Інформаційна технологія виявлення дублікатів програмного коду

Виконав:

студент групи 1-КН-18М

Никуляк А.В.

Керівник:

к.т.н проф.

Месюра В.

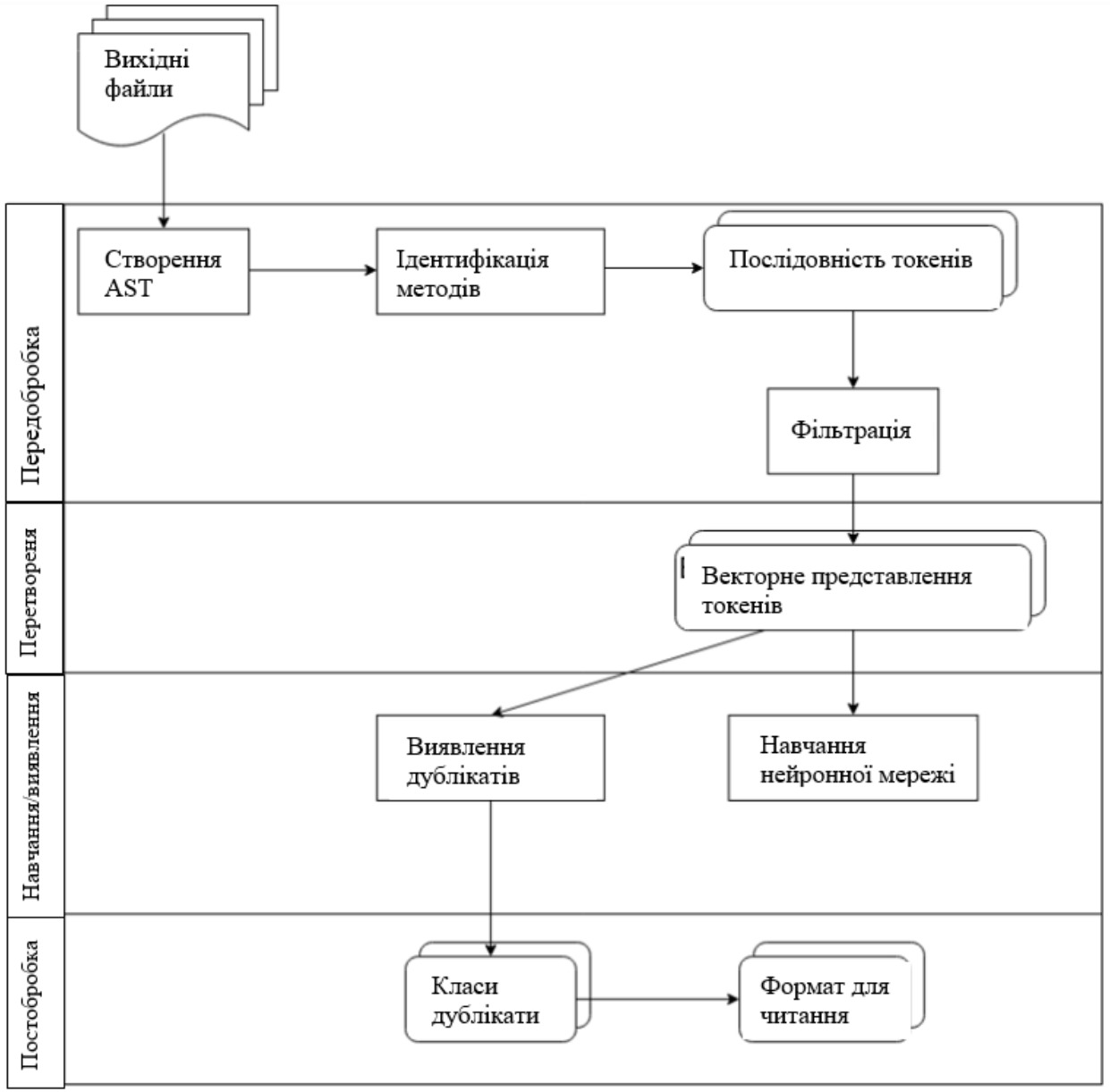
Інформаційна технологія виявлення дублікатів програмного коду

- **Метою** магістерської кваліфікаційної роботи є підвищення точності виявлення дублікатів програмного коду за рахунок розробки інтелектуального методу
- **Об’єкт дослідження** – процес виявлення дублікатів програмного коду.
- **Предмет дослідження** – методи виявлення дублікатів програмного коду та їх властивості.

Типи дубілікатів

Вихідний код	Тип-1	Тип-2
int main() { int x = 1; int y = x + 5; return y; }	int main() { int x = 1; int y = x + 5; return y;//output }	int func2() { int p = 1; int q = p + 5; return q; }

Тип-3	Тип-4
int main() { int s = 1; int t = s + 5; t = t/++s; return t; }	int func4() { int n = 5; return ++n; }

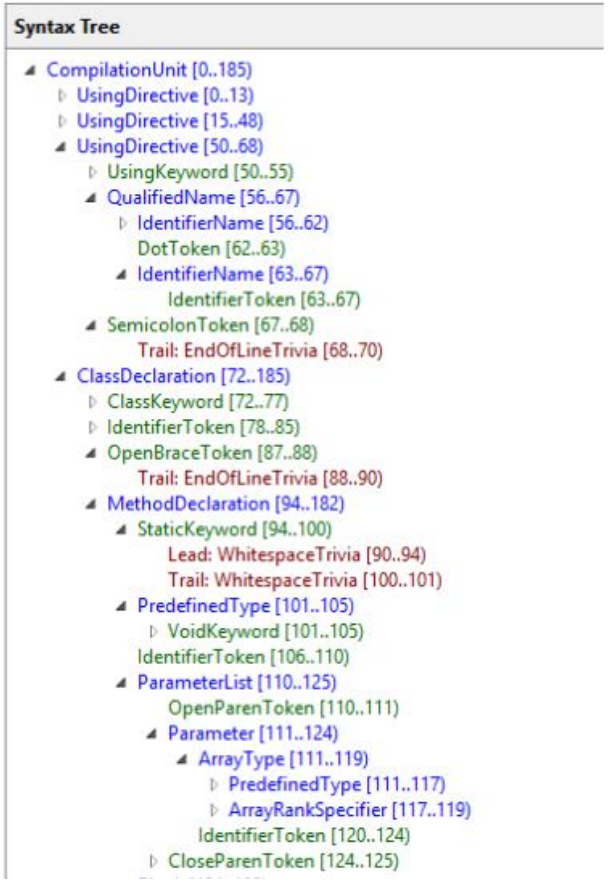

Вихідні файли
Передобробка
Створення AST
Ідентифікація методів
Послідовність токенів
Фільтрація
Перетворення
Векторне представлення токенів
Навчання/виявлення
Виявлення дублікатів
Навчання нейронної мережі
Постобробка
Класи дублікати
Формат для читання

Абстрактне синтаксичне дерево

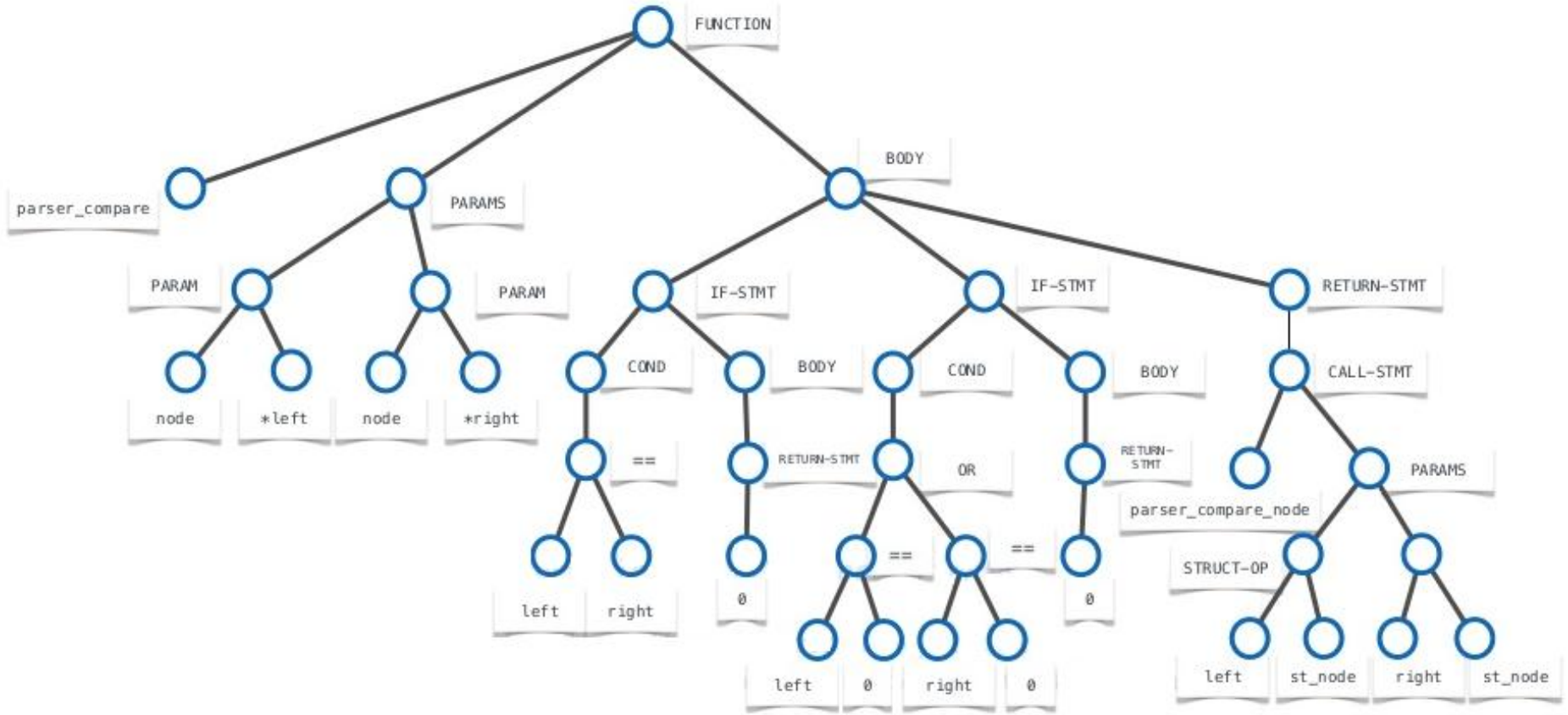
```
static int
parser_compare(PyST_Object *left, PyST_Object *right)
{
    if (left == right)
        return (0);

    if ((left == 0) || (right == 0))
        return (-1);

    return (parser_compare_nodes(left->st_node, right->st_node));
}
```



Абстрактне синтаксичне дерево



Рекурентні нейронні мере́жі

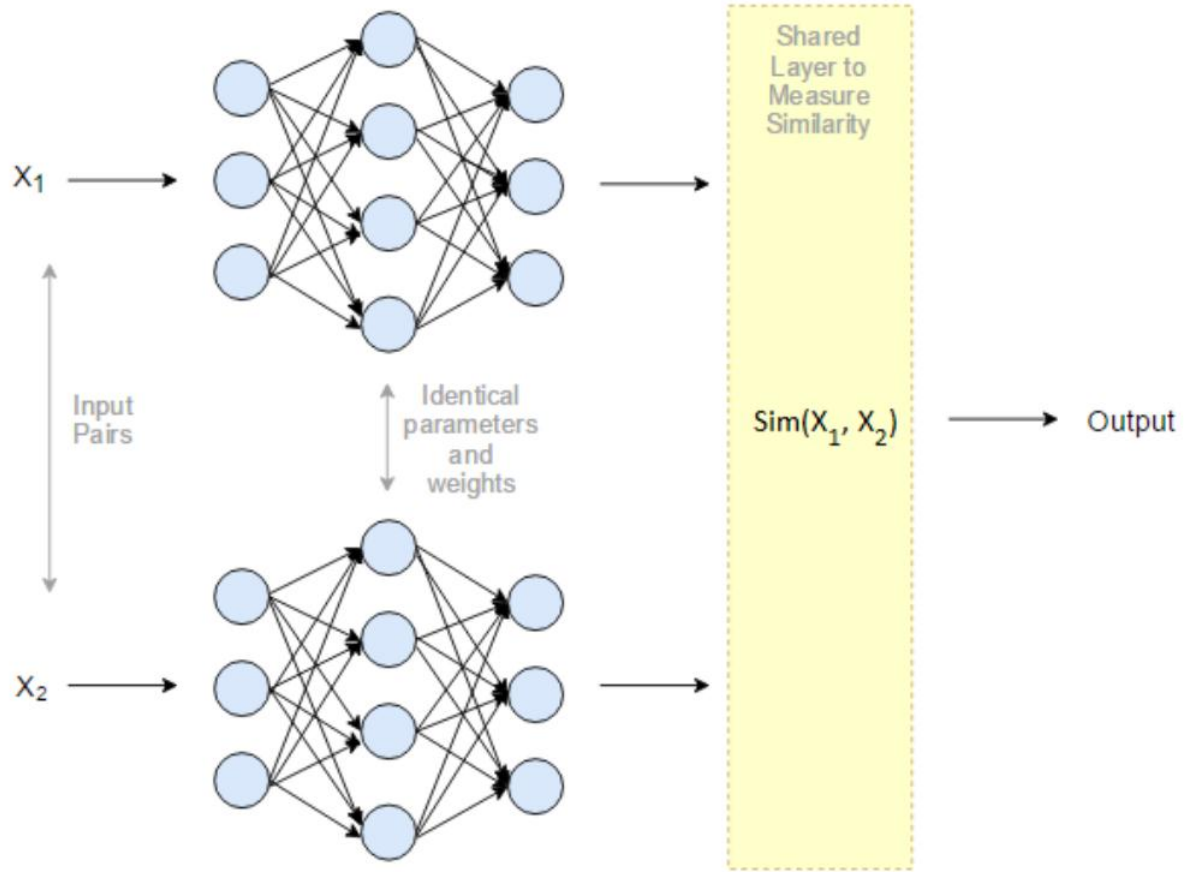
- **Рекуре́нтні нейро́нні мере́жі** — це тип нейронних мереж, в яких є зворотний зв'язок. При цьому під зворотним зв'язком мається на увазі зв'язок від логічно найвіддаленішого елементу до менш віддаленого. Наявність зворотних зв'язків дозволяє запам'ятовувати і відтворювати цілі послідовності реакцій на один стимул.

Цільова функція нейронної мережі

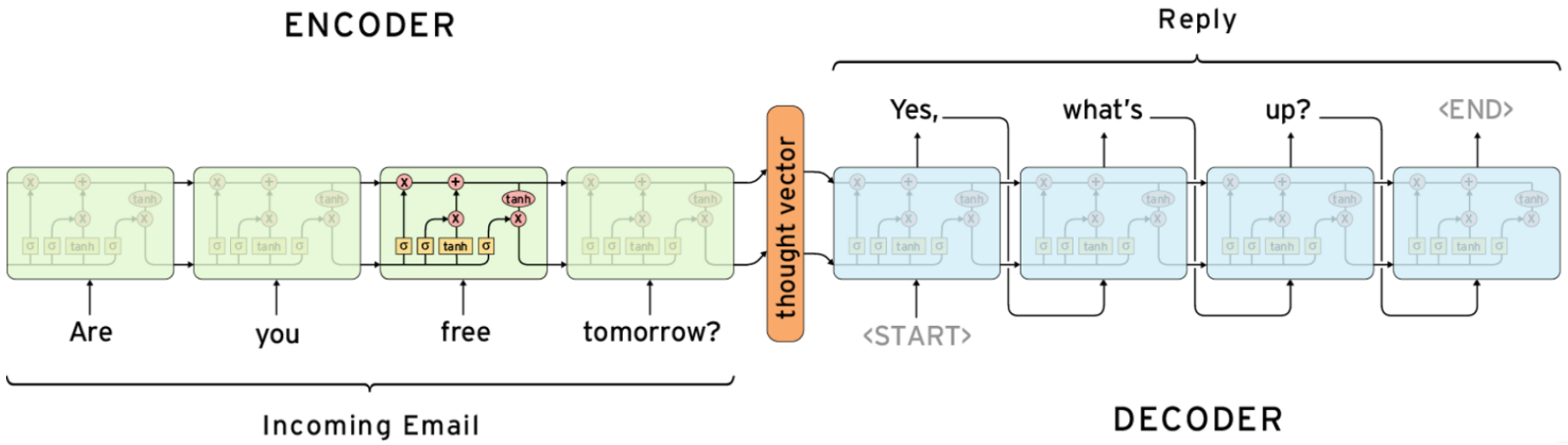
$$L(w) = \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{2N}\sum_{(i,j)\in D} y_{ij} d_{ij}^2 + (1 - y_{ij})\max(1 - d_{ij}^2, 0) \qquad (2.9)$$

- де w - ваги нейронної мережі,
- D - набір навчальних пар,
- $d^2{}_{ij}$ - квадратична відстань
- l_2 відстань між i і j послідовностями (розраховане між двома останніми шарами сіамської ШНМ),
- $y_{ij} \in 0, 1$ – як і було розглянуто раніше, значення відповідає за схожість або відмінність послідовностей.

Архітектура нейронної мережі

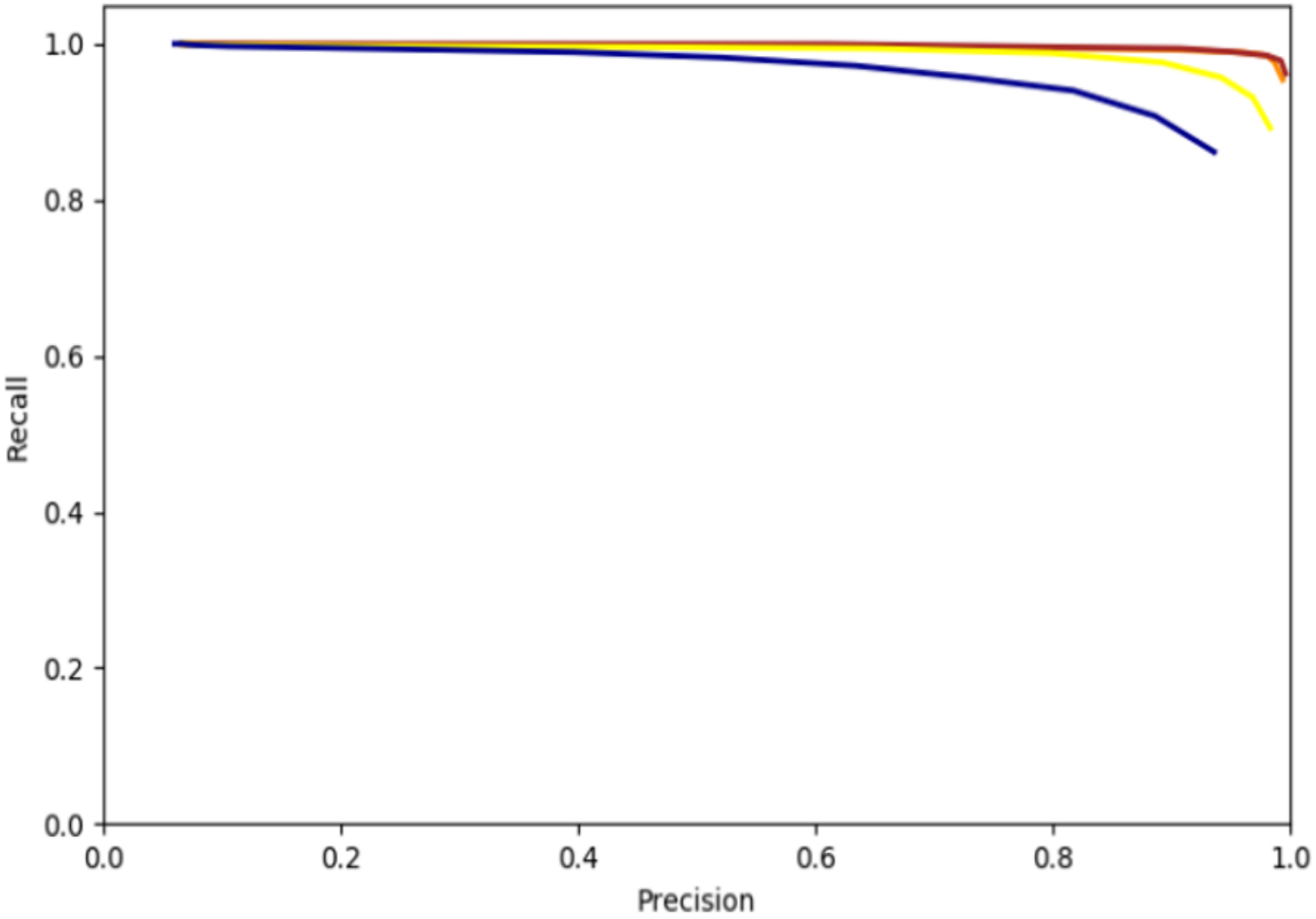


Модель Sequence-to-Sequence

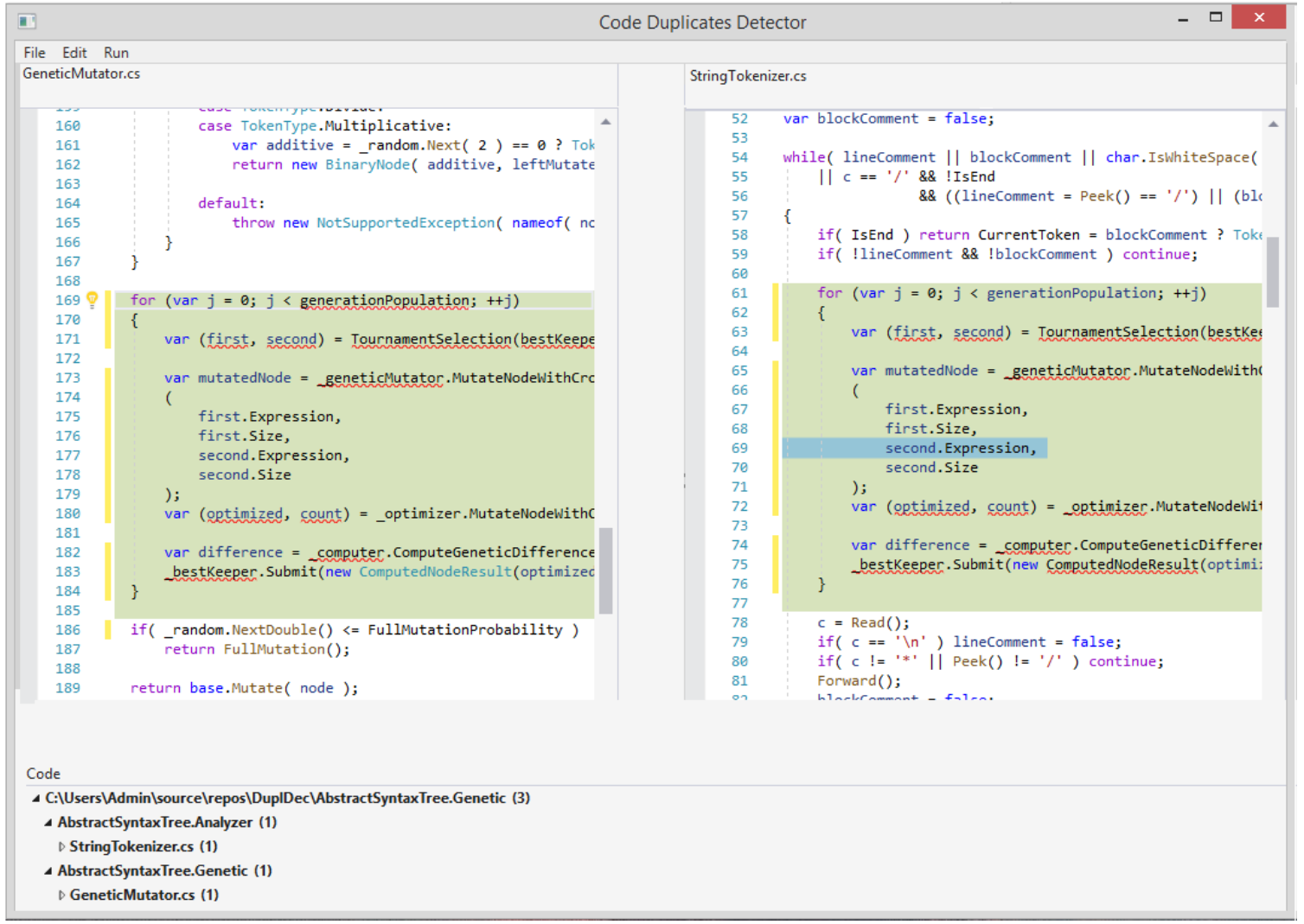


Результати тестування методів виявлення дублікатів

	BigCloneBench			OJClone		
Метод	P	R	F_1	P	R	F_1
Deckard	0.83	0.91	0.86	0.79	0.86	0.82
CDLH	0.85	0.80	0,82	0.71	0.67	0.68
SourcererCC	0.76	0.82	0.78	0.63	0.74	0.68
AST+RNN	0.94	0.96	0.95	0.94	0.95	0.94


1.0
0.8
0.6
0.4
0.2
0.0
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Precision

Програмна ралізація



Висновки

- Підвищення точності виявлення на 12%
- Підвищення повноти виявлення на 10%

в порівнянні з існуючими програмними реалізаціями методів

Дякую за увагу