One of the challenges that we are currently facing is how to identify the language of the file. File extensions are a good way of identifying the language that has been used but there are some situations when a developer uses a non-standard file extension, such as the name of the component it relates to or doesn’t add a file extension at all. This makes it hard, in these cases, to know which parser should be used to gather the static metrics.

Your challenge is carrying out a POC to investigate the options a round using a predictive model to identify the language a file is written in, by just looking at the contents of the file itself. For the POC we are looking for an algorithm that can identify whether a file contains java or not. The approach taken should allow you to increase the number of distinct languages covered.

We recommend that you source sample data from open source repositories such as GitHub, there is no need at this stage to be sourcing millions of files. You can take the below table as a reference for the possible sources of the files and a possible number:

File Type # Files Source

Java 1,000 https://github.com/google/guava

C 1,000 https://github.com/torvalds/linux

Java Script 1,000 https://github.com/nodejs/node

Python 1,000 https://github.com/python/cpython

The ideal completed submission would contain:

\* an outline of the data was sourced (we would like to see data exploration if possible)

\* detailed steps on the feature creation.

\* a simple approach using a ML algorithm of your preference,

\* an outline of what your next steps would be in tackling this problem, given more time.

more interested in seeing how you would approach this type of problem rather than seeing a perfect predictive model. You should attach all the code you have written (R/Python) and a brief summary of your findings. For the summary, we would like to see:

\* How you approach the problem and your cognitive thinking process to solve it

\* Visualisation of the data exploration (if you have it) in Tableau/R/Python