Capstone Project Proposal

Domain Background

I am a mid-career worker within IT infrastructure domain. I seek career transition into a AI/ML role in this domain.

Logs play an important role in the maintenance of systems. Every day a vast number of logs are generated by the live systems. When a problem occurs, engineers need to examine recorded logs to gain insights into the problem. I'm using AWS Sagemaker to apply ML log analysis in this project.

Problem Statement

It is very time consuming for engineers to diagnose large quantity of logs through manual examination of the logs.

- There is a large number of recurrent issues reflected by the logs, leading to a lot of redundant effort in examining logs and diagnosing the previously known problems.
- Log messages are also highly diverse. Not all log messages are equal in their importance for problem identification – some log messages appear in both normal and failure scenarios, while some log messages only appear in failed scenarios and are more likely to be related to the failures. It is thus challenging for engineers to effectively identify and differentiate various service problems through examining a large number of highly diverse log.

Solution Statement

The amount of time that engineers spent on diagnosing logs can be reduced by directing their attention to only logs that are abnormal.

Assuming past ground truth data is available on which logs are normal and abnormal, this project explores the use of supervised and unsupervised approaches to detecting abnormal logs.

Datasets and Inputs

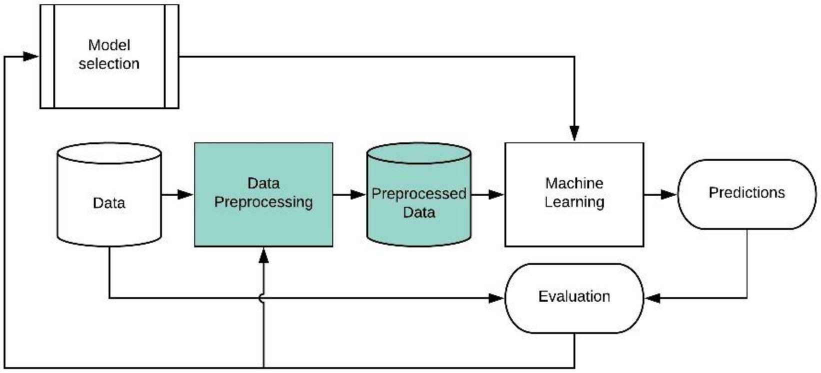
Dataset to be used:

- Logs: HDFS_100k.log_structured.csv
- Labels: abnormal_labels.csv

Project Design

Udacity's Machine Learning Nanodegree programme course syllabus had mainly been focused on how to use AWS Sagemaker in applying machine learning techniques, not so into machine learning techniques. For this project, there is yet new AWS Sagemaker areas not covered in Udacity course, where I would need to learn in order to deploy custom ML algorithms. Hence significant effort will required here in deployment using AWS Sagemaker.

The general workflow for both supervised and unsupervised approach in this project will be same as follows:



1) Supervised approach to aid in abnormal logs detection

   Given that ground truth data on abnormal events is available, I will use supervised learning algorithms from either AWS Sagemaker (Linear Learner eg.) or Sklearn libraries to detect abnormal events.

2) Unsupervised approach to aid in abnormal logs detection

   in AWS Sagemaker platform, I will apply the use of a custom library called LogCluster from loglizer to explore unsupervised approach of clustering to aid in detection of abnormal logs. This clustering approach from Microsoft research as illustrated below is based on agglomerative hierarchical clustering,

Benchmark & Evaluation Metrics

It is assumed in this project there that by reducing number of false positives, the quantity of logs that engineers need to analyze is reduced. The main evaluation metric to be used will be precision. Recall metric will also be used.

There isn’t a benchmark model available for the dataset used in this project. The approaches to be implemented (supervised, supervised) will be compared against each other using the evaluation metrics. In production scenario, it is unlikely that ground truth data on abnormal logs labels is available. The more possible approach is the unsupervised learning approach. Hence the unsupervised learning approach of clustering used in this project will serve as the reference (ie. supervised learning approach will be compared to)