

**SISTEM DETEKSI PLAGIASI MENGGUNAKAN
ALGORITME FREQUENCY BASED HASHING-S PADA
FILE PDF**

SKRIPSI



**Oleh
Thamrin Auliya
180210117**

**FAKULTAS TEKNIK
PROGRAM STUDI TEKNIK INFORMATIKA
UNIVERSITAS PUTERA BATAM
TAHUN 2021/ 2022**

**SISTEM DETEKSI PLAGIASI MENGGUNAKAN ALGORITME
FREQUENCY BASED HASHING-S PADA FILE PDF**

SKRIPSI

**Untuk memenuhi salah satu syarat
memperoleh gelar sarjana**



**Oleh
Thamrin Auliya
180210117**

**FAKULTAS TEKNIK
PROGRAM STUDI TEKNIK INFORMATIKA
UNIVERSITAS PUTERA BATAM
TAHUN 2021/ 2022**

SURAT PERNYATAAN ORISINALITAS

Yang bertanda tangan dibawah ini penulis:

Nama : Thamrin Auliya
NPM : 180210117
Fakultas : Teknik dan Komputer
Program studi : Teknik Informatika

Menyatakan bahwa **“Skripsi”** yang dibuat dengan judul:

“SISTEM DETEKSI PLAGIASI MENGGUNAKAN ALGORITME FREQUENCY BASED HASHING-S PADA FILE PDF”

Ini adalah karya sendiri dan bukan “duplikasi” dari karya orang lain. Sejauh yang penulis tahu dalam teks skripsi ini tidak ada karya ilmiah atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang disebutkan dalam teks ini dan disebutkan dalam sumber referensi kutipan. Apabila terdapat didalam naskah Skripsi ini dapat dibuktikan terdapat unsur unsur PLAGIASI, saya bersedia naskah Sikripsi ini digugurkan dan gelar akademik yang saya peroleh dibatalkan, serta diproses sesuai dengan peraturan perundangundang yang berlaku.

Demikian pernyataan ini saya buat dengan sebenarnya tanpa ada paksaan dari siapapun.

Batam, 8 Agustus 2022



Thamrin Auliya
180210117

**SISTEM DETEKSI PLAGIASI MENGGUNAKAN
ALGORITME FREQUENCY BASED HASHING-S
PADAFILE PDF**


SKRIPSI

**Untuk memenuhi salah satu syarat
memperoleh gelar sarjana**

**Oleh
Thamrin Auliya
180210117**

**Telah disetujui pembimbing pada
tanggal seperti yang tertera dibawah
ini**

Batam, 8 agustus 2022



Cosmas Eko Suharyanto, S.Kom., M.MSI

Pembimbing

ABSTRAK

Di era komputasi digital sekarang ini, telah terjadi pertumbuhan produksi data digital yang sangat tinggi. Dikutip dari laman forbes.com, ada 2,5 quintillion bytes data yang diproduksi setiap hari. Salah satu data digital tersebut adalah tugas mahasiswa yang dikumpulkan melalui media Google Classroom. Dari data mahasiswa yang dikumpulkan tersebut, banyak tugas yang terindikasi plagiasi dengan tugas mahasiswa yang lainnya. Jika dilakukan analisis secara manual, maka akan memakan waktu yang lama dan sangat melelahkan. Untuk efisiensi pemanfaatan waktu dan sumber daya, dibutuhkan proses penyaringan yang memiliki kemampuan untuk menghitung tingkat plagiasi dari setiap tugas mahasiswa. Metode Approximate Matching merupakan metode yang paling sering digunakan untuk menemukan kesamaan diantara data yang dibandingkan dengan menetapkan skor kesamaan. Pada penelitian ini algoritme Approximate Matching yang digunakan adalah algoritme Frequency Based Hashing-S. Kelebihan algoritme ini adalah aman terhadap serangan aktif dan memiliki tingkat akurasi 98% untuk format data terkompresi. Aplikasi ini bermanfaat bagi dosen yang akan memeriksa tugas mahasiswa yang banyak. Dengan adanya aplikasi ini, dosen hanya perlu memeriksa tugas mahasiswa yang tidak plagiasi saja sehingga mengurangi jumlah tugas mahasiswa yang akan diperiksa secara manual.

Kata Kunci: Frequency Based Hashing-S, Sistem Plagiasi, Sitem Penyaringan Data

ABSTRACT

In today's digital computing era, there has been a very high growth of digital data production. Quoted from the forbes.com page, there are 2.5 quintillion bytes of data created every day. One of the digital data is student assignments collected through Google Classroom media. From the student data collected, many assignments indicated plagiarism with other student assignments. If the analysis is done manually, it will take a long time and be very tiring. For efficient use of time and resources, a screening process is needed that has the ability to calculate the plagiarism level of each student's assignment. The Approximate Matching method is the method most often used to find similarities between the data being compared by setting a similarity score. In this study, the Approximate Matching algorithm used is the Frequency Based Hashing-S algorithm. The advantages of this algorithm are that it is safe against active attacks and has a 98% accuracy rate for compressed data formats. This application is useful for lecturers who will check a lot of student assignments. With this application, lecturers only need to check student assignments that are not plagiarized, thereby reducing the number of student assignments that will be checked manually.

Keywords: Frequency Based Hashing-S, Plagiarism System, Data Filtering System

KATA PENGANTAR

Segala puji dan syukur penulis panjatkan Kehadirat Tuhan Yang Maha Esa atas kuasa dan limpahan rahmat-Nya sehingga penulis dapat menyelesaikan proposal yang merupakan salah satu syarat untuk menyelesaikan program studi strata satu (S1) pada program studi Teknik Informatika Universitas Putera Batam.

Dengan segala keterbatasan, penulis menyadari pula bahwa proposal ini tidak akan terwujud tanpa bantuan, bimbingan, dan dorongan dari berbagai pihak. Untuk itu dengan segala kerendahan hati, penulis menyampaikan ucapan terimakasih kepada:

1. Ibu Dr. Nur Elfi Husda, S.Kom., M.SI. selaku Rektor Universitas Putera Batam.
2. Bapak Welly Sugianto, S.T., M.M. selaku Dekan Fakultas Teknik dan Komputer.
3. Ketua program studi teknik informatika Bapak Andi Maslan, S.T.,M.SI
4. Bapak Cosmas Suharyanto, S.Kom., M.Kom, selaku pembimbing skripsi pada program Stusi Teknik Informatika Universitas Putera Batam.
5. Kepada orang tua tercinta atas curahan kasih sayang, doa, nasihat, serta pesan yang disampaikan kepada penulis sehingga penulis tetap memiliki semangat juang dalam penyelesaian proposal ini.
6. Serta semua pihak yang tidak dapat saya dapat penulis sebutkan satu persatu yang telah membantu dalam penyusunan skripsi ini. yang telah membantu penyelesaian skripsi ini yang tidak dapat penulis sebutkan satu persatu

Batam, 8 Agustus 2022

Thamrin Auliya

DAFTAR ISI

SURAT PERNYATAAN ORISINALITAS	i
ABSTRAK	iv
ABSTRACT	v
KATA PENGANTAR	v
DAFTAR ISI	viii
DAPRAR TABEL	xii
DAPRAR GAMBAR	xiv
BAB 1	1
1.1. Latar Belakang	1
1.2. Identifikasi Masalah	4
1.3. Batasan Masalah	4
1.4. Rumusan Masalah	5
1.5. Tujuan Penelitian	6
1.6. Manfaat Penelitian	6
BAB II	8
2.1 Dasar Teori	8
2.2.1. Kriptografi	8
2.2.2. Hashing	8
2.2.3. Approximate Matching.....	9
2.2.4. Optical Character Recognition	9
2.2.5. Cosine Similarity.....	10
2.2.6. Frequency Based Hashing	10
2.2 Variabel (Indikator Masalah)	15
2.2.1 Pemeriksaan Tugas Siswa Secara Manual Kurang Efektif.....	15
2.2.2 Metode Pencocokan Fungsi Hash Tradisional Memiliki	

Keterbatasan 16

2.2.3 Algoritme Approximate Matching Rentan Terhadap Serangan Aktif 17

2.3 Perangkat Lunak Pendukung..... 19

2.3.1 XAMPP..... 20

2.3.2 Sublime 23

2.3.3 Penelitian Terdahulu 24

BAB III..... 28

3.1 Desain Penelitian 28

3.2 Perancangan Sistem 31

3.4.1 Perancangan Data dan UML..... 32

3.4.2 Perancangan Database 46

3.4.3 Perancangan Antarmuka..... 48

3.3 Perhitungan Manual 50

3.3.1 Digest 1 50

3.3.2 Digest 2 68

3.3.3 Cosin Similarity..... 86

3.4 Lokasi dan Jadwal Penelitian..... 86

3.4.1 Lokasi Penelitian 86

3.4.2 Jadwal Penelitian 86

BAB IV..... 88

4.1 Hasil Penelitian 88

4.1.1 Web Service Dapat Melakukan Proses Tambah Dataset..... 88

4.1.2 Web Service Dapat Melakukan Prose Tambah Datatest 90

4.1.3 Web Service Dapat Menampilkan Dataset..... 91

4.1.4 Web Service Dapat Menampilkan Nilai Skor Kesamaan..... 92

4.1.5	Pengujian Kinerja Sistem	94
4.1.6	Pengujian Akurasi Perhitungan Similarity	104
4.1.7	Pengujian Integritas Data.....	116
4.2	Pembahasan	124
4.2.1	Web Service Dapat Melakukan Proses Tambah Dataset.....	124
4.2.2	Web Service Dapat Melakukan Proses Tambah Datatest	125
4.2.3	Web Service Dapat Menampilkan Dataset.....	125
4.2.4	Web Service Dapat Menampilkan Nilai Similarity.....	125
4.2.5	Pengujian Kinerja Sistem	126
4.2.6	Pengujian Akurasi Perhitungan Similarity	130
4.2.7	Pengujian Integritas Data.....	130
BAB V	132
5.1	Kesimpulan	132
5.2	Saran.....	134
DAPTAR PUSTAKA	135
LAMPIRAN	138
Lampiran 1. Daftar riwayat hidup	138
Lampiran 2. Surat Keterangan Izin Penelitian	139
Lampiran 3. Hasil Turnitin Skripsi	140
Lampiran 4. Hasil Turnitin Jurnal	140
Lampiran 5. Letter Of Acceptance(LOA)	141

DAPTAR TABEL

Tabel 3. 1 Tabel Dataset.....	32
Tabel 3. 2 Tabel Datatestz39	
Tabel 3. 3 Tabel Struktur kosin_proses_dataset.....	47
Tabel 3. 4 Tabel Struktur kosin_proses_datatest	47
Tabel 3. 5 Tabel Struktur kosin_proses_tf	47
Tabel 3. 6 Jadwal Penelitian.....	87
Tabel 4. 1 Skenarion Pengujian Tambah Dataset	88
Tabel 4. 2 Pengujian Tambah Datatest.....	90
Tabel 4. 3 Pengujian Menampilkan Dataset.....	91
Tabel 4. 4 Pengujian Menampilkan Nilai Skor kesamaan	92
Tabel 4. 5 Pengujian Kinerja Sistem.....	94
Tabel 4. 6 Pengujian Waktu Eksekusi Sistem.....	95
Tabel 4. 7 Data Untuk Pengujian Kinerja Sistem	96
Tabel 4. 8 Waktu Eksekusi Proses Unggah Dataset.....	97
Tabel 4. 9 Waktu Eksekusi Proses Unggah Datatest	99
Tabel 4. 10 Waktu Eksekusi Proses Perhitungan Nilai Skor Kesamaan.....	102
Tabel 4. 11 Pengujian Akurasi Perhitungan Nilai Similarity.....	104
Tabel 4. 12 Data Pengujian Akurasi Perhitungan Similarity	106
Tabel 4. 13 Perhitungan Nilai GS	113
Tabel 4. 14 Proses <i>Confusion Matriks</i>	115
Tabel 4. 15 <i>Confusion Matriks</i>	115

Tabel 4. 16 Skenario Pengujian Integritas Data	117
Tabel 4. 17 Nilai <i>Chunk</i> dan <i>Rolling Hash</i> Teks 1.....	117
Tabel 4. 18 Nilai <i>Chunk</i> dan <i>Rolling Hash</i> Teks 2.....	117
Tabel 4. 19 Hasil Pengujian Penambahan Dataset	125
Tabel 4. 20 Hasil Pengujian Penambahan Datatest	125
Tabel 4. 21 Hasil Pengujian Menampilkan Dataset	125
Tabel 4. 22 Hasil Pengujian Menampilkan Nilai Skor kesamaan	126
Tabel 4. 23 Hasil Pengujian Kinerja Sistem	126
Tabel 4. 24 Hasil Pengujian Akurasi Perhitungan Nilai Kesamaans	130
Tabel 4. 25 Hasil Pengujian Integritas Data.....	130

DAPTAR GAMBAR

Gambar 2. 1 Tahapan Algoritme Frequency Based Hashing-S	11
Gambar 3. 1 Diagram Alir Penelitian.....	28
Gambar 3. 2 Diagram alir OCR PyPDF2	41
Gambar 3. 3 Diagram alir FbHash-S.....	42
Gambar 3. 4 Diagram alir Cosine Similarity.....	43
Gambar 3. 5 Sequence diagram unggah dataset.....	45
Gambar 3. 6 Sequence diagram perhitungan nilai similarity	46
Gambar 3. 7 Tampilan home.....	48
Gambar 3. 8 Tampilan tambah datateset	49
Gambar 3. 9 Tampilan tambah datatest.....	49
Gambar 3. 10 Tampilan hasil perhitungan nilai similarity.....	50
Gambar 4. 1 Tampilan tabel kosin_proses_dataset	89
Gambar 4. 2 Tampilan tabel kosin_proses_tf.....	90
Gambar 4. 3 Tampilan tabel kosin_proses_datatest.....	91
Gambar 4. 4 Tampilan halaman dashboard.....	92
Gambar 4. 5 Tampilan halaman form hasil perhitungan similarity	94
Gambar 4. 6 Perhitungan Nilai AM	113
Gambar 4. 7 Grafik perbandingan rata rata waktu eksekusi program berdasarkan ukuran dan jumlah <i>chunk</i> pada unggah dataset.....	127
Gambar 4. 8 Grafik perbandingan rata rata waktu eksekusi program berdasarkan ukuran data dan jumlah <i>chunknya</i> pada unggah datatest	128
Gambar 4. 9 Grafik rata rata waktu eksekusi program pada saat proses	

perhitungan nilai skor kesamaan 129