Project Proposal

Shrey Anand, Michael Clifford, AIOps, AICoE, Red Hat

Goals

Please suggest an overarching question and 5 or more smaller questions that the team should answer through this project. You can also include a list of non-goals to further define the scope of the project.

As data scientists we want to know what is the best way to process log data for machine learning tasks like anomaly detection, event correlation, error classification, etc. Logs are machine generated text excerpts that record the events of a system or an application. They are a critical component of software operations; and are often used by human experts for manually performing root cause analysis and troubleshooting. With systems growing more complex, the number of logs have also increased and automated analysis has become key. The first step to address this problem is to leverage deep learning and NLP to parse the semi-structured log files. The current approach is to write a parser using regular expressions based on the developer's knowledge of the log's template format. However, tailored parsers for specific log files are not scalable and can not be applied to larger domains with many different types of log files. Therefore, modern software monitoring tools need an automated way to learn and perform this task. We think this is a great problem for implementing a machine learning based approach.

Questions:

1. What are the current best practices vs state of the art for log parsing?
2. What are the current best practices vs state of the art for log encoding for ML tasks?
3. Do the encoding methods vary based on the downstream ML task?
4. Does the data that we aim to analyze (Kubernetes and Openshift CI data) have any particularities that would prevent it from using the methods discovered above or require a custom solution?
5. Can we develop or extend a set of tools for log parsing and encoding to accommodate a number of downstream ML tasks?

Data

Projects should contain a data cleaning / processing, or gathering / scraping component. Please indicate the sources of data available and how this data should be obtained.

There are two sources of data that can be used for this project. Loghub is an open source resource that contains a number of machine log datasets which can be used for initial prototyping and validation / benchmarking. This would be a good place for the team to start developing and experimenting with, but is already rather well preprocessed.

The target data for this project will be log files from the Kuberentes and Openshift CI platforms stored in a public google cloud storage instance. This data will require students to develop a number of data cleaning, processing, gathering and scraping components to complete this project.

Suggested Methodology

Optional. Here you can suggest techniques / methods that you would like the students to use to answer the questions above.

Part of this project will involve the students performing some background research themselves to determine techniques and methods to implement. However, the content in "Related Work" below can serve as a strong place to get started.

Related Work

Here you can link related work or past work that students should read before starting the project.

Literature survey, application, and comparison of current log parsers.

1. Log parser
2. Zebrium bayesian inference (blog 1, 2)
3. IBM Drain
4. Original Drain

Limitations

1. Limited Patterns Identified
    a. It's possible that log templates vary per user drastically enough that machine learning algorithms cannot successfully assign most.
    b. While there can be general templates found, it may not be able to tell us much if each log has a specific format
2. Too general templates
    a. More of a time limitation
    b. Time may not permit us to tweak the algorithm long enough to find specific (useful) log templates.

        i. Instead we may only find the general commonalities, which may not be useful when looking at the goal of the project.
3. Too much data
    a. There is a lot of data out there regarding logs
    b. It is essential we limit the scope of our data to a useful size where the results can be obtainable and meaningful
4. Peculiarities
    a. It is possible that the datasets we are aiming to solve have custom logs that wouldn't allow us to use existing libraries/resources
    b. May be problematic if we have to custom build too much
5. Data cleaning
    a. There will be a lot of data that we need to sift through
    b. While it needs to be as clean as possible, time is a limited resource
        i. Must clean efficiently and ensure to keep essential data intact