

**IDENTIFIKASI TINGKAT KEMIRIPAN DOKUMEN TEKS
MENGUNAKAN FUNGSI HASH PADA ALGORITMA
WINNOWER DAN PATTERN RECOGNITION PADA
ALGORITMA RATCLIFF/OBERSHELP**

SKRIPSI

**YUSUF KARIM RAMBE
NIM. 0701172069**



**PROGRAM STUDI ILMU KOMPUTER
FAKULAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SUMATERA UTARA
MEDAN
2022**

**IDENTIFIKASI TINGKAT KEMIRIPAN DOKUMEN TEKS
MENGUNAKAN FUNGSI HASH PADA ALGORITMA
WINNOWER DAN PATTERN RECOGNITION PADA
ALGORITMA RATCLIFF/OBERSHELP**

SKRIPSI

Diajukan unntuk Memenuhi Syarat Mencapai Gelar Sarjana Komputer

**YUSUF KARIM RAMBE
NIM. 0701172069**



**PROGRAM STUDI ILMU KOMPUTER
FAKULAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SUMATERA UTARA
MEDAN
2022**

PERSETUJUAN SKRIPSI

Hal : Surat Persetujuan Skripsi

Lamp : -

Kepada Yth,

Dekan Fakultas Sains dan Teknologi

Universitas Islam Negeri Sumatera Utara Medan

Assalamu'alaikum Wr. Wb.

Setelah membaca, meneliti, memberikan petunjuk, dan mengoreksi serta mengadakan perbaikan, maka kami selaku pembimbing berpendapat bahwa skripsi saudara,

Nama	: Yusuf Karim Rambe
NIM	: 0701172069
Program Studi	: Ilmu Komputer
Judul Skripsi	: Identifikasi Tingkat Kemiripan Dokumen Teks Menggunakan Fungsi Hash Pada Algoritma Winnowing dan Pattern Recognition Pada Algoritma Ratcliff/Obershelp

Dapat disetujui untuk segera dimunaqasyah-kan. Atas perhatiannya kami ucapkan terima kasih.

Medan, 15 Agustus 2022

17 Muharam 1444 H

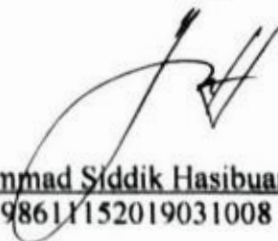
Komisi Pembimbing

Pembimbing Skripsi I,



Abdul Halim Hasugian, M.Kom
NIDN. 0427038801

Pembimbing Skripsi II,



Muhammad Siddik Hasibuan, M.Kom
NIP. 198611152019031008

SURAT PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan di bawah ini:

Nama : Yusuf Karim Rambe
NIM : 0701172069
Program Studi : Ilmu Komputer
Judul Skripsi : Identifikasi Tingkat Kemiripan Dokumen
Teks Menggunakan Fungsi Hash Pada
Algoritma Winnowing dan Pattern
Recognition Pada Algoritma
Ratcliff/Obershelp

Menyatakan bahwa skripsi ini adalah hasil karya saya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing disebutkan sumbernya. Apabila di kemudian hari ditemukan plagiat di dalam skripsi ini, maka saya bersedia menerima sanksi pencabutan gelar akademik yang saya peroleh dan sanksi lainnya sesuai dengan peraturan yang berlaku.

Medan, 15 Agustus 2022



Yusuf Karim Rambe
NIM. 0701172069

UNIVERSITAS ISLAM NE
SUMATERA UTARA MEDAN

ABSTRAK

Perkembangan teknologi yang sudah sangat maju ini dalam bidang komputer, sekarang semua dokumen sudah berubah menjadi file softcopy yang mana bisa diakses lewat komputer ataupun lewat smartphone masing-masing. Sehingga banyak yang melakukan penjiplakan dan tidak dicantumkan sumbernya sehingga terjadi pelanggaran hak cipta. Plagiarisme dapat dipahami sebagai tindakan mengambil pernyataan atau mencuri ide orang lain, dalam hukum positif, hak paten dan kekayaan intelektual sudah ada diatur dalam undang-undang. Oleh karena itu untuk memudahkan para guru dan dosen dalam bidang akademik, dibutuhkan pengidentifikasian pada dokumen-dokumen tersebut agar diketahui apakah dokumen tersebut termasuk plagiarisme dengan mendeteksi kemiripan teks antar dokumen, maka keaslian dalam tiap dokumen atau karya tulis bisa tetap terjaga keasliannya. Pengidentifikasian akan dilakukan dengan algoritma Winnowing dan Ratcliff/Obershelp. Algoritma Winnowing adalah algoritma yang digunakan untuk melakukan proses pengecekan kesamaan kata (document fingerprinting) untuk mengidentifikasi penjiplakan. Ratcliff/Obershelp juga untuk mendeteksi adanya kemiripan namun dalam prinsip kerjanya dilakukan pengembalian nilai yang dapat digunakan sebagai persentase dalam menunjukkan kesamaan dua string dengan memperhitungkan jumlah karakter yang terdapat pada kedua string tersebut. Kedua algoritma ini dapat digunakan dalam mengidentifikasi tingkat kemiripan dokumen teks.

Kata Kunci : Plagiarisme, Dokumen, Winnowing, Ratcliff/Obershelp

UNIVERSITAS ISLAM NEGERI
SUMATERA UTARA MEDAN

ABSTRACT

The development of this very advanced technology in computers era, all document have turned into softcopy file, which can be accessed with computer or smartphone. So many people commit plagiarism by not listed the source and resulting copyright infringement. The plagiarism can be interpreted as an action that takes statements of theft other people's ideas. In law, patents and intellectual property rights are already regulated In law. Therefore to make it easier for teachers and lecturers in academic. Identification of document is needed to know whether the document includes plagiarism by detecting the similarity of text between documents. Then the authenticity in each document can be maintained its authenticity. The process of identification are using Winnowing and Ratcliff/Obershelp algorithm. Winnowing algorithm is algorithm used to perform the process of checking word similarity (document fingerprint) to identify plagiarism of document. Ratcliff/Obershelp is also used to detect similarities, but it works by returning value that can be used as a percentage to show the similarity of two strings by calculating the number of characters contained in two strings. Both of these algorithm can be used to identify percentage of similarity of text document

Key Word : Plagiarism, Document, Winnowing, Ratcliff/Obershelp



UNIVERSITAS ISLAM NEGERI
SUMATERA UTARA MEDAN

KATA PENGANTAR



Assalamu'alaikum wr. wb.

Alhamdulillah rabbil'alamiin, puji syukur kehadiran Allah Swt, yang telah melimpahkan rahmat, hidayah serta karunia-Nya, sehingga penyusun dapat menyelesaikan Skripsi dalam rangka untuk memenuhi salah satu syarat untuk memperoleh gelar Sarjana pada Universitas Islam Negeri Sumatera Utara.

Shalawat serta salam semoga selalu tercurahkan kepada Rasulullah Saw. beserta keluarga, para sahabat dan pengikutnya termasuk kita semua yang senantiasa menantikan syafa'atnya kelak dihari akhir.

Skripsi ini disusun berdasarkan hasil dari bimbingan dengan dosen pembimbing dan penelitian yang telah penyusun lakukan. Penyusun menyadari bahwa penyusunan Skripsi ini tidak akan berjalan lancar tanpa dukungan dari berbagai pihak. Oleh karena itu, penyusun mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. Syahrin Harahap, MA, selaku Rektor Universitas Islam Negeri Sumatera Utara.
2. Bapak Dr. Mhd. Syahnan, MA, selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Sumatera Utara.
3. Bapak Ilka Zufria, M.Kom. selaku Ketua Jurusan Ilmu Komputer Fakultas Sains dan Teknologi Universitas Islam Negeri Sumatera Utara.
4. Bapak Abdul Halim Hasugian, M.Kom, selaku pembimbing I yang telah berkontribusi membantu penulis dalam memberikan ide, saran, kritik, dan bimbingannya kepada penulis selama penulis mengerjakan proposal skripsi ini.
5. Bapak Muhammad Siddik Hasibuan, M.Kom, selaku Dosen Pembimbing II yang juga telah berkontribusi membantu penulis dalam memberikan ide, saran, kritik, dan bimbingannya kepada penulis selama penulis mengerjakan proposal skripsi ini.

6. Bapak Rakhmat Kurniawan. R, M.Kom. selaku Sekretaris Jurusan Ilmu Komputer Fakultas Sains dan Teknologi Universitas Islam Negeri Sumatera Utara.
7. Kepada orang tua saya yang telah memberi dukungan dan motivasi sehingga saya bisa menyelesaikan Skripsi ini.
8. Serta seluruh teman-teman mahasiswa dan semua pihak yang telah memberikan dukungan sehingga penulis dapat menyelesaikan Skripsi ini. Penyusun menyadari bahwa Skripsi ini masih banyak terdapat kekurangan,

Oleh karena itu kritik dan saran yang membangun sangat diharapkan untuk kebaikan dikemudian hari. Penulis juga berharap semoga semua program kerja yang telah terlaksana bermanfaat bagi penyusun, mahasiswa/i dan seluruh pihak terkait. Amin. Wassalamu'alaikum wr. wb.

Medan, 15 Agustus 2022

Penulis



YUSUF KARIM RAMBE
NIM.0701172069

UNIVERSITAS ISLAM NEGERI
SUMATERA UTARA MEDAN

DAFTAR ISI

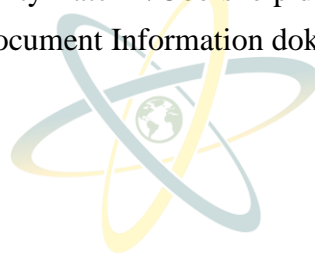
ABSTRAK	i
ABSTRACT	ii
KATA PENGANTAR.....	i
DAFTAR ISI.....	v
DAFTAR GAMBAR.....	vii
DAFTAR TABEL	ix
DAFTAR LAMPIRAN	x
BAB 1 PENDAHULUAN	Error! Bookmark not defined.1
1.1 Latar Belakang	2
1.2 Rumusan Masalah	3
1.3 Batasan Masalah.....	3
1.4 Tujuan Penelitian.....	4
1.5 Manfaat Penelitian.....	4
BAB II TINJAUAN PUSTAKA.....	5
2.1 Machine Learning.....	5
2.2 Teks Mining	5
2.2.1 Pengertian Plagiarism.....	6
2.2.1.1 Tipe-Tipe Plagiarism.....	7
2.2.1.2 Metode Pendeteksian Plagiarism	7
2.3 Preprocessing	9
2.4 Algoritma WInnowing.....	10
2.4.1 Metode K-Gram.....	13
2.4.2 Metode hash.....	13
2.4.3 Window.....	14
2.4.4 Fingerprint Dokumen.....	14
2.4.5 Jaccard's Similarity Coeficient.....	15
2.5 Algoritma Ratcliff/Obershelp.....	18
2.6 UML (Unified Modelling Language).....	18
2.6.1 Use Case Diagram	18
2.6.2 Activity Diagram	20
2.6.3 Class Diagram.....	21
2.7 Python.....	22

2.8 Penelitian yang Relevan	23
BAB III METODE PENELITIAN	28
3.1 Waktu dan Jadwal Pelaksanaan.....	28
3.2 Bahan dan Penelitian	28
3.2.1 Bahan penelitian	28
3.2.2 Alat penelitian.....	28
3.2.2.1 Perangkat keras	29
3.2.2.2 Perangkat lunak.....	29
3.3 Analisis Sistem	30
3.3.1 Perencanaan	33
3.3.2 Teknik Pengumpulan data	33
3.3.3 Analisa Kebutuhan.....	33
3.3.3.1 Metode Analisis	34
3.3.3.2 Hasil Analisis	34
3.3.3.3 Kebutuhan perangkat lunak.....	34
3.3.3.4 Kebutuhan Perangkat Keras.....	34
3.3.4 Flowchart Sistem	35
3.3.5 Pengujian	38
3.3.6 Penerapan/penggunaan	39
BAB IV HASIL DAN PEMBAHASAN	40
4.1 Pembahasan	40
4.1.1 Analisis Metode Yang Digunakan.....	40
4.1.2 Analisis Data.....	42
4.1.3 Representasi Data	42
4.2 Kemiripan Teks secara Manual.....	43
4.3 Pengujian Sistem.....	57
4.3.1 Pengujian Kemiripan Menggunakan Teks yang Berbeda.....	57
4.3.2 Pengujian Kemiripan Menggunakan Teks yang Sama	77
4.3.3 Hasil Pengujian	82
BAB V KESIMPULAN DAN SARAN	84
5.1 Kesimpulan.....	84
5.2 Saran	84
DAFTAR PUSTAKA	Error! Bookmark not defined.

DAFTAR GAMBAR

Gambar	Judul Gambar	Halaman
2. 1	Klasifikasi metode pendeteksi plagiarisme	9
2. 2	Hasil K-Gram terhadap teks	13
3. 1	Diagram sistem.....	30
3. 2	Skema aliran data	31
3. 3	Skema algoritma Winnowing.....	32
3. 4	Skema algoritma ratcliff/obershelp	32
3. 5	Flowchart perancangan sistem bagian 1	36
3. 6	Flowchart perancangan sistem bagian 2	37
4. 1	Pembentukan k-gram manual pada dokumen pertama	45
4. 2	Pembentukan k-gram manual pada dokumen kedua.....	45
4. 3	Perhitungan nilai hash manual pada dokumen pertama	50
4. 4	Perhitungan nilai hash manual pada dokumen kedua	53
4. 5	Pembentukan w-gram manual pada dokumen pertama	54
4. 6	Pembentukan w-gram manual pada dokumen kedua.....	55
4. 7	Pemilihan fingerprint manual pada dokumen pertama	55
4. 8	Pemilihan fingerprint manual pada dokumen kedua.....	56
4. 9	Output pemanggilan teks dari docx.....	59
4. 10	Output preprocessing pada dokumen yang berbeda.....	60
4. 11	Output pembentukan k-gram.....	61
4. 12	Output program pembentukan nilai hash	62
4. 13	Output pembentukan w-gram.....	63
4. 14	Output pemilihan fingerprint dari kedua dokumen	64
4. 15	Output hasil winnowing dengan jaccard similarity coefficient.....	64
4. 16	Output total keseluruhan karakter	65
4. 17	Output hasil pencarian kesamaan kata dari kedua dokumen	66
4. 18	Output hasil similarity dengan Rafcliff/Obershelp	66
4. 19	Output hasil dari fitur Get Document Information	68
4. 20	Properties information dokumen 1	72
4. 21	Properties information dokumen 2.....	72
4. 22	Properties information dokumen 3.....	73
4. 23	Properties information dokumen 4.....	73

4. 24	Properties information dokumen 5.....	74
4. 25	Properties information dokumen 6.....	74
4. 26	Properties information dokumen 7.....	75
4. 27	Properties information dokumen 8.....	75
4. 28	Properties information dokumen 9.....	76
4. 29	Properties information dokumen 10.....	76
4. 30	Output pemanggilan teks dari docx yang sama.....	80
4. 31	Output hasil similarity winnowing dokumen yang sama.....	81
4. 32	Output hasil similarity Ratcliff/Obershelp dokumen yang sama	81
4. 33	Output fitur Get Document Information dokumen yang sama	82



UNIVERSITAS ISLAM NEGERI
SUMATERA UTARA MEDAN

DAFTAR TABEL

Tabel	Judul Tabel	Halaman
2. 1	Contoh Pengambilan Substring.....	17
2. 2	Simbol – Simbol Use Case Diagram.....	20
2. 3	Simbol – Simbol Activity Diagram	20
2. 4	Simbol – Simbol Class Diagram	20
3. 1	Jadwal Pelaksanaan penelitian	26
4. 1	ascii untuk angka, nomor dan simbol.....	43
4. 2	hasil percobaan pertama dengan dokumen yang berbeda.....	69
4. 3	hasil percobaan kedua dengan dokumen yang berbeda	70
4. 4	hasil percobaan ketiga dengan dokumen yang berbeda	71
4. 5	hasil output fitur Get Document Information	77
4. 6	hasil percobaan pertama dengan dokumen yang sama	82
4. 7	hasil percobaan kedua dengan dokumen yang sama.....	83
4. 8	hasil percobaan ketiga dengan dokumen yang sama.....	84
4. 9	Tabel keseluruhan hasil pengujian dokumen.....	86

DAFTAR LAMPIRAN

Lampiran	Judul Lampiran
1.	Source Code
2.	Kartu Bimbingan Skripsi
3.	Curriculum Vitae



UNIVERSITAS ISLAM NEGERI
SUMATERA UTARA MEDAN