INSTITUCIÓN UNIVERSITARIA PASCUAL BRAVO

ARQUITECTURA DE BIG DATA

TRABAJO FINAL

El presente trabajo final se realizará en grupos de máximo 2 personas

1. Se le ha asignado la tarea de construir una tabla de recomendación de películas según las 1.000.209 valoraciones anónimas de aproximadamente 3.900 películas creado por 6.040 usuarios de MovieLens en el año 2000. Descargue el conjunto de datos desde el siguiente link: https://grouplens.org/datasets/movielens/1m/. En el archivo README encontrará la descripción de cada variable de cada archivo, tenga en cuenta que necesita los tres archivos.

i) En una base de datos de Mongo cree tres colecciones llamadas ratings, movies y users. Cada una contendrá los datos de los archivos con mismo nombre.

ii) Cree un script en python que sea capaz conectarse a mongo, lea cada archivo y lo ingeste línea a línea (cada línea como un documento) a su respectiva colección en Mongo. Eg: en movies, tenga en cuenta que cada fila del archivo se considera como un único documento dentro de mongo, por lo tanto, si el archivo movies tiene n elementos (filas) se van a ingestar n documentos a la colección movies.

iii) Investigue como consultar las colecciones creadas desde pyspark

2. Usando el conjunto de datos del punto anterior cargados como RDD, construya una API usando Flask que acceda a la RDD y devuelva los datos según las siguientes especificaciones:

i) Tres ENDPOINTS que acepten como parámetros "mes", "año", y "genero"; filtres la RDD según estos parámetros y devuelva la información necesaria, tenga en cuenta que no necesariamente tienen que aparecer los tres parámetros en simultaneo.

a. Un ENDPOINT llamado RATE_TOP20 que devuelva el top 20 de las películas con mejor rating según el filtro aplicado.

b. Un ENDPOINT llamado RATE_BOTTOM20 que devuelva las 20 películas con los peores ratings según el filtro aplicado.

c. Un ENDPOINT llamado COUNT_TOP 20 que devuelva el top 20 de las películas más vistas según el filtro aplicado.

ii) Un ENDPOINT llamado MOVIE que acepte el nombre de una película y devuelva la información de esta.

iii) Un ENDPOINT llamado LISTBYGENDER que acepte un género de película (Action, Adventure, Comedy, etc) y devuelva por genero elegido las 5 películas con más vistas y las 5 películas con mejor calificación promedio.

iv) Un ENDPOINT llamado SUGGEST que acepte dos parámetros género (F o M) y la edad (1, 18, 25, etc) y con esto filtre las películas que fueron calificadas por este grupo de personas y devuelva el top 20 de las películas con mejores ratings para el género y rango de edad especificados.

3. Replique los dos puntos anteriores como un proyecto aparte para el dataset movielens 32 millones https://grouplens.org/datasets/movielens/32m/ tomando en cuenta solo los filtros por película y no la parte de personas, es decir debe de ser capaz de filtrar por año de la película o por género.