УДК 004.855.5

Анализ исходного кода методами машинного обучения.

Магда И. А. федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»

Научный руководитель – Фильченков А.А., к.ф.-м.н., доцент университета ИТМО

Введение

Сегодня анализ исходного кода используется для выявления ошибок, выявления потенциальных уязвимостей, обучения, выявления закладок, постановки задачи для рефакторинга. Он может выполняться в ручном или автоматическом режиме. Если говорить о ручном режиме, то в основном речь идет о классическом обзоре коде, который служит для поиска ошибок, выработки рекомендаций по улучшению кода, а также обучению новых программистов. Автоматизированные методы в основном используются для выявления ошибок, уязвимостей и закладок. Для этих целей применяются инструменты статического анализа кода.

Для анализа исходного кода предлагается использование методом машинного обучения, основанного на представлении текстов исходного кода, как непрерывно распределенных векторов. Основная идея представить код, как коллекцию путей в абстрактном синтаксическом дереве (AST – Abstract syntax tree) и агрегирование данных путей в вектор фиксированной длины, который может быть использован для предсказания семантических свойств исходного кода.

Цель работы

Целью данной работы является разработка алгоритма анализа исходного кода, с использованием методов машинного обучения и векторного представления исходного кода.

Результаты

В данной работе реализован алгоритм векторного представления исходного кода для языка программирования Java, вектора которого могут быть использованы для предсказания семантических свойств исходного кода. Дальнейшие исследования могут быть направлены на улучшения качества получаемых вектором исходного кода и практического применения алгоритма векторного представления.

Список литературы

1. Ахо А., Ульман Дж. Теория синтаксического анализа, перевода и компиляции. Том 1. Синтаксический анализ. – М. : Мир, 1978..
2. code2vec: Learning Distributed Representations of Code [Электронный ресурс] // URL https://arxiv.org/abs/1803.09473v2
3. Learning to Represent Programs with Graphs [Электронный ресурс] // URL https://arxiv.org/abs/1711.00740v1
4. Syntax and Sensibility: Using language models to detect and correct syntax errors [Электронный ресурс] // URL http://softwareprocess.es/pubs/santos2018SANER-syntax.pdf

5. code2seq: Generating Sequences from Structured Representations of Code [Электронный ресурс] // URL https://arxiv.org/abs/1808.01400
6. Tree2Tree Neural Translation Model for Learning Source Code Changes [Электронный ресурс] // URL https://arxiv.org/pdf/1810.00314.pdf