FG Software Engineering
Leibniz Universität Hannover
Welfengarten 1
3. Stock, Flur G

Bachelor-Arbeit

- zu vergeben -

Anwendung von Ausgewählten Machine Learning und Natural Language Processing Verfahren um sicherheitsrelevanten Programmcode aus Stackoverflow zu Exportieren

Hintergrund

Stackoverflow [1] ist eine Quelle über welche Entwickler Source Code Fragmente finden können, welche Sie häufig direkt in Anwendungen einbetten. Teilweise gibt es zu Programmcode der von Nutzern verbreitet wird, sicherheitsbezogene Anmerkungen und Diskussionen. Auf diese Weise werden bekannte Schwachstellen in Softwareartefakten erwähnt und mögliche Patches oder Exploits dieser vorgeschlagen. Diese Beiträge beinhalten somit Code und natürlich sprachliche Texte.
Um diese und weitere sicherheitsrelevante Source Code Elemente in Stackoverflow zu finden, identifizieren und herunterzuladen zu können, existiert am Fachgebiet bereits ein Tool, welches diesen Code herunterlädt und diesen die die drei Typen: Schwachstellen-, Patch und Exploit-Code unterscheidet. Für die Unterscheidung in diese Typen, werden Machine Learning Ansätze mit Natural Language Processing (NLP) in Kombination verwendet. Aktuell funktioniert die Suche nach diesen StackOverflowbeiträgen schlüsselwortbezogenen, welches auf eine Suche mit nutzen von den bereits implementierten Klassifikatoren angepasst werden soll. Hierbei sollte nicht nur Code in die Klassifizierer eingebunden werden, sondern ebenfalls natürlich sprachliche Texte der StackOverflow posts. Um diese optimal zu verarbeiten, liefern state-of-the-art Ansätze des NLP hierzu geeignete Verfahren wie z.B. Bag-of-Words, Stemming, N-grams, etc.
[1] Stackoverflow: https://stackoverflow.com

Aufgabe

Im Rahmen dieser Arbeit soll ein (Java)-Tool adaptiert werden, das Machine Learning und Natural Language Processing (NLP) Ansätze kombinieren soll, um sicherheitsrelevanten Code von Strackoverflow herunterzuladen. Dabei soll NLP genutzt werden um die natürlich Sprachlichen Texte optimal zu nutzen um über die Sicherheitsrelevanz der in diesen Zusammenhang stehenden Code Fragmenten zu ermitteln. Die als sicherheitsrelevant klassifizierten Code Fragmente sollen mit einheitlicher Struktur inkl. einer Datenbank mit Metainformationen und Verweisen auf Ablagepfad, in einem Verzeichnis abgelegt werden. Um die Effizienz der verwendeten Ansätze zu messen, sollen diese mit geeigneten Verfahren evaluiert werden. Das Ergebnis der Anwendung soll in einem Verzeichnis abgelegt werden, welches über eine Datenbank mit Metadaten verfügt.
Weitere Anforderungen:
- Literaturrecherche zu verwandten Arbeiten
- Implementieren einer Java-Anwendung (Gut strukturierter und kommentierter Code)
- Manuelle Adaptierung bei fehlerhaften Einträge (Aufwandsmessung)

Betreuer: M. Sc. Fabien P. Viertel, fabien.viertel@inf .uni-hannover.de, Raum G307
Prüfer: Prof. Dr. Schneider **Beginn**: Ab sofort möglich