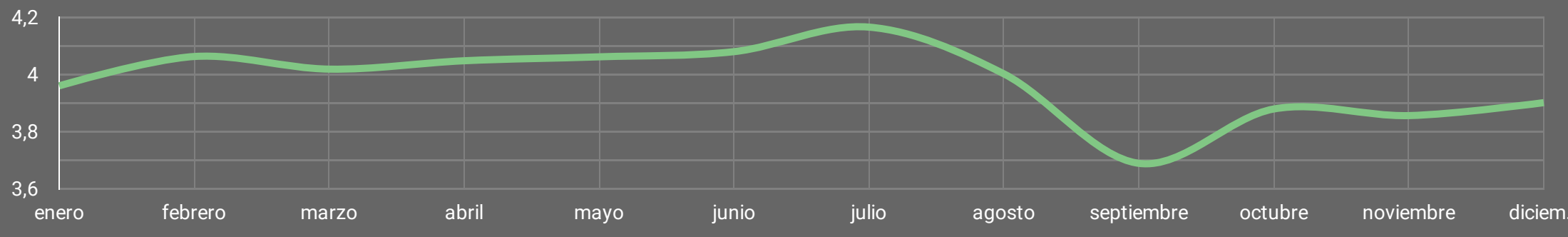
NPS
22,3

rating
3,97

¿Cómo fue la experiencia del usuario?



Rating a través del tiempo



CADS Data Solutions

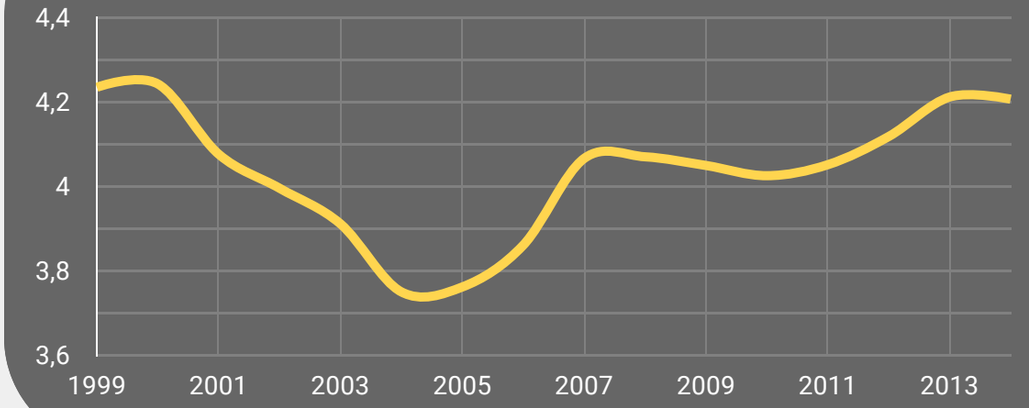
amazon

Elegí la categoría: Electronics (1)

Elegí el producto

Nombre del producto	Imagen	Promedio rating
Google Chromecast HDMI Streaming Media Player		3,97
SanDisk Ultra 64GB MicroSDXC Class 10 UHS Memory Card		4,54

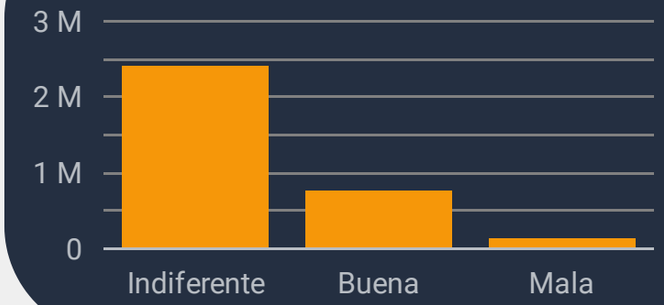
Rating a través de los años



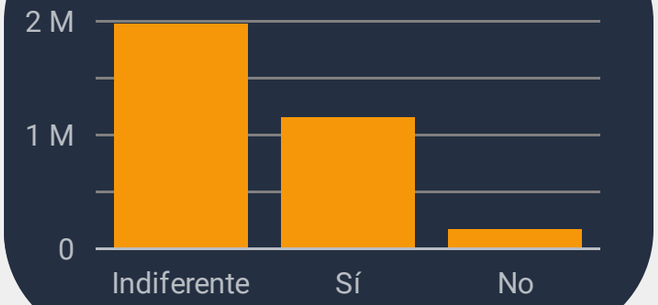
Cantidad de reseñas

3 M

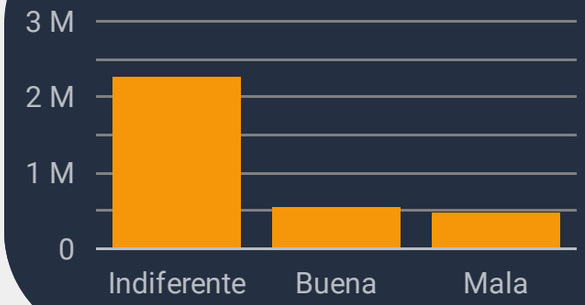
¿Cómo es la calidad de este producto?



¿Es fácil de usar este producto?



Relación calidad - precio



CADS Data Solutions

¡Muchas gracias por su atención!

CADS
Data Solutions

¿Hay alguna pregunta que nos quieras hacer?

¿Qué te pareció nuestra presentación?

Dejanos tu FeedBack! 💗