Bot Or Not?

For any online company, web traffic is a vital source of information. By using the logs our servers generate, we have source material to find answers to a wide array of questions about our business. You can think of analysing the logs to optimise sites and pages based on user behaviour or to detect and mitigate imminent threats.

Depending on the analysis, we are interested in either traffic generated by real people, or bot traffic: non-human traffic (NHT).

In this assignment you will help us identify the type of traffic we see coming onto our servers.

The data consists of a sample of log lines, labeled as bot traffic or not. Note: as web traffic contains very sensitive information, we provide you a simulated set.

Assignment: better model or app, up to you

It's up to you which emphasis you choose for this assignment. Either go crazy with the model and significantly improve the performance baseline or focus your attention on deploying the model. *Make sure your choice highlights your set of skills and the role you apply for.*

We're the kind of company that believes work and life are important, so please do not spend excessive amounts of time solving this assignment. We only want to assess your skill level (and we hope you have some fun as well).

Training data, a pre-trained model and some simple features are available and supplied in a compressed file that was supplied with this document.

Option 1: Improving the model

Based on the data provided, build a better model that can help us identify bot traffic coming from search engines (NHT-search), other bot traffic (NHT-other) and human traffic (HT). You can choose how you want to use the data to improve the labeling and prediction and are not limited to the choice we made for the baseline model and features.

You will take us through your notebook and code during the interview. Please pay attention to these aspects:

- We are interested in how you approach and break down the problem
- Make a clear choice for the type of model you use, and optimize for that specific type
- Clearly explain why and how you constructed your features
- Clearly explain what tuning/optimization you applied and why
- Supply us with some evaluation metrics you think are relevant to the performance of your model and explain how we should interpret them

Option 2: Improve & deploy the application

Take or quickly improve the model to classify NHT in the data we provided *and* deploy this model into an application, namely a REST API that can serve predictions on whether traffic is NHT or not. Please note the dataset contains more classes than just NHT | not-NHT.

Consider this:

- The application should take in a log line as per the dataset and
- Return a class prediction for the input in a simple response
- Choose one performance optimization to apply to the API performance. Apply it, and tell us why you chose that one and what the improvement was
- Expose technical performance of components in the app
- Package and send your code, we'd like to see a working version that we can run locally.
- Offer some API UI and Documentation (like swagger etc., optional)
- Consider setting up a free server somewhere (e.g. Heroku) to demo your app to us during the interview (optional)
- You can choose any language and framework for this task