Anteproyecto:
Sistema de recomendación para un Marketplace

Pablo Saldarriaga-Aristizabal, *Nicolás Prieto-Escobar*,
Victoria Álvarez-Restrepo, *Karen Velásquez-Moná*, *Ana Urán-González*
Maestría en Ciencia de los Datos y Analítica

Universidad EAFIT
Medellín – Colombia

1. Pregunta de investigación y objetivos

1.1. Pregunta de investigación

¿Cómo emplear reseñas y características de los productos de un marketplace para generar recomendaciones de otros artículos a usuarios, en presencia de grandes volúmenes de datos?

En este trabajo, se implementará un sistema de recomendación para artículos de Amazon. Este modelo le sugerirá a cada usuario los productos que es probable que le interesen, considerando tanto las reseñas que le ha dado a otros artículos que ha adquirido como también las características propias de los productos. El modelo será construido usando Spark y desplegado en Amazon EMR, de forma tal que utilice técnicas adecuadas para ser usado con grandes cantidades de datos, para que sea efectivamente escalable al gran catálogo de productos que tiene Amazon. Adicionalmente, se probarán y compararán varias técnicas para poder obtener el mejor rendimiento posible, es decir, para realizar sugerencias adecuadas de productos (tanto existentes como nuevos) para los usuarios.

1.2. Objetivos

General:
Crear sistemas de recomendación en un contexto de Big Data que permitan sugerir productos afines a los intereses de los usuarios en un marketplace.

Específicos:

- Implementar un flujo de trabajo básico usando Amazon EMR con Pyspark y Python para el sistema de recomendación.
- Comparar versiones adaptadas para Big Data de sistemas de recomendación con técnicas baseline (media, mediana, etc.) y sistemas de filtrado colaborativo (factorización de matrices con ALS, inclusión de sesgos en el modelo, etc.).
- Evaluar otro tipo de modelos para resolver la problemática de dar recomendaciones en presencia de usuarios o productos nuevos (inicio en frío).

2. Metodología de investigación.

Durante este proyecto, se seguirá la metodología CRISP-DM. Primero se realizará el entendimiento del problema y la revisión de los datos con PySpark usando Amazon EMR. Luego, se prepararán los datos de los productos y las reseñas para crear distintas versiones de modelos de recomendación, pasando primero por técnicas baseline (como el promedio de los ratings), luego por técnicas clásicas de la literatura (factorización de matrices con ALS e inclusión de los sesgos) y abordando problemáticas más especializadas (inicio en frío). Se evaluarán los modelos creados para elegir el mejor en cada caso y se configurarán estos modelos para que puedan dar recomendaciones para todos los usuarios.
Los primeros modelos que se revisarán emplean estadísticos básicos de los ratings para predecir las calificaciones restantes (promedio por artículo o por usuario, media global, mediana, etc.). Estos modelos básicos no sólo definen un punto de partida para comparar resultados, sino que podrían ser un paso necesario en el manejo de nuevos usuarios.
Posteriormente, aplicaremos técnicas de filtrado colaborativo, las cuales se basan en el uso de información de las preferencias de los usuarios, cuyos intereses sobre un artículo están evaluados con base en las calificaciones asignadas por el mismo usuario, y las recomendaciones se realizan mirando usuarios con comportamientos similares (Smirnov A V., 2014). En particular, en filtrado colaborativo se usarán técnicas basadas en factorización de matrices con el método ALS (Alternating Least Squares), ya que tiene un enfoque que descompone la matriz inicial usuario/artículo en factores de usuario y de artículo de menor dimensión, permitiendo reducir el tamaño de la matriz y generar beneficios en la paralelización requerida para el tratamiento de grandes volúmenes de datos a través de Spark (Panigrahi *et al.*, 2016). Adicionalmente, se tratarán técnicas más avanzadas, como el manejo de los sesgos (bias) que se pueden producir (Boratto *et al.*, 2019), los cuales permiten detectar, por ejemplo, que hay usuarios que suelen dar ratings mucho más altos que el usuario promedio, y tener en cuenta esto en el modelo.
Por otro lado, se explorarán también modelos para la problemática del inicio en frío ("cold start"), la cual ocurre cuando es necesario recomendar productos a un cliente completamente nuevo o cuando nuevos productos entran al conjunto de datos. En este caso, se tendrán en cuenta también modelos de content-based filtering, los cuales están basados en características propias de los productos diferentes al rating, como por ejemplo, su precio o su categoría (Bianchi *et al.*, 2017).

3. Descripción de los datos.

El dataset seleccionado es Amazon Customer Reviews, éste contiene más de 150 millones de reseñas registradas entre 1995 y 2015 por diferentes usuarios frente a sus compras en Amazon y se encuentran clasificadas en 46 categorías[1]. El conjunto de datos se encuentra almacenado en un bucket público de *S3* de Amazon y la información a utilizar está disponible en formatos TSV (la totalidad de los archivos comprimidos ocupa cerca de 32 GB).

[1] `https://s3.amazonaws.com/amazon-reviews-pds/readme.html`

En cuanto a la información relacionada a los datos de las reseñas de Amazon, la tabla 1 presenta una descripción de los campos que ésta contiene.

Tabla 1: Descripción del conjunto de datos

Variable	Tipo de dato	Descripción
marketplace	string	Código del país del usuario que escribe la reseña
customer_id	string	Identificador de usuario que escribe la reseña
review_id	string	Identificador único de la reseña
product_id	string	Identificador único del producto que recibe la reseña
product_parent	string	Otro identificador del producto
product_title	string	Nombre del producto
product_category	string	Amplia categoría de productos
star_rating	int	Calificación del producto de 1 a 5
helpful_votes	int	Número de votos de Reseña útil"que recibe la reseña
total_votes	int	Número total de votos
vine	string	Reseña escrita por un usuario del Vine program
verified_purchase	string	Compra realizada en Amazon
review_headline	string	Título de reseña
review_body	string	Reseña
review_date	bigint	Fecha de la reseña
year	int	Año de la reseña

El total de las reseñas corresponden a 22 millones de artículos únicos adquiridos por casi 80 millones de clientes. No obstante, al analizar la información de las reseñas disponibles en las diferentes categorías de productos, se evidencia un desbalance importante: prendas de vestir, zapatos, libros (físicos y digitales), artículos para el hogar y productos deportivos concentran casi la mitad de los artículos reseñados.
Además, se evidencia que el mínimo de los promedios de las calificaciones por categoría es de 3.5 y la máxima calificación promedio por categoría es de 4.7, mientras que el promedio de los promedios de los ratings por categoría corresponde a 4.2, por lo que, en general, los productos tienen reseñas positivas. También se menciona que las categorías con mayor cantidad de reseñas por producto corresponden a videojuegos, aplicaciones móviles y software.
Del conjunto de datos descrito, se obtuvo una muestra para ser explorada y tener una primera aproximación de si efectivamente la información disponible permite realizar recomendaciones de productos en el marketplace de Amazon. Esta muestra se obtuvo de la siguiente manera: se consideraron únicamente las reseñas de la categoría de productos musicales (aproximadamente 4 millones de registros), del conjunto de datos resultante, se eliminaron aquellas reseñas asociadas a los clientes que, en esa categoría, hayan realizado menos de 80 reseñas, dejando así el conjunto de datos en aproximadamente 3 mil registros. Paso seguido, con la información resultante se realiza un modelo inicial de sistema de recomendación (modelo SVD del paquete Surprise[2] de Python). Para la creación de este modelo, se realiza la partición del conjunto de datos en entrenamiento y prueba con una proporción de 75 % y 25 % respectivamente.
Al entrenar el modelo y realizar la predicción en el conjunto de prueba, se obtiene un MSE de 0.6985. Como punto de comparación, se creó un baseline para el sistema de

[2] `http://surpriselib.com/`

recomendación, en el cual la predicción del rating corresponde al promedio global de los ratings en el conjunto de entrenamiento, con esto se obtiene en el conjunto de prueba un MSE de 0.96341, lo cual representa un error mucho mayor al obtenido por el modelo creado. Así, usando esta muestra de los datos, se ve que es posible explicar los ratings de los productos de un marketplace como el de Amazon con la información disponible.
Adicional a la información descrita, en un repositorio de un profesor asociado de la Universidad de California[3], se encuentra información relacionada a los reviews de Amazon entre 1996 y 2018. En particular, hay un archivo que contiene metadatos de los productos de Amazon y contiene el nombre del producto, código (id del producto), precio, entre otras variables, las cuales permiten enriquecer la información ya mencionada. Esta nueva información será utilizada para abordar el problema de inicio en frío para nuevos productos. El archivo comprimido de metadata tiene un tamaño de 11 GB y contiene información de 15 millones de productos.

4. Plan (diagrama Gantt o Pert).

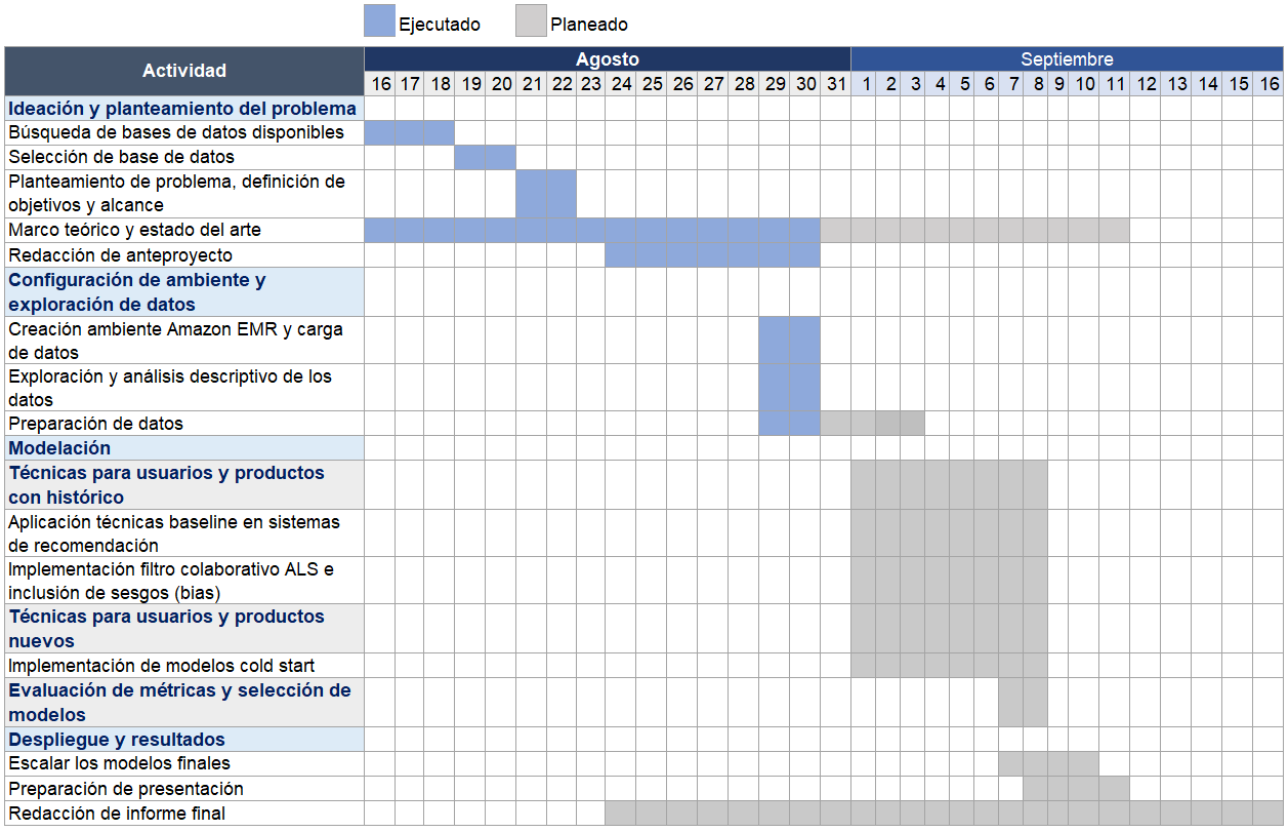

Figura 1: Diagrama Gantt para el desarollo del proyecto

[3] `http://jmcauley.ucsd.edu/data/amazon/`

5. Implicaciones éticas.

La naturaleza de los algoritmos de recomendación consiste en dar forma a las preferencias de los usuarios y guiar sus elecciones individuales y sociales, lo que conlleva en si mismo varios cuestionamientos éticos interesantes. Para este proyecto, se identificaron las siguientes implicaciones éticas (Milano, 2020):

- El riesgo latente de que la información del usuario sea filtrada o empleada indebidamente para violar su privacidad.
- Los impactos negativos que puedan generarse en las partes involucradas (productores-usuarios) que implique la quiebra de unos, junto con la falta de transparencia y explicabilidad en las recomendaciones hechas a los usuarios.
- Exponer al usuario a recomendaciones con contenido dañino u ofensivo según sus preferencias culturales, ideológicas y morales.

6. Aspectos legales y comerciales.

Este proyecto brinda a los E-commerce, almacenes de cadena y grandes superficies una herramienta para ayudar a sus clientes a tomar, de manera informada, una decisión ágil de compra, permitiendo optimizar tiempo y dinero tanto a las empresas como a los consumidores. Adicionalmente, los proveedores y productores tendrán un impacto directo en sus ingresos al considerarse que un producto altamente recomendado puede tener mayor oportunidad de compra. Sin embargo, el diseño de sistemas de recomendación trae grandes desafíos desde el punto de vista jurídico y legal debido a que sus resultados dependen de la información suministrada por otros usuarios, lo que podría llevar erróneamente a difamaciones o tergiversaciones que representen un daño reputacional para una marca o una exposición indebida de los clientes (Burgess, 2011).

Referencias

Bianchi, Mattia, Cesaro, Federico, Ciceri, Filippo, Dagrada, Mattia, Gasparin, Alberto, Grattarola, Daniele, Inajjar, Ilyas, Metelli, Alberto Maria, & Cella, Leonardo. 2017. Content-based approaches for cold-start job recommendations. *Pages 1–5 of: Proceedings of the Recommender Systems Challenge 2017.*

Boratto, Ludovico, Fenu, Gianni, & Marras, Mirko. 2019. The effect of algorithmic bias on recommender systems for massive open online courses. *Pages 457–472 of: European Conference on Information Retrieval.* Springer.

Burgess, Stephen, Sellitto Carmine Cox Carmen. 2011. Trust perceptions of online travel information by different content creators: Some social and legal implications. *Page 221–235 of: Information Systems Frontiers.* Springer.

Milano, Silvia, Floridi Luciano. 2020. Recommender systems and their ethical challenges. *In: AI SOCIETYl*.

Panigrahi, Sasmita, Lenka, Rakesh Ku, & Stitipragyan, Ananya. 2016. A Hybrid Distributed Collaborative Filtering Recommender Engine Using Apache Spark. *Pages 1000–1006 of: Procedia Computer Sciencel*. Sciencedirect.

Smirnov A V., Shilov NG, Ponomarev A V. Kashevnik AM Parfenov VG. 2014. Group context-aware recommendation systems. *Pages 325–334 of: Scientific and Technical Information Processing*. Springer.