Assignment #2: Log Analysis for Anomaly Detection

Qianyu Ouyang

Part#1: comparing current log parsing methods

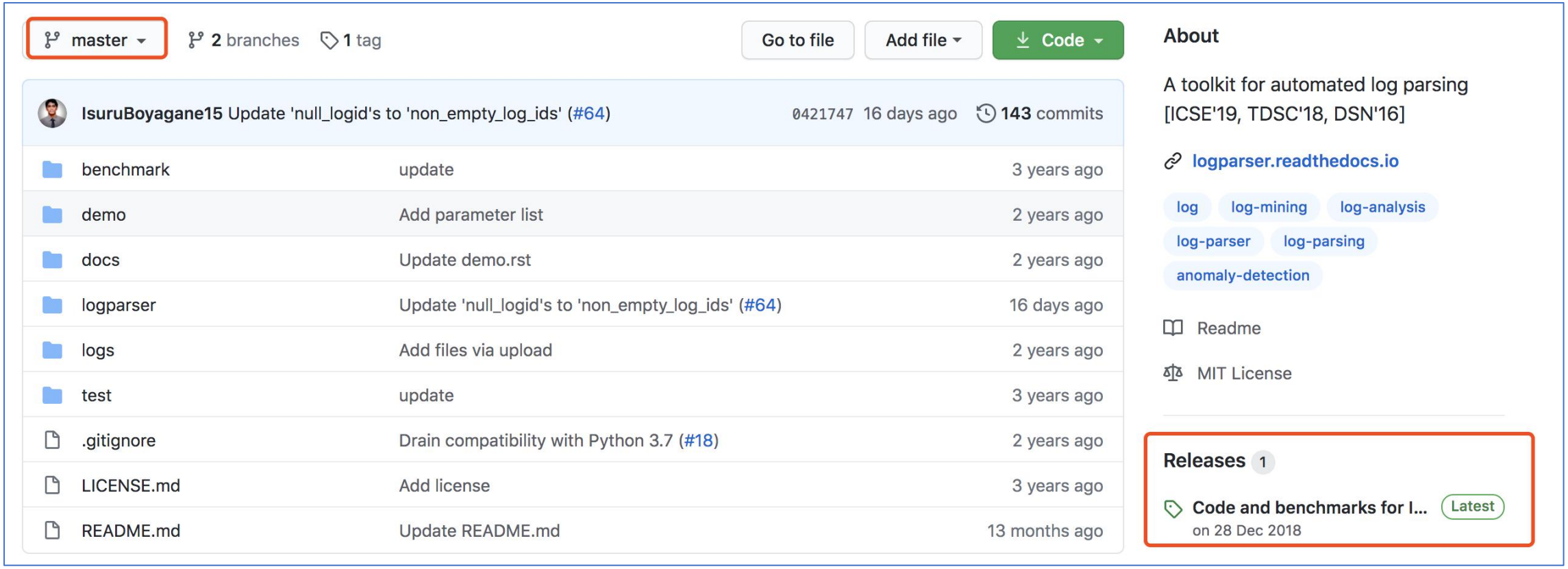
- Learn the DSN 16 paper (or original paper for these log parsing methods) and do the following jobs using the given datasets:

  1. Learn **four** log parsing algorithms: **LogSig**, **IPLom**, **SLCT** and **LKE**
  2. Use toolkit to run four log parsing algorithms on five datasets (BGL, HDFS, HPC, Proxifier, Zookeeper).
  3. Plot **runtime, F-score**, **RandIndex**[3] (a metrics for evaluation clustering) with bar charts when four algorithms are parsing logs.
  4. Display your template files.
  5. Describe your own experience or findings in doing log pasrsing. For example, advantages and disadvantages of these algorithms.

[3] https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-

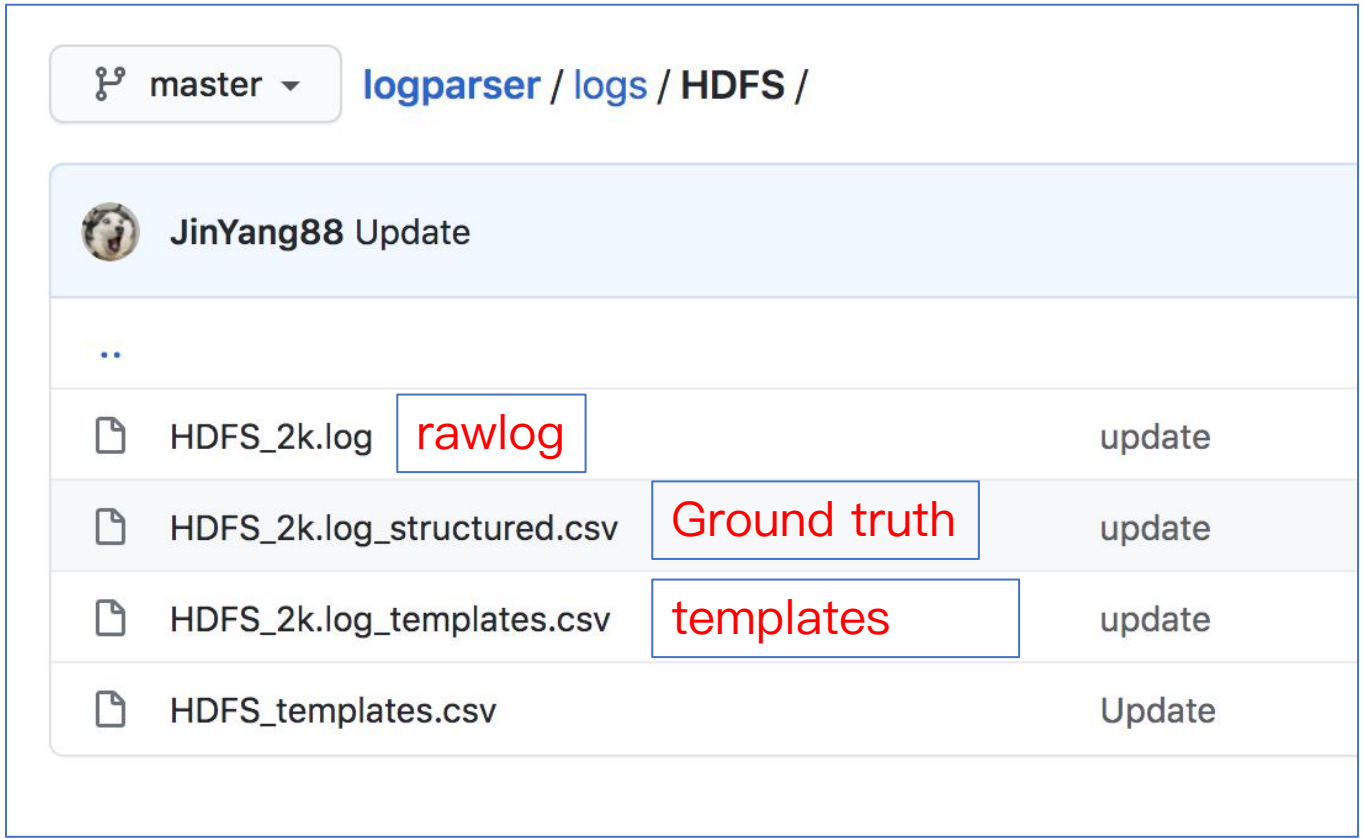
Part#1: comparing current log parsing methods

- Toolkit version



Part#1: comparing current log parsing methods

- Dataset:
  - https://github.com/logpai/logparser/tree/master/logs
  - Five types of logs (BGL, HDFS, HPC, Proxifier, Zookeeper), each type has 2000 logs



Part#2: comparing anomaly detection methods

- Learn the ISSRE paper (or original paper for these anomaly detection methods) and do the following jobs using the **HDFS logs**.

  - Learn three **unsupervised** anomaly detection models: **Invariants Mining, PCA and Log Clustering.**

  - Choose a log parsing method (mentioned in part1) to parse HDFS logs, and use this toolkit to run **two of three** anomaly detection models (Considering Log Clustering suffers from high computational complexity, you needn't run Log Clustering).

  - Plot **precision, recall, F-score ,runtime** with bar charts when these models detecting anomaly.

  - When running **Invariants Mining**, add some codes and display **three relationships** (please check the paper for more information), *e.g.*, n(A) = n(B), where n(*) represents the number of logs which belong to corresponding template *. And explain why, *e.g.*, template A is "Interface *, changed state to down", while template B is "Interface *, changed state to up".

  - Describe your own experience or findings in doing those jobs.

Part#2: comparing anomaly detection methods

- QA
  - Toolkit：
    - https://github.com/logpai/loglizer
  - Dataset:
    - https://www.dropbox.com/s/akef557hnla0h9v/ANM-data.zip?dl=0
    - https://cloud.tsinghua.edu.cn/f/c8806b4c81ee45afa03c/?dl=1
  - LogParser:
    - You should choose a LogParser (4 methods in part1) to parse the log data according to its performance in the part1
  - LogCluster:
    - You don't need to run logCluster in part2.

When you finish this assignment, you only need to **submit a zip file , which includes template files in part#1 and an assignment report**.

Project

- The 12-hour test data is be generated in the server every 12 hours.
- You can use consumer.py to read the data from kafka directly.
- And you can use the submit function in consumer.py to send your answers to the server. So you don't have to build the docker container.

- NOTICE: you should use the consumer.py on the web-learning site.
- The final test will start at Dec.16 0:00 (beijing time zone)

Project

- http://81.70.98.179:8000/standings/show/

rank	group name	score
1	学堂路车神	170
2	Veritaserum	0
3	The Anomalies	0
4	study group	0
5	MSSherlock	0
6	meow meow	0
7	Learning Failure	0
8	flower group	0
9	DANM!	0
10	ANM小组	0
11	ANMG	0

Q&A