Tool: Applying ML to log analysis for anomalies detection

Team
(**C**) valeryk2 (Valery Kreidenko)

What is the function of the tool
The tool is a thematic continuation of the technology review. The tool will allow to train and evaluate performance of different machine learning models on the datasets of the labeled log files.

Who will benefit from such a tool
SREs, data scientists and researchers who would like to experiment and test different ML models for anomalies detection from the log files.

Does this kind of tools already exist
Some work has been already done in this area to support the research brought in the technology review though not sure what exactly since the code isn't publicly available.

What existing resources can be used
We'll use the approach which was researched and suggested by the authors of https://hal.laas.fr/hal-01576291/document.
We'll try to acquire some real existing datasets from the real system.

What techniques/ algorithms will be used
The current plan is:

- to use *word2vec* or other NLP technique(s) to map text data into the ML models consumable data
- we'll implement at least 2 supervised ML methods
- per performance evaluation
    - for effectiveness we'll use F-measure and maybe others
    - for efficiency we'll compute model training and classification times for models comparison
- for validation we plan to implement 10-fold or other method
- finally, we'll try to apply OOP design patterns to make this tool extensible for easy plugging of more ML models, etc.
- implementation language: Python + its ML libraries

How the usefulness of the tool will be demonstrated
We'll run at least 2 models of the tool on the same data and we'll evaluate and compare the relative performance in terms of the effectiveness and efficiency. We hope to demo this tool on the real dataset, otherwise we'll synthesize some.

Timeline
This is estimated as a ~3 weeks project.