Dataset contains multiple Android malware families; each family has a number of sample. Each sample is described by two files, **.json** file, that describes the actions of a sample, analysed by **CopperDroid**, and **.syscall** file. The **.syscall** file, mainly contains the system calls initiated or made by the malware sample.

In order to understand the basic behaviour of a specific family, a python implementation is produced, which basically, takes the **.json** file as an input and sort its contents in alphabetic order, and then start comparing each family sample with others related to it, until a match is found. The result would significantly summarise the behaviour of a certain family. Thus, the first step in getting the feature vector, which will be determined using Machine Learning concepts, and then fed to Clustering Algorithm.