Tecnológico de Monterrey

Análisis de grandes volúmenes de datos (Gpo 10)

Equipo 24

Análisis de caso - Ventajas y limitaciones de diferentes *frameworks* de procesamiento de *Big Data*

Luis Salomón Flores Ugalde - A00817435
Oscar Israel Lerma Franco - A01380817
Alejandro Guzmán Chávez - A01795398

25/05/2024

En esta entrega es necesario realizar un reporte donde se enlisten los siguientes aspectos:

- Realiza la implementación de al menos un algoritmo de recomendación avanzado (por ejemplo, factorización matricial, enfoques basados en aprendizaje profundo). La evidencia se debe poner en el repositorio GitHub del equipo.
- Identifica y justifica las métricas de evaluación utilizadas para evaluar el desempeño de los sistemas de recomendación, con el proyecto elegido por equipo.
- Resultados y análisis. Enlista al menos 3 recomendaciones donde se muestran los resultados obtenidos del punto 1 (implementación del algoritmo). La evidencia se debe poner en el repositorio GitHub del equipo.

Criterio	Valor
Portada con datos completos de los integrantes	5
Descripción del algoritmo de recomendación avanzado elegido.	25
Identificación y justificación de métricas de evaluación utilizadas para evaluar el desempeño de los sistemas de recomendación (que aplican al proyecto elegido por el equipo)	35
Experimentación con al menos 1 algoritmo de recomendación básico.	35
Total	100

Implementación de algoritmo de recomendación:

Algoritmo de recomendación *item-item*

Debido al formato de la información *steam_games*[1], en la base de datos:

steam_games_df

	publisher	genres	app_name	title	url	release_date	tags	discount_price	reviews_url	specs	price	early_access	id	developer	sentiment	metascore
0	Kotoshiro	[Action, Casual, Indie, Simulation, Strategy]	Lost Summoner Kitty	Lost Summoner Kitty	http://store.steampowered.com/app/761140/Lost_...	2018-01-04	[Strategy, Action, Indie, Casual, Simulation]	4.49	http://steamcommunity.com/app/761140/reviews/?...	[Single-player]	4.99	False	761140	Kotoshiro	NaN	NaN
1	Making Fun, Inc.	[Free to Play, Indie, RPG, Strategy]	Ironbound	Ironbound	http://store.steampowered.com/app/643980/Ironb...	2018-01-04	[Free to Play, Strategy, Indie, RPG, Card Game...	NaN	http://steamcommunity.com/app/643980/reviews/?...	[Single-player, Multi-player, Online Multi-Pla...	Free To Play	False	643980	Secret Level SRL	Mostly Positive	NaN
2	Poolians.com	[Casual, Free to Play, Indie, Simulation, Sports]	Real Pool 3D - Poolians	Real Pool 3D - Poolians	http://store.steampowered.com/app/670290/Real_...	2017-07-24	[Free to Play, Simulation, Sports, Casual, Ind...	NaN	http://steamcommunity.com/app/670290/reviews/?...	[Single-player, Multi-player, Online Multi-Pla...	Free to Play	False	670290	Poolians.com	Mostly Positive	NaN
3	彼岸领域	[Action, Adventure, Casual]	弹炸人2222	弹炸人2222	http://store.steampowered.com/app/767400/2222/	2017-12-07	[Action, Adventure, Casual]	0.83	http://steamcommunity.com/app/767400/reviews/?...	[Single-player]	0.99	False	767400	彼岸领域	NaN	NaN

(El dataset contiene reseñas públicas de la plataforma de venta de videojuegos Steam. Cada reseña incluye el distribuidor, genero, título, Nombre del juego, URL de la reseña, etiquetas relevantes (como genero o subgenero), fecha de estreno, descuento, URL del juego en Steam, Especificaciones (Multi o Solo un jugador), precio en dólares, si es early access, id del juego, desarrollador, sentimiento (reseña positiva o negativa), y puntaje metascore.

Se concluyó que el sistema de recomendación más apropiado en este segundo avance es *item-item*. Nótese que existen múltiples reseñas por usuario, y cada videojuego posee múltiples reseñas. Si se desea realizar recomendaciones al usuario de videojuegos a consumir, estas deberían estar basadas en *juegos* ya experimentados, *más* que en todo el historial del usuario (*user-user)*.

Item-Item funciona de la siguiente manera:

- Cada elemento (reseña), puede ser representado por un vector, (por ejemplo: convertir atributos relevantes de cada reseña, en nuestro caso generos y tags, como matrices booleanas).
- Por cada elemento, calcula la *similitud vectorial* con cada otro elemento en el dataset.
- La métrica para lo anterior escogida es similitud de Coseno. Se basa en el coseno del ángulo entre dos vectores (c.e., *A* y *B)* en un espacio multidimensional. Entre más pequeño es el ángulo, mayor es la similitud entre los vectores, y se calcula de la siguiente forma:

  $$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta$$

  El producto punto euclidiano de dos vectores ó

[1] Obtenidos de Kang W., McAuley J.

2

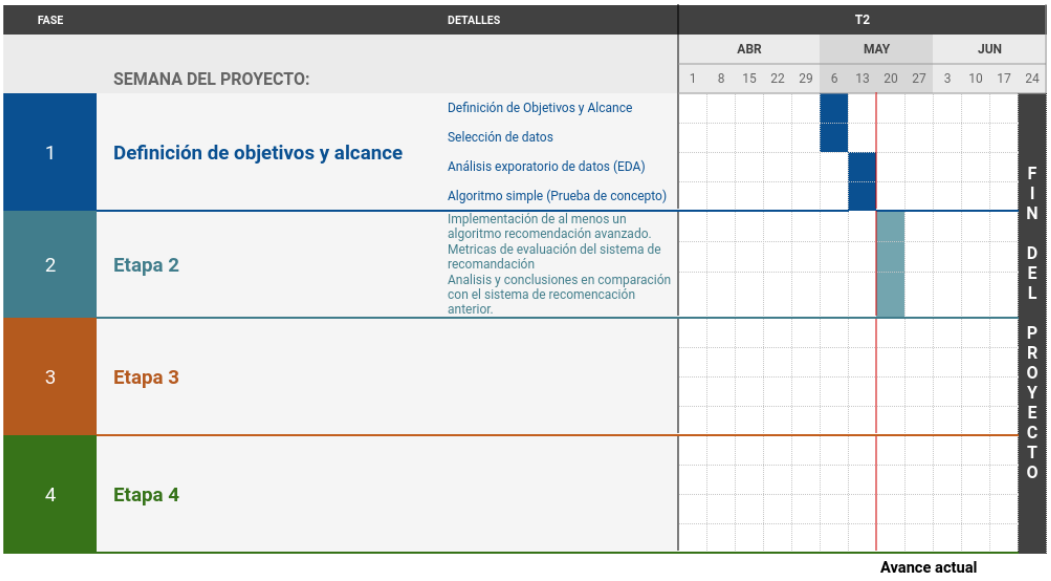
$$S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}},$$

En este ejercicio, utilizamos la función de scikit-learn de cosine_similarity para generar la matriz de similitud de tipo (n, n). Esta matriz nos permitirá extraer juegos que sean parecidos al videojuego de interés.

- Generar la matriz de similitud, donde cada entrada (i, j) refleja la similitud del elemento i con el elemento j.
- Generar la recomendación, dado un elemento que el usuario ha interactuado, identificar los elementos más similares dentro de la matriz y listarlos por similitud.

Cronograma del Avance 2

TÍTULO DEL PROYECTO	Implementación de modelo de recomendación	EMPRESA	INSTITUTO TECNOLÓGICO DE MONTERREY
RESPONSABLE DEL PROYECTO	Equipo 24	FECHA	26/05/24



Métricas de evaluación

Entre las métricas de precisión utilizadas:

[2] Wikimedia, https://en.wikipedia.org/wiki/Cosine_similarity

- Precisión@K

  La *precisión a 5* mide la proporción de elementos recomendados entre los 5 primeros que son relevantes.

  Valor obtenido: 0,17696472518457754

  Interpretación: Por término medio, el 17,7% de los artículos de las 5 primeras recomendaciones son pertinentes. Esto significa que por cada 5 artículos recomendados, aproximadamente 0.89 son relevantes.

  Justificación: Métrica sencilla para obtener qué elementos recomendados poseen valor para el usuario. (El problema es que este solo muestra relevancia binaria, o es útil o no lo es, no existen grados de relevancia.)

- Recall@K

  Recall a 5 mide la proporción de elementos relevantes que se han recuperado en las 5 primeras recomendaciones.

  Valor: 0,6471694450955116

  Interpretación: Esto significa que de todos los elementos relevantes, aproximadamente el 64,7% se encuentran en las 5 primeras recomendaciones.

  Justificación: Poner a prueba la habilidad del sistema para obtener los elementos más relevantes de todas las opciones posibles. (Nota: esto no refleja la posición de los elementos recomendados.)

- MAP@K

  La precisión media a 5 (MAP@5) mide la media de las puntuaciones medias de precisión para cada usuario, considerando las 5 primeras recomendaciones.

  Valor: 0,3076029304530125

  Interpretación: La relevancia del ranking en el orden de recomendación actual es aproximadamente del 30,8%.

  Justificación: Mostrar la calidad del orden de recomendaciones, útil para medir la calidad de las primeras recomendaciones que el sistema arroja al usuario (Es importante que la primera opción sea la más relevante). Pero igual que Presision@k, este solo posee relevancia binaria. Esta métrica combina la precisión y *recall* teniendo en cuenta el orden de las recomendaciones y la posición de los elementos relevantes.

- NDCG@K

Normalized Discounted Cumulative Gain at 5 (NDCG@5) mide la calidad de clasificación de las recomendaciones teniendo en cuenta la posición de los elementos relevantes en el top 5. (Esta operación es más compleja, y requiere del cálculo de *puntajes de relevancia* por elemento.)

Valor: 0,4110959049914195

Interpretación: La calidad de clasificación de las recomendaciones es 41,1% de la clasificación ideal (IDCG) en la que todos los elementos relevantes se encuentran en los primeros puestos.

Justificación: Parecido al anterior, medir la relevancia de las recomendaciones por su ranking, en comparación a un ranking ideal (donde las recomendaciones están mejor ordenadas con su relevancia).

- MRR@K

  El rango recíproco medio a 5 (MRR@5) mide la media de los rangos recíprocos del primer elemento relevante en las 5 primeras recomendaciones.

  Valor: 0,38881460213289576

  Interpretación: El rango recíproco medio es de aproximadamente 38,9%. Esto significa que, en promedio, el primer elemento relevante aparece en torno a la posición 2,57 en las 5 primeras recomendaciones (1 / 0,38881460213289576 ≈ 2,57).

  Justificación: En conjunto con las respuestas anteriores, se busca encontrar en qué lugar realmente aparece la mejor recomendación para el usuario, esta métrica es directamente relevante al objetivo de recomendar al usuario la mejor opción en primer lugar.

Resultados y análisis

Con los valores obtenidos anteriormente, observamos:

- **Alto *recall*, baja precisión**: El sistema recupera una gran proporción de elementos relevantes (*recall* de 64.7%), pero también incluye muchos irrelevantes. Identifica “bien” los elementos potencialmente relevantes, pero tiene dificultades para filtrarlos con precisión.
- **MAP y NDCG medios**: Los valores mediados (no tan altos ni bajos) sugieren que, aunque el sistema identifica elementos relevantes, su

ordenación podría mejorarse. Los elementos relevantes no se sitúan sistemáticamente en los primeros puestos.

- **MRR medio**: la presencia de elementos relevantes cerca de la parte superior de la lista es “relativamente buena” (con un 38.9%), pero se debe enfocar su mejora.

Conclusión

El sistema de recomendación *item-item*, mostró indicios de progreso relevante en comparación con el sistema de recomendación anterior (SVD simple):

Precisión@5(SVD)[3]	**Precisión@5(item-item)**
0.0407844	0.1769647251

Sin embargo, para futuros pasos:

- **Mejorar la precisión (énfasis en MRR):** Centrarse en perfeccionar el algoritmo de recomendación para distinguir mejor entre artículos relevantes e irrelevantes. La mejora en la ingeniería de características (tanto en el preprocesamiento como el uso de más atributos que el género o tags) o/y el uso de modelos más sofisticados podrían ayudar.
- **Equilibrar Recall y Precisión**: Aunque un *recall* alto es bueno, es crucial equilibrarse con la precisión. Unas recomendaciones demasiado amplias (baja precisión) pueden diluir la satisfacción del usuario, donde se busca la mejor recomendación en la más alta posición.

GITHUB:

Notebook directo:
https://github.com/VF1Gimure/MNA_GVD24/blob/main/Proyecto_Avance_2_24.ipynb

GIT: https://github.com/VF1Gimure/MNA_GVD24/tree/main

[3] Vease: https://github.com/VF1Gimure/MNA_GVD24/blob/main/Proyecto_Avance_1%2324.ipynb

Bibliografía

1. Marabelli, M; Saunders, C; y Wiener, M. (2020). Big-data business models: A critical literature review and multiperspective research framework. 35(1), 66-91. https://journals.sagepub.com/doi/reader/10.1177/0268396219896811.
2. Kang W., McAuley J. (*ICDM*, 2018) **Self-attentive sequential recommendation.** UC San Diego pdf