PROYECTO INTEGRADOR FINAL

Lineamientos generales

Data Science – Digital House

2022

GRUPO 3

INTEGRANTES:

Guglielmi, Felix

Rasia, Martín

Rodriguez Elorza, Carlos

Bakken, Louise

Alvarez Hurtado, Juan José

1. ***Tema de Investigación***

El tema elegido es el análisis de las preferencias de los usuarios que visualizan contenido en la plataforma de streaming Flow con el objetivo de brindar sugerencias de otros títulos, tanto a ellos como a otros usuarios.

2. ***Antecedentes del tema***

Es de conocimiento general la proliferación en los últimos años de las plataformas de contenido digital audiovisual (Ej: Plataforma de Netflix en 2007, Amazon Prime 2006). Dichas plataformas han intentado retener y aumentar la cantidad de usuarios proponiendo contenido que sea atractivo a los más diversos gustos. Por este motivo, las plataformas han incorporado numerosos títulos, episodios, programas, etc. que atiendan la necesidad y deseo de los consumidores, especialmente en sus momentos de ocio.

De manera frecuente, las personas accedían a las plataformas para visualizar un contenido específico, sea porque habían investigado previamente, porque eran influenciados por recomendaciones de terceros (amigos, familiares, etc.), o porque habían visto una publicidad al respecto.

No obstante ello, y en línea con los últimos avances de las plataformas de streaming, sería posible también fomentar una potencial demanda adicional de otros títulos a través de herramientas de recomendaciones específicas. Una implementación de dichos sistemas sería a través de la identificación y clasificación de grupos de usuarios en función del contenido visualizado, que posteriormente al ser correlacionado con una base de datos del contenido disponible, permitan reconocer títulos apropiados para los distintos grupos.

En particular, la plataforma de streaming de Flow permite acceder públicamente a la información para poder ensayar un prototipo de herramienta de recomendaciones. Flow ofrece las siguientes descargas:

- datasets con información histórica de visualizaciones por cuenta de cliente y usuario específico (donde se identifica el título o denominación del contenido que se visualizó);
- dataset con la oferta de títulos disponible y las características de los mismos (por ej: año de emisión, país de origen) y su categorización (drama, comedia, documental, etc.).

3. ***Aporte esperado***

El objetivo principal es, dado un usuario que ha visualizado previamente una serie de contenido determinado, realizar sugerencias de contenidos a visualizar que resulten de su potencial interés.

4. ***Disponibilidad de datos e infraestructura***

La información de base para el desarrollo del trabajo está constituida por los dos datasets:

- **Dataset 1**: conteniendo la información principal de los ID de cuenta, usuario y contenido visualizado. Incluye también información adicional referida a tipo de dispositivo desde el cual se accedió, hora de inicio y fin de la reproducción.

	customer_id	account_id	device_type	asset_id	tunein	tuneout	resume
2750315	82636	102343	STB	19668.0	2021-02-22 13:45:00.0	2021-02-22 13:47:00.0	0
2543824	77824	96738	STB	4360.0	2021-02-07 00:37:00.0	2021-02-07 01:45:00.0	1
3621980	111044	25047	STB	11013.0	2021-03-09 17:23:00.0	2021-03-09 17:53:00.0	1
3275991	98527	15485	STB	10478.0	2021-02-19 14:14:00.0	2021-02-19 14:18:00.0	1
621778	21956	43750	STB	14337.0	2021-03-03 18:39:00.0	2021-03-03 19:35:00.0	0
2355274	72909	91399	STB	12111.0	2021-01-14 10:15:00.0	2021-01-14 10:17:00.0	0

- **Dataset 2**: base de datos del contenido disponible en la plataforma de streaming, con detalle del ID del contenido y características/descripción del mismo.

	asset_id	content_id	title	reduced_title	episode_title	show_type	released_year	country_of_origin	category
4257	16334	86.0	11/25 - Bienvenidos a bordo	Bienvenidos_25-11	25 Noviembre 2020	TV	2020.0	AR	Entretenimiento/Competencia
6244	8456	929.0	T:2 Ep:16 Crudo y sin censura	Crudo_y_sin__T2_E16	Episodio 16	TV	2008.0	US	Documental/Animales
32203	17037	48.0	T:2 Ep:28 Bia	Bia_T2_E28	Episodio 28	TV	2020.0	AR	Infantil/Comedia
15790	26245	2034.0	T:2 Ep:01 Run Coyote Run	T:2 Ep:01 Run Coyote Run	La mafia china	Serie	2018.0	US	Comedia/Aventura
28024	32182	139.0	01/02 - TN Internacional	TN_Internacio_02-01	2 Enero 2021	TV	2021.0	AR	Interés General/Periodístico
5155	1529	385.0	Ep:021 Separadas	Separadas_UY_E021	El camino	Serie	2020.0	AR	Romance/Drama

En línea con el objetivo trazado para el proyecto, se describen a continuación en términos generales las etapas que se contemplan dentro del sistema que será implementado:

1. **LIMPIEZA**: Análisis preliminar y limpieza de los datasets, eliminación de outliers, generación de la información base para la implementación de los algoritmos.

2. **DATA WRANGLING**: Ajuste de los datasets para poder utilizar las herramientas de generación de modelos.

3. **IMPLEMENTACIÓN**: Utilización de la librería LightFM para la generación de recomendaciones.

4. **EVALUACIÓN:** Revisión de los resultados de predicción por usuario, utilizando métricas de salubridad.

5. ***Plan de trabajo / Cronograma tentativo***