Title: Investigating the Efficacy of Unstructured Text Analysis for Failure Detection in Syslog

Authors: Katy Felkner (USRC), Elisabeth Moore (HPC-DES, LANL)

Abstract: Each node of a supercomputer produces a detailed log of its operation, called a syslog. It is impossible for system administrators to review all syslog data produced by the thousands of compute nodes associated with a single HPC machine. However, analysis of these logs to detect and predict failures is crucial to maintaining the health of supercomputers. The majority of prior work using machine learning to study syslog has relied heavily on the semi-structured nature of system logs, and there has been less work in examining syslogs as unstructured, purely textual natural language data. We show that treating syslog output as unstructured natural language text without regard for numeric variables does not perform well, and that researchers must exploit the structure within syslog data to produce more useful results. In order to extract features from syslog text, we employ several popular word embeddings and then cluster both word and message level vectors using K-Means and DBSCAN. Finally, we prepared a dataset for supervised learning by aggregating the syslog into 15-minute time windows and extracting the distribution of clusters within that window. Our predictive models performed achieved a relatively low maximum AUC of .59 using a gradient-boosted random forest. This performance barely out-performs random guessing, but does suggest the presence of signal that could be amplified in future work. We also make available our datasets generated using a virtual compute cluster to simulate failures. We conclude that the incorporation of domain knowledge into predictive models, as well as the use of numerical features and structural information in syslog data, rather than a unilateral application of natural language processing techniques must be crucial to build deployable, trustworthy tools.

LA-UR #: LA-UR-20-25858