

Приложение Б

(Справочное)

Сравнительный обзор информационных источников

Таблица 2.1 — Обзор источников

Название работы	Авторы, год публикации	Методы, использованные в работе	Описание данных	Достигнутая точность классификации	Язык программирования
Using classification techniques to determine source code authorship [6]	B. Pellin, 2008	АСТ, SVM	4 схожие программы, 2 автора	67 — 88 %	Java
Source code authorship attribution using n-grams [7]	S. Burrows, S. Tahaghoghi, 2007	N-граммы	Выборка из 1640 файлов исходного кода и 100 авторов	67 %	C
Identifying Authorship by Byte-Level N-Grams: The Source Code Author Profile (SCAP) Method [5]	G. Frantzeskou, E. Stamatatos, S. Gritzalis, 2007	Составление профиля программиста на основе статистических метрик, подсчет отклонения от профиля	Не указано	88 % для C++, 100 % для Java	Java, C++
Application of information retrieval techniques for source code authorship attribution [9]	S. Burrows, A. Uitdenboger, T. Urpin, 2009	N-граммы, рейтинговые схемы	100 авторов, классифицировались по 10, 1579 программных файлов	77 %	C

Продолжение таблицы 2.1

Название работы	Авторы, год публикации	Методы, использованные в работе	Описание данных	Достигнутая точность классификации	Язык программирования
De-anonymizing Programmers via Code Stylometry [10]	A. Caliskan-Islam, R. Harang, A. Liu, F. Yamaguchi, 2015	Статистический подсчет признаков, нечеткие АСТ	250 авторов, 1600 файлов	94 — 98 %	C/C++, Python
Git Blame Who?: Stylistic Authorship Attribution of Small, Incomplete Source Code Fragments [12]	A. Caliskan-Islam, E. Dauber, R. Harang, R. Greenstadt, 2017	Калибровочные кривые, нечеткие АСТ, классифика-тор Random Forest	Некомпилируемые неполные образцы кода с ресурса GitHub	70 — 100 %	C/C++