



Tecnológico de Monterrey

Análisis de grandes volúmenes de datos (Gpo 10)

Equipo 24

Análisis de caso - Retos éticos del análisis de Big
Data

Luis Salomón Flores Ugalde - A00817435

Oscar Israel Lerma Franco - A01380817

Alejandro Guzmán Chávez - A01795398

17/05/2024

Análisis de caso - Retos éticos del análisis de Big Data	1
Plan de proyecto de acuerdo a la industria elegida por el equipo.	4
Industria de entretenimiento en videojuegos	4
Cronograma del Avance 1	5
Justificación de selección del conjunto de datos utilizados	5
Descripción de los pasos del preprocesamiento	5
Bibliografía	6

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

- Genera un plan de proyecto de acuerdo con la industria elegida por tu equipo en la actividad 2.2 y detalla el plan del proyecto con su cronograma.
- Justifica la selección del conjunto de datos utilizado y describe los pasos de preprocesamiento.
- Realiza al menos un ejercicio de exploración inicial y análisis del conjunto de datos de la industria elegida (la evidencia se debe poner en el repositorio GitHub del equipo).
- Programa al menos un 1 algoritmo de recomendación básico con el conjunto de datos elegido (la evidencia se debe poner en el repositorio GitHub del equipo).

Criterio	Valor
Portada con datos completos de los integrantes	5
Descripción general del plan del proyecto y el cronograma correspondiente solamente al avance 1 de proyecto, en los siguientes avances se estará actualizando.	20
Descripción del conjunto de datos utilizado y sus pasos de preprocesamiento.	25
Exploración inicial y análisis del conjunto de datos.	25
Experimentación con al menos 1 algoritmo de recomendación básico.	25
Total	100

Plan de proyecto de acuerdo a la industria elegida por el equipo.

Industria de entretenimiento en videojuegos

El proyecto busca desarrollar un sistema de recomendación donde pretende predecir si un usuario disfrutará un juego en específico. Al analizar muestra de reviews, identificamos patrones clave e insights para dar información personalizada en las recomendaciones.

El objetivo principal es *demostrar que la analítica con big data puede mejorar la experiencia de los jugadores al permitir el descubrimiento de posibles juegos que no se estén en su consideración inmediata.*

En seguida describiremos el primer avance junto a posibles futuras mejoras y cambios:

Como primer avance se aplica un sistema muy básico de recomendaciones basadas únicamente en la opinión de otros usuarios en términos básicos de 'sí se recomienda' y 'no se recomienda' (como 1 y 0 respectivamente).

Para esto primero utilizaremos descomposición en valores singulares (SVD) tomando en cuenta únicamente este valor singular. Debemos destacar que, por el momento, estaremos utilizando los datos de reviews de Australia, esto para reducir el tamaño y poder después escalar con todos los datos del mercado global.

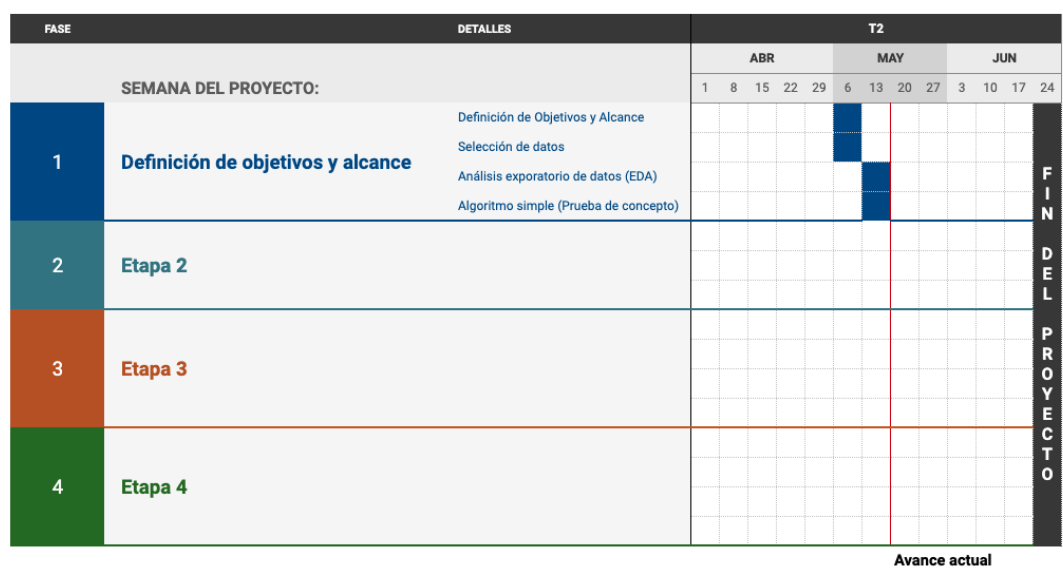
Mencionamos que primero probaremos con el sistema de esta manera. Pero a futuras mejoras, tenemos mucha más información disponible aparte de la mencionada anteriormente. El set de datos cuenta con reviews escritas por los usuarios y también tenemos el catálogo completo de juegos, los cuales también tiene descripciones en diferentes maneras, cómo tags, géneros y texto, entre otros. La cantidad de información disponible es numerosa y debería ser implementada por secciones. Esto nos lleva a cada mejora disponible a futuro, puesto que podemos adentrarnos a la información desde varios puntos. Estos pueden ser por más tags presentes o incluso aplicando técnicas en conjunto de procesamiento de lenguaje natural.

Dada la naturaleza de publicación de videojuegos en esta plataforma, es necesario tomar en cuenta que uno puede o no tener un 'publisher' oficial, y esto lleva a cambios que se tienen que manejar adecuadamente o de forma equivalente e información que beneficia a algunas y debería ser utilizada pero es entendible que la gran mayoría no cuente con ella por su naturaleza.

Cada mejora debería intentar implementar una sección de información disponible más a la anterior y así, al final obtendremos un modelo de recomendación superior el cual toma en cuenta factores diferentes.

Cronograma del Avance 1

TÍTULO DEL PROYECTO	Implementación de modelo de recomendación	EMPRESA	INSTITUTO TECNOLÓGICO DE MONTERREY
RESPONSABLE DEL PROYECTO	Equipo 20	FECHA	17/05/24



Justificación de selección del conjunto de datos utilizados

El conjunto de datos utilizados provienen de STEAM. El dataset cuenta con 7,793,069 reviews, 2,567,538 usuarios, y 32,135 juegos. Este es el paquete completo. Nosotros en este primer avance solo utilizaremos la información de aquí correspondiente a Australia (usuarios y reviews). Esto para poder hacer pruebas y mejorar el modelo antes de aplicarlo al set de datos completos. Aparte que, si tomamos los datos completos de forma inmediata, algunas de las mejoras en otros avances tendrían que tomar o quitar entradas que no estén en inglés. Con el set de datos Australianos tenemos un control superior que nos permitirá después aplicar cambios necesarios y más certeros en secciones específicas.

Descripción de los pasos del preprocesamiento

Primero, decidimos limpiar el json inicial. Esto para evitar usar código peligroso de python como “eval”.

Volvimos a recrear el json de forma limpia para poder leerlo de manera simple con pd.read_json (lines) en futuras exploraciones.

De manera inicial hicimos esto con los juegos, usuarios australianos y reviews australianas.

Después de obtener esto, aún tenemos que extraer las reviews individuales con el propósito de aplicar esto a svd.

Transformamos el 'recommend' que tiene su información en True y False a 1 y 0. y En esta misma columna, eliminamos las entradas erróneas que contienen NaN (el sistema de steam no debería tener estas si son publicadas, por lo que asumimos que son errores (también tiene NaN en todas las otras entradas por lo que confirmamos que es ruido). La cantidad de estos no llega a ser ni siquiera un 0.01% de los datos por lo que decidimos eliminarlos.

En este paso, contamos con más información adicional de la review pero al momento, en este primer avance no tenemos uso ni una forma específica de trabajar en ella por lo que decidimos columnas como 'funny', 'helpful', 'posted', 'last_edited', 'user_url'. Algunas de estas puede que sean utilizadas después, tentativamente diríamos que la columna opcional "helpful" tiene más capacidad de ayudarnos a seleccionar información adecuada para procesamiento de tipo NLP a futuro.

Pero de momento, nos quedamos únicamente con "user_id", "recommend" y "item_id" para esta primera aplicación.

En el preprocesamiento, y hablando de item_id. Para tener nombres con los datos que tenemos. Quitamos de los datos los reviews de juegos que no existen dentro de nuestro dataset de juegos. Esto únicamente para evitar dar resultados únicamente con un ID y no con el título. Del dataset de reviews terminamos eliminando alrededor de mil entradas. Este paso puede ser más opcional si así lo deseamos. O bien con el id podemos aumentar nuestro dataset adquiriendo directamente los datos faltantes, por el momento dejamos estas entradas por fuera.

Por ahora el set de juegos solo ocupa limpiar sus tipos para poder hacer la conexión después con el set de reviews cuando ya tengamos las recomendaciones iniciales.

Conclusión

Al escogerse los datos de usuarios de Australia como dataset reducido, utilizamos atributos con "user_id", "recommend" y "item_id" para esta primera aplicación, para implementar un simple algoritmo SVD básico, de recomendación colaborativa de usuarios con recomendaciones positivas similares de juegos. En el análisis de procesamiento, sobresale una discrepancia un sesgo positivo del atributo *recomendado*, casi 88% (52473) de los registros muestra *SI* en recomendación.

Dada la simplicidad del modelo, este ofrece un rating predictivo de 0.04, sin embargo, es capaz de obtener resultados consistentes por búsqueda de usuario.

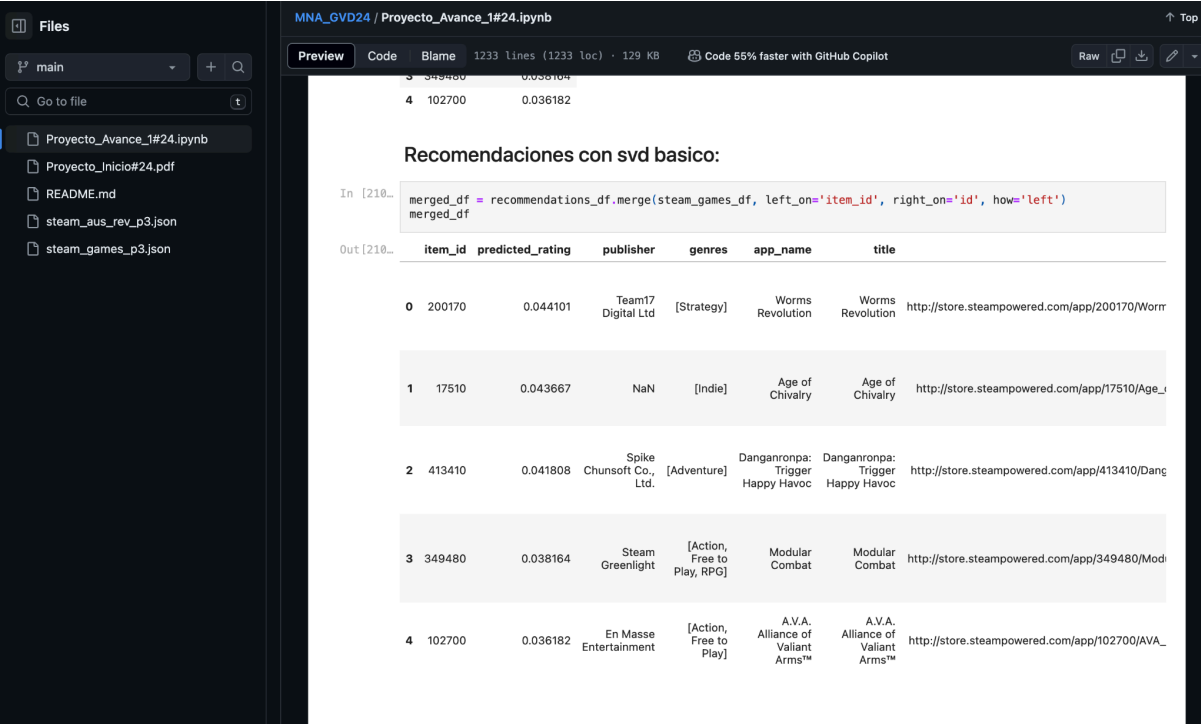
Para próximos pasos, se debería realizar la limpieza de datos para obtener relaciones dentro del dataset al momento de poseer relaciones iniciales. Y obtener dimensionalidad en casos como género, *publisher*, fecha de estreno. Además de la implementación de algoritmos más relevantes (ex. kNN) y probar con dataset de mayor volumen (ex. reseñas globales) .

GITHUB:

Notebook directo:

https://github.com/VF1Gimure/MNA_GVD24/blob/main/Proyecto_Avance_1%2324.ipynb

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



The screenshot shows a Jupyter Notebook titled "MNA_GVD24 / Proyecto_Avance_1#24.ipynb". The left sidebar displays a file explorer with the following files: "main", "Proyecto_Avance_1#24.ipynb", "Proyecto_inicio#24.pdf", "README.md", "steam_aus_rev_p3.json", and "steam_games_p3.json". The main area shows the notebook content, which includes a code cell and its output.

Recomendaciones con svd basico:

In [210]:

```
merged_df = recommendations_df.merge(steam_games_df, left_on='item_id', right_on='id', how='left')
merged_df
```

Out [210]:

	item_id	predicted_rating	publisher	genres	app_name	title
0	200170	0.044101	Team17 Digital Ltd	[Strategy]	Worms Revolution	Worms Revolution http://store.steampowered.com/app/200170/Worms_Revolution
1	17510	0.043667	NaN	[Indie]	Age of Chivalry	Age of Chivalry http://store.steampowered.com/app/17510/Age_of_Chivalry
2	413410	0.041808	Spike Chunsoft Co., Ltd.	[Adventure]	Danganronpa: Trigger Happy Havoc	Danganronpa: Trigger Happy Havoc http://store.steampowered.com/app/413410/Danganronpa_Trigger_Happy_Havoc
3	349480	0.038164	Steam Greenlight	[Action, Free to Play, RPG]	Modular Combat	Modular Combat http://store.steampowered.com/app/349480/Modular_Combat
4	102700	0.036182	En Masse Entertainment	[Action, Free to Play]	A.V.A. Alliance of Valiant Arms™	A.V.A. Alliance of Valiant Arms™ http://store.steampowered.com/app/102700/A.V.A._Alliance_of_Valiant_Arms

Bibliografía

1. Kang W., McAuley J. (ICDM, 2018) **Self-attentive sequential recommendation**. UC San Diego [pdf](#)
2. Sandhu, A. K. (2021). Big data with cloud computing: Discussions and challenges. Big Data Mining and Analytics, 5(1), 32-40.
<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9663258> [Links to an external site.](#)