Progress Report: Applying ML to log analysis for anomalies detection

Progress made so far:

- *Deepening knowledge in the area.* The technology review in the related area as well as some video-courses on NLP and Leveraging NLP and Word Embeddings in ML.
- *Dependencies for the project.* Python libraries: *spaCy* (text parsing), *pandas*, *numpy* (data manipulation), *gensim* (word2vec embedding), *sklearn* (ML models, validations, …)
- *Installing and configuring dev env.*
- *Breaking down the tasks and creating the backlog.*

Backlog:

- *(in background) Data acquisition.* Work to obtain real log files
- *Skeleton.* Build a skeleton of the steps w/stubs (+ some tests)
- *1st implementation.* Implement *word2vec* embedding w/small synthetic log files; kFold of the first ML algorithm (TBD); implement perf. reporting (e.g. accuracy, runtime)
- *Refactor to OOP.* Should enable an easy extension at least with a different ML algorithm; (optionally embeddings as well)
- *Extend the implementation with at least one more model.* Ideally should be able to run a comparison of models (details TBD)
- *Run the tool on the real log files.* Run and compare models performance
- *Documentation.*

[Optional]

- *Dockerization.* For more convenient tool usage
- *Non-functional requirements.* Look at the performance optimization
- *Hyperparameters tuning.*

Challenges so far:

- *Real data acquisition.* If no real data achieved the fallback would be to use a synthetically generated dataset
- *Env tech issues.* Some issue w/scapy corpus load – troubleshooting in progress…