Task Objective

Design and implement two distinct news article recommendation algorithms that leverage a user's past reading behavior. The algorithms should consider:

- User's reading history
- User's expressed interests
- Popularity of news articles
- Relevance of articles to the user's current location

The developed algorithms must be capable of real-time recommendations and suitable for deployment in a production environment.

Dataset

The dataset is provided as a zip file containing four subfolders: User, Event, Training set (Content Schema), and Test Set (Content Schema).
Drive Location :
https://drive.google.com/file/d/1t2-eMkqLfQkoox-UPOux4VN9eH5rXC48/view?usp=sharing

Alternatively , data is also available at following **GCS** location:

- Device Data : gs://nis-interview-task-data/devices
- Event Data : gs://nis-interview-task-data/event
- Testing Content : gs://nis-interview-task-data/testing_content
- Training Content : gs://nis-interview-task-data/training_content

Instructions

1. **Data Gathering:** Acquire the news dataset.
2. **Data Preparation:** Clean and prepare the dataset for analysis.
3. **Data Querying (EDA Analysis):**
    a. Write and execute SQL queries in your chosen tool to analyze the user, event, and content data.
    b. Your queries should demonstrate your ability to:
        i. Filter and aggregate data based on specific criteria.
        ii. Join tables to combine data from different sources.
        iii. Perform calculations and transformations on the data.
    c. Present your SQL queries and the resulting output for evaluation.
4. **Feature Identification:** Explore the dataset to identify features relevant for the recommendation algorithms.
5. **Dataset Splitting:** Split the dataset into training and test sets. The test dataset should contain new content IDs/hash IDs intended for recommendation.
6. **Algorithm Development:** Build two distinct recommendation algorithms.
7. **Performance Evaluation:** Evaluate the performance of both algorithms.

8. **A/B Testing & Analysis:** Conduct A/B testing for both algorithms and provide a comparative analysis of their pros and cons.

Deliverables

- **Code:** The source code for both recommendation algorithms.
- **EDA Report** : A detailed report summarizing the key findings from the EDA, including performance KPI , data visualizations and statistical analysis.
- **Algorithm Evaluation Report:** A report evaluating the performance of the algorithms, including final recommendations for the test set. Ideally, this report should provide the top 50 recommended content items for users and their corresponding ranks from the test set.

Evaluation Criteria

- **Relevance:** The recommendation algorithm must effectively recommend news articles that are relevant to user interests.
- **Real-time Capability:** The recommendation algorithm must be able to provide recommendations in real time.
- **Data Analysis Skills:** Demonstrates a strong understanding of data analysis techniques and the ability to extract meaningful insights from the data.
- **Statistical Knowledge:** Applies statistical concepts and methods effectively to analyze the data.
- **Problem-Solving:** Demonstrates problem-solving skills by identifying and addressing data challenges.
- **Communication Skills:** Clearly communicates findings and insights through well-structured reports and visualizations.

Data Schema

- **User**

Field Name	Description Data Type
deviceid	The unique user identifier (string)
platform	The user's operating system (string)
os_version	The version of the user's operating syst…
model	Device model (string)
networkType	(string)
district	(string)
lastknownsubadminarea	User's city (string)
language_selected	(string)
created_datetime	The timestamp when the user first activ…
app_updated_at	(string)
last_active_at	(string)

- **Content**

  Data is split into two:

  1. For training: details of news acted upon by devices.
  2. For testing: news inventory available for users in the future.

Field Name	Description Data Type
hashid	The unique identifier for the content (stri…
title	(string)
content	(string)
newsType	Values: 1) VIDEO_NEWS: Full page vid…
author	(string)
categories	The broader labels (internal) of the cont…
hashtags	The topic, if any, of the content (string)
newsDistrict	(string)
createdAt	The timestamp when news was publish…
updatedAt	The timestamp when news was update…
newsLanguage	(string)
sourceName	The source of the content (string)

- **Event**

Field Name	Value	Description	Data Type
deviceId		The unique user identifier	string
event_type	TimeSpent-Front	Generated when user finishes viewing news content	string
	TimeSpent-Back	User clicks on summary content & proceeds to viewing full news from the source	
	News Shared		
	News Bookmarked	Add content to favorites	
	Search Relevancy Option Selected	Search news with keyword User registers the topic or categories as green (interested), yellow, red (not interested)	
	News Unbookmarked	Remove content from favorites	
eventTimestamp		Unix timestamp when the event took place	string
hashId		The unique identifier for the content	string
categoryWhenEventHappened		The unique identifier of the (sub)scene where the event took place. Example: Homepage, options tab, search tab, etc.	string

cardViewPosition		The page number (of the content) where the event took place. For example: (X = page_number) Users swipe X times to like the content.	string
overallTimeSpent		The time user spent viewing the content in seconds	string
searchTerm		Keyword provided for search. Relevant only for event_type='Search'	string
relevancy_color		Color selected by user to show interest. Relevant only for event_type='Relevancy Options Selected'	string
relevancy_topic		Topic or category selected by user to show interest. Relevant only for event_type='Relevancy Options Selected'	string
state		Related to location of user when content was viewed	string
locality		Related to location of user when content was viewed	string
district	Related to location of user when content was viewed		