Tecnológico de Monterrey

Maestría en Inteligencia Artificial Aplicada (MNA-V)

Trimestre Abr - Jul 2024

Análisis de Grandes Volúmenes de Datos

Avance - Sistema de Recomendación

Equipo #11

Daniel Guzmán Ávila	**A00781387**
Julio Cesar Ruiz Marks	
Alicia Sanchez Carmona	**A01652134**

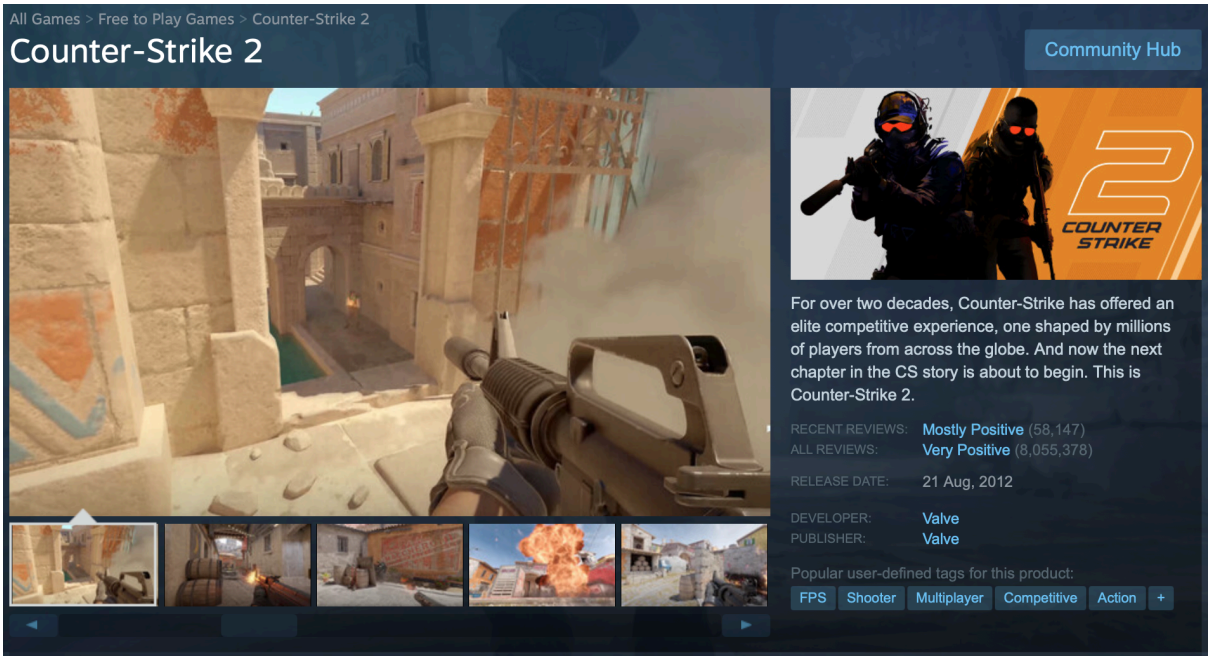
Profesor Titular: Dr. Nestor Velasco

19.05.24

Descripción del Proyecto

El proyecto consistirá en analizar un set de datos con reseñas de la plataforma Steam y experimentar con algoritmos básicos de recomendación. La plataforma Steam vende videojuegos de PC de manera completamente digital a todo el que disponga de una conexión a internet. Esta plataforma permite que desarrolladores pequeños puedan publicar juegos por su cuenta y vender directamente a usuarios finales compitiendo con desarrolladores corporativos. Además, publican periódicamente rankings de juegos por diferentes métricas como cantidad de usuarios en línea (la cual es importante para los juegos como servicio) y ventas brutas (la suma de recaudación de los ingresos por servicio y de las ventas unitarias) para toda la plataforma.

Por ejemplo, la página del juego # 1 en ingresos globales muestra la siguiente portada:



Steam implementa un sistema de calificación donde de manera abierta muestran la valoración general del público y los usuarios pueden acceder a los textos de miles de reseñas. Abajo de la descripción del juego, aparecen las reviews recientes y las que han aparecido a lo largo de toda la vida del juego junto con una etiqueta que va desde “Overwhelmingly positive” hasta “Overwhelmingly negative” dependiendo del puntaje global de los jugadores. Estas características hacen posible que se generen muchos datos sobre los gustos y tendencias de los usuarios.

Cronograma preliminar

El desarrollo de un sistema de recomendación implica una planificación y una ejecución detallada para asegurar que las etapas del proyecto se completen eficazmente y dentro del tiempo establecido. Por lo que se realizó un cronograma para organizar, gestionar el tiempo y los recursos a utilizar. Este proyecto se centrará en la creación de un sistema de recomendación utilizando un conjunto de datos con reseñas de la plataforma de juegos Steam, proporcionando recomendaciones a los usuarios.
El cronograma se encuentre dentro de la siguiente liga:
Sistema de recomendación○○

Descripción del conjunto de datos a analizar

El conjunto de datos a analizar es es steam reviews dataset, adicionalmente el equipo mostrará los pasos de preprocesamiento de ser necesarios en el caso de encontrar errores de origen. La descripción de los datos es:

```
*  date_posted : Fecha de la reseña
* funny: Cuantos jugadores piensan que la reseña es chistosa
* helpful: cuantos jugadores piensan que la reseña es útil
* hours_played: cuantas horas jugó antes de publicar la review
*  is_early_acces_review:  si  la  review  fue  publicada  en  un  programa
early access
* recommendation: Si recomienda el juego o no
* review: la reseña
* title: nombre del juego
```

Preprocesamiento

En la etapa de ED se identificó que hay muchos usuarios que dejan review jugando 3 horas o menos. Por la experiencia del equipo una persona podría emitir una opinión del juego siempre y cuando lo haya jugado, y 3 horas podría ser un plazo de confianza para eliminar a usuarios jugando con el sistema:

```
df = df[df['hour_played'] > 3]
```

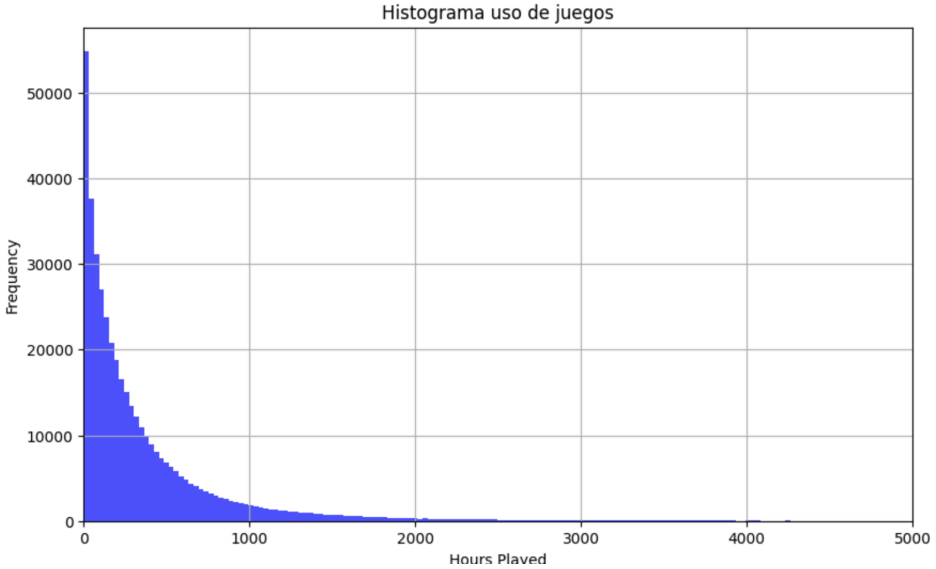
Este paso elimina del dataset toda review que haya sido dejada sin haber jugado por lo menos tres horas. Esto elimina muchas reviews vacías y reduce el riesgo del review bombing, es decir, cuando los usuarios dejan en masa reviews buenas o malas sin haber jugado el juego realmente. El sistema de steam permite dejar reviews siempre y cuando el usuario haya comprado el juego pero esto no garantiza que las borra si el usuario compra el juego, deja una review e inmediatamente solicita el reembolso.

Más adelante en el análisis nos dimos cuenta que Steam también ofrece apps y no solamente juegos, pero decidimos no buscar un preprocesamiento para eliminarlos porque no hay una lista fácilmente disponible que nos diga cuales son apps y cuales son juegos.

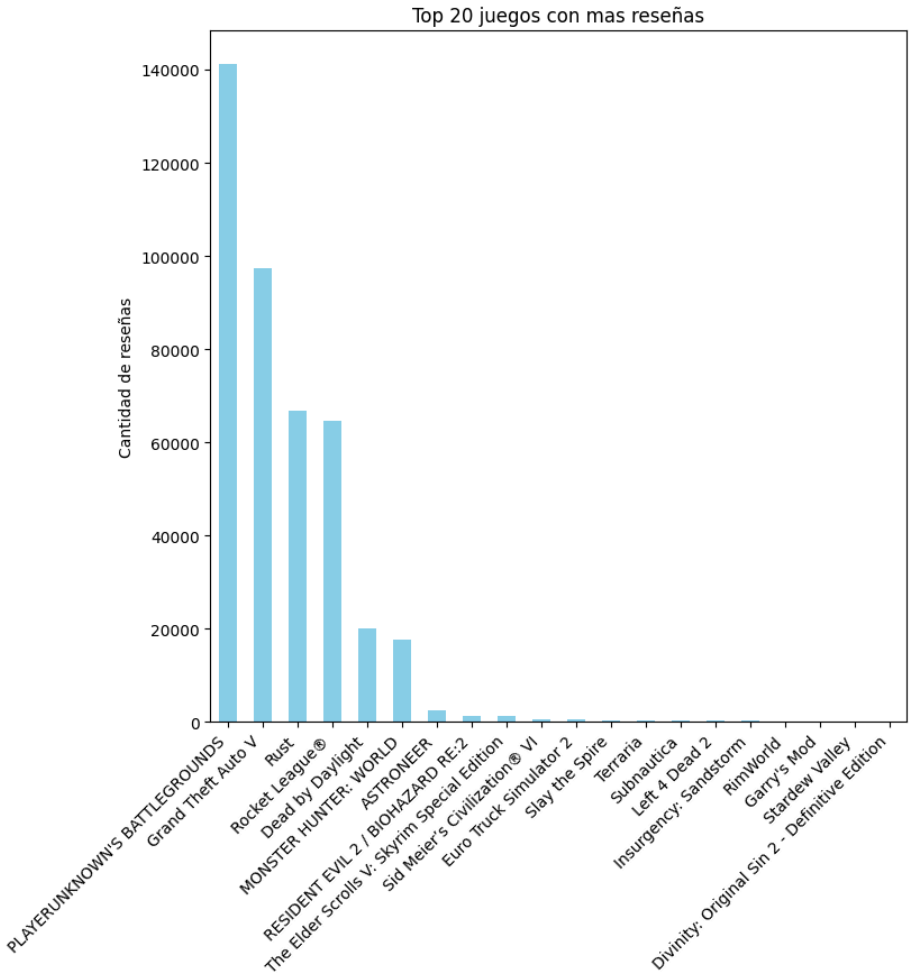
Análisis Exploratorio (EDA) Inicial

A continuación se describe el EDA aplicado al conjunto de datos. OBSERVACIÓN: Esto es un resumen, favor de buscar todo el código en el github del grupo para mayor referencia.

1) Distribución de reviews por hora jugada



2) Top de Juegos con Mayor cantidad de reseñas



3) Los 20 juegos con mayor proporción de reseñas positivas

ACE COMBAT™ 7: SKIES UNKNOWN	100.000000
Foundation	100.000000
Wallpaper Engine	100.000000
Tom Clancy's Rainbow Six® Siege	100.000000
Tannenberg	100.000000
Subnautica: Below Zero	100.000000
Subnautica	100.000000
Stardew Valley	100.000000
RimWorld	100.000000
Kenshi	100.000000
GOD EATER 3	100.000000
My Time At Portia	100.000000
Don't Starve Together	100.000000
Euro Truck Simulator 2	100.000000
Beat Saber	100.000000
Cold Waters	100.000000
Warhammer 40,000: Mechanicus	100.000000
Expansion - Hearts of Iron IV: Man the Guns	100.000000
Factorio	100.000000
Terraria	99.612403
Left 4 Dead 2	99.530516
RESIDENT EVIL 2 / BIOHAZARD RE:2	99.404762
Slay the Spire	99.227799
Garry's Mod	97.014925
ASTRONEER	96.869984
Insurgency: Sandstorm	95.673077
Human: Fall Flat	95.394737
Rocket League®	91.798995
Divinity: Original Sin 2 - Definitive Edition	90.860215
Overcooked! 2	90.000000

Demostración de un algoritmo de recomendación

Para generar que el algoritmo pueda recomendar una lista de títulos similares al registro de interés existen diferentes opciones, sin embargo, como el campo de recomendación es binario (Lo recomiendan o no) se pensó en utilizar SVD, pero el dataset no tiene un Steam_ID asociado ni el género por lo que a estas altura no es sencillo usar SVD, coseno o un sistema item-item con base en las características. A pesar de las limitaciones del dataset se puede utilizar el promedio bayesiano para hacer un top de juegos más recomendables de la plataforma.

```python
# Se aplica un promedio bayesiano para darle prioridad a los juegos que tienen mas reviews positivas considerando las totales
# Vemos que Ace Combat y Foundation a pesar de que parecen juegos buenos tienen pocas reviews
# Ahora los mas populares con criterio bayesiano son Rocket League, Astroneer y RE:2
# Con el promedio bayesiano, de los juegos con avg_rating simple de 100% solo aparece Subnautica que ahora tiene un bayesiano de 80%

df['binary_recommendation'] = df['recommendation'].apply(lambda x: 1 if x == 'Recommended' else 0)

ratings_per_game = df.groupby('title')['binary_recommendation'].agg(['count', 'mean'])
ratings_per_game.columns = ['N', 'avg_rating']

m = ratings_per_game['avg_rating'].mean()
C = ratings_per_game['N'].mean()

ratings_per_game['bayesian_avg'] = (C*m + ratings_per_game['N'] * ratings_per_game['avg_rating']) / (C + ratings_per_game['N'])

sorted_games_by_bayesian_average = ratings_per_game.sort_values(by='bayesian_avg', ascending=False)

print(sorted_games_by_bayesian_average.head(20))
```

Los resultados obtenidos son:

title	N	avg_rating	bayesian_avg
Rocket League®	64675	0.917990	0.903705
ASTRONEER	2492	0.968700	0.835523
RESIDENT EVIL 2 / BIOHAZARD RE:2	1344	0.994048	0.823673
Rust	66887	0.812385	0.810652
Euro Truck Simulator 2	477	1.000000	0.807857
Terraria	258	0.996124	0.803039
Slay the Spire	259	0.992278	0.802950
Subnautica	243	1.000000	0.802820
Left 4 Dead 2	213	0.995305	0.802043
RimWorld	203	1.000000	0.801933
Stardew Valley	199	1.000000	0.801844
MONSTER HUNTER: WORLD	17590	0.803752	0.801621
Garry's Mod	201	0.970149	0.801213
Factorio	167	1.000000	0.801127
Insurgency: Sandstorm	208	0.956731	0.801032
Don't Starve Together	158	1.000000	0.800925
Wallpaper Engine	144	1.000000	0.800609
Dead by Daylight	20109	0.801532	0.800256
Human: Fall Flat	152	0.953947	0.799997
Divinity: Original Sin 2 - Definitive Edition	186	0.908602	0.799636

Como conclusión, se sugiere incrementar el tamaño del dataset agregando datos por lo menos hasta 2023 y agregando de ser posible Steam_ID anonimizado y el género del juego.

Repositorio GitHub:
https://github.com/aliciaschz/Equipo_11/blob/main/Avance1_Sistema_Recomendacion.ipynb

Bibliografía

- *Steam Reviews Dataset*. (2019, April 3). Kaggle. https://www.kaggle.com/datasets/luthfim/steam-reviews-dataset
- *Counter-Strike 2 en Steam*. (n.d.). https://store.steampowered.com/app/730/CounterStrike_2/?l=spanish