Anteproyecto:
Sistema de Recomendación

Integrantes:
Juliana Ochoa Ramírez - Código: 201910048228
Juan Esteban Torres Marulanda - Código: 201910052228
Andrés Franco Zapata - Código: 201910043228

Minería de datos para grandes volúmenes de información
Maestría en ciencias de los datos y analítica
Universidad EAFIT

Docentes:
Tomas Olarte Hernandez, Santiago Cortés, Santiago Hernández

23 de marzo de 2020

Pregunta de Investigación y Objetivo

De acuerdo con los temas vistos en el curso de Minería de Datos, el objetivo del proyecto consiste en la implementación de un sistema de recomendación para un conjunto de datos masivos mediante el uso de Apache Spark. Para el desarrollo del proyecto se utilizará la base de datos "Amazon Customer Reviews" disponible en https://s3.amazonaws.com/amazon-reviews-pds/readme.html, la cual contiene un alto volumen de datos de las evaluaciones y opiniones de los usuarios de Amazon acerca de diferentes productos disponibles para la venta en la plataforma. Específicamente se busca utilizar Apache Spark para la implementación de una técnica reconocida de sistemas de recomendación y realizar el despliegue sobre altos volúmenes de datos en la nube a través de AWS Educate.

Los avances en la ejecución del proyecto se encuentran en el siguiente repositorio github https://github.com/franco18/TrabajoFinal_ST1803_MINERIA

Metodología de investigación

Para el desarrollo del proyecto se implementarán las técnicas de "Collaborative filtering" para sistemas de recomendación, las cuales buscan a través de la información dada por los usuarios en las revisiones, identificar patrones similares de ratings entre usuarios que permitan estimar predicciones o similitudes para un determinado usuario.

Principalmente se aplicará la categoría de tipo "model-based" utilizando las herramientas equipadas en spark para el aprendizaje de factores latentes como es el algoritmo alternating least squares (ALS) para la construcción de un modelo (Koren Y, 2009) que permita la predicción de los ratings no provistos por los usuarios; sin descartar la posibilidad de utilizar las técnicas de la categoría "memory-based" enfocadas en las similitudes entre usuarios sin estimar parámetros.

Adicionalmente se buscará mejorar el sistema de recomendación implementando una técnica hibrida (Leung CWK, 2006) combinando rating con análisis de texto sobre los comentarios de los usuarios. Para calibrar el sistema de recomendación vamos a realizar cross validation para determinar el mejor modelo.

Descripción de los datos

Amazon provee una descripción de la información suministrada en la base de datos indicando el contenido de cada columna en relación con los reviews de los usuarios (ver Figura 1), adicionalmente se puede acceder separadamente a la base de datos por categoría de producto. Para el proyecto se estima trabajar en memoria entre 10 - 20 Gb, lo que equivale aproximadamente a la información de 10 a 15 categorías de productos.

```
DATA COLUMNS:
marketplace       - 2 letter country code of the marketplace where the review was written.
customer_id       - Random identifier that can be used to aggregate reviews written by a single author.
review_id         - The unique ID of the review.
product_id        - The unique Product ID the review pertains to. In the multilingual dataset the reviews
                    for the same product in different countries can be grouped by the same product_id.
product_parent    - Random identifier that can be used to aggregate reviews for the same product.
product_title     - Title of the product.
product_category  - Broad product category that can be used to group reviews
                    (also used to group the dataset into coherent parts).
star_rating       - The 1-5 star rating of the review.
helpful_votes     - Number of helpful votes.
total_votes       - Number of total votes the review received.
vine              - Review was written as part of the Vine program.
verified_purchase - The review is on a verified purchase.
review_headline   - The title of the review.
review_body       - The review text.
review_date       - The date the review was written.

DATA FORMAT
Tab ('\t') separated text file, without quote or escape characters.
First line in each file is header; 1 line corresponds to 1 record.
```

Figura 1: Descripción de los datos

Se realizó la primera exploración con la base de datos de los reviews de la categoría

'Sports' utilizando Spark para determinar los tipos de datos (ver Figura 2), posteriormente se realiza el filtro de los campos *customer-id, review-id, product-id, product-title, product-category, review-headline, review-body y review-date* como variables requeridas para la implementación de las técnicas seleccionadas (ver Figura 3).

```
root
 |-- marketplace: string (nullable = true)
 |-- customer_id: string (nullable = true)
 |-- review_id: string (nullable = true)
 |-- product_id: string (nullable = true)
 |-- product_parent: string (nullable = true)
 |-- product_title: string (nullable = true)
 |-- product_category: string (nullable = true)
 |-- star_rating: string (nullable = true)
 |-- helpful_votes: string (nullable = true)
 |-- total_votes: string (nullable = true)
 |-- vine: string (nullable = true)
 |-- verified_purchase: string (nullable = true)
 |-- review_headline: string (nullable = true)
 |-- review_body: string (nullable = true)
 |-- review_date: string (nullable = true)
```

Figura 2: Datos cargados en Spark

customer_id	review_id	product_id	product_title	product_category	star_rating	review_headline	review_body	review_date
48945260	R1WBPB8MDCCN8F	B012P7UPSM	Chicago Blackhawk...	Sports	5	LOVE IT. 6 stars!	Bought this last ...	2015-08-31
5782091	R32M0YEWV77XG8	B001GQ3VHG	Copag Poker Size ...	Sports	5	Shipped fast.	These are the bes...	2015-08-31
45813853	RR8V7WR27NXJ5	B008VS8M58	Baoer 223 5.56x45...	Sports	1	Good idea if it w...	It looks good, th...	2015-08-31
1593730	R1MHO5V9Z932AY	B005F06F4U	All Terrain Tackl...	Sports	5	Five Stars	Great jig	2015-08-31
29605511	R16PD71086BD2V	B010T4IE2C	Swim Cap - 3 Pack...	Sports	5	Great quality sil...	I love swimming i...	2015-08-31

only showing top 5 rows

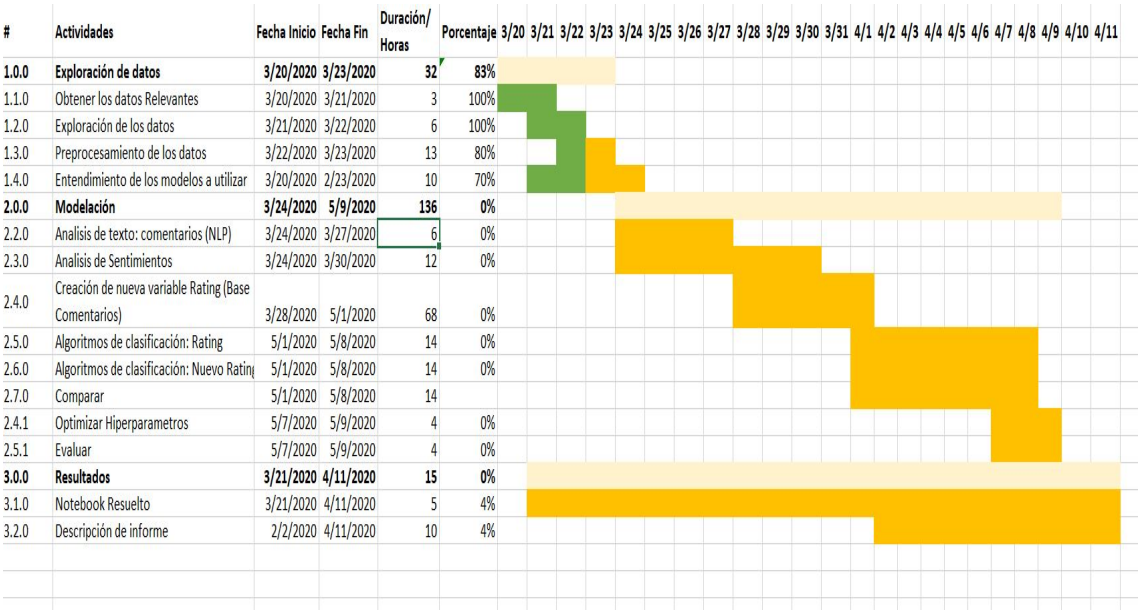
Figura 3: Estructura de datos a utilizar

Por último, se realiza un proceso de limpieza y configuración de los datos para probar en un dataset de muestra la implementación de una recomendación (ver notebook *TrabajoFinal.ipynb* en repositorio github).

Con este proceso confirmamos que tenemos la información requerida para la ejecución del proyecto.

Plan (diagrama Gantt o Pert)

Se incluye el diagrama de Gantt (ver Figura 4) y adicionalmente se adjunta el archivo de Excel para tener mejor visión.



#	Actividades	Fecha Inicio	Fecha Fin	Duración/ Horas	Porcentaje
1.0.0	**Exploración de datos**	**3/20/2020**	**3/23/2020**	**32**	**83%**
1.1.0	Obtener los datos Relevantes	3/20/2020	3/21/2020	3	100%
1.2.0	Exploración de los datos	3/21/2020	3/22/2020	6	100%
1.3.0	Preprocesamiento de los datos	3/22/2020	3/23/2020	13	80%
1.4.0	Entendimiento de los modelos a utilizar	3/20/2020	2/23/2020	10	70%
2.0.0	**Modelación**	**3/24/2020**	**5/9/2020**	**136**	**0%**
2.2.0	Analisis de texto: comentarios (NLP)	3/24/2020	3/27/2020	6	0%
2.3.0	Analisis de Sentimientos	3/24/2020	3/30/2020	12	0%
2.4.0	Creación de nueva variable Rating (Base Comentarios)	3/28/2020	5/1/2020	68	0%
2.5.0	Algoritmos de clasificación: Rating	5/1/2020	5/8/2020	14	0%
2.6.0	Algoritmos de clasificación: Nuevo Ratin	5/1/2020	5/8/2020	14	0%
2.7.0	Comparar	5/1/2020	5/8/2020	14	
2.4.1	Optimizar Hiperparametros	5/7/2020	5/9/2020	4	0%
2.5.1	Evaluar	5/7/2020	5/9/2020	4	0%
3.0.0	**Resultados**	**3/21/2020**	**4/11/2020**	**15**	**0%**
3.1.0	Notebook Resuelto	3/21/2020	4/11/2020	5	4%
3.2.0	Descripción de informe	2/2/2020	4/11/2020	10	4%

Figura 4: Diagrama Gantt

Implicaciones éticas

Las implicaciones éticas para este caso, estarían muy relacionadas con el manejo de los datos que las personas brindan en la plataforma y como se busca que sea algo que las personas sientan y no inducido para dar gusto algún interés en particular o manipular los intereses de los que participan. Acá se busca es como a través de la data encontramos patrones que relacionen el gusto entre las personas (Milano, Taddeo, y Floridi, 2019).

Aspectos legales y comerciales

Es una aplicación que le puede servir a almacenes de cadena para recomendar ofertas según los hábitos de consumo de los clientes para fidelizar al consumidor con nuevos productos y servicios, con el fin de estar en la cotidianidad de los clientes. Es muy importante el tratamiento de los datos y el cumplimiento de la regulación jurídica que establezca cada estado sobre los datos sensibles de clientes, dado que información obtenida sobre las calificaciones que las personas brinden podría exponer la integridad de los clientes.

Referencias

Koren Y, V. C., Bell R. (2009). Matrix factorization techniques for recommender systems. *IEEE Computer*, *42*(8), 30-37.

Leung CWK, C. F., Chan SCF. (2006). Integrating collaborative filtering and sentiment analysis. *ECAI 2006 Worshop on Recommender Systems, Riva del Garda, Italy*, 62–66.

Milano, S., Taddeo, M., y Floridi, L. (2019). Recommender systems and their ethical challenges. *Available at SSRN 3378581*.