

KLASTERISASI MULTILEVEL SOURCE CODE C BERBASIS REPRESENTASI AST DAN DISASSEMBLY UNTUK DETEKSI POTENSI PLAGIARISME

AHMAD NUR ROHIM



**PROGRAM SARJANA ILMU KOMPUTER
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026**

@Hak cipta milik IPB University

IPB University



IPB University
— Bogor Indonesia —

Perpustakaan IPB University

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Klasterisasi Multilevel *Source Code* C Berbasis Representasi AST dan Disassembly untuk Deteksi Potensi Plagiarisme” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Juli 2026

Ahmad Nur Rohim
G6401221001



ABSTRAK

AHMAD NUR ROHIM. Klasterisasi Multilevel *Source Code C* Berbasis Representasi AST dan Disassembly untuk Deteksi Potensi Plagiarisme. Dibimbing oleh AZIZ KUSTIYO dan MUSHTHOFA.

Plagiarisme kode sumber merupakan permasalahan penting dalam pembelajaran pemrograman karena kemiripan kode dapat muncul melalui penyalinan langsung maupun modifikasi struktur program. Penelitian ini mengusulkan pendekatan klasterisasi multilevel untuk mendeteksi indikasi plagiarisme kode sumber C dengan mengombinasikan representasi Abstract Syntax Tree (AST) dan disassembly. AST digunakan untuk merepresentasikan struktur sintaksis program, sedangkan disassembly digunakan untuk merepresentasikan struktur semantik program. Dataset yang digunakan terdiri atas tiga dataset kode sumber C, yaitu dua dataset sintetis dan satu dataset organik dari platform online judge. Ekstraksi fitur dilakukan menggunakan TF-IDF, reduksi dimensi menggunakan Singular Value Decomposition (SVD). Kombinasi fitur dilakukan menggunakan concatenation, baik tanpa Canonical Correlation Analysis (CCA) maupun dengan tambahan CCA. Proses klasterisasi dilakukan menggunakan DBSCAN dan dievaluasi dengan Rand Index (RI), Silhouette Score, serta Davies-Bouldin Index (DBI). Hasil penelitian menunjukkan bahwa representasi AST memperoleh rata-rata RI, Silhouette Score, dan DBI masing-masing sebesar 0.7537, 0.6570, dan 0.7390, sedangkan representasi disassembly meningkat menjadi 0.8320, 0.7362, dan 0.6584. Multilevel tanpa CCA menghasilkan RI tertinggi sebesar 0.8618, sementara multilevel dengan CCA menghasilkan kualitas internal terbaik dengan Silhouette Score 0.8781 dan DBI 0.4115. Hasil tersebut menunjukkan bahwa representasi multilevel mampu meningkatkan efektivitas klasterisasi kode sumber.

Kata kunci: Abstract Syntax Tree, DBSCAN, disassembly, klasterisasi multilevel, plagiarisme kode sumber.

ABSTRACT

AHMAD NUR ROHIM. *Multilevel Clustering of C Source Code Based on AST and Disassembly Representations for Potential Plagiarism Detection*. Supervised by AZIZ KUSTIYO and MUSHTHOFA.

Source code plagiarism is a significant concern in programming education, as code similarity can result from both direct copying and structural modification. This study presents a multilevel clustering approach for detecting signs of plagiarism in C source code by combining Abstract Syntax Tree (AST) and disassembly representations. The AST captures the syntactic structure of the code, while the disassembly representation captures the semantic structure of the code. The dataset used in this study consists of three C source code datasets, two synthetic datasets and one organic dataset collected from an online judging platform. Feature extraction was conducted using TF-IDF, and dimensionality reduction was achieved through Singular Value Decomposition (SVD). Feature combination was



performed using concatenation, both without Canonical Correlation Analysis (CCA) and with the addition of CCA. For the clustering process, we utilized the DBSCAN algorithm and evaluated the results using the Rand Index (RI), Silhouette Score, and Davies-Bouldin Index (DBI). The findings show that AST achieved average RI, Silhouette Score, and DBI values of 0.7537, 0.6570, and 0.7390, respectively, while disassembly achieved 0.8320, 0.7362, and 0.6584. The multilevel approach without CCA produced the highest RI of 0.8618, whereas the multilevel approach with CCA yielded the best internal cluster quality, with a Silhouette Score of 0.8781 and a DBI of 0.4115. These results indicate that multilevel representation enhances the effectiveness of source code clustering.

Keywords: Abstract Syntax Tree, DBSCAN, disassembly, multilevel clustering, source code plagiarism.

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2026
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.



KLASTERISASI MULTILEVEL SOURCE CODE C BERBASIS REPRESENTASI AST DAN DISASSEMBLY UNTUK DETEKSI POTENSI PLAGIARISME

AHMAD NUR ROHIM

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana pada
Program Studi Ilmu Komputer

**PROGRAM SARJANA ILMU KOMPUTER
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tim Penguji pada Ujian Skripsi:
Dr. Toto Haryanto, S.Kom, M.Si



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Judul Skripsi : Klasterisasi Multilevel *Source Code* C Berbasis Representasi AST
dan Disassembly untuk Deteksi Potensi Plagiarisme

Nama : Ahmad Nur Rohim
NIM : G6401221001

@Hak cipta milik IPB University

Disetujui oleh

Pembimbing 1:
Dr. Aziz Kustiyo S.Si., M.Kom.

Pembimbing 2:
Dr. Mushthofa S.Kom., M.Sc.

Diketahui oleh

Ketua Program Sarjana Ilmu Komputer:
Dr. Sony Hartono Wijaya, S.Kom., M.Kom.
19810809 200812 1 002

Tanggal Ujian:
2 Juli 2026

Tanggal Lulus:



PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan dengan baik. Penelitian ini disusun dengan judul “Klasterisasi Multilevel *Source Code C* Berbasis Representasi AST dan Disassembly untuk Deteksi Potensi Plagiarisme”.

Penulis menyadari bahwa terselesaikannya karya ilmiah ini merupakan hasil dari dukungan, bimbingan, arahan, serta bantuan yang diberikan oleh berbagai pihak. Oleh karena itu, penulis menyampaikan penghargaan dan ucapan terima kasih yang sebesar-besarnya kepada:

1. Bapak Dr. Aziz Kustiyo S.Si., M.Kom. dan Bapak Dr. Mushtofa S.Kom., M.Sc. selaku dosen pembimbing yang telah meluangkan waktu, tenaga, dan pikiran dalam memberikan arahan, masukan, serta bimbingan selama proses penelitian dan penulisan karya ilmiah ini.
2. Bapak Dr. Toto Haryanto, S.Kom, M.Si selaku dosen penguji, atas saran, kritik, dan masukan yang membangun demi penyempurnaan karya ilmiah ini.
3. Seluruh dosen dan sivitas akademika yang telah memberikan ilmu, bimbingan, serta dukungan kepada penulis selama menempuh pendidikan di Institut Pertanian Bogor.
4. Bapak, Ibu, dan Adik tercinta yang selalu memberikan doa, semangat, dan dukungan yang tiada henti kepada penulis dalam setiap tahap perjalanan studi.
5. Teman-teman TLEverse, Yudo, dan Zaky yang telah menjadi rekan seperjuangan dalam berbagai kompetisi selama masa studi serta memberikan dukungan dan motivasi kepada penulis. Tak lupa, Putra, Uno, Adid, Yuda, Naya, Berlin, Wisnu, Habib, dan Fari yang sudah menjadi rekan seperjuangan selama masa studi di Ilmu Komputer.
6. Seluruh teman Ilkomerz 59 yang tidak dapat disebutkan satu-persatu yang telah memberi semangat dan berjuang bersama-sama selama perkuliahan.

Besar harapan penulis agar karya ilmiah ini dapat memberikan manfaat, baik sebagai referensi bagi penelitian selanjutnya maupun sebagai kontribusi dalam pengembangan ilmu pengetahuan.

Bogor, Juli 2026

Ahmad Nur Rohim



DAFTAR ISI

DAFTAR TABEL	xi
DAFTAR GAMBAR	xi
DAFTAR LAMPIRAN	xi
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan	2
1.4 Manfaat	2
1.5 Ruang Lingkup	3
TINJAUAN PUSTAKA	4
2.1 Plagiarisme Kode Sumber	4
2.2 AST	4
2.3 Disassembly	5
2.4 TF-IDF	5
2.5 Cosine Similarity	6
2.6 DBSCAN	6
2.7 Rand Index	6
2.8 Silhouette Coefficient	7
2.9 Davies-Bouldin Index	7
2.10 Singular Value Decomposition	8
2.11 Canonical Correlation Analysis	8
METODE	10
3.1 Data Penelitian	10
3.2 Tahapan Penelitian	10
3.3 Peralatan Penelitian	15
HASIL DAN PEMBAHASAN	16
4.1 Pengumpulan Data dan Praproses	16
4.2 Ekstraksi Fitur dan Pembentukan Vektor	16
4.3 Klasterisasi Abstract Syntax Tree	20
4.4 Klasterisasi Disassembly	20
4.5 Klasterisasi Multilevel (Kombinasi AST dan Disassembly)	22
4.6 Perbandingan Antar Metode	24
SIMPULAN DAN SARAN	29
5.1 Simpulan	31
5.2 Saran	31
DAFTAR PUSTAKA	33
LAMPIRAN	36
RIWAYAT HIDUP	42

DAFTAR TABEL

1	Ringkasan dataset	16
2	Hasil evaluasi klasterisasi AST	20
3	Hasil evaluasi klasterisasi Disassembly	21
4	Hasil evaluasi klasterisasi multilevel tanpa CCA	22
5	Hasil evaluasi klasterisasi multilevel dengan CCA	23

DAFTAR GAMBAR

1	Tahapan penelitian	10
2	Ilustrasi kombinasi vektor	13
3	Contoh hasil visualisasi AST	17
4	Perbandingan rata-rata metrik evaluasi klasterisasi per metode dan dataset	24
5	Perbandingan kualitas klasterisasi berdasarkan Rand Index	25
6	Perbandingan kualitas klasterisasi berdasarkan Silhouette Score	26
7	Perbandingan kualitas klasterisasi berdasarkan Davies-Bouldin Index	27
8	Perbandingan banyak klaster hasil klasterisasi	28
9	Perbandingan waktu antar metode	29

DAFTAR LAMPIRAN

1	Detail anggota klaster <i>ground truth</i> setiap dataset	37
2	Contoh fitur TF-IDF AST	39
3	Contoh fitur TF-IDF Disassembly	40
4	Tautan repositori dan berkas pendukung penelitian	41