УДК 004.853

ИССЛЕДОВАНИЕ ПОДХОДОВ К ВЕКТОРИЗАЦИИ ИСХОДНОГО КОДА

Н.А. Мурычева, студент, Университет ИТМО, Санкт-Петербург
Научный руководитель – Т.А. Брыксин, JetBrains Research*, Санкт-Петербург*

Машинное обучение является эффективным инструментом для анализа данных, в том числе и для анализа текстов. Решаются такие задачи, как машинный перевод, синтез и генерация речи, описание изображений и др. В свою очередь существует большое количество задач в области программного обеспечения, которые потенциально можно решить или они уже решаются при помощи машинного обучения. Примерами таких задач являются деобфускация кода, выявление плагиата, автоматическая генерация документации, обнаружение ошибок в коде и т.д.

Несмотря на большой интерес исследователей к этой области, до сих пор нет общепризнанного механизма препроцессинга исходного кода, который бы сопоставлял входным данным соответствующие им вектора, как, например, для естественных языков это можно сделать, применяя набор алгоритмов word2vec. Существующие модели представления кода отличаются значительным разнообразием: программа описывается как последовательность токенов [1] или последовательность вызовов API [2], рассматриваются построенные по исходному коду абстрактное синтаксическое дерево [3] или графы определенного типа [4], обрабатывается промежуточное представление кода [5].

Целью работы является сравнение существующих моделей представления кода применительно к задаче определения авторства кода.

Критериями выбора моделей для дальнейшего анализа являлись хорошие результаты их применения, а также разнообразие подходов. В конечном итоге было выбрано четыре модели [3, 5, 6, 7], каждая из которых удовлетворяет всем описанным ранее критериям.

В качестве данных для обучения и тестирования использовались командные проекты с открытым исходным кодом. Рассматривались части проектов, подходящие под особенности выбранных моделей и удовлетворяющие условию однозначного определения автора.

Результатом работы является анализ и сравнение выбранных моделей на основе задачи определения авторства.

Литература:

1. M. White, C. Vendome, M. Linares-V´asquez, D. Poshyvanyk, Toward deep learning software repositories: IEEE/ACM 12th Working Conference on Mining Software Repositories, 2015
2. T. D. Nguyen, A. T. Nguyen, T. N. Nguyen, Mapping API elements for code migration with vector representations: International Conference on Software Engineering, 2016.
3. X. Hu, Y. Wei, G. Li, and Z. Jin., CodeSum: Translate Program Language to Natural Language: 2017.
4. V. Raychev, M. Vechev, and A. Krause. Predicting program properties from "big code": 42Nd Annual ACM SIGPLANSIGACT Symposium on Principles of Programming Languages, 2015.

5. M. Tufano, C. Watson, G. Bavota, M. Di Penta, M. White, and D. Poshyvanyk, Deep learning similarities from different representations of source code, Mining Software Repositories, 2018.
6. M. Allamanis, M. Brockschmidt, M. Khademi, Learning to represent programs with graphs: International Conference on Learning Representations, 2018.
7. T. Ben-Nun, A. Sh. Jakobovits, T. Hoefler, Neural Code Comprehension: A Learnable Representation of Code Semantics, 2018.