INSTITUCIÓN UNIVERSITARIA PASCUAL BRAVO

ARQUITECTURA EN BIG DATA

SEGUIMIENTO No.2

Para los dos primeros puntos use un notebook sea de Jupyter o de Colab. Descargue el dataset "genoma_2021.zip" del siguiente link: https://grouplens.org/datasets/book-genome/, considere solo los archivos "ratings.json" y "metadata.json", en la archivo "readme.txt" encontrara información de cada uno de estos archivos. Una vez descomprima el archivo .zip ubique los archivos en:

- .\book_dataset\raw\ratings.json
- .\book_dataset\raw\metadata.json

1. Inicialice Spark y cargue los datos del archivo "ratings.json" a una RDD llamado "ratings". Usando el archivo metadata.json cree un diccionario llamado "titulos", como variable broadcast de spark, que contenga como "key" el id del libro (id_item) y como "valor" una lista con los siguientes elementos: El titulo (title), los autores (authors), y el año
2. Liste el (los) nombre(s) de(l) autor(es) del libro y el nombre del libro con mejor rating, en caso de ser más de un libro listelos todos. Haga lo mismo para el libro con peor rating.

Para el siguiente ejercicio use Flask y construya una APP que acepte los siguientes ENDPOINTS con sus especificaciones, entregue un archivo llamado app.py

3. Cree un ENDPOINT llamada “book_by_rating” que acepte dos parámetros llamado "rating_min" y “rating_max”, que acepte dos numero entre 1 y 5, y devuelva listado los títulos y autores de los libros que estén dentro de esta calificación promedio, tenga en cuenta que rating_min siempre tiene que ser menor a rating_max.
4. Cree un ENDPOINT que se llame “book_by_author” que acepte un parámetro llamado “author” y devuelva una lista con los nombres, el año y la calificación de los libros del autor del parámetro. Tenga presente que puede existir un libro con más de un autor, si uno de esos autores coincide con el parámetro este libro debería entrar en la lista. También considere hacer la comparación interna (dentro del ENDPOINT) de los nombres en minúscula y formateando el espacio en blanco con un “+” ejemplo si el autor fuese John Green, el parámetro debe de pasarse como “…?author=John+Green”, el nombre puede estar en mayúsculas, minúsculas o capitalizado como en este ejemplo
5. Cree un ENDPOINT que se llame “book_by_year” que acepte un parámetro llamado “year” y devuelva una lista con los nombres y la calificación de los mejores 10 libros de ese año. Tenga en cuenta que para este ENDPOINT los promedios de los libros deben de ser calculados considerando solo los libros filtrados por el valor del parámetro year (NO DE FORMA GENERAL).
6. Cree un ENDPOINT que se llama “get_book” que acepte los parámetros year, rating_min, rating_max y author; y haga la consulta como se indica en los 3 puntos anteriores considerando todos los parámetros posibles y la ausencia de múltiples de ellos. Tenga en cuenta que este ENDPOINT en especifico tiene distintas variantes según el parámetro y combinación de parámetros que se ingrese, toda combinación debe de tener una salida coherente en dado caso de que la consulta no se pueda hacer.

A continuación, se presentan 8 de las queries que se usaran para el testing y que la app mínimamente debe de funcionar bien o tener una salida coherente con el error

1. htttp://127.0.0.1:5000/book_by_rating?rating_min=4.2

   [
   Genesis, Bernard Beckett
   ...
   ]

2. htttp://127.0.0.1:5000/book_by_author?book_by_author=fulanito+de+tal

   "El autor Fulanito de tal no registra libros publicados en nuestra base de datos"

3. htttp://127.0.0.1:5000/book_by_author?book_by_author=SCOTT+SNYDER

   [
   Batman, Volume 1: The Court of Owls, 2002, 4.0
   ...
   ]

4. htttp://127.0.0.1:5000/book_by_year

   "Especifique un año en el parámetro "year", eg: year=2005

5. htttp://127.0.0.1:5000/book_by_year?year=2.002

   [
   Batman, Volume 1: The Court of Owls, 4.7
   ...
   ]

6. htttp://127.0.0.1:5000/get_book?year=2,002

   El mismo resultado que generaría la query htttp://127.0.0.1:5000/book_by_year?year=2002

7. htttp://127.0.0.1:5000/get_book?author=Scott+snyder&year=2002

   [
   Batman, Volume 1: The Court of Owls,
   <Otro título del autor del año 2002>,
   ...
   ]

8. htttp://127.0.0.1:5000/get_book?rating_min=4.0&rating_max=3.8

   "Especifique un valor de 'rating_max' mayor al valor de 'rating_min"