Data Science Task for the SLIIT Datathon, Sri Lanka: January 2023

Background and Problem Statement: Wiley is a leading publisher of research and delivers research-related content to customers. To improve our services, we need to be able to accurately report content usage. To this end, being able to discern between unidentified scripted behavior and legitimate human usage is particularly important. In the current task, you are to build a means of identifying in a list of server-side events those that are suspicious of being non-human generated.

Evaluation Criteria: This is an open-ended task since we do not have any reliable direct user monitoring. As such, the evaluation will focus on three aspects:

1. Problem and data modeling, specifically, feature engineering. This is a paramount and nontrivial step for this type of problem.
2. Model evaluation and selection. No matter how you choose to approach the problem, it is likely you will experiment with multiple models (ML, stochastic, or traditional rule-based system), so it is important to have a clear take on how to choose among your options.
3. Results and reporting. Being able to explain and elaborate upon what you did is an important and necessary component of Data Science work. The means of accomplishing this is up to you: a well-documented notebook, a slide presentation, a dashboard, or just amply commented and extremely readable source code, anything that works, as long as you can properly convey your message.

Dataset: The provided file contains data extracted from the server logs. The data has been further processed to remove personally identifying and proprietary information. For example, the ip addresses contained therein while structurally valid, may or may not exist in the real world. We also intentionally left in some inconsistencies due to all kinds of vagaries in the data collection pipelines. Due to file size concerns, a sample of one day's worth of data has been selected maintaining the statistical profile of the original data but including all the relevant events for the present entities.

event_id: a unique identifier for each record. You may use it to report back your results; for example, as a table containing event_id and crawler_flag (as classified by your model).

session_id: a server-side user session identifier. One session will typically have multiple events. A session identifier is generated even if the user does not authenticate against the application (a lot of our content is open to general access).

auth_session_id: an auth session is created when the user authenticates with the application. It is quite common to have multiple session_id values for the same auth_session_id. A null auth_session_id corresponds to sessions without authentication.

transaction_date: an event timestamp in the server-side time zone.

event_type_id: an event type identifier for the current event, each event type corresponds to a different user-driven action.

product_id: is a product or article identifier. It identifies uniquely each piece of content the user has access to and the target of the current event.

title_product_id: a title of journal identifier. Multiple product_id values correspond to the same title_product_id. Generally, a title/journal will correspond to a well-defined area of research.

ip_address: the originating ip address of the request as seen by the application. As mentioned above, they have been processed to remove the personally identifying information. To retain some of the location information we added country_id (see below).

country_id: a numeric code for the country; events having the same country_id have originated in the same country.