ZUSAMMENFASSUNG

Log-Dateien dokumentieren das Verhalten einer Anwendung oder eines Systems, was deren Analyse zu einem Schlüsselfaktor für die Sicherheit, Stabilität und Nutzbarkeit eines Systems macht. Dabei gibt es in vielen Applikationen in den Log-Dateien Fehler oder Warnungen, bei denen kein Handlungsbedarf besteht. Die hohe Anzahl solcher Meldungen im Vergleich zur geringen Menge an tatsächlichen Fehlern führt dazu, dass sich die Analyse sehr aufwändig gestaltet. Diese Problemstellung findet sich auch bei der ORDIX AG bei einem Mircoservice für interne Zwecke wieder. Ziel dieser Arbeit war deshalb die automatisierte Fehlererkennung basierend auf den Log-Dateien.

Zu diesem Zweck musste zunächst eine geeignete Methode zur Vorverarbeitung entwickelt werden, um aus den unstrukturierten Log-Dateien Features zu gewinnen, die für den Einsatz von Machine-Learning-Modellen zur Fehlererkennung verwendet werden können. Hierfür wurde aus einer Log-Meldung der konstante Teil, der sogenannte Log-Key, extrahiert.

Anschließend wurde die Fehlererkennung betrachtet und zu diesem Zweck diverse Supervised-, Semi-Supervised- und Unsupervised-Methoden untersucht. Dabei hat sich gezeigt, dass die Unsupervised-Verfahren nicht für den Anwendungsfall dieser Arbeit geeignet sind, sondern in erster Linie für Anwendungen, in denen die Daten mittels numerischer Werte verglichen und so Fehler identifiziert werden können. Im Rahmen dieser Arbeit war dies jedoch nicht gegeben, weshalb jene Verfahren in der Umsetzung nicht berücksichtigt wurden.

Für die Evaluierung wurden die nicht erkannten Fehler sowie die Meldungen, die inkorrekterweise als Fehler klassifiziert wurden, berücksichtigt. Die Supervised-Verfahren erzielten sehr gute Ergebnisse, insbesondere durch den Einsatz des Decision Tree konnten fast alle Meldungen korrekt klassifziert werden. Im Bereich der Semi-Supervised-Fehlererkennung konnten mittels eines Long Short-Term Memory alle Fehler erkannt werden. Das Modell wies jedoch, im Vergleich zu den Supervised-Methoden, eine höhere Anzahl an Meldungen auf, die inkorrekt als Fehler klassifiziert wurden.

Es wurde gezeigt, dass sowohl der Supervised- als auch der Semi-Supervised-Ansatz für eine automatisierte Fehlererkennung auf Basis der Log-Dateien des Microservice eingesetzt werden kann. Bei den Modellen des Supervised Learnings besteht der Aufwand dabei in der manuellen Vergabe der Labels, während er bei der Semi-Supervised-Fehlererkennung in der manuellen Analyse der Modellergebnisse liegt.

ABSTRACT

Log files protocol a system's behavior and are therefore a valuable source of information about the security, stability, and usability of a system. Nevertheless, many applications display errors or warnings in their log files, even though there no action is required. The large number of such messages compared to the few actual errors leads to a very time-consuming need of manual error message screening. This problem is also found at the ORDIX AG at a microservice application. For this reason, the goal of this thesis was an automated error detection based on the log files.

For this purpose a suitable preprocessing method was first needed to extract features from the unstructured logs to then be used for machine learning models to detect the errors. Thus, the log key, which is the constant part of a log message, was extracted.

For the error detection various methods for supervised, semi-supervised and unsupervised error detection were examined. However, unsupervised models proved to be unsuitable for the context of this work since they primarily rely on analysis of numerical values. As this was not given in the context of this work, those procedures were not considered for implementation.

For the evaluation, the number of detected errors and the messages that were incorrectly classified as errors were taken into account. The supervised methods achieve very good results, in particular the decision tree, which classified almost all messages correctly. In the area of semi-supervised anomaly detection, all errors could be detected using a long short-term memory, but a larger number of messages was incorrectly classified as errors compared to the supervised methods.

It has been shown that both, the supervised and the semi-supervised approach, are suitable for automated error detection based on the microservice's log files. For the supervised models, the effort is in the manual labelling, while for the semi-supervised error detection, it is in the manual analysis of the model results.