https://github.com/Datathon2021/Recomendador

El desafío: desarrollar un sistema de recomendación para predecir nuevos contenidos, no previamente vistos, son más probables a ser elegidos para ver en Flow por un grupo de usuarios en base a su historial de visualizaciones.

La personalización de las plataformas digitales de contenidos audiovisuales es un aspecto central en la calidad de la experiencia de los usuarios. Dada la amplia oferta de posibilidades, los consumidores tienden a valorar las recomendaciones sobre qué contenidos ver que sean acordes a sus preferencias.
Partiendo de las visualizaciones de un conjunto de usuarios de Flow correspondientes a un trimestre del 2021, el desafío consiste en predecir qué nuevas visualizaciones tuvo cada perfil durante el mes siguiente.

Links importantes ❗

- Campus Party Digital Edition - Argentina
- Registro Datathon Telecom - Recomendador Flow
- Entregables
- Bases y condiciones

Datasets 📂

Para el desarrollo del recomendador, se proveen los siguientes conjuntos de datos:

- train.csv
- metadata.csv
- iso_3166_1.json

Estos archivos pueden ser descargados desde aquí.

train.csv

Este dataset contiene los registros de visualizaciones de contenidos de Flow del formato video on demand (VOD), correspondiente a una muestra aleatoria de más de 113 mil perfiles. A continuación, se detalla el diccionario de variables de esta tabla:

- customer_id: código de identificación de cada cliente de Flow (puede tener asociados uno o más account_id)
- account_id: código de identificación de cada perfil de Flow (se corresponde con un único customer_id)
- device_type: indica el tipo de dispositivo desde el que se efectuó la visualización. Las categorías posibles son:
    - CLOUD: cliente web
    - PHONE: teléfono celular
    - STATIONARY: smart TV
    - STB: set-top box, el decodificador Flow
    - TABLET
- asset_id: código de identificación de cada activo (video) disponible en la plataforma
- tunein: fecha y hora de inicio de cada visualización
- tuneout: fecha y hora de finalización de cada visualización

- resume: variable dummy que indica si se reanuda un consumo anterior del mismo asset_id

Así se ven algunos registros de esta tabla:

customer_id	account_id	device_type	asset_id	tunein	tuneout	resume
14758	37750	PHONE	17473	2021-02-14 20:53:00	2021-02-14 21:31:00	0
14759	37751	STB	12589	2021-03-25 22:05:00	2021-03-25 22:08:00	0
14760	37752	STB	24534	2021-01-15 15:35:00	2021-01-15 17:06:00	1
14760	37752	STB	32059	2021-01-30 10:22:00	2021-01-30 10:41:00	0
14760	37752	STB	29982	2021-01-30 10:41:00	2021-01-30 12:28:00	0

metadata.csv

Contiene la metadata asociada a cada uno de los contenidos. Las variables incluidas son:

- asset_id: código de identificación de cada activo (video) disponible en Flow
- content_id: código de identificación que agrupa los distintos asset_id asociados a un mismo contenido (por ejemplo, cada episodio de una misma serie tiene su propio asset_id, mientras que la serie se identifica con un content_id único)
- title: título
- reduced_title: título reducido
- episode_title: título del episodio (válido para contenidos episódicos, como las series)
- show_type: tipo de show - las categorías son autorreferenciales con excepción de “Rolling”, que indica que se trata de una serie incompleta, y “Web”, la cual remite a contenidos pensados íntegramente en formato digital (series web) -
- released_year: año de lanzamiento
- country_of_origin: país de origen – expresado con el código de dos letras propio del estándar internacional de normalización ISO 3166 -
- category: categoría o género al que pertenece el contenido - puede haber una o más -
- keywords: palabras clave o tags asociadas al contenido - puede haber una o más -

- description: descripción (sinopsis)
- reduced_desc: descripción (sinopsis) reducida
- cast_first_name: nombre y apellido de los actores y actrices principales
- credits_first_name: nombre y apellido del director o directora
- run_time_min: duración total, expresada en minutos
- audience: audiencia target
- made_for_tv: variable dummy que indica si el contenido fue hecho para TV
- close_caption: variable dummy que indica si el contenido posee subtítulos
- sex_rating: variable dummy que indica si el contenido tiene escenas de sexo explícitas
- violence_rating: variable dummy que indica si el contenido tiene escenas de violencia explícitas
- language_rating: variable dummy que indica si el contenido posee lenguaje que puede ser considerado ofensivo o inapropiado
- dialog_rating: variable dummy que indica si el contenido posee diálogos que pueden ser considerado ofensivos o inapropiados
- fv_rating: variable dummy que indica si el contenido tiene rating de FV, que corresponde a público infantil con violencia ficticia
- pay_per_view: variable dummy que indica si se trata de un alquiler
- pack_premium_1: variable dummy que indica si se trata de un contenido exclusivo del pack premium 1
- pack_premium_2: variable dummy que indica si se trata de un contenido exclusivo del pack premium 2
- create_date: fecha de creación del activo
- modify_date: fecha de modificación del activo
- start_vod_date: fecha desde la cual el activo se encuentra disponible en la plataforma
- end_vod_date: fecha de finalización de la disponibilidad del activo en la plataforma

iso_3166_1.json

Este archivo constituye un diccionario de los códigos de nombres de países propios del estándar internacional de normalización ISO 3166. En este json, las claves corresponden a los códigos de dos letras y los valores, a los respectivos nombres de cada país.
Siempre que lo consideren relevante, los participantes podrán incorporar cualquier otra información externa a este conjunto de datos y metadatos.

Evaluación ☑

Se deberá entregar un listado de veinte (20) nuevos contenidos, no previamente vistos, de cada uno de los usuarios que se encuentren en la base de entrenamiento.
Algunas consideraciones:

- La evaluación se realizará al nivel de cada perfil (distinguibles por su account_id) y por grupo de contenidos (es decir, se busca predecir los distintos content_id de las visualizaciones, pero no los asset_id).
- Los listados de predicciones no deben incluir contenidos que los usuarios ya hayan consumido al menos una vez durante el período del set de train. Nos interesa que el sistema de recomendación pueda sugerir títulos que resulten novedosos, no que sea reiterativo en relación al historial de visualizaciones de cada perfil.

- En línea con el punto anterior, al comparar las predicciones contra las visualizaciones reales del set de test, excluiremos tanto los contenidos que ya hayan sido vistos por cada perfil en el set de train como las nuevas incorporaciones del catálogo VOD de Flow del mes del set de test.
- El set de test es privado, lo cual significa que los participantes no tendrán acceso a él.

Métrica

Las predicciones serán evaluadas utilizando la métrica de mean average precision (MAP). MAP es una medida estándar para comparar la lista completa de los veinte contenidos recomendados con los contenidos efectivamente consumidos. Dado un conjunto de Q recomendaciones, MAP constituye la media de la precisión promedio (average precision) de cada recomendación:

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}$$

donde Q es la cantidad de recomendaciones y AP(q) es la precisión promedio de la recomendación q.

La precisión promedio (AP) es el promedio de los valores de precisión en todos los rangos donde se encuentran los contenidos efectivamente consumidos, considerados como ítems relevantes:

$$AP = \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{cantidad\ de\ items\ relevantes}$$

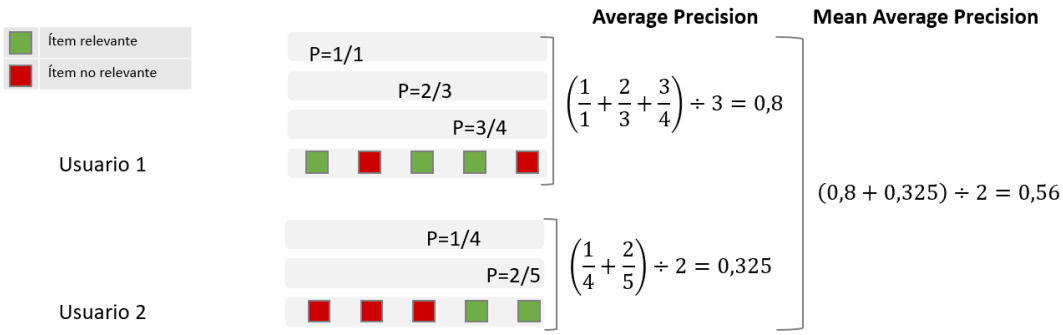
donde n es la cantidad de elementos recomendados, P(k) es la precisión alcanzada hasta la posición k de la lista y rel(k) es una función indicadora que vale 1 cuando el ítem de la posición k resulta relevante (es decir, es un contenido efectivamente consumido), y 0 en caso contrario. Si todos los contenidos vistos en el set de test se encuentran ordenados consecutivamente entre las primeras recomendaciones para un usuario en particular, la AP tendrá un valor de 1. En el extremo opuesto, si no hay ningún contenido relevante entre las recomendaciones, la AP valdrá 0.

Para obtener la MAP, los valores de AP se promedian aritméticamente sobre el conjunto total de recomendaciones.

Esta métrica se basa en las siguientes premisas:

- Relevancia del contenido: se considera relevante un contenido que se haya visto al menos una vez en el set de test (excluyendo los contenidos ya vistos en train y las novedades del catálogo del mes al que corresponden los datos de testeo)
- Posición en la lista: el orden en que se presentan las recomendaciones influye en el score obtenido (es deseable que los ítems relevantes se encuentren al inicio y no al final de la lista)

A modo de ejemplo:



Formula: https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52
Ejemplo: https://medium.com/swlh/rank-aware-recsys-evaluation-metrics-5191bba16832

En el directorio de ejemplos hay una notebook con el código para replicar los resultados que se observan en la ilustración.

Baseline

A modo de baseline, el valor de MAP que se debe superar es 0,014. Este valor se obtiene con un recomendador naïve que simplemente recomienda los veinte contenidos más populares durante el trimestre de train (en términos de la cantidad de perfiles distintos que los vieron al menos una vez), filtrando previamente los contenidos ya vistos. Para ingresar al leaderboard de la competencia se deberá obtener una MAP mayor a este baseline.

Entregables

Recomendaciones

Las recomendaciones deben enviarse en un archivo csv con la siguiente estructura:

- Una columna que identifica a cada perfil con su código de identificación (account_id).
- Otra columna con una lista de los veinte contenidos recomendados, identificados cada uno con su propio content_id y ordenados de mayor a menor relevancia predicha para cada perfil. La lista debe definirse con corchetes ([]), y los elementos dentro de ella deben estar separados por coma (,).

Ej: 123, [1,2,3,4,5, …, 20].
Los valores del csv deberán estar separados por comas, y las variables no deben tener encabezado. Para garantizar la concordancia de las predicciones con el set de testeo, los registros del csv deberán estar ordenados por account_id, de menor a mayor.
En el directorio de ejemplos proporcionamos un archivo de envío de muestra para ilustrar el formato esperado de cada submit.

Código

El procesamiento de los datos y metadatos provistos (y los datos adicionales que los participantes deseen incorporar), el desarrollo del sistema de recomendación y la generación de las recomendaciones para cada perfil deberá hacerse con un programa de código reproducible y abierto (open source). Los participantes podrán trabajar con el lenguaje y las herramientas de programación que deseen mientras los programas cumplan con estos dos requisitos. Asimismo, el código deberá contar con una mínima documentación que explique el paso a paso.

Tanto las recomendaciones como el código (comprimido en caso de tratarse de múltiples archivos) deberán enviarse a través de este formulario. Sólo se aceptará una carga diaria (máximo) por participante.