Обнаружение плагиата в программном коде

Веселов Иван Дмитриевич
научный руководитель: Брыксин Тимофей Александрович

СПб АУ НОЦНТ РАН

20 февраля 2017

Проблема плагиата

Современные решения

- SPLaT
- JPlag
- MOSS
- SIM
- ...

Цели, задачи

- Исследовать область
- Попробовать применение машинного обучения
- Создать утилиту, помогающую обнаруживать заимствованный код

Выбранный подход

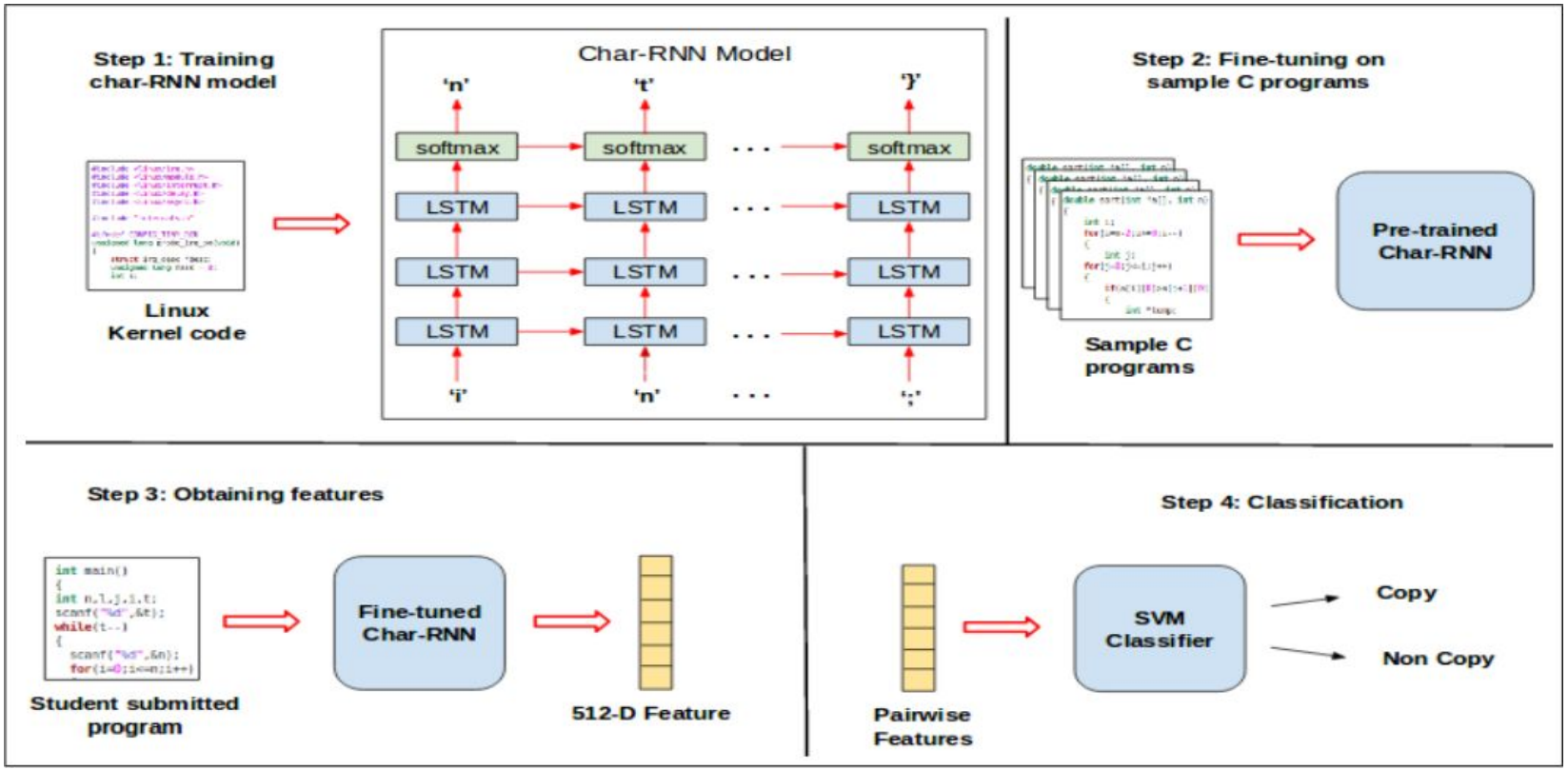
Plagiarism Detection in Programming Assignments Using Deep Features

(2017)

Jitendra Yasaswi, Suresh Purini, C. V. Jawahar

International Institute of Information Technology, Hyderabad

Предлагаемый подход



Реализация

Технологии:

- Keras
- Scikit-Learn

https://github.com/ml-in-programming/deep-features-for-plagiarism-detection

Генерация данных

Преобразования:

- Перемешивание объявлений
- Переименование идентификаторов

Технологии:

- JavaParser
- IntelliJ Platform

Генерация данных

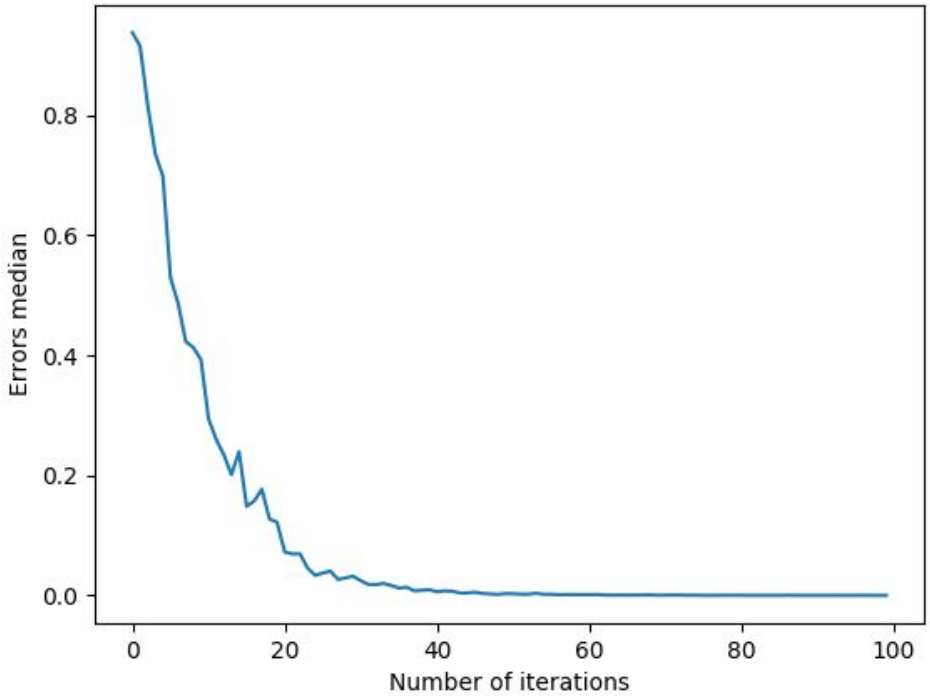
```
class Example {
    void foo() {
        int value = 42;
        bar(value);
    }

    void bar(int arg) {}
}
```

```
class Vamue {
    void atg(int vclue) {
    }

    void mbin() {
        int gpo = 42;
        atg(gpo);
    }

}
```

Генерация данных

- GitHub Java Corpus
- Собрано 12 652 исходных файлов
- Сгенерированы 1 053 копии

Валидация


Errors median
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
Number of iterations

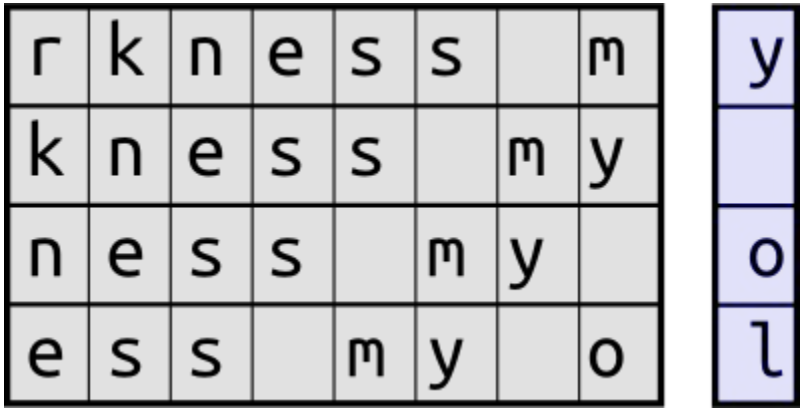
Результаты

- Построена модель Char-RNN + SVM
- Собрано большое количество Java файлов
- Сгенерированы синтетический данные для обучения

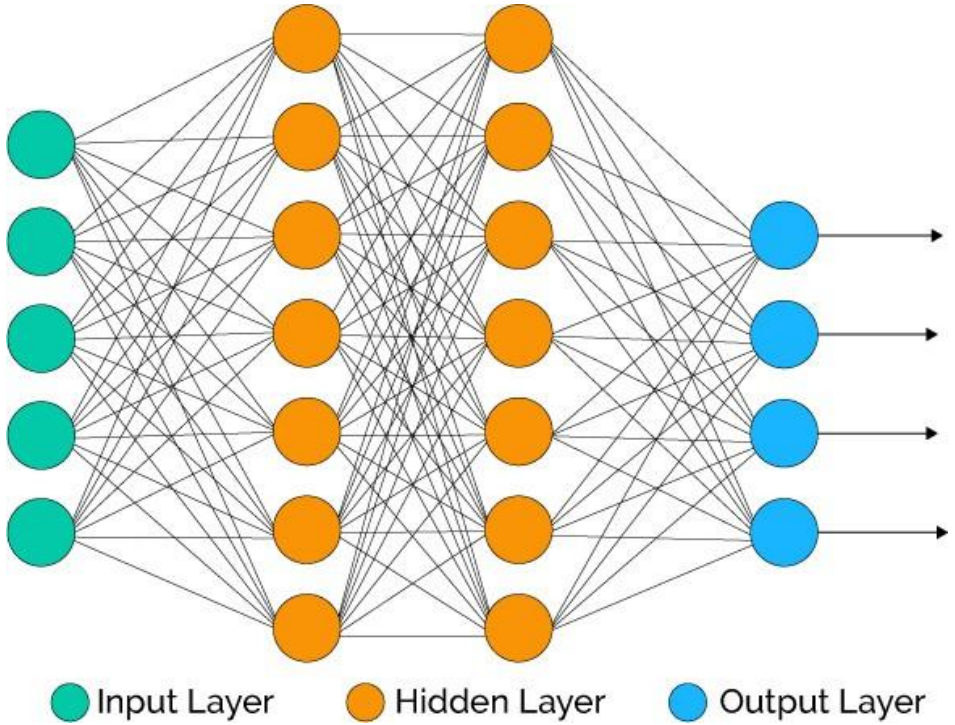
Возможное дальнейшее развитие

- Обучить модель на производительном компьютере
- Улучшить генерацию данных
- Создать на её основе утилиту

Char-RNN



Извлечение признаков



SVM