Tarea 3

Introducción a la Ciencia de Datos
Grupo 16

María Luciana Martínez
4.421798-2
luchymd89@gmail.com

Lucía Lemes
5.127.425-6
lucia.lemes@fing.edu.uy

Introducción

En el presente trabajo se utilizó una colección de productos de la tienda en línea de Amazon [1] y las distintas reseñas realizadas por los usuarios de la misma. Los datos se obtuvieron de la página de Computación e Ingeniería de la Universidad San Diego [2], hay dos opciones disponibles, el conjunto de datos completo (con 233,1 millones de reseñas) y un subconjunto, para uso experimental (con 75,26 millones de reseñas), en el cual cada usuario y producto tiene cinco reseñas (5-core). El objetivo del trabajo es poder realizar recomendaciones de productos a los usuarios, teniendo en cuenta las valoraciones de reseñas y rating, así como los patrones de compras de otros usuarios.

Formato del dataset

El conjunto de datos incluye reseñas (usuario, producto, texto, puntuación, votos), metadatos de cada producto (descripción, categoría, precio, marca, imágenes del producto, lista con productos similares, lista con productos comprados en conjunto, detalles del producto como color, tamaño, etc.). Los archivos se encuentran en formato json, donde cada línea corresponde a una reseña. En la Figura 1 se muestran dos ejemplos de reseña con algunos campos importantes.


```
{
"image": ["https://images-na.ssl-images-amazon.com/images
/I/71eG75FTJJL._SY88.jpg"],
"overall": 5.0,
"vote": "2",
"verified": True,
"reviewTime": "01 1, 2018",
"reviewerID": "AUI6WTTT0QZYS",
"asin": "5120053084",
"style": {
        "Size:": "Large",
        "Color:": "Charcoal"
        },
"reviewerName": "Abbey",
"reviewText": "I now have 4 of the 5 available colors of this
shirt... ",
"summary": "Comfy, flattering, discreet--highly recommended!",
"unixReviewTime": 1514764800
}
```

```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "vote": 5,
  "style": {
    "Format:": "Hardcover"
  },
  "reviewText": "I bought this for my husband who plays the piano.
He is having a wonderful time playing these old hymns.  The music  is
at times hard to read because we think the book was published for
singing from more than playing from.  Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```


- **reviewerID** - Identificador de quién realizó la reseña.
- **asin** - Identificador del producto.
- **reviewerName** - Nombre de quién realizó la reseña.
- **vote** - Votos de la revisión.
- **style** - un diccionario de los metadatos del producto, por ejemplo, "Formato" es "Tapa dura".
- **reviewText** - Texto de la reseña.
- **overall** - Puntuación de la reseña.
- **summary** - Resumen de la reseña.
- **unixReviewTime** - Fecha de la reseña (formato unix).
- **reviewTime** - Fecha de la reseña (formato raw).
- **image** - Imágenes subidas por los usuarios luego de recibir el producto.

Figura 1: Ejemplos de reseñas en el dataset. Cada reseña guarda información referida al usuario, el producto adquirido y sus detalles, un texto con la reseña y un resumen de la misma, así como el puntaje final de evaluación dado para el producto (*rating*).

En la Figura 2 se muestra un ejemplo de los metadatos del producto, con la descripción del mismo y otros campos relevantes.


```
{
  "asin": "0000031852",
  "title": "Girls Ballet Tutu Zebra Hot Pink",
  "feature": ["Botiquecutie Trademark exclusive Brand",
              "Hot Pink Layered Zebra Print Tutu",
              "Fits girls up to a size 4T",
              "Hand wash / Line Dry",
              "Includes a Botiquecutie TM Exclusive hair flower
bow"],
  "description": "This tutu is great for dress up play for your
little ballerina. Botiquecute Trade Mark exclusive brand. Hot Pink
Zebra print tutu.",
  "price": 3.17,
  "imageURL": "http://ecx.images-amazon.com/images
/I/51fAmVkTbyL._SY300_.jpg",
  "imageURLHighRes": "http://ecx.images-amazon.com/images
/I/51fAmVkTbyL.jpg",
  "also_buy": ["B00JHONN1S", "B002BZX8Z6", "B00D2K1M3O",
"0000031909", "B00613WDTQ", "B00D0WDS9A", "B00D0GCI8S", "0000031895",
"B003AVKOP2", "B003AVEU6G", "B003IEDM9Q", "B002R0FA24", "B00D23MC6W",
"B00D2K0PA0", "B00538F5OK", "B00CEV86I6", "B002R0FABA", "B00D10CLVW",
"B003AVNY6I", "B002GZGI4E", "B001T9NUFS", "B002R0F7FE", "B00E1YRI4C",
"B008UBQZKU", "B00D103F8U", "B007R2RM8W"],
  "also_viewed": ["B002BZX8Z6", "B00JHONN1S", "B008F0SU0Y",
"B00D23MC6W", "B00AFDOPDA", "B00E1YRI4C", "B002GZGI4E", "B003AVKOP2",
"B00D9C1WBM", "B00CEV8366", "B00CEUX0D8", "B0079ME3KU", "B00CEUWY8K",
"B004FOEEHC", "0000031895", "B00BC4GY9Y", "B003XRKA7A", "B00K18LKX2",
"B00EM7KAG6", "B00AMQ17JA", "B00D9C32NI", "B002C3Y6WG", "B00JLL4L5Y",
"B003AVNY6I", "B008UBQZKU", "B00D0WDS9A", "B00613WDTQ", "B00538F5OK",
"B005C4Y4F6", "B004LHZ1NY", "B00CPHX76U", "B00CEUWUZC", "B00IJVASUE",
"B00GOR07RE", "B00J2GTM0W", "B00JHNSNSM", "B003IEDM9Q", "B00CYBU84G",
"B008VV8NSQ", "B00CYBULSO", "B00I2UHSZA", "B005F50FXC", "B007LCQI3S",
"B00DP68AVW", "B009RXWNSI", "B003AVEU6G", "B00HSOJB9M", "B00EHAGZNA",
"B0046W9T8C", "B00E79VW6Q", "B00D10CLVW", "B00B0AVO54", "B00E95LC8Q",
"B00GOR92SO", "B007ZN5Y56", "B00AL2569W", "B00B608000", "B008F0SMUC",
"B00BFXLZ8M"],
  "salesRank": {"Toys & Games": 211836},
  "brand": "Coxlures",
  "categories": [["Sports & Outdoors", "Other Sports", "Dance"]]
}
```


- **asin** - Identificador del producto.
- **title** - Nombre del producto.
- **feature** - Características del producto.
- **description** - Descripción del producto.
- **price** - Precio en dólares del producto.
- **imageURL** - Url de la imágen del producto.
- **imageURLHighRes** - Url de la imágen en alta resolución del producto.
- **related** - Productos relacionados:
    - **also_bought** - también comprado,
    - **also_viewed** - también visto,
    - **bought_together** - comprado en conjunto,
    - **buy_after_viewing** - comprado luego de visto.
- **salesRank** - Información del ranking de ventas.
- **brand** - Marca.
- **categories** - Categorías a las que pertenece el producto.
- **tech1** - Primera tabla de detalles técnicos del producto.
- **tech2** - Segunda tabla de detalles técnicos del producto.
- **similar** - Tabla de productos similares.

Figura 2: Ejemplo de instancia de un producto en el dataset. El mismo incluye el código del producto, su título, características, descripciones, categorías a las que pertenece, imágenes del mismo, la tienda que lo vende y a qué precio, así como su rango de ventas dentro de la categoría. Adicionalmente, se incluye información de los códigos de productos que fueron comprados o vistos en conjunto por los usuarios.

Amazon Fashion	reviews (883,636 reviews)	metadata (186,637 products)
All Beauty	reviews (371,345 reviews)	metadata (32,992 products)
Appliances	reviews (602,777 reviews)	metadata (30,459 products)
Arts, Crafts and Sewing	reviews (2,875,917 reviews)	metadata (303,426 products)
Automotive	reviews (7,990,166 reviews)	metadata (932,019 products)
Books	reviews (51,311,621 reviews)	metadata (2,935,525 products)
CDs and Vinyl	reviews (4,543,369 reviews)	metadata (544,442 products)
Cell Phones and Accessories	reviews (10,063,255 reviews)	metadata (590,269 products)
Clothing Shoes and Jewelry	reviews (32,292,099 reviews)	metadata (2,685,059 products)
Digital Music	reviews (1,584,082 reviews)	metadata (465,392 products)
Electronics	reviews (20,994,353 reviews)	metadata (786,868 products)
Gift Cards	reviews (147,194 reviews)	metadata (1,548 products)
Grocery and Gourmet Food	reviews (5,074,160 reviews)	metadata (287,209 products)
Home and Kitchen	reviews (21,928,568 reviews)	metadata (1,301,225 products)
Industrial and Scientific	reviews (1,758,333 reviews)	metadata (167,524 products)
Kindle Store	reviews (5,722,988 reviews)	metadata (493,859 products)
Luxury Beauty	reviews (574,628 reviews)	metadata (12,308 products)
Magazine Subscriptions	reviews (89,689 reviews)	metadata (3,493 products)
Movies and TV	reviews (8,765,568 reviews)	metadata (203,970 products)
Musical Instruments	reviews (1,512,530 reviews)	metadata (120,400 products)
Office Products	reviews (5,581,313 reviews)	metadata (315,644 products)
Patio, Lawn and Garden	reviews (5,236,058 reviews)	metadata (279,697 products)
Pet Supplies	reviews (6,542,483 reviews)	metadata (206,141 products)
Prime Pantry	reviews (471,614 reviews)	metadata (10,815 products)
Software	reviews (459,436 reviews)	metadata (26,815 products)
Sports and Outdoors	reviews (12,980,837 reviews)	metadata (962,876 products)
Tools and Home Improvement	reviews (9,015,203 reviews)	metadata (571,982 products)
Toys and Games	reviews (8,201,231 reviews)	metadata (634,414 products)
Video Games	reviews (2,565,349 reviews)	metadata (84,893 products)

Figura 3: Categorías de ventas disponibles en el dataset. Al lado de cada categoría se incluye la cantidad de reviews asociados a productos en estas categorías, así como la cantidad de productos (metadata) asignados.

Los productos tienen asignadas diferentes categorías las cuales permiten clasificarlos y agruparlos; este dato se encuentra en el atributo *categories* del metadata del producto. Las categorías corresponden a las listadas en la Figura 3.

Métodos propuestos

De acuerdo a [4], se pueden identificar seis grandes tipos de recomendaciones. Este trabajo buscó centrarse en aquellas que fueran más accesibles teniendo en cuenta los datos con los que se contaba. Entonces, se pudieron identificar las siguientes recomendaciones a explorar con los datos obtenidos:

- **Intereses de los usuarios:** Para recomendar un producto es útil considerar las compras anteriores que ha realizado el usuario, así como las calificaciones que ha dado. También se puede tener en cuenta los productos que ha buscado y su localización geográfica. Para el dataset actual, sólo se contaba con las reseñas hechas por cada usuario sobre un producto como “prueba de compra”.
- **Productos frecuentemente comprados en conjunto:** Recomendar productos que suelen comprarse juntos le brinda más información al usuario de lo que está adquiriendo. Para el dataset mencionado en [2], se contó con información de productos comprados en conjunto con otros (tag related del metadata), por lo que éste tipo de recomendación fue una opción a explorar.

Se buscó, entonces, recomendar productos que pudieran ser de interés a un usuario basándose en las reseñas realizadas por este y otros usuarios en distintos productos. Adicionalmente, se planificó explorar un recomendador que identificara productos asociados o comprados en conjunto, y complementara la información dada por las reseñas.

Limpieza de datos

Como primer paso fue necesario decidir cuáles campos serían de utilidad en la toma de decisiones y cuáles no debían tenerse en cuenta para el análisis, y procesar estos campos para asegurar que no existían incongruencias.

Tal como advirtieron los autores del dataset, fue posible encontrar distintos problemas de calidad, principalmente asociados a la atomicidad de los productos o las reseñas. Dado que Amazon asigna la misma reseña a más de un producto cuando los considera iguales, era de esperar que varias de las reseñas se encontraran duplicadas en el dataset. Otras consideraciones a tener en cuenta, especialmente al utilizar el dataset reducido, fue si todas las claves de producto y claves de usuario tenían un producto o reseña asignado (por ejemplo, podría suceder que una clave de producto se encuentre en el listado de “also_buy” pero no tenga una instancia de producto asociada). Se recomienda, en caso de encontrar tales casos, en un principio descartarlos.

Respecto de los campos a utilizar, dado que interesa realizar recomendaciones basados en ratings, así como en productos similares, se utilizarán los siguientes campos de cada documento del dataset:

- De *reviews* se conservaron los ID de usuario y producto, así como su puntuación (“overall”). Opcionalmente podría incluirse el texto dado en “summary” o en “reviewText”.
- De *metadata* se planeó usar el identificador del producto y los códigos listados en “also_buy”, “bought_together” y opcionalmente “also_viewed”. En una instancia

posterior se podría evaluar incluir la categoría del producto (ya que productos de la misma categoría podrían adquirirse en conjunto para una sola compra), así como la lista de productos similares del tag “similar”.

Extracción de características

El procesamiento de cada clase (reseñas y productos) implicó la extracción de características de los datos obtenidos. Dado que no se tuvo, en primera instancia, campos con texto, bastó con considerar la relación entre productos, usuarios y puntuaciones, representado en una matriz esparsa de tamaño número de usuarios X número de productos, donde cada celda guardaba el puntaje asignado.
Para el entrenamiento se considera un conjunto de test comprendido entre el 60% y el 80% de los datos, mientras que en test se emplea el porcentaje restante. Se entrena con los datos de train y se evalúa el desempeño del modelo pasándole una parte de los datos de test cortando algunas columnas con productos (subconjunto de observaciones) mientras se le pide que prediga el comportamiento de cada usuario de test sobre los productos restantes (subconjunto de test) [6].
Es preferible realizar un muestreo estratificado entre las distintas categorías y reviews. Para verificar que los conjuntos de test con respecto al de entrenamiento estén equilibrados se podría observar de manera analítica por medio de una tabla y/o de forma gráfica utilizando un gráfico de barras que muestre la distribución para cada conjunto.
Adicionalmente, se podría evaluar la aplicación de técnicas de submuestreo, para equiparar la cantidad de productos y reseñas en cada categoría de producto.
De acuerdo a [6] y [7], se pueden aplicar métodos de reducción de dimensionalidad basados en factorización de matrices (SVD o PCA) o autoencoders (mediante redes neuronales).

Modelos de aprendizaje supervisado

Las consultas a resolver seguían lineamientos similares a los problemas de recomendación basados en filtros colaborativos. Las técnicas de filtrado colaborativo tienen como objetivo completar las entradas faltantes de una matriz de asociación de usuario-producto. El enfoque de filtrado colaborativo se basa en la idea de que las mejores recomendaciones provienen de personas que tienen gustos similares. En otras palabras, utiliza el histórico de las calificaciones de productos de usuarios con gustos similares para predecir cómo alguien calificaría un elemento.
De acuerdo a [8] y [9], los abordajes más habituales en filtros colaborativos fueron análisis de vecindario, mediante modelos como el K-vecinos más cercanos (KNN). Adicionalmente, en [9] se planteó el uso de otros modelos (supervisados o no) como Clustering Bayesiano o Máquinas de soporte vectorial (SVM).
Considerando lo anterior, se planteó implementar recomendadores para los tipos mencionados a partir de modelos KNN y SVM (especialmente los dados por la librería *surprise*), comparando sus resultados a partir del error cuadrático medio (RMSE de *sklearn.metrics.mean_squared_error*) así como de las matrices de confusión resultantes al evaluar cada modelo. La última es especialmente útil para poder visualizar de manera rápida cuál modelo arroja mejores resultados, teniendo en cuenta, además, los valores de las variables F1, sensibilidad, precisión, exactitud y DCG sobre el subconjunto de test [6].

Las evaluaciones se realizan aplicando, además, algoritmos de validación cruzada (también incluidos en el paquete *surprise*).

Referencias

[1] Amazon, https://www.amazon.com/.

[2] Amazon review data 2018, disponible en: https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/

[3] Ni, J., Li, J., McAuley, J. (2019) Justifying recommendations using distantly-labeled reviews and fine-grained aspects. *Empirical Methods in Natural Language Processing (EMNLP)*

[4] Amazon's recommendation algorithm drives 35% of its sales, https://evdelo.com/amazons-recommendation-algorithm-drives-35-of-its-sales/

[5] Consumer trust relies heavily on reviews and brands honesty, https://www.insiderintelligence.com/content/consumer-trust-relies-heavily-on-reviews-and-brand-honesty

[6] Kordìk, P. (2019, December 15). Machine learning for Recommender Systems - Part 1 (algorithms, evaluation and cold start). Extraído de: https://medium.com/recombee-blog/machine-learning-for-recommender-systems-part-1-algorithms-evaluation-and-cold-start-6f696683d0ed

[7] Ajitsaria, A. (2022, August 18). *Build a recommendation engine with collaborative filtering*. Real Python. Extraído de: https://realpython.com/build-recommendation-engine-collaborative-filtering/#algorithms-based-on-k-nearest-neighbours-k-nn

[8] Ekstrand, M. D. (2011). *Collaborative Filtering Recommender Systems. Foundations and Trends® in Human–Computer Interaction, 4(2), 81–173.* doi:10.1561/1100000009

[9] Shah, K., Salunke, A., Dongare, S., & Antala, K. (2017). *Recommender systems: An overview of different approaches to recommendations. 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).* doi:10.1109/iciiecs.2017.8276172