Reto 2: YouTube Social Analysis REHASH

Objetivo

El objetivo de este reto es poner en práctica los conceptos aprendidos en clase acerca de las estructuras de datos **tablas de hash**; así como, sobre algoritmos de ordenamiento y búsquedas eficientes de información. Y adicionalmente que el estudiante utilice adecuadamente el ambiente de desarrollo (VS Code, Git y GitHub).

Fecha Límite de Entrega

07 de abril, 11:59 p.m.

Contexto

YouTube es un sitio web de origen estadounidense dedicado a compartir videos. Presenta una variedad de clips de películas, programas de televisión y vídeos musicales, así como contenidos amateurs como videoblogs y YouTube Gaming, entre otros. Fue creado por Chad Hurley, Steve Chen y Jawed Karim, en febrero de 2005 y, en octubre de 2006 fue adquirido por Google Inc. a cambio de 1650 millones de dólares y ahora opera como una de sus filiales. Es el sitio web de su tipo más utilizado en internet logrando hoy en día cerca de dos mil millones de usuarios registrados activos, tiene versiones locales en más de 100 países, presenta videos en más de 80 idiomas, y cada minuto se suben más de 500 horas de contenido a la plataforma.

YouTube mantiene una lista de los vídeos que son tendencia en la plataforma. Según la revista Variety "para determinar los vídeos que son tendencia durante el año, YouTube utiliza una combinación de factores que incluyen la medición de las interacciones de los usuarios (número de visualizaciones, comparticiones, comentarios y me gusta). Hay que tener en cuenta que no se trata de los vídeos más vistos en general durante el año natural". Los vídeos más vistos en la lista de tendencias de YouTube son los musicales (como el famoso y viral "Gangam Style"), las actuaciones de famosos y/o realities, y los vídeos virales aleatorios producidos por algún usuario.

A través del análisis de los datos generados por YouTube es posible realizar estudios sociales que permitan identificar comportamientos de la población en general o específicamente como respuesta a algunos eventos o determinados momentos del año. Así como entender los factores que podrían afectar la popularidad de un vídeo en la plataforma, entre muchos otros estudios.

El tema de este reto está relacionado con el análisis de los datos de videos que fueron tendencia en YouTube en diversos lugares del mundo.

Fuente de Datos

A continuación, se presenta una descripción de la fuente de datos que se utilizará en el reto. Los datos están contenidos en el archivo `videos_all.csv` y `category_id.csv` que se pueden descargar del aula unificada en Bloque Neón.

El archivo `videos_all.csv`, contiene un conjunto de datos que corresponde a un registro diario de los vídeos de YouTube que son tendencia en algunos países del mundo (Alemania, Canadá, Corea del Sur, EE.UU., Francia, Gran Bretaña, India, Japón, México, y Rusia), con hasta 200 vídeos de tendencia listados por día.

Los datos incluyen entre otros datos: el título del vídeo, el título del canal, la hora de publicación, las etiquetas, las visualizaciones, los me gusta y los no me gusta, la descripción y el recuento de comentarios. Así mismo, incluyen un campo `category_id`; para recuperar el nombre de las categorías de un vídeo específico este debe buscarse en el archivo `category_id.csv`.

Los datos originales fueron tomados del siguiente enlace: https://www.kaggle.com/datasnaek/youtube-new.

Trabajo Propuesto

Parte 1 – Configuración Repositorio

1. Cree en GitHub un repositorio llamado R2_202110.
2. Cree el README del repositorio donde aparezcan los nombres y códigos de los miembros del equipo de trabajo.
3. Realice el procedimiento para crear el directorio en su computador de trabajo para que relacione este directorio con el repositorio remoto que acaba de crear.
4. Descargue los datos descritos en la sección fuentes de datos y cópielos en la carpeta **data** del repositorio local.

Parte 2 – Desarrollo

Para responder a los requerimientos presentados más adelante, usted deberá cargar la información del archivo de videos y de categorías entregado; es importante anotar que solo es permitido leer una vez la información del archivo.

Al final de la carga hay que reportar los siguientes datos:

- El Total de registros de videos cargados del archivo.
- El Total de registros de categorías cargados del archivo.

Así mismo, con la finalidad de mejorar las complejidades y tiempos de respuesta de las consultas, con respecto al Reto 1, en este Reto **se debe** implementar **como estructura principal,** para la resolución de cada requerimiento propuesto, **tablas de hash**.

Parte 3 – Desarrollo de la solución a los requerimientos

1. Requerimiento 1 (Equipo de Trabajo).

El equipo de análisis quiere conocer cuáles son los **n** videos con más views que son tendencia en un determinado **país**, dada una **categoría** específica.

Para dar respuesta a este requerimiento el equipo de desarrollo debe recibir como entrada la siguiente información:

- category_name.
- country.
- Número de videos que quiere listar (n).

Y como respuesta debe presentar en consola la siguiente información:

- trending_date.
- Title.
- cannel_title.
- publish_time.
- Views.
- Likes.
- Dislikes.

Ejemplo: Se quieren conocer los **tres** videos con más **views** de la categoría **music** que fueron tendencia en **Canadá**.

trending_date	title	channel_title	publish_time	views	likes	dislikes
18.13.05	Childish Gambino - This Is America (Official Video)	ChildishGambinoVEVO	2018-0506T04:00:07.000Z	98938809	3037318	161813
18.12.05	Childish Gambino - This Is America (Official Video)	ChildishGambinoVEVO	2018-0506T04:00:07.000Z	85092067	2735961	140711
18.23.05	BTS (ë°©íƒ„ì†Œë…„ë‹¨) 'FAKE LOVE' Official MV	ibighit	2018-05-18T09:00:02.000Z	80738011	5053338	165854

2. Requerimiento 2 (Estudiante A).

El equipo de análisis quiere conocer **cuál** es el video que más días ha sido trending para un **país** específico.

Para dar respuesta a este requerimiento el equipo de desarrollo debe recibir como entrada la siguiente información:

- Country.

Y como respuesta debe presentar en consola la siguiente información:

- title.
- cannel_title.
- country.
- número de días que estuvo como tendencia.

Ejemplo: Se quieren conocer el video con más días de tendencia en **Canadá**.

title	channel_title	country	Días
Marvel Studios' Avengers: Infinity War Official Trailer	Marvel Entertainment	canada	8

3. **Requerimiento 3 (Estudiante B)**

El equipo de análisis quiere conocer **cuál** es el video que más días ha sido trending para una **categoría** específica.

Para dar respuesta a este requerimiento el equipo de desarrollo debe recibir como entrada la siguiente información:

- category_name.

Y como respuesta debe presentar en consola la siguiente información:

- title.
- cannel_title.
- category_id.
- número de días que estuvo como tendencia.

Ejemplo:

Se quieren conocer el video con más días de tendencia de la categoría **music**.

title	channel_title	category_id	Días
Childish Gambino - This Is America (Official Video)	ChildishGambinoVEVO	10	92

4. **Requerimiento 4 (Equipo de Trabajo)**

El equipo de análisis quiere conocer cuáles son los **n** videos diferentes con más **likes** dado un país y un **tag** específico (es decir el tag debe ser la frase/palabra completa). El tag debe ser considerado sin importar si está escrito en mayúsculas o minúsculas (e.g. Venom debe ser considerado igual a venom).

Para dar respuesta a este requerimiento el equipo de desarrollo debe recibir como entrada la siguiente información:

- Tag dentro de Tags.
- country.
- Número de videos que quiere listar (n).

Y como respuesta debe presentar en consola la siguiente información:

- title.
- cannel_title.
- publish_time.
- views.
- likes.
- dislikes.
- tags.

Ejemplo:

Se quieren conocer los **3** videos de **Canadá** con más **likes** con el tag **2018**.

title	channel_title	publish_time	views	likes	dislikes	tags
Eminem - River ft. Ed Sheeran	EminemVEVO	2018-02-14T15:00:06.000Z	65582241	1495054	39195	Eminem|"River"|"Aftermath/Shady/Interscope"|"Rap"|"River video"|"River official video"|"River music video"|"River video Eminem Ed Sheeran"|"River official video Eminem Ed Sheeran"|"River music video Eminem Ed Sheeran"|"Eminem 2017"|"Eminem 2018"|"New Eminem"|"New Eminem 2017"|"New Eminem 2018"|"Eminem Ed Sheeran"|"Eminem Ed Sheeran Collab"|"Eminem River"|"Ed Sheeran River"|"Eminem Ed Sheeran River"|"River ft Ed Sheeran"|"Revival"|"Eminem Revival"|"Revival 2017"|"all"
Dua Lipa - IDGAF (Official Music Video)	Dua Lipa	2018-01-12T12:00:07.000Z	47060428	1484766	41849	dua lipa|"idgaf"|"dl1"|"I don't give a fuck"|"dua idgaf"|"dua lipa official"|"dua new video"|"dua lipa video"|"dua leepa"|"warner bros records"|"warner Bros"|"so i cut you off"|"dua lipa i dont give a fuck"|"Pop"|"Dance"|"2018"|"Debut Album"|"I Don't Give A F"|"Dueling Duas"|"Dueling Lipas"|"Double Dua"|"Double Lipa"|"Dueling Dua Lipas"|"Double Dua Lipas"
VENOM - Official Trailer (HD)	Sony Pictures Entertainment	2018-04-24T03:45:03.000Z	53071887	1243479	44414	Venom|"Venom Movie"|"Venom (2018)"|"Marvel"|"Marvel Comics"|"Planet of the Symbiotes"|"Eddie Brock"|"Tom Hardy"|"Ruben Fleischer"|"Spider-man"|"Spider-man: Homecoming"|"Michelle Williams"|"Jenny Slate"|"Riz Ahmed"|"Spider-man Spinoff"|"We Are Venom"|"Peter Parker"|"Sony Pictures Entertainment"|"film"|"movie"|"official"|"official venom movie trailer"|"official trailer"|"sony pictures venom"

TIP: Se recomienda que todos los datos del campo **tags** sean procesados y convertidos a minúsculas.

Parte 4 – Análisis de resultados

Realice un análisis en términos de rendimiento (tiempo de ejecución y consumo de memoria) para cada uno de los requerimientos, de la resolución del Reto 2 en relación con la solución dada por el grupo para el Reto 1.

TIP: Tomen como guía para este punto las herramientas y los análisis realizados en los laboratorios 4, 5, y 7.

Genere un archivo en formato **PDF** para la entrega de este análisis y guárdelo en la carpeta **Docs** de su repositorio

Entrega

1. Para hacer la entrega del taller usted debe agregar los usuarios de los monitores y su profesor a su repositorio Github.
2. Si No da acceso a su repositorio a los monitores y al profesor, el taller **NO** podrá ser calificado.
3. Recuerden que cualquier documento solicitado durante las actividades debe incluirse en el repositorio GIT y que solo se calificara hasta el último **COMMIT** realizado dentro de las fechas límites.