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. Answered: What are the current best practices vs state of the art for log parsing?
2. Answered :What are the current best practices vs state of the art for log encoding for ML tasks?
3. Answered: 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.
        ii. It is also possible for us to create an algorithm that will train on features and formate on a small number of selective files.

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
    c. Due to the size and variation, we should only work with a small number of selective files.
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
        i. drain3
    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.
    c. Maybe setting up a standard/procedure to clean up data will be a potentially good solution.
6. Data limitations
    a. We are at the whim of existing data, meaning we have to hope it'll be workable.
    b. We'll attempt with one data set first then try others, the hope being that we can find a pattern in the original data to work with on others.
7. Achieving project goal
    a. This is a challenging and giant project, one that is not realistic for 5 people with limited knowledge to finish by the end of the semester.
    b. We are treating this project as a research-based project, meaning we aim to find some sort of pattern/insight of templating logs, but cannot guarantee an implementable product for Red Hat day-to-day.
        i. Working with a small number of logs. Identify the features of them and train over those features are more feasible to do.
    c. Our findings should allow a full-time group to implement effectively.
8. working
    a. Constantine having problems logging in to openshift
    b. Google colab is hard to work on it at the same time.
    c. It is also hard to understand other people's code without comments

Code & file

Everything is in this file
https://drive.google.com/drive/folders/1r5D7INeDUS6yz2UhktdzSB_ZhU3HYju_?usp=sharing

Meeting notes & Tasks

1st meeting, 02/17/2021
Agenda: Meet & Greet.
To do:

1. Looking at "build-log.txt" https://gcsweb-ci.apps.ci.l2s4.p1.openshiftapps.com/gcs/origin-ci-test/logs/canary-release-openshift-origin-installer-e2e-aws-4.5-cnv/

2. Looking at https://testgrid.k8s.io and https://prow.ci.openshift.org
    a. Understanding CI/CD environment where the logs represent
        i. Continuous integration and continuous deployment/delivery
3. BU resources: Repository, Course schedule

2nd meeting, 02/25/2021
Agenda:

1. Provide greater project context (Michael)
2. Questions about log data (Team)
3. confirm bi-weekly meeting
4. Slack Channel (Hong)

Deliverable 1:

1. Web scraping (Kyle, due 02/26/2021)
    a. BeautifulSoup for web scraping
2. Analysis on the log data. Trying to find a framework. API (Ningxiao, Parker, Tianze, Hong, due 02/28/2021)
    a. Identify limitations with data and potential risks of achieving project goals.
        i. Determining what information is important to collect and analyze will be difficult. Specific keywords can have different meanings depending on the context of the log.
    b. https://colab.research.google.com/drive/1izeOB105SWF9cvhJD8VoOWegTeGTb8Ie?usp=sharing
    c. https://gcsweb-ci.apps.ci.l2s4.p1.openshiftapps.com/gcs/origin-ci-test/logs/canary-release-openshift-origin-installer-e2e-aws-4.5-cnv/1347679157491863552/build-log.txt
    How should we determine what information is useful in such logs. What is the meaning of the number of the + sign.

3. Answer questions (Hong due 02/28/2021)
    a. What are the current best practices vs state of the art for log parsing?
        i. Current practices:
            1. SolarWinds Security Event Manager is a log analysis tool for Windows that provides a centralized log monitoring experience.
            2. Datadog

3. Papertrail is a log analyzer for Windows that automatically scans through your log data.
4. State of the art
    a. https://ieeexplore.ieee.org/abstract/document/8804456
    b. https://ieeexplore.ieee.org/abstract/document/8067504
    c. https://ieeexplore.ieee.org/abstract/document/8416368

b. What are the current best practices vs state of the art for log encoding for ML tasks?
    i. Best Practice
        1. https://dev.splunk.com/enterprise/docs/developapps/addsupport/logging/loggingbestpractices/
        2. https://tech.churchofjesuschrist.org/wiki/Encoding_(Internationalization_best_practices)
        3. https://www.dnsstuff.com/5-best-practices-for-c-logging-for-it-pros
    ii. State of the art
        1. https://www.sciencedirect.com/science/article/pii/S0920410519310083?casa_token=G-J2pguf0YYAAAAA:Az4ldxBE5GkE0qPYg7r_-5KOcli4fZcweO-POtFuyIvTdVP4BeCTOC2CzDNouRAAlTvWqF9w4S8
        2. https://ieeexplore.ieee.org/abstract/document/8322600
        3. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A666765 0&dswid=7073

3rd meeting, 03/11/2021
Where we are coding:
https://oauth-openshift.apps.zero.massopen.cloud/oauth/authorize?response_type=code&redirect_uri=https%3A%2F%2Fjupyterhub-opf-jupyterhub.apps.zero.massopen.cloud%2Fhub%2Foauth_callback&client_id=system%3Aserviceaccount%3Aopf-jupyterhub%3Ajupyterhub-hub&state=eyJzdGF0ZV9pZCI6ICI0MDc2YmM5MmQ3N2I0OWY3OTk0MDFhYzU4N2MxZmYzNiIsICJuZXh0X3VybCI6ICIvaHViLyJ9&scope=user%3Ainfo
Methodology to try:
Drain3
https://www.operate-first.cloud

Deliverable 2: due 03/24/2021

1. Collect and pre-process a secondary batch of data
    code
2. Refine the preliminary analysis of the data performed in PD1
    a. code
3. Answer another key question
    Do the encoding methods vary based on the downstream ML task?
    - No, these methods do not vary because the logs we are dealing with contain the same structure, and once parsed, they are ready to be processed in the drain3 algorithm we are using to cluster the logs.

4. Refine project scope and list of limitations with data and potential risks of achieving project goal
    a. Included in the limitations section.
5. Submit a PR with the above report and modifications to the original proposal