Sistemas de Recomendación
Filtros Colaborativos

Integrantes:

Juliana Ochoa Ramírez - Código: 201910048228

Juan Esteban Torres Marulanda - Código: 201910052228

Andrés Franco Zapata - Código: 201910043228

Objetivo

El objetivo del presente trabajo es desarrollar un sistema de recomendaciones para un conjunto de datos masivos a partir de técnicas de filtros colaborativos y el respectivo despliegue en la nube a través de AWS Educate.

Base de datos utilizadas

Amazon Customer Reviews: https://s3.amazonaws.com/amazon-reviews-pds/readme.html

```
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.
```

Categorías Utilizadas:

1. Prototipo baja escala:
   - amazon_reviews_us_Digital_Video_Games_v1_00.tsv.gz .
   - Peso: 160 MB

2. Altos Volúmenes:
   - amazon_reviews_us_Sports_v1_00.tsv.gz.
   - Peso: 5.1 GB

Datos

Amazon Customer Reviews

Exploración Limpieza

Análisis Descriptivo

Datos Pre procesados

Separación datos

Cluster:
1 master
2 slaves

aws educate

CO

Sistema de Recomendación:

Collaborative Filtering Model-based

Training 80%

Test 20%

Latent Factor Model ALS-WR

Adding Biases

Optimización hiperparámetros

Validación Cruzada

Spark

```
+-----------+-------------+--------------------+------------------+
|customer_id|product_new_id|       product_title|             score|
+-----------+-------------+--------------------+------------------+
|   51023203|         6705|Funky Lab Rat [On...|5.4666571051560435|
|   51023203|         7477|     Brain Challenge|5.4666571051560435|
|   51023203|         3723|Guilty Gear Xrd SIGN| 5.441901150413612|
|   51023203|         1564|PixelJunk Monster...| 5.430506709285835|
|   51023203|         7068|Aqua Slider [Down...| 5.411696735092262|
|   51023203|         6679|  Ninja Gaiden Sigma| 5.390872719951728|
+-----------+-------------+--------------------+------------------+
```

Sistema de recomendación

Modelo entrenado

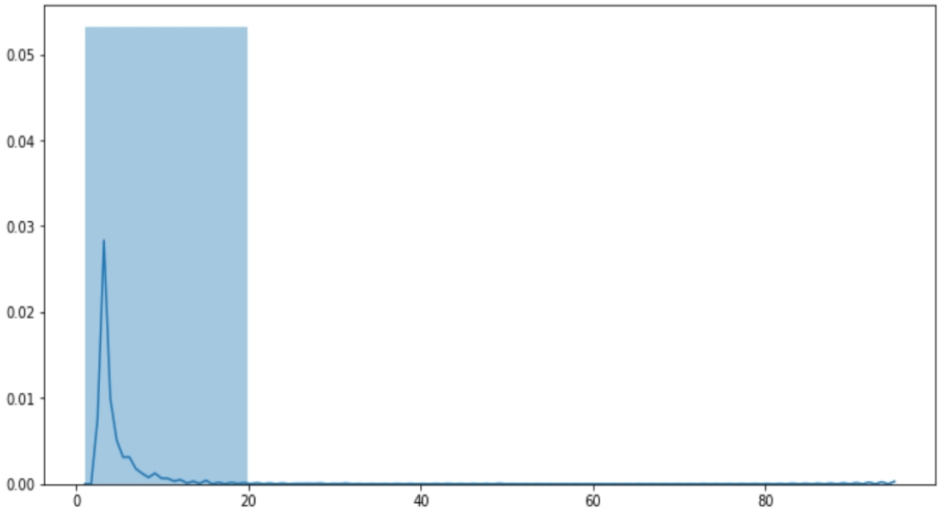
Predicciones ratings

Evaluación modelo ***RMSE***

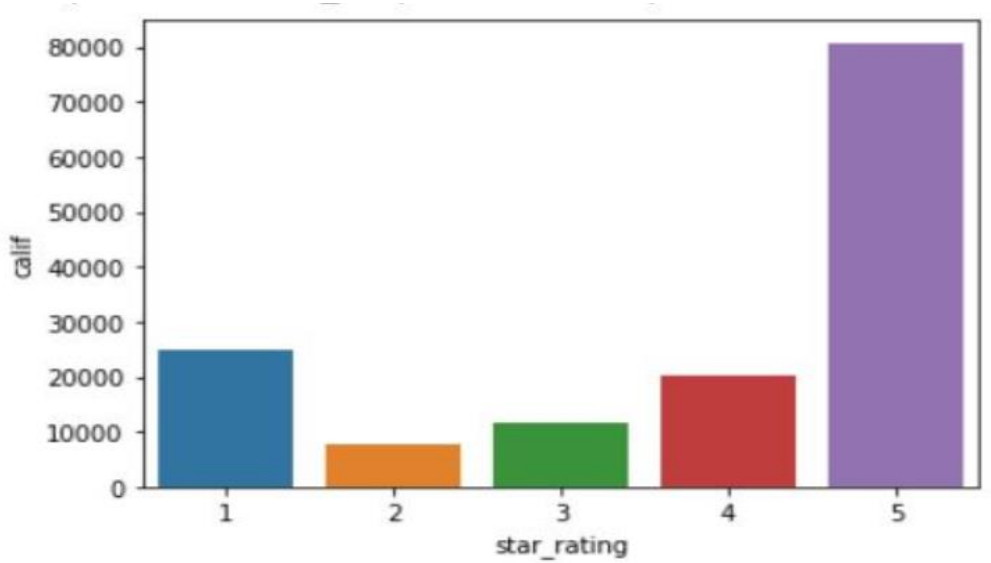
Digital_Video_Games

113.405 usuarios y 7.948 productos

Distribución Items Calificados por Usuarios



Calificación por usuarios

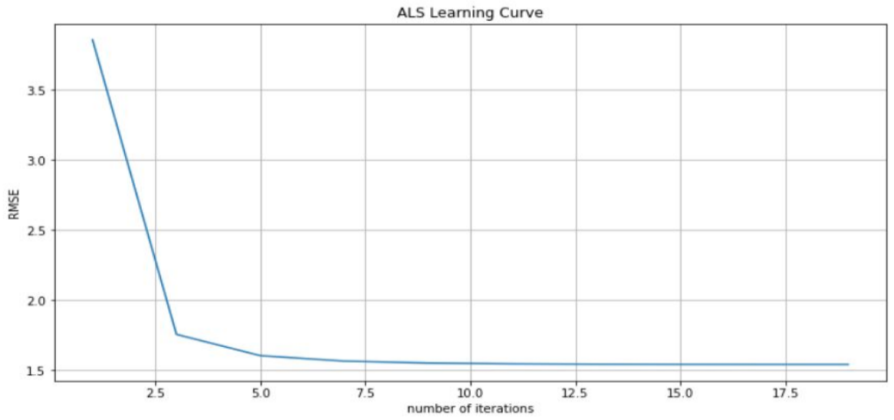


Sports

2.818.178 usuarios y 1.046.129 productos

Digital_Video_Games

ranks = [2,4,6,8,10]
reg_params = [0.1,0.2,0.3,0.4,0.5]



	RMSE		
	ALS default	ALS tuning	ALS - Biases tuning
train	0.17	0.6	0.47
test	1.82	1.54	0.99
rank	10	2	6
regParam	0.1	0.4	0.4
maxIter	10	20	20
variación		-15.38%	-35.71%

Recomendación Digital_Video_ Games

customer_id	product_new_id	product_title	score
51023203	6705	Funky Lab Rat [On...	5.4666571051560435
51023203	7477	Brain Challenge	5.4666571051560435
51023203	3723	Guilty Gear Xrd SIGN	5.441901150413612
51023203	1564	PixelJunk Monster...	5.430506709285835
51023203	7068	Aqua Slider [Down...	5.411696735092262
51023203	6679	Ninja Gaiden Sigma	5.390872719951728

Sports

	RMSE	
	ALS Tuning	ALS-Biases Tuning
Test	1.5957	0.8414
Rank	2	8
regParam	0.4	0.5
maxIter	20	20
Variacion		-47.3%

Recomendación Sports

customer_id	product_new_id	product_title	score
48480929	787393	tasc Performance ...	6.051005361784318
48480929	575707	New York Yankees ...	6.049028096903184
48480929	676013	VIEW Swimming Gea...	6.044649658907273
48480929	585590	XXIO 8i Hybrid Ri...	6.0333239121326
48480929	950635	Storm Tropical Br...	6.02743920515857
48480929	794567	DeFeet Woolie Boo...	6.021140484321931

Conclusiones

- La construcción del modelo desde un inicio debe contemplar la escalabilidad, este elemento es determinante para el éxito del modelo en producción (despliegue en altos volumenes).

- Para nuestro proyecto mientras más queríamos mejorar el modelo requeríamos de mayor cómputo, lo cual era complejo sin la ayuda de Spark.

- La optimización de los hiperparámetros y la regularización jugaron un papel muy importante en el sistema de recomendación final, lo que nos muestra que la paralelización y métodos para hacer eficiente el computo son no negociables en proyectos de altos volúmenes de datos si se quiere obtener una buena generalización.

- El método de factores latentes aplicando el algoritmo de ALS-WR nos mostró la importancia de tener en cuenta métodos iterativos que contemplen la paralelización ante problemas de alta dimensionalidad, en nuestro caso se logra dar manejo a los problemas de dimensionalidad y de falta de información de la matriz de ratings (sparse).

- Tuvimos mejores avances en la construcción de la metodología cuando trabajamos en un prototipo de menos datos, dado que intentar modificar la estructura tardaba mucho en AWS, por lo cual nos pareció mas óptimo una vez definimos la arquitectura de aprendizaje escalarlo en AWS con una base de datos ya más significativa.

Bibliografía

- Aggarwal, C. C. (2016). Recommender systems: The textbook.Springer.
- John S. Breese, D. H. . C. K. (1998).
- Empirical analysis of predictive algorithms forcollaborative filtering.In Proceedings of the Fourteenth conference on Uncertaintyin artificial intelligence (UAI’98).Koren Y, V. C., Bell R. (2009).
- Matrix factorization techniques for recommendersystems.IEEE Computer,42(8), 30-37.Töscher A, J. M. (2008).
- The bigchaos solution to the netflix prize 2008.ATT Labs -Research.Xiaoyuan Su, T. M. K. (2009).
- A survey of collaborative filtering techniques.Advancesin Artificial Intelligence archive