

**INTEGRASI TF-IDF DAN ALGORITMA *COSINE*
SIMILARITY UNTUK DETEKSI TINGKAT
KEMIRIPAN JUDUL PENELITIAN
TEKNIK INFORMATIKA**

Oleh

SITTI MEAUTIAH DEVI AMELIA JAN

T3119082

SKRIPSI

**Untuk memenuhi salah satu syarat ujian
guna memperoleh gelar Sarjana**



**PROGRAM SARJANA
TEKNIK INFORMATIKA
UNIVERSITAS ICHSAN GORONTALO
GORONTALO
2023**

PERSETUJUAN SKRIPSI

**INTEGRASI TF-IDF DAN ALGORITMA *COSINE*
SIMILARITY UNTUK DETEKSI TINGKAT
KEMIRIPAN JUDUL PENELITIAN
TEKNIK INFORMATIKA**

Oleh

SITTI MEAUTIAH DEVI AMELIA JAN

T3119082

SKRIPSI

Untuk memenuhi salah satu syarat ujian
guna memperoleh gelar Sarjana
Program Studi Teknik Informatika,
ini telah disetujui oleh Tim Pembimbing
Gorontalo,2023

Pembimbing I

Pembimbing II



Irvan Abraham Salihi, M.Kom

NIDN : 0928028101



Sudirman S. Panna, M.Kom

NIDN : 0924038205

PENGESAHAN SKRIPSI

INTEGRASI TF-IDF DAN ALGORITMA *COSINE SIMILARITY* UNTUK DETEKSI TINGKAT KEMIRIPAN JUDUL PENELITIAN TEKNIK INFORMATIKA

Oleh

SITTI MEAUTIAH DEVI AMELIA JAN

T3119082

Diperiksa oleh Panitia Ujian Strata Satu (S1)

Universitas Ichsan Gorontalo

1. Ketua Penguji
Asmaul Husna Nasrullah, M.Kom

2. Anggota
Sudirman Melangi, M.Kom

3. Anggota
Yusrianto Malago, S.Kom, M.Kom

4. Anggota
Irvan Abraham Salihi, M.Kom

5.. Anggota
Sudirman S. Panna, M.Kom

Mengetahui

Dekan Fakultas Ilmu Komputer



Irvan Abraham Salihi, M.Kom
NIDN. 0928028101

Ketua Program Studi



Sudirman S Panna, M.Kom
NIDN. 0924038205

PERNYATAAN SKRIPSI

Dengan ini saya menyatakan bahwa:

1. Karya tulis (Skripsi) saya ini adalah asli dan belum pernah diajukan untuk mendapat gelar akademik (Sarjana) baik di Universitas Ichsan Gorontalo maupun di perguruan tinggi lainnya.
2. Karya tulis (Skripsi) saya ini adalah murni gagasan, rumusan, dan penelitian saya sendiri tanpa bantuan pihak lain, kecuali arahan dari Tim Pembimbing.
3. Dalam karya tulis (Skripsi) saya ini tidak terdapat karya atau pendapat yang telah dipublikasikan oranglain, kecuali secara tertulis dicantumkan sebagai acuan/sitasi dalam naskah dan dicantumkan pula dalam daftar pustaka.
4. Pernyataan ini saya buat dengan sesungguhnya dan apabila dikemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka saya bersedia menerima sanksi akademik berupa pencabutan gelar yang telah diperoleh karena karya tulis ini, serta sanksi lainnya sesuai dengan norma-norma yang berlaku di Universitas Ichsan Gorontalo.

Gorontalo, April 2023

Yang membuat Pernyataan



Sitti Meautiah Devi Amelia Jan

ABSTRACT

SITTI MEAUTIAH DEVI AMELIA JAN. T3119082. INTEGRATION OF TF- IDF AND COSINE SIMILARITY ALGORITHMS FOR DETECTION OF SIMILARITY LEVELS OF INFORMATICS ENGINEERING RESEARCH TITLES

Plagiarism is still common, especially in the academic field. Plagiarism is done especially when completing a final project. To avoid plagiarism, it is necessary to detect the similarity of the final project title. Similarity detection is committed to checking the similarity of documents with other documents. One solution to check/detect document similarity quickly and precisely is the application of the Cosine Similarity algorithm and TF-IDF weighting. This study aims to find the performance of the Cosine Similarity algorithm in detecting the level of similarity of final assignments or research. By utilizing text preprocessing consisting of case folding, tokenizing, stop-word removal, and stemming which is then carried out the calculation of TF-IDF weighting and similarity value using cosine similarity to get a similarity percentage value. The results of the experiments in this study are in the form of a similarity level of the final project title being compared. The confusion matrix testing obtains an average accuracy value of 89.7%, precision of 72.4%, and recall of 94.6%.



Keywords: *Similarity, TF-IDF, Cosine Similarity*

ABSTRAK

SITTI MEAUTIAH DEVI AMELIA JAN. T3119082. INTEGRASI TF-IDF DAN ALGORITMA COSINE SIMILARITY UNTUK DETEKSI TINGKAT KEMIRIPAN JUDUL PENELITIAN TEKNIK INFORMATIKA

Penjiplakan/plagiarisme masih sering terjadi, terutama dalam dunia akademik. Penjiplakan banyak dilakukan terutama saat menyelesaikan tugas akhir atau skripsi. Untuk menghindari kegiatan plagiat perlu dilakukan dengan mendeteksi kemiripan atau kesamaan judul TA. Deteksi Similaritas dilakukan untuk memeriksa kesamaan dokumen dengan dokumen lain. Salah satu solusi untuk memeriksa / mendeteksi kemiripan dokumen secara cepat dan tepat adalah dengan penerapan algoritma *Cosine Similarity* dan pembobotan TF-IDF. Penelitian ini bertujuan untuk mengetahui kinerja dari algoritma *Cosine Similarity* dalam mendeteksi tingkat kemiripan tugas akhir atau penelitian. Dengan memanfaatkan preprocessing teks yang terdiri dari case folding, tokenizing, stopword removal, dan stemming yang selanjutnya dilakukan perhitungan pembobotan TF-IDF dan nilai kemiripan menggunakan *cosine similarity* sehingga mendapatkan nilai persentase kemiripan. Hasil dari percobaan pada penelitian ini adalah berupa tingkat kemiripan dari judul tugas akhir yang dibandingkan, dengan pengujian *confusion matrix* diperoleh nilai *accuracy* memiliki rata-rata sebesar 89.7%, *precision* 72.4%, dan *recall* 94.6%.

Kata kunci: *Similarity*, TF-IDF, *Cosine Similarity*

KATA PENGANTAR

Alhamdulillah, Penulis dapat menyelesaikan skripsi ini dengan judul **“INTEGRASI TF-IDF DAN ALGORITMA *COSINE SIMILARITY* UNTUK DETEKSI KEMIRIPAN JUDUL PENELITIAN TEKNIK INFORMATIKA”**, untuk memenuhi salah satu syarat Ujian Akhir guna memperoleh gelar Sarjana Komputer pada Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Ichsan Gorontalo.

Penulis menyadari sepenuhnya bahwa skripsi ini tidak mungkin terwujud tanpa bantuan dan dorongan dari berbagai pihak, baik bantuan moril maupun materil. Untuk itu, dengan segala keikhlasan dan kerendahan hati, penulis mengucapkan banyak terimakasih dan penghargaan yang setinggi-tingginya kepada:

1. Mohammad Ichsan Gaffar, SE.M. Ak, selaku Ketua Yayasan Pengembangan Ilmu Pengetahuan dan Teknologi (YPIPT) Ichsan Gorontalo;
2. Dr. Abdul Gaffar La Tjokke, Msi, selaku Rektor Universitas Ichsan Gorontalo;
3. Irvan Abraham Salihi, M.Kom, selaku Dekan Fakultas Ilmu Komputer Universitas Ichsan Gorontalo sekaligus selaku Pembimbing I yang telah memberikan bimbingan dan arahan kepada penulis untuk menyelesaikan skripsi ini;
4. Sudirman Melangi, M.Kom, selaku Wakil Dekan I Bidang Akademik Fakultas Ilmu Komputer Universitas Ichsan Gorontalo;
5. Irma Surya Kumala Idris, M.Kom, selaku Wakil Dekan II Bidang Administrasi Umum dan Keuangan Fakultas Ilmu Komputer Universitas Ichsan Gorontalo;
6. Sudirman S. Panna, M.Kom, selaku Ketua Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Ichsan Gorontalo sekaligus

pembimbing II, yang telah memberikan bimbingan dan arahan kepada penulis untuk menyelesaikan skripsi ini;

7. Bapak dan Ibu Dosen Universitas Ichsan Gorontalo yang telah mendidik dan mengajarkan berbagai disiplin ilmu kepada penulis;
8. Teristimewa kepada keluarga tercinta, Ayah saya H. Adijbsardjana Hasan Jan, ST.MBA, Ibu saya Fitrya Rauf, SE,AK, Kakak saya Sitti Rachmah Puspita Sari Jan,dan Adik saya Moh. Azkha Raffa Jan atas segala kasih sayang, jerih payah, doa restu, serta telah mendukung penulis mencapai cita-cita untuk menjadi seorang sarjana;
9. Rekan-rekan seperjuangan Fikom Angkatan 2019, Empat Serangkai Hizrawati Yahya, dan Putri Siti Salsabilla Ibrahim, Siti Endang Utami Taha, Teman-teman saya Axel Nicho Jacobus, Ismail A. Muhamat, Samuel Takaalumang, Grace Roslin Pasumah, Melinda Sikape, Ciwi-ciwi Squad, Maya Elisa Rivai, kepada Kucing saya Nyawong, dan Harry Styles beserta anggota grup One Direction yang telah memberikan bantuan, dukungan, dan motivasi kepada penulis dalam menyelesaikan skripsi ini;
10. Kepada semua pihak yang telah membantu dalam penyelesaian skripsi ini tak sempat penulis sebutkan satu-persatu.

Semoga Allah SWT melimpahkan balasan atas jasa-jasa mereka kepada kami. Penulis menyadari sepenuhnya bahwa apa yang telah dicapai ini masih jauh dari kesempurnaan dan masih banyak terdapat kekurangan. Oleh karena itu, penulis sangat mengharapkan adanya kritik dan saran yang konstruktif. Akhirnya penulis berharap semoga hasil yang telah dicapai ini dapat bermanfaat bagi kita semua, Amiin.

Gorontalo, April 2023

Sitti Meautiah Devi Amelia Jan

DAFTAR ISI

HALAMAN JUDUL	i
PERSETUJUAN SKRIPSI.....	ii
PENGESAHAN SKRIPSI.....	iii
PERNYATAAN SKRIPSI	iv
<i>ABSTRACT</i>	v
ABSTRAK.....	vi
KATA PENGANTAR.....	vii
DAFTAR ISI	ix
DAFTAR GAMBAR.....	xii
DAFTAR TABEL	xiii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Identifikasi Masalah	3
1.3 Rumusan Masalah	4
1.4 Tujuan Penelitian.....	4
1.6 Manfaat Penelitian.....	4
BAB II LANDASAN TEORI	5
2.1 Tinjauan Studi	5
2.2 Tinjauan Pustaka	6
2.2.1 <i>Similarity</i> (Kemiripan).....	6
2.2.2 <i>Plagiarisme</i>	7
2.2.3 <i>Data Mining</i>	8
2.2.4 <i>Text Mining</i>	8
2.2.5 <i>Preprocessing</i>	9

2.2.6	<i>Term Frequency-Inverse Document Frequency (TF-IDF)</i>	10
2.2.7	<i>Cosine Similarity</i>	11
2.2.8	<i>Confusion Matrix</i>	12
2.3	Kerangka Pikir.....	14
BAB III	METODE PENELITIAN	15
3.1	Jenis, Metode, Subjek,Objek, Waktu dan Lokasi Penelitian.....	15
3.2	Alur Metodologi Penelitian	15
3.3	Pengumpulan Data	16
3.4	Pemodelan <i>Cosine Similarity</i>	16
3.4.1	<i>Dataset</i>	16
3.4.2	<i>Data train</i>	18
3.4.3	<i>Data testing</i>	18
3.4.4	<i>Preprocessing</i>	18
3.4.5	Pembobotan <i>TF-IDF</i>	19
3.4.6	<i>Cosine Similarity</i>	19
3.4.7	Evaluasi.....	19
BAB IV	HASIL PENELITIAN	20
4.1	Pengumpulan Data	20
4.2	Model yang diusulkan	22
4.2.1	Tahap <i>Preprocessing</i>	24
4.2.2	Pembobotan <i>TF-IDF</i>	30
4.2.3	<i>Cosine Similarity</i>	36
4.2.4	Menghitung kemiripan dokumen.....	42
4.3	Implementasi <i>Cosine</i> pada <i>Python</i>	45
4.4	Evaluasi Model.....	46

BAB V	PEMBAHASAN PENELITIAN	51
5.1	Pembahasan Model.....	51
BAB VI	PENUTUP PENELITIAN	53
6.1	Kesimpulan.....	53
6.2	Saran.....	53
DAFTAR PUSTAKA		
LAMPIRAN		
Hasil Turnitin		
Listing Program		
Dataset		
Daftar Riwayat Hidup		
Surat Penelitian		
Surat Keterangan Penelitian		
Surat Keterangan Bebas Plagiasi		
Surat Keterangan Bebas Pustaka		

DAFTAR GAMBAR

Gambar 2.1 : <i>Preprocessing Text</i>	10
Gambar 2.2 : Kerangka Pikir	14
Gambar 3.1: Alur Metodologi Penelitian	15
Gambar 3.2 : Pemodelan <i>Cosine Similarity</i>	16
Gambar 4.2 : <i>Preprocessing</i>	24
Gambar 4.3 : <i>Case Folding</i>	28
Gambar 4.4 : <i>Tokenizing</i>	29
Gambar 4.5 : <i>Stopword Removal/Filtering</i>	29
Gambar 4.6 : <i>Stemming</i>	30
Gambar 4.7 : Pembobotan Dokumen.....	30
Gambar 4.8 : Hasil TF-IDF pada Python.....	36
Gambar 4.9 : Pengukuran Kemiripan dengan <i>Cosine Similarity</i>	37
Gambar 4.10 : Kemiripan Dokumen dengan <i>Cosine Similarity</i>	45
Gambar 5.1 : Grafik Hasil Pengujian	52

DAFTAR TABEL

Table 2.1 : <i>State of Art</i>	5
Table 2.2 : Perhitungan Uji Kemiripan.....	12
Table 3.1: Sampel <i>Dataset</i>	17
Table 4.1: Hasil Pengumpulan Data	20
Table 4.2: Hasil <i>Preprocessing</i>	25
Table 4.3 : <i>Term</i> yang dihasilkan	27
Table 4.4 : <i>Term Frequency</i> (TF)	31
Table 4.5: <i>Document Frequency</i> (DF).....	32
Table 4.6 : <i>Hasil Perhitungan Inverse Document Frequency</i> (IDF).....	33
Table 4.7: Hasil Pembobotan Kata (TF-IDF)	35
Table 4.8: Hasil Perkalian Q dan D	38
Table 4.9: Hasil Perhitungan Akar Tiap Dokumen	40
Table 4.10: Pengurutan Tingkat Kemiripan	45
Table 4.11: Kata Kunci.....	46
Table 4.12: Hasil Pengujian tiap <i>query</i>	47
Table 4.13: Hasil Evaluasi.....	48
Table 4.14: Visualisasi Confussion Matrix	48

BAB I

PENDAHULUAN

1.1 Latar Belakang

Dalam suatu instansi akademik, penelitian tugas akhir bagi mahasiswa merupakan sks wajib pada setiap program studi diseluruh perguruan tinggi. Tugas akhir ialah suatu hal yang sudah tidak asing lagi dikalangan para mahasiswa semester akhir, karena banyaknya alumni, proses memperoleh judul menjadi lebih sulit. Untuk itu penelitian ini masih berhubungan dengan judul skripsi[1]. Penjiplakan atau plagiarisme sering terjadi terutama dalam dunia akademik, terlebih pada perguruan tinggi atau universitas. Plagiarisme dilakukan untuk mempermudah tugas dengan teknik copy paste tanpa harus memahami atau mendalami materi yang dimaksud, akibatnya banyak plagiarisme yang terjadi terutama saat menyelesaikan tugas akhir atau skripsi[2]. Unsur plagiarisme dalam sebuah penelitian berasal dari judul maupun isi. Bila tidak dideteksi, TA yang bertambah dari tahun ke tahun dapat menyebabkan penelitian yang sudah ada akhirnya diteliti kembali. Judul TA merupakan bagian penting dalam sebuah penulisan TA, karena berfungsi untuk menggambarkan pokok bahasan dan argumentasi dalam skripsi atau TA[3]. Salah satu cara untuk mengurangi plagiarisme atau penjiplakan adalah dengan pencegahan dan deteksi dini. Pencegahan berarti menjaga dan mencegah agar plagiarisme tidak dilakukan. Hal ini harus dilakukan terutama dalam sistem pendidikan[2]. Untuk menghindari kegiatan plagiat perlu diketahui apakah judul TA yang dibuat sudah pernah dilakukan sebelumnya dengan mendeteksi kemiripan atau kesamaan judul TA. Deteksi similaritas dilakukan untuk memeriksa kesamaan dokumen dengan dokumen lain[3].

Penelitian tentang kemiripan judul telah banyak dilakukan, agar dapat mengetahui tingkat kemiripan sebuah judul diperlukan metode khusus yang dapat menangani tingkat kemiripan tersebut. Beberapa metode yang sudah pernah digunakan untuk pendeteksian kemiripan ialah *Smith Waterman*, *Single Hierarchical*, *TF-IDF*, dan *Cosine Similarity*.

Uswatun Hasanah, dkk [1] menerapkan algoritma *Smith Waterman* untuk mengetahui persentase kemiripan judul skripsi, dan mendapatkan hasil berupa skor kemiripan dan persentase tingkat kemiripan judul skripsi. Hanya saja metode ini tidak cocok digunakan untuk perbandingan dokumen dengan data yang lebih besar karena akan memerlukan waktu yang agak lama untuk mendapatkan hasil pendeteksian dari dokumen yang dibandingkan. Selanjutnya Rizal Tjut Adek dan Miftahul Jannah [4] melakukan penelitian serupa dengan menggunakan *Single Linkage Hierarchical* yang menghasilkan nilai kecocokan pada tiap kategori sebesar 82% hingga 100% untuk tingkat kemiripan tinggi dan 11,11% untuk nilai kecocokan rendah. Akan tetapi, masih terjadi kekeliruan hasil pencocokan antara data uji dan data *training*. Pembentukan *cluster*-nya juga masih kurang optimal karena algoritma ini hanya mengenali jarak tanpa memperhatikan makna yang terkandung dari tiap atribut. Arif dan Ipam Fuddiana [3] menggunakan *TF-IDF* pada sistem pencarian similaritas. Dalam penerapannya, ditemukan bahwa *TF-IDF* masih belum akurat dalam mendeteksi kemiripan karena belum bisa mendeteksi kata kunci yang mengandung karakter spasi, selain itu metode ini juga tidak mampu mendeteksi dua kata yang sebenarnya masih satu kesatuan. Muhammad Azmi [2] mencoba menggunakan *Cosine Similarity* yang dipadukan dengan *TF-IDF* untuk tingkat plagiasi, memperoleh hasil tingkat kemiripan yang diperoleh dari perbandingan antara nilai kemiripan pada tahapan yang menggunakan *stemming* dan yang tidak. Hasil dengan menggunakan *stemming* menghasilkan nilai kemiripan lebih besar dibandingkan dengan tanpa *stemming*.

Dari hasil penelitian yang telah banyak dilakukan, algoritma yang akan digunakan pada penelitian ini adalah *Cosine Similarity* dengan mengoptimalkan metode pembobotan dari *TF-IDF*. Pembobotan *TF-IDF* digunakan untuk mencari representasi nilai tiap dokumen dalam koleksi. Sedangkan *Cosine Similarity* diharapkan dapat menangani permasalahan pada penelitian ini. Metode ini, tergolong baik digunakan pada analisis teks untuk pengukuran kesamaan dokumen. Sehingga dapat membantu mengetahui performa model yang dirancang untuk menjamin kualitas model usulan sebelum diimplementasikan. Metode *Cosine similarity* ialah metode dengan pendeteksian secara cepat dan tepat [5]. *Cosine*

similarity merupakan algoritma yang memiliki pengaruh besar terhadap nilai *similarity*[6]. Algoritma ini memanfaatkan kata kunci pada sebuah dokumen sebagai bentuk pengukuran tingkat kesamaan yang dinyatakan dalam bentuk *vector*[7]. Hal itu dikarenakan metode cosine similarity yang merupakan alat ukur untuk menghitung nilai cosines sudut antar dua vector. Sehingga diketahui berapa banyak jumlah kemiripan teks yang dibandingkan. Kelebihan dari algoritma *cosine similarity* adalah mampu mendeteksi tingkat kemiripan, dengan akurasi yang tinggi. Adapun kelemahan algoritma ini yaitu tidak dapat mendeteksi makna (sinonim) dari setiap kata yang dibandingkan[8]. Pada penelitian yang dilakukan oleh [6] menggunakan algoritma *cosine similarity* untuk klasifikasi kemiripan topik judul, diperoleh sebanyak 26 dari 30 data, menghasilkan persentase akurasi sebesar 86,66%. Adapun penelitian serupa yang dilakukan oleh [9] memperoleh nilai *precision* sebesar 88% dan nilai *recall* sebanyak 76.1%. dengan kelebihan daripada metode *cosine similarity*, maka digunakanlah metode ini pada penelitian deteksi kemiripan tugas akhir.

Berdasarkan berbagai pemaparan diatas, maka peneliti tertarik untuk melakukan penelitian dengan judul **“Integrasi TF-IDF Dan Algoritma Cosine Similarity Untuk Deteksi Tingkat Kemiripan Judul Penelitian Teknik Informatika.”** Penelitian ini berfokus pada langkah penerapan perhitungan algoritma *Cosine Similarity* yang diharapkan dapat memberikan hasil yang terbaik dalam mendeteksi kemiripan judul secara efektif. Sehingga kedepannya dapat mencegah terlebih mengurangi tingkat kemiripan tugas akhir.

1.2 Identifikasi Masalah

1. Tugas Akhir mahasiswa yang terus mengalami penambahan dari tahun ke tahun dapat mengakibatkan adanya judul yang mirip bahkan kemungkinan sama. Dengan demikian diperlukan pencegahan dini untuk dapat mengatasi hal tersebut.
2. Dalam menyelesaikan masalah dibutuhkan pendekatan komputasi atau metode dengan hasil akurasi tinggi dan cepat. Guna membantu memudahkan proses pengukuran kemiripan.

1.3 Rumusan Masalah

Berdasarkan identifikasi masalah, maka rumusan masalahnya yaitu “Bagaimana kinerja dari algoritma *cosine similarity* dalam mendeteksi tingkat kemiripan judul tugas akhir mahasiswa teknik informatika ?”

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah, tujuan penelitian ini adalah “Untuk mengetahui kinerja algoritma *cosine similarity* dalam mendeteksi tingkat kemiripan judul tugas akhir mahasiswa teknik informatika.”

1.5 Manfaat Penelitian

1. Manfaat Teoritis : Memberikan masukan bagi perkembangan ilmu pengetahuan dan teknologi, khususnya pada bidang ilmu komputer, yaitu berupa uji coba metode cosine similarity.
2. Manfaat Praktis : Sumbangan pemikiran, karya, bahan pertimbangan atau solusi yang dapat dijadikan salah satu dasar acuan yang tepat dan akurat dalam melakukan pendeteksian kemiripan dokumen dengan penerapan metode cosine similarity, agar dapat memberikan kontribusi untuk peneliti selanjutnya dalam pengembangan sistem, maupun metode.

BAB II

LANDASAN TEORI

2.1 Tinjauan Studi

Adapun beberapa penelitian terdahulu yang terkait tentang deteksi kemiripan dan penggunaan metode *cosine similarity*, seperti dibawah ini :

Table 2.1 : State of Art

NO	PENELITI	JUDUL	TAHUN	METODE	HASIL
1	Uswatun Hasanah, Ramdan Sastra, Fitriyani Umar	Deteksi Kemiripan Judul Menggunakan Algoritma <i>Smith Waterman</i>	2020	<i>Smith Waterman</i>	Dari hasil uji coba nilai persentase tertinggi adalah pasangan D1 dan D1 dengan perolehan skor 68 atau 100% mirip, skor ini membuktikan bahwa karakter D1 dan D1 bernilai sama lebih banyak (match) dibandingkan dengan yang bernilai tidak sama (mismatch). Sehingga dapat disimpulkan bahwa dari ketiga sampel yang paling mirip adalah sampel 1.

NO	PENELITI	JUDUL	TAHUN	METODE	HASIL
2	Arif Amrulloh, Ipam Fuaddina Adam	Sistem Pencarian Similaritas Judul Tugas Akhir Menggunakan Metode TF- IDF	2021	<i>Term Frequency- Invers Document Frequency</i>	Dari hasil ujicoba menggunakan kata kunci yang sudah ditentukan diperoleh hasil pembobotan dan ditemukan similaritas judul tugas akhir sebanyak 99 data dari total 384 judul tugas akhir.
3	Muhammad Azmi	Analisis Tingkat Plagiasi Dokumen Skripsi Dengan Metode Cosine Similarity dan Pembobotan TF-IDF	2022	<i>Cosine Similarity dan Pembobotan Term Frequency- Invers Document Frequency</i>	Hasil pengujian kemiripan pada tahapan yang menggunakan (1) <i>stemming</i> diperoleh kemiripan sedang 5 dan kemiripan ringan 2 serta tanpa <i>stemming</i> didapatkan hasil jumlah ringan 3 dan jumlah sedang 4. Hasil <i>stemming</i> menunjukkan tingkat kemiripan yang lebih tinggi daripada tanpa <i>stemming</i> .

2.2 Tinjauan Pustaka

2.2.1 *Similarity* (Kemiripan)

Similarity merupakan suatu hasil analisis terhadap suatu karya ilmiah guna menentukan apakah karya ilmiah tersebut mengandung unsur

plagiarisme atau tidak[10]. Terdapat 3 macam teknik penentuan nilai kemiripan dari suatu dokumen antara lain: [8]

1. *Distance-based similarity measure*, adalah pengukuran tingkat kesamaan berdasarkan jarak.
2. *Feature-based similarity measure*, menghitung tingkat kesamaan dalam bentuk feature yang diperbandingkan. *Feature-based Similarity* sering digunakan untuk melakukan pengklasifikasian atau pattern matching pada gambar maupun teks.
3. *Probabilistic-based measure*, penghitungan tingkat kemiripan dengan membandingkan dua set objek dalam bentuk probability. Seperti *Kullback Leibler Distance* dan *Posterior Probability*.

2.2.2 Plagiarisme

Plagiarisme adalah penjiplakan atau pengambilan karangan dari orang lain dan dijadikan seolah-olah menjadi karangan milik sendiri[11]. Plagiat tergolong sebagai tindak pidana karena mencuri karangan orang lain. Dalam dunia Pendidikan, plagiat akan menerima hukuman berat[12]. Sebagaimana dijelaskan dalam peraturan kementerian No. 17 Tahun 2010 tentang plagiarisme. Berdasarkan peraturan tersebut, terdapat beberapa ciri-ciri dari sebuah dokumen yang mengandung plagiarisme, antara lain:

- a. Mengutip atau mengulang dan mencuri gagasan orang lain tanpa merujuk mencantumkan, atau menyebutkan sumber atau pemilik gagasan.
- b. Menggunakan sumber gagasan atau pernyataan orang lain tanpa memberi catatan kaki orang atau daftar pustaka.
- c. Melakukan kutipan tak langsung dari sumber gagasan secara tertulis tanpa melakukan *parafrase*.
- d. Dua tulisan yang berjudul dan berisi sama, maka yang keluar belakangan merupakan hasil plagiat.

- e. Menyalin seluruh hasil karya orang lain dan diakui sebagai tulisan sendiri dan dipublikasikan tanpa menyatakan sumber yang memadai.

Dalam penentuan seberapa besar terjadinya praktek plagiat. Plagiarisme dikelompokkan menjadi 3 bagian yaitu:[11]

1. <30% dikategorikan sebagai plagiarisme ringan.
2. 30-70% dikategorikan sebagai plagiarisme sedang.
3. >70% dikategorikan sebagai plagiarisme berat.

2.2.3 *Data Mining*

Dalam buku *Data Mining*, mendefinisikan *data mining* sebagai proses yang menggunakan satu atau lebih teknik pembelajaran komputer untuk menganalisis dan mengekstrak data (informasi) secara otomatis. Data mining merupakan proses identifikasi *trend*/pola yang diinginkan agar dikenali oleh perangkat tertentu untuk dapat memberikan analisa data yang berguna dan berwawasan yang kemudian dapat dipelajari dengan lebih teliti. [13].

2.2.4 *Text Mining*

Text Mining adalah pengukuran data teks yang bersumber dari dokumen untuk dilakukan analisa lebih lanjut seperti mencari kesamaan kata, dan lainnya. *Text mining* terkadang digunakan sebagai penerapan konsep *data mining* dengan menganalisis data tekstual, untuk mendapatkan informasi[12]. Singkatnya, *text mining* merupakan analisis untuk menggali informasi dari teks dengan mengadopsi proses-proses dan juga menggunakan teknik *data mining*[6]. *Text mining* mempunyai tugas khusus, yaitu pengkategorisasian teks (*text categorization*) dan pengelompokan teks (*clustering text*)[14].

2.2.5 Preprocessing

Preprocessing bertujuan untuk mendapatkan *dataset* yang dapat diolah dengan cepat dan tepat[14]. Selain itu, fungsi dari *processing* atau *text processing* adalah mengubah data tidak terstruktur menjadi data terstruktur. Tahapan awal dari *text mining* adalah mempersiapkan teks menjadi data terstruktur untuk diproses pada tahap berikutnya[15]. Berikut tahapan dari *preprocessing*:

1. *Case Folding*

Proses ini merupakan proses pertama yang dilakukan pada rangkaian tahapan preprocessing. Pada proses ini dilakukan perubahan pada seluruh huruf kapital menjadi huruf kecil, A-Z menjadi a-z.

2. *Tokenizing*

Proses ini adalah proses pemisahan kalimat dalam dokumen menjadi penggalan kata yang kemudian dijadikan sebagai *string* input (*token*), untuk dianalisis lebih lanjut pada proses selanjutnya.

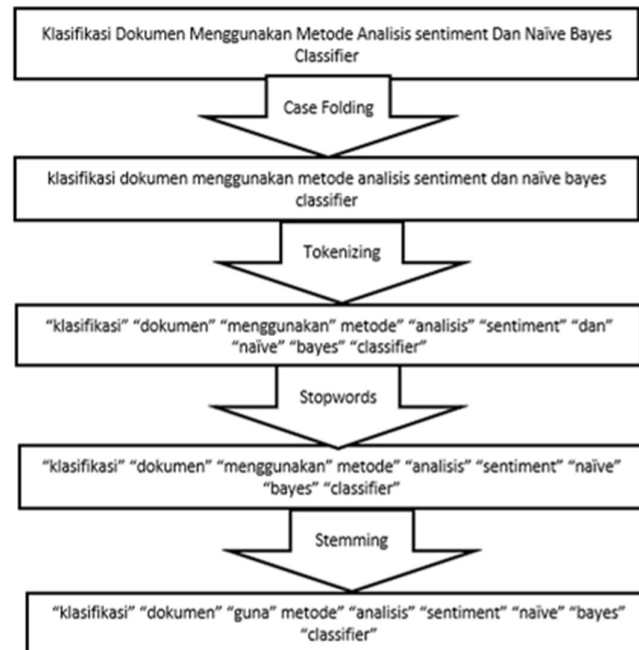
3. *Stopword/Filtering Stopwords Removal*

Proses ini merupakan proses penghilangan kata tidak penting atau kata *stopword* dari hasil *tokenizing*, sehingga yang tersisa hanya kata yang dianggap penting atau keyword.

4. *Stemming*

Proses ini adalah tahap pencarian *root* kata pada setiap kata hasil *filtering*. Pada tahap ini dilakukan proses perubahan kata berimbuhan menjadi kata dasar.

Secara umum, contoh preprocessing teks dapat direpresentasikan sebagai berikut:



Gambar 2.1 : Preprocessing Text

2.2.6 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF merupakan metode penghitungan bobot setiap kata dan menjadi suatu algoritma paling umum digunakan pada natural language processing dalam information retrieval[12]. Metode ini mencakup dua konsep penghitungan bobot, yaitu kemunculan frekuensi kata tertentu dalam sebuah dokumen dan frekuensi inverse dokumen yang mengandung kata tersebut. Jumlah kemunculan kata di dalam dokumen[3]. Dalam *TF-IDF* perhitungan mengenai nilai statistik merupakan penggambaran evaluasi seberapa pentingnya kata demi kata dalam sebuah dokumen. Perhitungan ini berfungsi untuk mendapatkan nilai bobot relevansi *term* dari sebuah dokumen terhadap keseluruhan data[6].

Agar memperoleh nilai *TF-IDF* dibutuhkan nilai *Idft* (*Inverse Document Frequency*) adalah pembobotan yang bertujuan untuk mengetahui seberapa

penting pengaruh *term* dari suatu dokumen terhadap dokumen lain[15].

Nilai *idf* dirumuskan dalam persamaan berikut:

$$idf_t = \log \left(\frac{n}{df_t} \right) \quad (1)$$

Hasil perhitungan *idf* dimasukkan kedalam persamaan *TF-IDF* berikut:

$$W_{dt} = tf_{d,t} \times idf_{d,t} \quad (2)$$

Dengan:

n = Total dokumen yang ada

df = Jumlah dokumen yang mengandung *term* i.

idf = *Inversed Document Frequency*

tf = Banyaknya kemunculan *term* i pada sebuah data

W_{dt} = Bobot term dalam sebuah data

2.2.7 Cosine Similarity

Cosine similarity diaplikasikan untuk mengukur kedekatan antara dua vektor[6]. Model ruang vektor untuk merepresentasikan nilai numerik dokumen sehingga kemudian dapat dihitung kedekatan antar dokumen. Kemiripan antar dokumen dihitung menggunakan suatu fungsi ukuran kemiripan (*similarity measure*)[14]. *Cosine similarity* dapat diimplementasikan untuk menghitung nilai kemiripan antar kalimat[6]. Untuk melakukan perhitungan kesamaan dokumen, persamaannya adalah:

$$Cos \propto \frac{X \cdot Y}{|X||Y|} = \frac{\sum_{i=1}^n X_i \times Y_i}{\sqrt{\sum_{i=1}^n (X_i)^2} \times \sqrt{\sum_{i=1}^n (Y_i)^2}} \quad (3)$$

Dengan:

X = vektor X

Y = vektor Y

$|X|$ = panjang vektor X

$|Y|$ = panjang vektor Y

Contoh implementasi rumus *Cosine Similarity* terhadap kemiripan dua buah kalimat:

A = The best Italian restaurant enjoy the best pasta

B = The best the best American restaurant

Pengaplikasiannya sebagai berikut:

Table 2.2 : Perhitungan Uji Kemiripan

No	Daftar Kata	Jumlah Kemunculan Kata	
		A	B
1	The	2	2
2	Best	2	2
3	Italian	1	0
4	American	0	1
5	Restaurant	1	1
6	Enjoy	1	0
7	Pasta	1	0

Vektor :

A : [2, 2, 1, 0, 1, 1, 1]

B : [2, 2, 0, 1, 1, 0, 0]

$$\begin{aligned}
 \text{Cos } \alpha &= \frac{X \cdot Y}{|X||Y|} = \frac{\sum_{i=1}^n X_i \times Y_i}{\sqrt{\sum_{i=1}^n (X_i)^2} \times \sqrt{\sum_{i=1}^n (Y_i)^2}} \\
 &= \frac{(2 \times 2) + (2 \times 2) + (1 \times 0) + (0 \times 1) + (1 \times 1) + (1 \times 0) + (1 \times 0)}{\sqrt{2^2 + 2^2 + 1^2 + 0^2 + 1^2 + 1^2 + 1^2} \times \sqrt{2^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 0^2}} \\
 &= \frac{(4) + (4) + (0) + (0) + (1) + (0) + (0)}{\sqrt{4 + 4 + 1 + 0 + 1 + 1 + 1} \times \sqrt{4 + 4 + 0 + 1 + 1 + 0 + 0}} \\
 &= \frac{9}{\sqrt{12} \times \sqrt{10}} = 0,821584
 \end{aligned}$$

2.2.8 Confusion Matrix

Pada implementasi algoritma *cosine similarity*, tingkat keberhasilan diukur dari seberapa besar kinerja algoritma tersebut bisa memecahkan masalah yang ada[16]. Metode *cosine similarity* dievaluasi dengan mencari nilai akurasi. Untuk itu, dalam melakukan perhitungan untuk mencari nilai akurasi data dilakukan pengujian menggunakan *confusion matrix*. Untuk menggunakan metode ini perlu dilakukan beberapa perhitungan dalam mencari nilai akurasi, presisi, dan *recall*[10].

Dalam *confusion matrix* beberapa persamaan yang digunakan untuk mengukur performa daripada metode antara lain:

- a. *Accuracy* menggambarkan seberapa akurat model dapat mengklasifikasikan dengan benar.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

- a. *Precision* menggambarkan tingkat keakuratan antara data yang diminta dengan hasil prediksi yang diberikan oleh model.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

- b. *Recall* menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi.

$$Recall = \frac{TP}{FN+TP} \quad (6)$$

Keterangan :

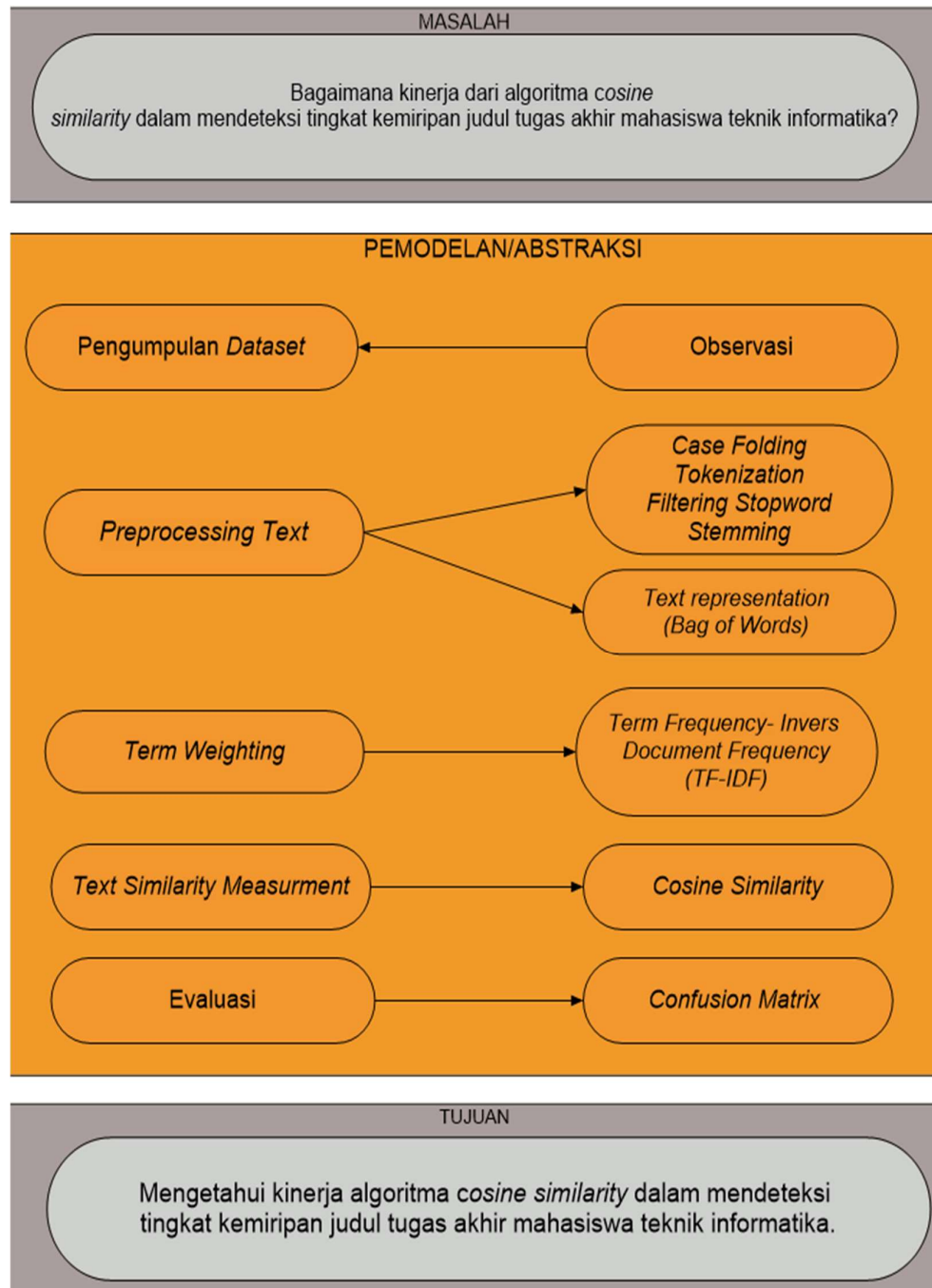
TP (*True Positif*) , adalah jumlah data benar dari data relevant.

TN (*True Negatif*), adalah jumlah data salah dari data relevant.

FP (*False Positif*), adalah jumlah data salah dari data tidak relevant.

FN (*False Negatif*), adalah jumlah data salah dari data tidak relevant.

2.3 Kerangka Pikir



Gambar 2.2 : Kerangka Pikir

BAB III

METODE PENELITIAN

3.1 Jenis, Metode, Subjek, Objek, Waktu dan Lokasi Penelitian

Dilihat dari tingkat penerapannya, penelitian ini merupakan penelitian terapan. Berdasarkan jenis informasi yang diolah, penelitian ini merupakan penelitian kuantitatif. Dari segi perlakuan terhadap data, maka ini adalah penelitian konfirmatori.

Penelitian ini bersifat eksperimen. Oleh karena itu penelitian ini merupakan penelitian eksperimental.

Subjek penelitian ini ialah deteksi kemiripan pada judul tugas akhir mahasiswa. Penelitian ini dimulai dari Oktober 2022 sampai dengan Februari 2023 yang berlokasi di kampus Universitas Ichsan Gorontalo.

3.2 Alur Metodologi Penelitian

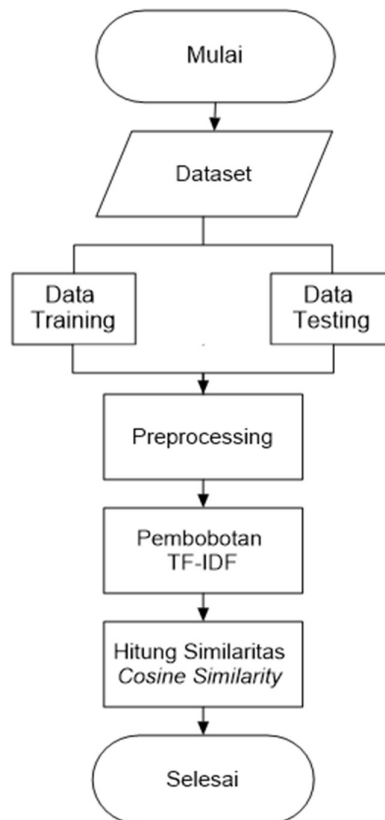


Gambar 3.1: Alur Metodologi Penelitian

3.3 Pengumpulan Data

Data Primer penelitian ini adalah data judul tugas akhir mahasiswa teknik informatika yang didapatkan langsung dari sumbernya yaitu bertempat di Fakultas Ilmu Komputer Universitas Ichsan Gorontalo, yang dikumpulkan dengan teknik observasi dan wawancara. Sedangkan data sekunder yaitu data judul-judul yang ada pada situs google scholar dikumpulkan dengan cara mengambil data-data yang diperoleh dari dokumen, literatur, atau buku-buku yang berhubungan dengan penelitian sebagai pendukung penelitian.

3.4 Pemodelan *Cosine Similarity*



Gambar 3.2 : Pemodelan *Cosine Similarity*

3.4.1 Dataset

Dataset yang digunakan dalam penelitian ini adalah data judul skripsi oleh mahasiswa teknik informatika yang diperoleh dari kantor program

studi (prodi) dalam bentuk file excel(.xls). data judul tugas akhir yang digunakan adalah judul tugas akhir 3 tahun terakhir yaitu tahun 2019/2020, 2020/2021, 2021/2022, dan beberapa data judul penelitian yang diperoleh dari situs google scholar dengan tahun yang sama.

Table 3.1: Sampel *Dataset*

No	Nim	Nama_Mhs	Judul
1	T3115144	Amin Dwi Satrio	Pengenalan Kematangan Buah Pisang Menggunakan Metode Gray Level Co-Occurance Matrix dan K-Nearest Neighbor.
2	T3116036	Ournia Shandy	Penerapan Metode GLCM (Gray Level Co-Occurance Matrix) dan KNN (K-Nearest Neighbor) Untuk Mendeteksi Tingkat Kematangan Buah Belimbing Bintang.
3	T3116101	Mohammad Rezky Bagus cahputra Pakaya	Pengenalan Tingkat Kematangan Buah Mangga Menggunakan Metode <i>GLCM</i> dan <i>KNN</i> .
4	T3116343	Fadlun Gonta	Deteksi Tingkat Kematangan Buah Pisang Gapi Menggunakan Metode K-Nearest Neighbor (KNN) Dan Gray Level Co-Occurance Matrix (GLCM).
5	T3116021	Dana Pratama Jambo	Pengenalan Ekspresi Wajah Secara Real Time Menggunakan Klasifikasi K-NN (K-Nearest Neighbor)
6	T3116307	Ikbal Iskandar	Prototype Kandang Ayam cerdas Berbasis Mikrokontroller Arduino Uno
7	T3116123	Rifai Gobel	Prototype Deteksi Suhu Tubuh Menggunakan Sensor GY-906MLXESF Berbasis Internet Of Things
8	T3116067	Mohammad Rizal Kasim	Pengenalan Buah Nanas Menggunakan Metode Gray Level Co-Occurance Matrix (GLCM) dan K-Nearest Neighbor (KNN).
.....			

352		T Susim	Pengolahan Citra untuk Pengenalan Wajah (Face Recognition) Menggunakan OpenCV
-----	--	---------	---

Sumber: Data Judul Tugas Akhir Fakultas Ilmu Komputer dan google scholar.

Terdapat total “ 352 “ data judul penelitian mahasiswa. Judul skripsi merupakan data yang akan digunakan sebagai data *training* maupun data *testing* dalam mengukur tingkat kemiripan judul.

3.4.2 Data train

Data training adalah data yang nantinya akan diolah untuk dilatih menggunakan algoritma *cosine similarity*. Data training yang digunakan berupa dokumen-dokumen judul tugas akhir mahasiswa teknik informatika Universitas Ichsan Gorontalo yang diperoleh, dengan kalimat yang berbeda-beda.

3.4.3 Data testing

Data testing yang digunakan merupakan dokumen judul penelitian teknik informatika yang bersumber dari situs google scholar, yang kemudian dokumen tersebut nantinya akan digunakan sebagai *query* atau dokumen pembandingan terhadap dokumen judul tugas akhir untuk mengetahui tingkat keberhasilan algoritma dalam mendeteksi kemiripan. Jadi, data testing nantinya akan lebih sedikit dibanding data training.

3.4.4 Preprocessing

Setelah melakukan pengumpulan dataset, tahap selanjutnya yang dilakukan sebelum data diolah, terlebih dahulu dilakukan preprocessing data. Hal ini dilakukan agar data asli berubah menjadi data siap olah untuk tahap selanjutnya. Adapun hal-hal yang dilakukan pada tahap ini yaitu *case folding*, *tokenizing*, *stopword*, dan *stemming*.

3.4.5 Pembobotan *TF-IDF*

Pada tahap ini akan dilakukan perubahan proses perubahan data. Yang tadinya bertipe data string berubah menjadi numerik dengan cara melakukan perhitungan pembobotan untuk mengetahui seberapa sering *term* kata yang muncul dengan menggunakan persamaan (1) dan (2).

3.4.6 *Cosine Similarity*

Hasil daripada pembobotan TF-IDF dimasukkan kedalam persamaan (3) untuk dihitung tingkat kemiripan antara dokumen tes dengan dokumen pembandingan menggunakan metode *cosine similarity*. Pada tahap ini akan dilakukan pendeteksian kemiripan dengan menggunakan metode *cosine similarity*. Metode *cosine similarity* adalah metode yang digunakan untuk menghitung *similarity* dua buah objek yang didasarkan pada *vector space similarity*[16]. Total data yang akan diolah nantinya menggunakan *tools python*, yang terdiri dari *natural language processing* dan *python sastrawi*. Selain itu juga hasil persentase dari setiap perbandingan dokumen akan digunakan untuk diklasifikasikan kedalam 2 kategori rendah dan tinggi.

3.4.7 Evaluasi

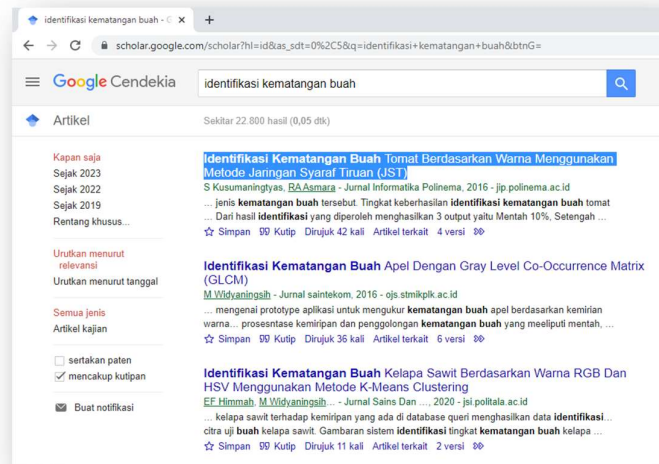
Evaluasi dilakukan untuk mengukur kinerja dari metode yang digunakan. Dalam mengukur seberapa baik metode yang digunakan, akan dilakukan dengan menggunakan *confusion matrix* pada seluruh data testing untuk dinilai tingkat akurasinya.

BAB IV

HASIL PENELITIAN

4.1 Pengumpulan Data

Dalam mengumpulkan data, data diambil dari kumpulan judul tugas akhir mahasiswa teknik informatika Universitas Ichsan Gorontalo yang sudah Selesai dan beberapa jurnal penelitian teknik informatika dari situs <https://scholar.google.com/>, yang akan dijadikan sebagai bahan uji coba perbandingan judul TA mahasiswa teknik informatika Universitas Ichsan Gorontalo. Metode pengumpulan data jurnal tersebut, dapat dilihat pada gambar dibawah ini:



Data judul jurnal diatas di *copy* terlebih dahulu kemudian disalin kedalam *file excel* untuk dimasukkan bersama *dataset* judul yang didapatkan agar selanjutnya dilakukan *preprocessing*.

Berdasarkan hasil pengumpulan data, diperoleh data sebagai berikut yang dapat dilihat pada tabel 4.1 dibawah ini:

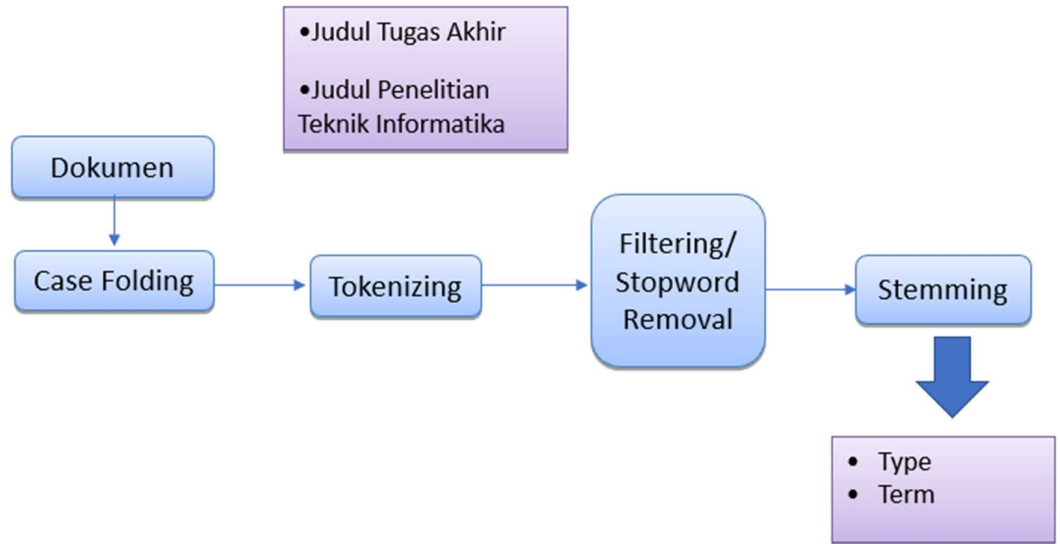
Table 4.1: Hasil Pengumpulan Data

No	Nim	Nama Mhs	Judul
1	T3115144	Amin Dwi Satrio	Pengenalan Kematangan Buah Pisang Menggunakan Metode Gray Level Co-Occurance Matrix dan K-Nearest Neighbor.

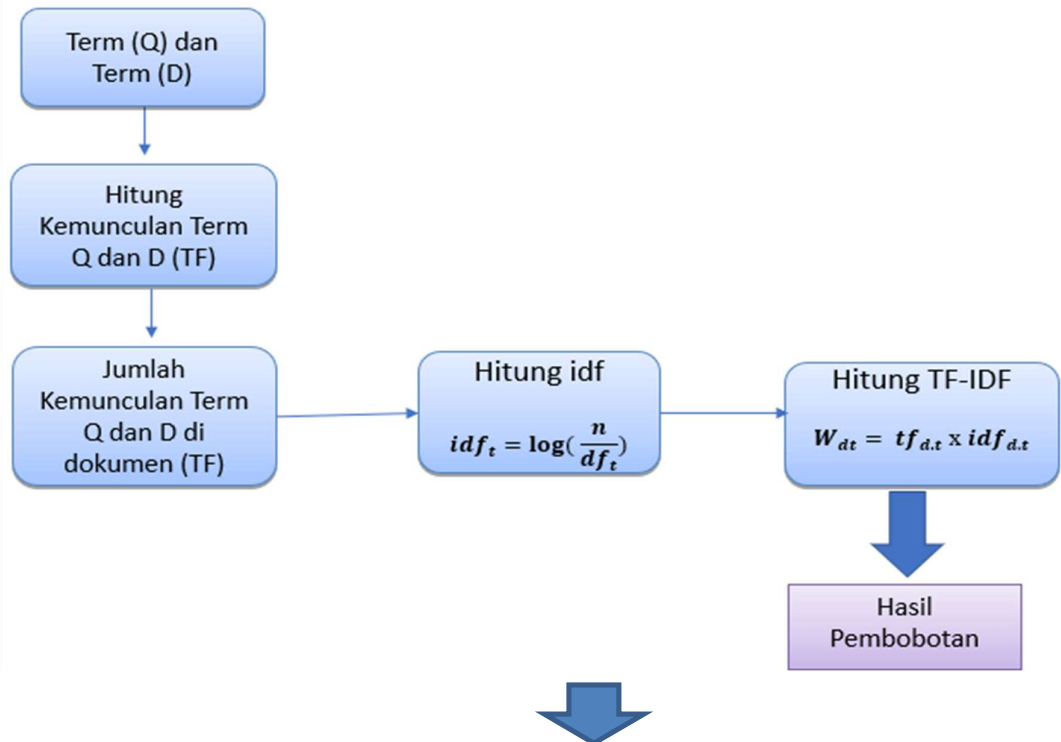
No	Nim	Nama_Mhs	Judul
2	T3116036	Ournia Shandy	Penerapan Metode GLCM (Gray Level Co-Occurance Matrix) dan KNN (K-Nearest Neighbor) Untuk Mendeteksi Tingkat Kematangan Buah Belimbing Bintang.
3	T3116101	Mohammad Rezky Bagus cahputra Pakaya	Pengenalan Tingkat Kematangan Buah Mangga Menggunakan Metode <i>GLCM</i> dan <i>KNN</i> .
4	T3116343	Fadlun Gonta	Deteksi Tingkat Kematangan Buah Pisang Gapi Menggunakan Metode K-Nearest Neighbor (KNN) Dan Gray Level Co-Occurance Matrix (GLCM).
5	T3116021	Dana Pratama Jambo	Pengenalan Ekspresi Wajah Secara Real Time Menggunakan Klasifikasi K-NN (K-Nearest Neighbor)
6	T3116307	Ikbal Iskandar	Prototype Kandang Ayam cerdas Berbasis Mikrokontroller Arduino Uno
7	T3116123	Rifai Gobel	Prototype Deteksi Suhu Tubuh Menggunakan Sensor GY-906MLXESF Berbasis Internet Of Things
8	T3116067	Mohammad Rizal Kasim	Pengenalan Buah Nanas Menggunakan Metode Gray Level Co-Occurance Matrix (GLCM) dan K-Nearest Neighbor (KNN).
.....			
352		T Susim	Pengolahan Citra untuk Pengenalan Wajah (Face Recognition) Menggunakan OpenCV

4.2 Model yang diusulkan

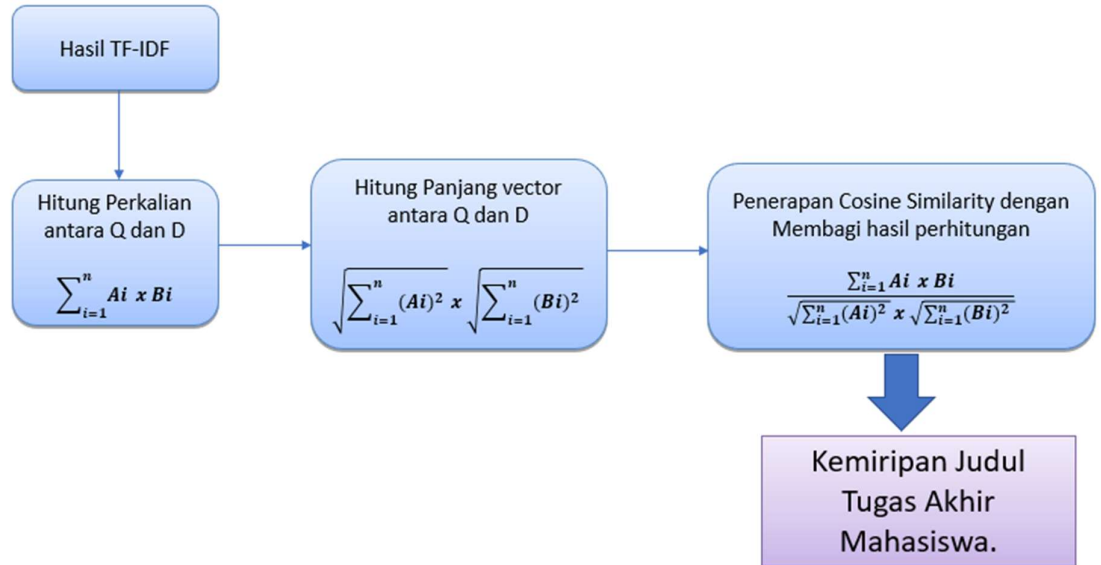
Preprocessing dengan Library Sastrawi



Pembobotan Dokumen dengan TF-IDF



Pengukuran Kemiripan Dokumen dengan Cosine Similarity



Evaluasi Model

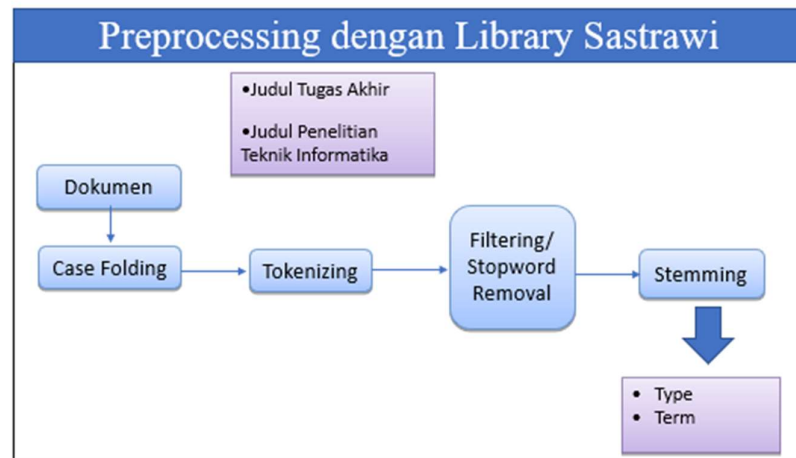
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{FN+TP}$$

4.2.1 Tahap *Preprocessing*

Preprocessing adalah tahapan seleksi dokumen yang akan dipakai untuk analisa berikutnya. Terdapat beberapa tahapan untuk memproses ialah sebagai berikut:



Gambar 4.1 : *Preprocessing*

a. *Case Folding* : *Case Folding* merupakan proses untuk mengkonversi teks ke dalam format huruf kecil (*lowercase*). Hal ini bertujuan untuk memberikan bentuk standar pada teks. Sebagai contoh “Meminjam” menjadi “meminjam”.

b. *Tokenizing* : *Tokenizing* atau disebut juga tahap *Lexical Analysis* adalah proses pemotongan kalimat teks menjadi bagian-bagian yang lebih kecil, yang disebut token. Sehingga menjadi suatu kata yang berdiri sendiri baik dalam bentuk perulangan maupun tunggal.

c. *Stopword Removal* : Penghilangan kata-kata tidak deskriptif atau kata *stopword* dari hasil *tokenizing*, untuk mengurangi waktu dan menghindari proses *list* yang panjang guna meningkatkan efisiensi, serta keefektifan dari perolehan informasi.

d. *Stemming* : *Stemming* adalah proses pengubahan bentuk kata menjadi kata dasar atau tahap mencari *root* kata dari tiap kata hasil *filtering*. Sebagai contoh kata meminjam, dipinjam, meminjamkan, akan distem ke kata dasarnya yaitu “pinjam”.

Table 4.2: Hasil *Preprocessing*

ID	Hasil Case Folding	Hasil Tokenizing	Hasil Stopword/Filtering	Hasil Stemming
q	identifikasi kematangan buah jambu biji dengan image processing menggunakan metode knn dan fitur ekstraksi glcm	[identifikasi], [kematangan], [buah] [jambu], [biji], [dengan], [image], [processing], [menggunakan], [metode], [knn], [dan], [fitur], [ekstraksi], [glcm].	identifikasi', 'kematangan', 'buah', 'jambu', 'biji', 'image', 'processing', 'menggunakan', 'metode', 'knn', 'fitur', 'ekstraksi', 'glcm'.	identifikasi, matang, buah, jambu, biji, image, processing, guna, metode, knn, fitur, ekstraksi, glcm.
D1	pengenalan kematangan buah pisang menggunakan metode gray level co-occurrence matrix dan k- nearest neighbor.	[pengenalan], [kematangan], [buah], [pisang], [menggunakan], [metode], [gray], [level], [co], [occurrence], [matrix], [dan], [k], [nearest], [neighbor].	pengenalan', 'kematangan', 'buah', 'pisang', 'menggunakan', 'metode', 'gray', 'level', 'co', 'occurrence', 'matrix', 'k', 'nearest', 'neighbor'.	kenal, matang, buah, pisang, guna, metode, gray, level,

ID	Hasil Case Folding	Hasil Tokenizing	Hasil Stopword/Filtering	Hasil Stemming
				co, occurance, matrix, k, nearest, neighbor.
D2	penerapan metode glcm (gray level co-occurrence matrix) dan knn (k-nearest neighbor) untuk mendeteksi tingkat kematangan buah belimbing bintang.	[penerapan], [metode], [glcm], [gray], [level], [co], [occurance], [matrix], [dan], [knn], [k], [nearest], [neighbor], [untuk], [mendeteksi], [tingkat], [kematangan], [buah], [belimbing], [bintang].	'penerapan', 'metode', 'glcm', 'gray', 'level', 'co', 'occurance', 'matrix', 'knn', 'k', 'nearest', 'neighbor', 'untuk', 'mendeteksi', 'tingkat', 'kematangan', 'buah', 'belimbing', 'bintang'.	terap, metode, glcm, gray, level, co, occurance, matrix, knn, k, nearest, neighbor, untuk, mendeteksi, tingkat, matang, buah, belimbing,

ID	Hasil Case Folding	Hasil Tokenizing	Hasil Stopword/Filtering	Hasil Stemming
				bintang.
D3	pengenalan tingkat kematangan buah mangga menggunakan metode <i>glcm</i> dan <i>knn</i> .	[pengenalan], [tingkat], [kematangan], [buah], [mangga], [menggunakan], [metode], [<i>glcm</i>], [dan], [<i>knn</i>].	'pengenalan', 'tingkat', 'kematangan', 'buah', 'mangga', 'menggunakan', 'metode', ' <i>glcm</i> ', ' <i>knn</i> '.	kenal, tingkat, matang, buah, mangga, guna, metode, <i>glcm</i> , <i>knn</i> .
..	..	..	..	..

Dibawah ini ialah term yang dihasilkan dari tahapan preprocessing dapat dilihat pada tabel berikut:

Table 4.3 : *Term* yang dihasilkan

Type	Term	Term yang diperoleh
guna,	guna	guna
metode,	metode	metode
knn,	knn	knn
fitur,	fitur	fitur
ekstraksi,	ekstraksi	ekstraksi
<i>glcm</i>	<i>glcm</i>	<i>glcm</i>
kenal,	kenal,	kenal,
matang,		pisang,
buah,		gray,

Type	Term	Term yang diperoleh
pisang,	pisang,	level,
guna,		co,
metode,		occurance,
gray,	gray,	matrix,
level,	level,	k,
co,	co,	nearest,
gray,		ekspresi,
..	..	..

Process Praprocesing Dengan Menggunakan Bahasa Pemrograman Python Menggunakan *library*:

1. *Streamlit* : *framework* berbasis *Python*
2. Sastrawi : *library preprocessing* Untuk Dokumen Berbahasa Indonesia

a) *Case Folding*

Masukkan Judul Tugas Akhir 😊

Identifikasi Kematangan Buah Jambu Biji Dengan Image Processing Menggunakan Metode KNN Dan Fitur Ekstraksi GLCM

Process

Case Folding:

```

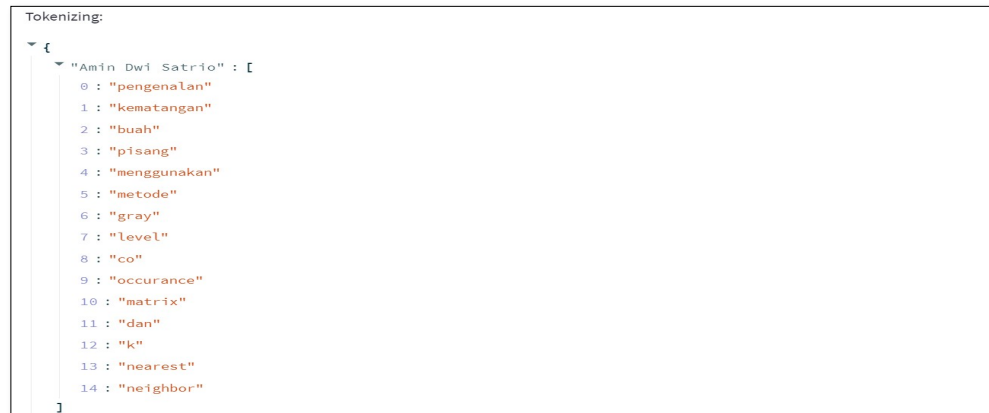
Amin Dwi Satrio" :
"pengenalan kematangan buah pisang menggunakan metode gray level co-occurance matrix dan k-nearest neighbor."
Ournia Shandy" :
"penerapan metode glcm (gray level co-occurance matrix) dan knn (k-nearest neighbor) untuk mendeteksi tingkat
kematangan buah belimbing bintang."
Mohammad Rezky Bagus cahputra Pakaya" :
"pengenalan tingkat kematangan buah mangga menggunakan metode glcm dan knn."

```

Gambar 4.2 : Case Folding

Case Folding merupakan proses untuk mengkonversi teks ke dalam format huruf kecil (*lowercase*). Hal ini bertujuan untuk memberikan bentuk standar pada teks.

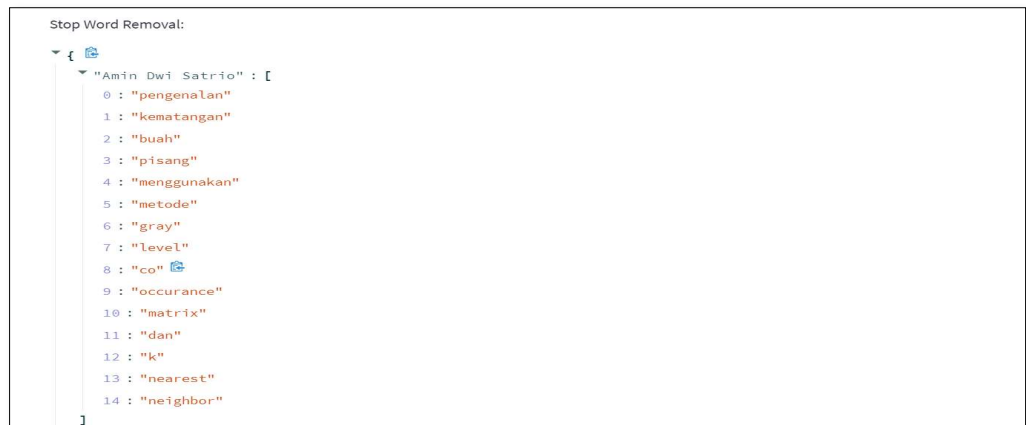
b) Tokenizing



Gambar 4.3 : Tokenizing

Tokenizing atau disebut juga tahap *Lexical Analysis* adalah proses pemotongan teks menjadi bagian-bagian yang lebih kecil, yang disebut token.

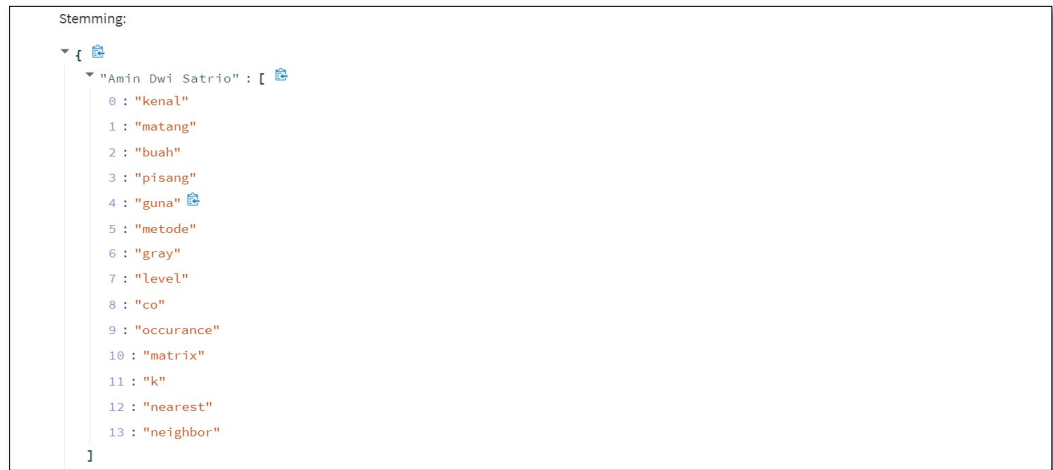
c) Stopword Removal/Filtering



Gambar 4.4 : Stopword Removal/Filtering

Penghilangan kata-kata tidak deskriptif atau kata *stopword* dari hasil *tokenizing*.

d) Stemming

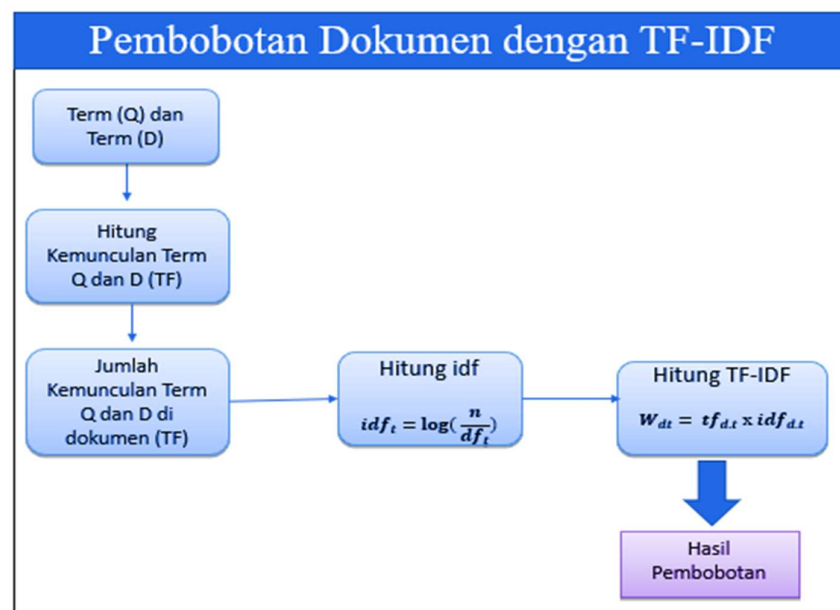


Gambar 4.5 : Stemming

Stemming adalah proses pengubahan bentuk kata menjadi kata dasar atau tahap mencari *root* kata dari tiap kata hasil *filtering*.

4.2.2 Pembobotan *TF-IDF*

Setelah dilakukan tahapan *preprocessing*, selanjutnya dilakukan perhitungan metode dengan menggunakan *TF.IDF*.



Gambar 4.6 : Pembobotan Dokumen

a) *Term Frequency*

Proses Pertama yang dilakukan ialah menghitung TF (*Term Frequency*) dari setiap dokumen dapat dilihat pada tabel dibawah ini:

Table 4.4 : *Term Frequency* (TF)

Term(t)	Q	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
identifikasi	1	0	0	0	0	0	0	0	0	0	0
image	1	0	0	0	0	0	0	0	0	0	0
processing	1	0	0	0	0	0	0	0	0	0	0
jambu	1	0	0	0	0	0	0	0	0	0	0
biji	1	0	0	0	0	0	0	0	0	0	0
fitur	1	0	0	0	0	0	0	0	0	0	0
ekstraksi	1	0	0	0	0	0	0	0	0	0	0
kenal	0	1	0	1	0	1	0	0	1	1	0
matang	1	1	1	1	1	0	0	0	0	0	0
buah	1	1	1	1	1	0	0	0	1	0	0
pisang	0	1	0	0	1	0	0	0	0	0	0
metode	1	1	1	1	1	0	0	0	1	0	1
gray	0	1	1	0	1	0	0	0	1	0	0
level	0	1	1	0	1	0	0	0	1	0	0
co	0	1	1	0	1	0	0	0	1	0	0
occurance	0	1	1	0	1	0	0	0	1	0	0
matrix	0	1	1	0	1	0	0	0	1	0	0
k	0	1	1	0	1	2	0	0	1	0	1
nearest	0	1	1	0	1	1	0	0	1	0	1
neighbor	0	1	1	0	1	1	0	0	1	0	1
bintang	0	0	1	0	0	0	0	0	0	0	0
mangga	0	0	0	1	0	0	0	0	0	0	0

Term(t)	Q	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
gapi	0	0	0	0	1	0	0	0	0	0	0
..	..	..	..	..	..	..	..	..	..	..	..

b) *Document Frequency (DF)*

Selanjutnya dilakukan tahapan dengan menghitung DF ialah dimana menghitung berapa kali dari setiap kata muncul pada dokumen, proses perhitungannya dapat dilihat pada tabel dibawah ini :

Table 4.5: Document Frequency (DF)

Term(t)	Doc(d)	df
identifikasi	Q	1
Image	Q	1
processing	Q	1
Jambu	Q	1
Biji	Q	1
Fitur	Q	1
ekstraksi	Q	1
Kenal	d1,d3,d5,d8,d9	5
matang	d1,d2,d3,d4	4
Buah	Q,d1,d2,d3,d4,d5,d8	7
pisang	d1,d4	2
metode	Q,d1,d2,d3,d4,d8,d10	7
Gray	d1,d2,d4,d8	4
Level	d1,d2,d4,d8	4
Co	d1,d2,d4,d8	4
occurance	d1,d2,d4,d8	4
matrix	d1,d2,d4,d8	4

Term(t)	Doc(d)	df
K	d1,d2,d4,d5,d8,d10	6
nearest	d1,d2,d4,d5,d8,d10	6
neighbor	d1,d2,d4,d5,d8,d10	6
Terap	d2	1
Glm	Q,d2,d3	3
Knn	Q,d2,d3,d4,d5,d8	6
deteksi	d2,d4,d7	3
tingkat	d2,d3,d4	3
belimbing	d2	1
bintang	d2	1
mangga	d3	1
Gapi	d4	1
ekspresi	d5	1
..	...	..

c) IDF (Invers Document Frequency)

Proses selanjutnya dilakukan perhitungan IDF (Invers Document Frequency) dengan menggunakan rumus dibawah ini untuk menghasilkan suatu nilai IDF :

$$idf_t = \log \left(\frac{n}{df_t} \right)$$

Proses perhitungan secara manual dapat dilihat pada tabel dibawah ini dari hasil perhitungan IDF seperti :

Table 4.6 : Hasil Perhitungan Inverse Document Frequency (IDF)

Term(t)	df	Idf
identifikasi	1	1,041
image	1	1,041

Term(t)	df	Idf
processing	1	1,041
jambu	1	1,041
biji	1	1,041
fitur	1	1,041
ekstraksi	1	1,041
kenal	5	0,342
matang	4	0,439
buah	7	0,196
pisang	2	0,74
metode	7	0,196
gray	4	0,439
level	4	0,439
co	4	0,439
occurance	4	0,439
matrix	4	0,439
k	6	0,263
nearest	6	0,263
neighbor	6	0,263
terap	1	1,041
glcm	3	0,564
knn	6	0,263
..	..	...

Selanjutnya ialah suatu proses perhitungan TF.IDF secara manual, prosesnya dapat dilihat pada bagian dibawah ini :

$$W_{dt} = tf_{dt} \times idf_{dt}$$

Pada tabel dibawah ini dapat dilihat hasil perhitungan dari TF.IDF ialah sebagai berikut :

Table 4.7: Hasil Pembobotan Kata (TF-IDF)

Q	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	...
1,041	0	0	0	0	0	0	0	0	0	0	..
1,041	0	0	0	0	0	0	0	0	0	0	..
1,041	0	0	0	0	0	0	0	0	0	0	..
1,041	0	0	0	0	0	0	0	0	0	0	..
1,041	0	0	0	0	0	0	0	0	0	0	..
1,041	0	0	0	0	0	0	0	0	0	0	..
1,041	0	0	0	0	0	0	0	0	0	0	..
0	0,342	0	0,342	0	0,342	0	0	0,342	0,342	0	..
0	0,439	0,4393	0,439	0,439	0	0	0	0	0	0	..
0,196	0,196	0,1963	0,196	0,196	0	0	0	0,196	0	0	
0	0,74	0	0	0,74	0	0	0	0	0	0	
0,196	0,196	0,1963	0,196	0,196	0	0	0	0,196	0	0,196	
0	0,439	0,4393	0	0,439	0	0	0	0,439	0	0	
0	0,439	0,4393	0	0,439	0	0	0	0,439	0	0	
0	0,439	0,4393	0	0,439	0	0	0	0,439	0	0	
0	0,439	0,4393	0	0,439	0	0	0	0,439	0	0	
0	0,439	0,4393	0	0,439	0	0	0	0,439	0	0	
0	0,263	0,2632	0	0,263	0,526	0	0	0,263	0	0,263	
0	0,263	0,2632	0	0,263	0,263	0	0	0,263	0	0,263	
0	0,263	0,2632	0	0,263	0,263	0	0	0,263	0	0,263	
0	0	1,0414	0	0	0	0	0	0	0	0	
0,564	0	0,5643	0,564	0	0	0	0	0	0	0	

Q	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	...
0,263	0	0,2632	0,263	0,263	0,263	0	0	0,263	0	0	
..	..	..	..	..	...	...	..	..	..	..	

d) Hasil Penerapan *TFIDF* pada *Python*

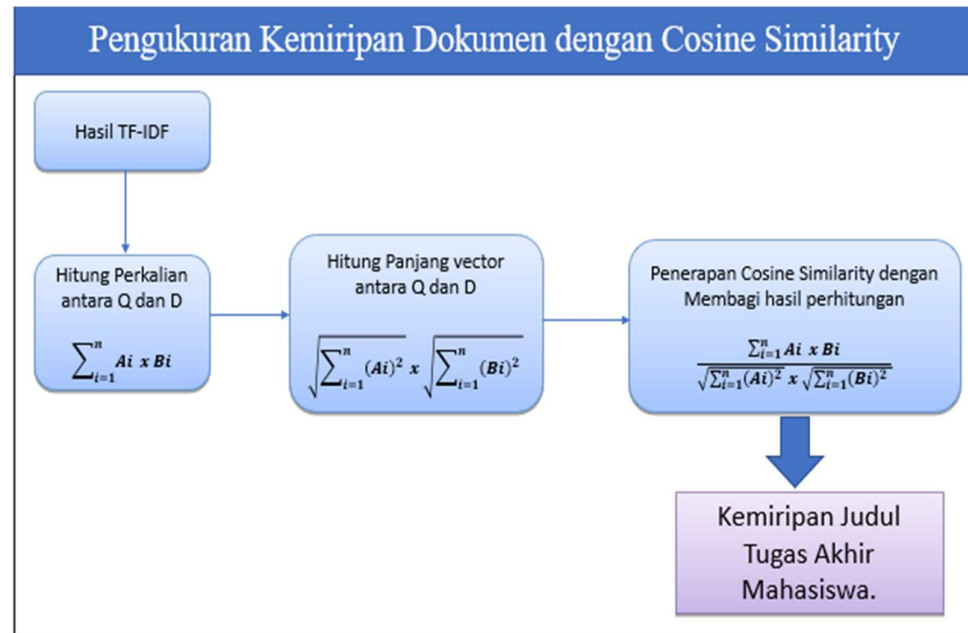
1. *Sklearn* : Library Machine Learning Untuk Membangun Model

	Amin Dwi Satrio	Ournia Shandy	Mohammad Re	Fadlun Gonta	Dana Pratama Jambo	Ikbai Iskandar	Rifai Gobel	Mohammad Rizal Kas
Amin Dwi Satrio	1	0.6145	0.367	0.7594	0.2247	0	0.0137	0.74
Ournia Shandy	0.6145	1	0.4288	0.7525	0.0936	0	0.0587	0.63
Mohammad Rezky Bagus	0.367	0.4288	1	0.456	0.1304	0	0.0161	0.41
Fadlun Gonta	0.7594	0.7525	0.456	1	0.1088	0	0.0718	0.66
Dana Pratama Jambo	0.2247	0.0936	0.1304	0.1088	1	0	0.0128	0.21
Ikbai Iskandar	0	0	0	0	0	1	0.1534	
Rifai Gobel	0.0137	0.0587	0.0161	0.0718	0.0128	0.1534	1	0.01
Mohammad Rizal Kasim	0.7413	0.6334	0.4138	0.6631	0.2158	0	0.0132	
Rafli Baid	0.0278	0.0124	0.0326	0.0236	0.012	0	0.011	0.02
Tridita Bidul	0.0086	0	0.0101	0.0073	0.008	0	0.0073	0.00

Gambar 4.7 : Hasil TF-IDF pada Python

4.2.3 Cosine Similarity

Setelah perhitungan TF.IDF dilakukan dan nilai TF.IDF diketahui, maka dilakukan tahapan selanjutnya ialah perhitungan metode cosine similarity yang dimana cosine similarity merupakan proses dimana untuk mengetahui tingkat kemiripan dokumen.



Gambar 4.8 : Pengukuran Kemiripan dengan *Cosine Similarity*

Berikut ialah langkah-langkah perhitungan manual algoritma Cosine Similarity:

1. Dikalikan nilai *term frequency* dengan nilai *inverse document frequency* tiap *term* dalam Q maupun D.
2. Dihitung hasil perkalian skalar masing-masing dokumen terhadap query key. Hasil perkalian dari setiap jawaban dengan query dijumlahkan (sesuai pembilang pada rumus di atas).
3. Dihitung hasil perkalian vektor tiap dokumen dan query.
4. Dihitung nilai Cosine similarity (nilai vektor beda antara D terhadap Q) dengan menggunakan rumus :

$$\text{Cos} \propto = \frac{X \cdot Y}{|X||Y|} = \frac{\sum_{i=1}^n X_i \times Y_i}{\sqrt{\sum_{i=1}^n (X_i)^2} \times \sqrt{\sum_{i=1}^n (Y_i)^2}}$$

Hasil proses dari perhitungan *cosine similarity* dapat dilihat pada tabel dibawah ini:

Table 4.8: Hasil Perkalian Q dan D

[illegible]

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
Total D0.Di	0,084	0,46	0,472	0,15	0,077	0	0,01	0,15	0,01	0,05

Pada tabel dibawah ini dapat dilihat proses dimana hasil perhitungan dari rumus *cosine similarity* menggunakan perhitungan akar:

Table 4.9: Hasil Perhitungan Akar Tiap Dokumen

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Q
0	0	0	0	0	0	0	0	0	0	1,084
0	0	0	0	0	0	0	0	0	0	1,084
0	0	0	0	0	0	0	0	0	0	1,084
0	0	0	0	0	0	0	0	0	0	1,084
0	0	0	0	0	0	0	0	0	0	1,084
0	0	0	0	0	0	0	0	0	0	1,084
0	0	0	0	0	0	0	0	0	0	1,084
0,117	0	0,117	0	0,12	0	0	0,12	0,12	0	0
0,193	0,193	0,193	0,193	0	0	0	0	0	0	0
0,038	0,039	0,038	0,038	0	0	0	0,04	0	0	0,038
0,548	0	0	0,548	0	0	0	0	0	0	0
0,038	0,039	0,038	0,038	0	0	0	0,04	0	0,038	0,038
0,193	0,193	0	0,193	0	0	0	0,19	0	0	0
0,193	0,193	0	0,193	0	0	0	0,19	0	0	0
0,193	0,193	0	0,193	0	0	0	0,19	0	0	0

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Q
0,193	0,193	0	0,193	0	0	0	0,19	0	0	0
0,193	0,193	0	0,193	0	0	0	0,19	0	0	0
0,069	0,069	0	0,069	0,28	0	0	0,07	0	0,069	0
0,069	0,069	0	0,069	0,07	0	0	0,07	0	0,069	0
0,069	0,069	0	0,069	0,07	0	0	0,07	0	0,069	0
0	1,085	0	0	0	0	0	0	0	0	0
0	0,318	0,318	0	0	0	0	0	0	0	0,318
0	0,069	0,069	0,069	0,07	0	0	0,07	0	0	0,069
0	0,318	0	0,318	0	0	0,32	0	0	0	0
0	0,318	0,318	0,318	0	0	0	0	0	0	0
0	1,085	0	0	0	0	0	0	0	0	0
0	1,085	0	0	0	0	0	0	0	0	0
0	0	1,084	0	0	0	0	0	0	0	0
0	0	0	1,084	0	0	0	0	0	0	0
0	0	0	0	1,08	0	0	0	0	0	0
0	0	0	0	1,08	0	0	0	0	0	0
0	0	0	0	1,08	0	0	0	0	0	0
0	0	0	0	1,08	0	0	0	0	0	0
0,008	0	0,008	0,008	0,01	0	0,01	0,01	0,01	0,008	0,008
0	0	0	0	1,08	0	0	0	0	0	0
0	0	0	0	0	0,55	0,55	0	0	0	0
0	0	0	0	0	1,08	0	0	0	0	0
0	0	0	0	0	0,55	0	0	0	0,548	0
0	0	0	0	0	1,08	0	0	0	0	0
0	0	0	0	0	1,08	0	0	0	0	0
0	0	0	0	0	1,08	0	0	0	0	0

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Q
	0	0	0	0	0	1,08	0	0	0	0	0
	0	0	0	0	0	1,08	0	0	0	0	0
	0	0	0	0	0	0	1,08	0	0	0	0
	0	0	0	0	0	0	1,08	0	0	0	0
	0	0	0	0	0	0	1,08	0	0	0	0
	0	0	0	0	0	0	1,08	0	0	0	0
	0	0	0	0	0	0	1,08	0	0	0	0
	0	0	0	0	0	0	1,08	0	0	0	0
	0	0	0	0	0	0	1,08	0	0	0	0
	0	0	0	0	0	0	1,08	0	0	0	0
	0	0	0	0	0	0	0	1,08	0	0	0
	0	0	0	0	0	0	0	0	1,08	0	0
	0	0	0	0	0	0	0	0	1,08	0	0
	0	0	0	0	0	0	0	0	1,08	0	0
	0	0	0	0	0	0	0	0	1,08	0	0
	0	0	0	0	0	0	0	0	1,08	0	0
	0	0	0	0	0	0	0	0	0	1,084	0
	0	0	0	0	0	0	0	0	0	1,084	0
	0	0	0	0	0	0	0	0	0	1,084	0
Total Kuadrat	2,113	5,721	2,183	3,785	6,03	7,6	9,54	2,53	5,54	4,052	8,057
Akar Kuadrat	1,454	2,392	1,478	1,945	2,46	2,76	3,09	1,59	2,35	2,013	2,839

4.2.4 Menghitung kemiripan dokumen

Dibawah ini merupakan hasil dari perhitungan *cosine similarity*:

Similarity Q dan D1

$$\text{Cosine Similarity} = D1/(\text{akar kuadrat Q} \times \text{akar kuadrat D1})$$

$$= 0,043$$

$$\text{Persentase} = 4\%$$

Similarity Q dan D2

$$\text{Cosine Similarity} = D2/(\text{akar kuadrat Q} \times \text{akar kuadrat D2})$$

$$= 0,388$$

$$\text{Persentase} = 39\%$$

Similarity Q dan D3

$$\text{Cosine Similarity} = D3/(\text{akar kuadrat Q} \times \text{akar kuadrat D3})$$

$$= 0,246$$

$$\text{Persentase} = 25\%$$

Similarity Q dan D4

$$\text{Cosine Similarity} = D4/(\text{akar kuadrat Q} \times \text{akar kuadrat D4})$$

$$= 0,103$$

$$\text{Persentase} = 10\%$$

Similarity Q dan D5

$$\text{Cosine Similarity} = D5/(\text{akar kuadrat Q} \times \text{akar kuadrat D5})$$

$$= 0,067$$

$$\text{Persentase} = 7\%$$

Similarity Q dan D6

$$\text{Cosine Similarity} = D6 / (\text{akar kuadrat } Q \times \text{akar kuadrat } D6)$$

$$= 0$$

$$\text{Persentase} = 0\%$$

Similarity Q dan D7

$$\text{Cosine Similarity} = D7 / (\text{akar kuadrat } Q \times \text{akar kuadrat } D7)$$

$$= 0,010$$

$$\text{Persentase} = 1\%$$

Similarity Q dan D8

$$\text{Cosine Similarity} = D8 / (\text{akar kuadrat } Q \times \text{akar kuadrat } D8)$$

$$= 0,084$$

$$\text{Persentase} = 8\%$$

Similarity Q dan D9

$$\text{Cosine Similarity} = D9 / (\text{akar kuadrat } Q \times \text{akar kuadrat } D9)$$

$$= 0,008$$

$$\text{Persentase} = 0,8\%$$

Similarity Q dan D10

$$\text{Cosine Similarity} = D10 / (\text{akar kuadrat } Q \times \text{akar kuadrat } D10)$$

$$= 0,035$$

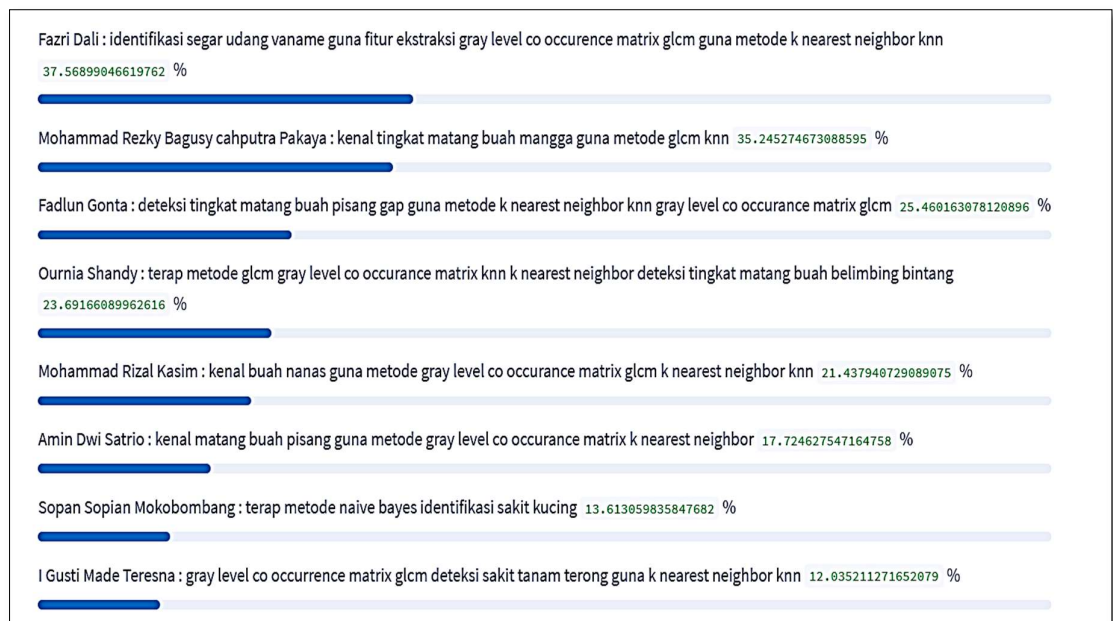
$$\text{Persentase} = 4\%$$

Selanjutnya pengurutan tingkat kemiripan yang dihasilkan tiap dokumen dengan query yang diujikan.

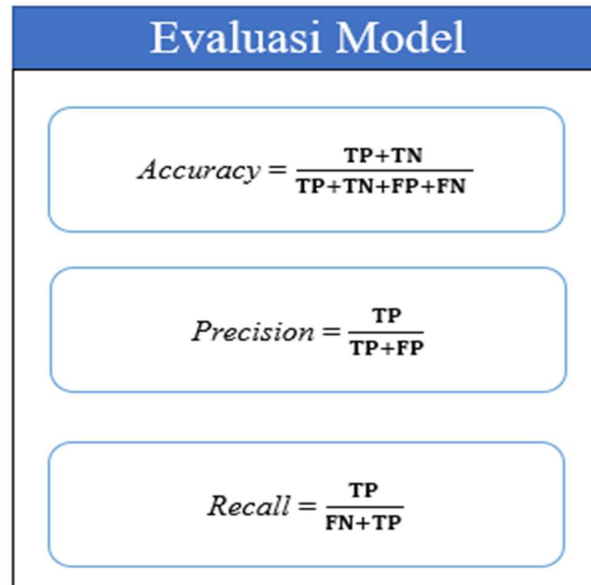
Table 4.10: Pengurutan Tingkat Kemiripan

Dokumen	Nilai Kemiripan
D2	39%
D3	25%
D4	10%
D8	8%
D5	7%
D1	4%
D10	4%
D7	1%
D9	0,08%
D6	0%

4.3 Implementasi *Cosine* pada *Python*

**Gambar 4.9 :** Kemiripan Dokumen dengan *Cosine Similarity*

4.4 Evaluasi Model



Evaluasi Model dilakukan untuk mencari perbedaan yang akan dijadikan perbandingan dari sistem yang telah dibuat. Pada Penelitian ini Evaluasi model menggunakan *Accuracy*, *Precision*, dan *Recall*. (*accuracy*) menyangkut perihal pengukuran kinerja dari sebuah metode yang digunakan, (*precision*) berkaitan dengan kemampuan sistem untuk tidak memanggil dokumen yang tidak relevan, sedangkan perolehan (*recall*) berhubungan dengan kemampuan sistem untuk memanggil dokumen yang relevan.

Evaluasi di lakukan sebanyak 5 kali dengan menggunakan *Term* dari kata kunci yang sering muncul yaitu:

Table 4.11: Kata Kunci

Pengujian	Kata Kunci
Ke-1	Prediksi
Ke-2	Kematangan Buah
Ke-3	Metode KNN
Ke-4	Rancang Bangun Game
Ke-5	Arduino Uno

Table 4.12: Hasil Pengujian tiap *query*

No.	Pengujian 1			Pengujian 2			Pengujian 3		
	Nm Mhs	Sim.	Rel.	Nm Mhs	Sim.	Rel.	Nm Mhs	Sim.	Rel.
1	Riton	30%	Ya	Moh.Rezky	43%	Ya	Moh.Rezky	40%	Ya
2	Zulkifli	27%	Ya	Sela	39%	Ya	Rizal	32%	Ya
3	Imelda	26%	Ya	Amin	35%	Ya	Fadlun	28%	Ya
4	Andre	25%	Ya	Rizal	34%	Ya	Dimas	27%	Ya
5	I Kadek	24%	Ya	Fadlun	30%	Ya	Ournia	26%	Ya
6	Amin	0%	Tidak	Dimas	29%	Ya	I Gusti	19%	Tidak
7	Ournia	0%	Tidak	Ournia	28%	Ya	Sela	12%	Tidak
8	Moh.Rezky	0%	Tidak	Moh.Kasim	16%	Tidak	Riton	12%	Tidak
9	Fadlun	0%	Tidak	Dana	0%	Tidak	La Ode	11%	Tidak
No.	Pengujian 4			Pengujian 5					
	Nm Mhs	Sim.	Rel.	Nm Mhs	Sim.	Rel.			
1	Imam	51%	Ya	Mirna	39%	Ya			
2	Mirna	35%	Ya	Ikbal	38%	Ya			
3	Alfaridzi	34%	Ya	Alfaridzi	38%	Ya			
4	Adnan	32%	Ya	Tahir	34%	Ya			
5	Zul	29%	Ya	Zul	33%	Ya			
6	Fadhila	18%	Tidak	Rahmad	27%	Ya			
7	Ridwan	17%	Tidak	Arisman	17%	Tidak			
8	Nuriman	14%	Tidak	Amin	0%	Tidak			
9	Rizki	13%	Tidak	Ournia	0%	Tidak			

Jumlah data yang digunakan adalah 35 data, dengan menggunakan Threshold 30% sebagai patokan dokumen tersebut relevan atau tidak.

Sehingga di dapatkan hasil evaluasi model:

Table 4.13: Hasil Evaluasi

q.	Confusion Matrix						
	TP	FP	FN	TN	Pre	Re	Acc
1	5	0	0	30	1	1	1
2	7	1	1	26	0,88	0,88	0,94
3	5	10	0	20	0,33	1	0,71
4	5	4	0	26	0,56	1	0,89
5	6	1	1	27	0,86	0,86	0,94
Rata-rata					72,4	94,6	89,7

Pada evaluasi model, digunakan jumlah data training pada *database* sebanyak 35 data. Dengan data uji atau query yang digunakan sebagai data pembandingan sebanyak 5 data. Didapatkan hasil akurasi sebesar 89.7%.

Table 4.14: Visualisasi Confussion Matrix

Confusion Matrix

	Rel.	Tidak Relevan
Di-Retrieve	TP	FP
tidak Retrieve	FN	TN

Perhitungan nilai akurasi, persisi dan recall sesuai dengan query masing-masing:

a) **Kata Kunci** : Prediksi

$$\begin{aligned}
 Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\
 &= \frac{5+30}{35} = 1,00 \\
 Precision &= \frac{TP}{TP+FP} \\
 &= \frac{5}{5+0} = 1,00
 \end{aligned}$$

$$\begin{aligned}
 Recall &= \frac{TP}{FN+TP} \\
 &= \frac{5}{0+5} = 1,00
 \end{aligned}$$

b) **Kata Kunci** : Kematangan Buah

$$\begin{aligned}
 Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\
 &= \frac{7+26}{35} = 0,94
 \end{aligned}$$

$$\begin{aligned}
 Precision &= \frac{TP}{TP+FP} \\
 &= \frac{7}{7+1} = 0,88
 \end{aligned}$$

$$\begin{aligned}
 Recall &= \frac{TP}{FN+TP} \\
 &= \frac{7}{1+7} = 0,88
 \end{aligned}$$

c) **Kata Kunci** : Metode KNN

$$\begin{aligned}
 Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\
 &= \frac{5+20}{35} = 0,71
 \end{aligned}$$

$$\begin{aligned}
 Precision &= \frac{TP}{TP+FP} \\
 &= \frac{5}{5+10} = 0,33
 \end{aligned}$$

$$\begin{aligned}
 Recall &= \frac{TP}{FN+TP} \\
 &= \frac{5}{0+5} = 1,00
 \end{aligned}$$

d) **Kata Kunci** : Rancang Bangun Game

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\ &= \frac{5+26}{35} = \mathbf{0,89} \end{aligned}$$

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ &= \frac{5}{5+4} = \mathbf{0,56} \end{aligned}$$

$$\begin{aligned} Recall &= \frac{TP}{FN+TP} \\ &= \frac{5}{0+5} = \mathbf{1,00} \end{aligned}$$

e) **Kata Kunci** : Arduino Uno

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\ &= \frac{6+27}{35} = \mathbf{0,94} \end{aligned}$$

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ &= \frac{6}{6+1} = \mathbf{0,86} \end{aligned}$$

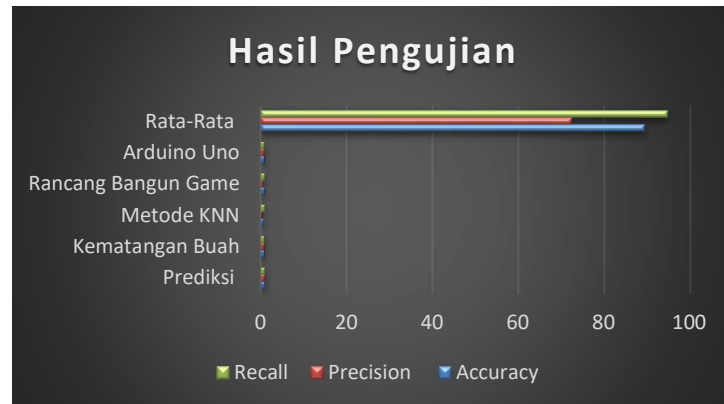
$$\begin{aligned} Recall &= \frac{TP}{FN+TP} \\ &= \frac{6}{1+6} = \mathbf{0,86} \end{aligned}$$

BAB V

PEMBAHASAN PENELITIAN

5.1 Pembahasan Model

- Data judul yang berhasil dikumpulkan pada penelitian ini yaitu sebanyak 352 data, yang terdiri dari data judul tugas akhir mahasiswa teknik informatika Universitas Ichsan Gorontalo dan data judul penelitian mahasiswa yang diperoleh dari situs *google scholar* dengan tahun yang sama 2019-2020,2020-2021,2021-2022.
- Percobaan pada penelitian ini dilakukan sebanyak dua kali, yakni dalam tahapan pengujian pertama pengukuran tingkat similaritas dilakukan menggunakan judul tugas akhir mahasiswa teknik informatika Universitas Ichsan Gorontalo dengan judul artikel jurnal pada *google scholar*, terlihat kemiripan cosine similarity rata-rata tingkat kemiripan dominan rendah. Sedangkan pada pengujian kedua pengukuran tingkat similaritas dilakukan dengan cara membandingkan tiap judul tugas akhir mahasiswa teknik informatika Universitas Ichsan Gorontalo yang pilih secara acak, diperoleh rata-rata tingkat kemiripan dominan tinggi.
- Faktor-faktor yang mempengaruhi tingkat kemiripan yaitu terletak pada tahapan preprocessing khususnya dalam proses *stemming*. *Stemming* menyebabkan kata yang dibandingkan teridentifikasi sama atau bahkan sebaliknya.
- Berdasarkan hasil proses evaluasi yang dilakukan terhadap data *testing* yang berjumlah 5 query pada 35 data dengan nilai *threshold* yang digunakan sebesar 0,3 atau 30%, dapat diperoleh nilai kinerja terhadap model dari algoritma cosine similarity yang digunakan yaitu nilai recall sebesar 94.6%, nilai precision sebesar 72.4%, dan nilai accuracy sebesar 89.7%
- Berikut ini merupakan tampilan grafik dari hasil pengujian dengan menggunakan 5 suku kata kunci:



Gambar 5.1 : Grafik Hasil Pengujian

Nilai rata-rata hasil pengujian diatas diperoleh dari hasil menjumlahkan tingkat *accuracy* pada masing-masing *query*, kemudian dibagi dengan jumlah *query* yang dijadikan persen. Sehingga dapat disimpulkan berdasarkan perolehan nilai kerja model tersebut menunjukkan bahwa penggunaan algoritma *cosine similarity* dapat bekerja dengan baik dalam mengukur tingkat kemiripan judul tugas akhir, untuk itu penggunaan metode tersebut dapat digunakan dalam membangun aplikasi terhadap bidang yang membutuhkan.

BAB VI

PENUTUP PENELITIAN

6.1 Kesimpulan

Berdasarkan hasil analisa dan pengujian yang telah dilakukan, maka dapat ditarik kesimpulan bahwa integrasi antara pembobotan TF-IDF dan metode Cosine Similarity dapat menjadi salah satu metode untuk mendeteksi tingkat kemiripan dari judul tugas akhir. Hal ini diperoleh setelah melakukan pengujian dengan membandingkan tiap judul tugas akhir Universitas Ichsan Gorontalo diperoleh rata-rata tingkat kemiripan yang dominan tinggi, dibandingkan dengan perbandingan antara judul artikel jurnal dan judul tugas akhir memiliki nilai tingkat kemiripan rata-rata dominan rendah, hal tersebut dikarenakan penggunaan *stemming* dalam tahap preprocessing yang dapat meningkatkan tingkat kemiripan suatu dokumen.

Pengukuran kinerja/evaluasi dilakukan dengan perhitungan manual, menggunakan *confusion matrix* diperoleh rata-rata akurasi hingga 89.7%. Angka tersebut diperoleh dengan menjumlahkan akurasi pada masing-masing *query* kemudian dibagi dengan jumlah *query* lalu dijadikan persen. Dengan demikian penerapan *Cosine Similarity* dalam mendeteksi kemiripan judul tugas akhir dapat memberikan akurasi yang tergolong dalam kategori baik. Hal ini dibuktikan dengan nilai *accuracy* 89.7%, *precision* 72.4%, dan *recall* 94.6%.

6.2 Saran

Dari hasil pengujian akurasi dapat dikatakan performa algoritma ini baik, karena mendekati angka 100%. Meskipun demikian, karena perhitungan pengujian dilakukan dengan manual sehingga ketika dilakukan perhitungan akurasi menghasilkan akurasi yang tinggi. Selain itu keterbatasan data yang digunakan dalam penelitian. Tentu masih banyak aspek pada penelitian ini yang dapat ditingkatkan lagi untuk membantu memperbaiki performa algoritma TF-IDF dan *Cosine Similarity*.

Untuk penelitian kedepannya disarankan agar meningkatkan kuantitas data khususnya data judul artikel jurnal baik dengan menambah, menggabungkan,

maupun menggantinya, dan dapat menggunakan ratio perbandingan antara data judul tugas akhir Universitas Ichsan Gorontalo dan judul artikel jurnal minimal 50:50. Selain itu penelitian ini dapat dikembangkan lagi dengan menggunakan penggabungan metode yang lebih bervariasi ataupun metode selain *Cosine Similarity*, sehingga dapat meningkatkan performa yang lebih baik lagi dan juga memiliki interface agar mudah digunakan.

DAFTAR PUSTAKA

- [1] U. N. Hasanah, R. Satra, and F. Umar, “Deteksi Kemiripan Judul Skripsi Menggunakan Algoritma Smith Waterman,” *Bul. Sist. Inf. dan ...*, vol. 1, no. 1, pp. 56–65, 2020, [Online]. Available: <http://jurnal.fikom.umi.ac.id/index.php/BUSITI/article/view/676>.
- [2] M. Azmi, “Analisis Tingkat Plagiasi Dokumen Skripsi Dengan Metode Cosine Similarity Dan Pembobotan Tf-Idf,” *Tek. Teknol. Inf. dan Multimed.*, vol. 2, no. 2, pp. 90–95, 2022, doi: 10.46764/teknimedia.v2i2.51.
- [3] A. Amrulloh and I. F. Adam, “Sistem Pencarian Similaritas Judul Tugas Akhir Menggunakan Metode TF-IDF,” vol. 7, no. 2, pp. 74–82, 2021, doi: 10.24014/coreit.
- [4] R. T. Adek, M. Jannah, J. B. No, and K. B. Indah, “MAHASISWA DENGAN MENGGUNAKAN.”
- [5] I. Abdullah and E. Aribowo, “Rancang Bangun Aplikasi Pengecekan Kemiripan Judul Skripsi Dengan Metode Cosine Similarity (Studi Kasus : Program Studi Teknik Informatika UAD),” vol. 6, no. 2, pp. 43–52, 2018.
- [6] N. Andriani and A. Wibowo, “Implementasi Text Mining Klasifikasi Topik Tugas Akhir Mahasiswa Teknik Informatika Menggunakan Pembobotan TF-IDF dan Metode Cosine Similarity Berbasis Web,” *Senamika*, no. September, pp. 130–137, 2021, [Online]. Available: <https://conference.upnvj.ac.id/index.php/senamika/article/view/1807%0Ahttps://conference.upnvj.ac.id/index.php/senamika/article/download/1807/1350>.
- [7] F. Septian, “Penerapan Algoritma Cosine Similarity untuk Deteksi Kesamaan Konten pada Sistem Informasi Penelitian dan Pengabd ...,” 2016.
- [8] F. B. Sejati, P. Hendradi, and B. Pujiarto, “KEMIRIPAN TEMA MENGGUNAKAN METODE COSINE SIMILARITY (Studi Kasus : Di Universitas Muhammadiyah Magelang),” vol. 2, no. 2, 2019.
- [9] A. Riyani, M. Zidny, and A. Burhanuddin, “Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen,” vol. 2, no. 1,

pp. 23–27, 2019.

- [10] D. Kurniadi, S. F. C. Haviana, and A. Novianto, “Implementasi Algoritma Cosine Similarity pada sistem arsip dokumen di Universitas Islam Sultan Agung,” *J. Transform.*, vol. 17, no. 2, p. 124, 2020, doi: 10.26623/transformatika.v17i2.1613.
- [11] A. Agustiawan, “Analisis Similarity/Kemiripan Artikel Jurnal Online Terbitan Tahun 2019-2020 Di ISI Yogyakarta,” *ABDI PUSTAKA J. Perpust. dan Kearsipan*, vol. 2, no. 1, pp. 29–43, 2022, doi: 10.24821/jap.v2i1.6984.
- [12] A. Millah and S. Nurazizah, “Perbandingan Penggunaan Algoritma Cosinus dan Wu Palmer untuk Mencari Kemiripan Kata dalam,” vol. 2, no. 1, pp. 15–25, 2017.
- [13] G. E. I. Kambey and Dkk, “Penerapan Clustering pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia,” *Penerapan Clust. pada Apl. Pendeteksi Kemiripan Dok. teks Bhs. Indones.*, vol. 15, no. 2, pp. 75–82, 2020.
- [14] B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, “Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity,” vol. 2, no. 1, pp. 306–312, 2018.
- [15] E. L. Amalia, A. J. Jumadi, I. A. Mashudi, and D. W. Wibowo, “Analisis Metode Cosine Similarity Pada Aplikasi Ujian Online Otomatis (Studi Kasus JTI POLINEMA),” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 2, p. 343, 2021, doi: 10.25126/jtiik.2021824356.
- [16] G. E. I. Kambey *et al.*, “Penerapan Clustering pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia,” vol. 15, no. 2, pp. 75–82, 2020.

LAMPIRAN

Hasil Turnitin



Similarity Report ID: oid:25211:33413005

PAPER NAME

SKRIPSI_T3119082_SITTI MEAUTIAH D.
A JAN.docx

AUTHOR

T3119082-Sitti Meautiah D.A Ja semuth
styles@gmail.com

WORD COUNT

8194 Words

CHARACTER COUNT

42016 Characters

PAGE COUNT

51 Pages

FILE SIZE

2.2MB

SUBMISSION DATE

Apr 1, 2023 11:55 PM GMT+8

REPORT DATE

Apr 2, 2023 12:00 AM GMT+8

● 9% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 9% Internet database
- 0% Publications database
- Crossref database
- Crossref Posted Content database
- 0% Submitted Works database

● Excluded from Similarity Report

- Cited material
- Small Matches (Less than 30 words)

Summary

9% Overall Similarity

Top sources found in the following databases:

- 9% Internet database
- 0% Publications database
- Crossref database
- Crossref Posted Content database
- 0% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	scribd.com Internet	5%
2	andi.ddns.net Internet	<1%
3	apps.dtic.mil Internet	<1%
4	nero.trunojoyo.ac.id Internet	<1%
5	repository.ittelkom-pwt.ac.id Internet	<1%
6	core.ac.uk Internet	<1%
7	ijasce.org Internet	<1%
8	repository.ub.ac.id Internet	<1%



jom.fti.budiluhur.ac.id
Internet

<1%

Listing Program

```
from nltk.tokenize import RegexpTokenizer

from sklearn.feature_extraction.text import CountVectorizer

from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.metrics.pairwise import euclidean_distances

from sklearn.metrics.pairwise import cosine_similarity

from Sastrawi.StopWordRemover.StopWordRemoverFactory import
StopWordRemoverFactory

from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

import pandas as pd

import streamlit as st

import plotly.express as px

from PIL import Image

import numpy as np

import time

import operator


st.set_page_config(page_title='Cek Kesamaan Judul Tugas Akhir')

image = Image.open('images/Ichsan.jpeg')

st.image(image)

st.subheader('Deteksi Kemiripian Judul Tugas Akhir')

st.subheader('Fakultas Ilmu Komputer Universitas Ichsan Gorontalo')

# st.markdown('Streamlit is **_really_ cool**.')

```

```

# st.markdown("This text is :red[colored red], and this is
** :blue[colored]** and bold.")

# st.markdown(":green[\sqrt{x^2+y^2}=1$] is a Pythagorean identity.
:pencil:")

### --- LOAD DATAFRAME

excel_file = 'Book1s.xlsx'

sheet_name = 'Sheet2'

df = pd.read_excel(excel_file,

                    sheet_name=sheet_name,

                    header=0)

df_keterangan.dropna(inplace=True)

keterangan = df['Keterangan'].unique().tolist()

keterangan_selection = st.sidebar.multiselect('Apa yang akan di
Cek:',

                                              keterangan,

                                              default=keterangan)

# mask = df['Keterangan'].isin(keterangan_selection)

mask = df['Keterangan'].isin(keterangan_selection)

number_of_result = df[mask].shape[0]

st.markdown(f'*Available Results: {number_of_result}*')

jumlahdata=df.Nama_Mhs.count()

dataset={}

```

```

for i in range(0,jumlahdata):

    dataset[df>Nama_Mhs[i]]=df.Judul_tugas_akhir[i]

    st.write(dataset)

st.sidebar.write(dataset)

tab1, tab2= st.tabs(["Cek Kesamaan","Pengujian Model"])

with tab1:

    txt = st.text_area('Masukkan Judul Tugas Akhir :sunglasses:',
    '')

    if st.button('Process'):

        dataset["q"]=txt

        Case Folding

        st.write('Hasil Case Folding')

        for k in dataset.keys():

            hasil_case_folding = dataset[k].lower()

            dataset[k] = hasil_case_folding

        st.write(dataset)

        Tokenizing

        st.write('Hasil Tokenizing')

        tokenizer = RegexpTokenizer(r"\w+")

        for k in dataset.keys():

            tokens = tokenizer.tokenize(dataset[k])

            dataset[k] = tokens

```

```
st.write(dataset)
```

Number Removal

```
st.write('Number Removal')
```

```
for k in dataset.keys():  
    tokens = []  
    for t in dataset[k]:  
        if t.isnumeric() == False:  
            tokens.append(t)  
    dataset[k] = tokens  
st.write(dataset)
```

Stopword Removal by Sastrawi

```
st.write('Hasil Stopword Removal')  
factory = StopWordRemoverFactory()  
stopword_list = factory.get_stop_words()  
for k in dataset.keys():  
    tokens = []  
    for t in dataset[k]:  
        if t not in stopword_list:  
            tokens.append(t)  
    dataset[k] = tokens  
st.write(dataset)
```

Stemming by Sastrawi

```
st.write('Hasil Stemming')

factory = StemmerFactory()

stemmer = factory.create_stemmer()

for k in dataset.keys():

    tokens = []

    for t in dataset[k]:

        tokens.append(stemmer.stem(t))

    dataset[k] = tokens

st.write(dataset)
```

Mengembalikan Format Dataset Awal

```
st.write('Hasil Preprocessing')

for k in dataset.keys():

    dataset[k] = " ".join(dataset[k])

st.write(dataset)
```

```
# document = ["One Geek helps Two Geeks", "Two Geeks help  
Four Geeks", "Each Geek helps many other Geeks at GeeksforGeeks"]
```

Create a Vectorizer Object

```
# vectorizer = CountVectorizer()
```

```
# vectorizer.fit(document)
```

```

# Frekuensi Kemunculan Kata

tf = CountVectorizer()

term_doc_matrix = tf.fit_transform(dataset.keys())

pd.DataFrame(term_doc_matrix.toarray(),
index=dataset.keys(), columns=tf.get_feature_names_out())

# Pembobotan TF.IDF

tfidf = TfidfVectorizer()

inverted_index = tfidf.fit_transform(dataset.values())

pd.DataFrame(inverted_index.toarray(), index=dataset.keys(),
columns=tfidf.get_feature_names_out())

# Perhitungan Cosine Similarity

st.write('Hasil Perhitungan Cosine Similarity')

cs = cosine_similarity(inverted_index, inverted_index)

df_cs = pd.DataFrame(cs, index=dataset.keys(),
columns=dataset.keys())

#df_cs

# Pengurutan Reranking Cosine Similarity

st.write('Hasil Pengurutan Tingkat Kemiripan Dokumen')

rank_cs = {}

for k in dataset.keys():

```

```

        if k != "q":

            rank_cs[k] = df_cs.at[k, 'q']

    top_rank_cs = dict(sorted(rank_cs.items(),
key=operator.itemgetter(1), reverse=True))

    tampil=pd.DataFrame(top_rank_cs.values(),
index=top_rank_cs.keys(),columns=["Cosine Similarity"])

    #tampil

    hasil = {}

    for j in top_rank_cs.keys():

        if (top_rank_cs[j] >= 1):

            top_rank_cs[j]=1.00

            nilai=top_rank_cs[j]*100

        else :

            nilai=top_rank_cs[j]*100

    st.write(j,':',dataset[j],nilai,'% ' )

    progress_text1 = dataset[j]

    my_bar1 = st.progress(top_rank_cs[j])

```

DATASET

No	Nim	Nama_Mhs	Judul	Tahun
1	T3115018	Eka Damayanti Madina	SPK Untuk Pemetaan Daerah Potensial Pengembangan Budidaya Perikanan Menggunakan Metode TOPSIS	2019/2020
2	T3115316	Miyanti Moo	Penerapan Data Mining Untuk Memprediksi Jumlah Penderita Penyakit Kanker Serviks Menggunakan Algoritma C.45	2019/2020
3	T3115130	Fitrawaty Dungi	Prediksi Harga Jual Tanah Kota Gorontalo Menggunakan Metode C4.5	2019/2020
4	T3115136	Chintia A. Beu	Klasifikasi Pemberian Dana Beasiswa Program Indonesia Pintar (PIP) Menggunakan Metode K-Nearest Neighbor	2019/2020
5	T3114219	Alfred I. Mohi	Sistem Cerdas Prediksi Kanker Payudara Menggunakan Gaussian Naïve Bayes	2019/2020
6	T3112128	Siti Ajari Utiahman	Sistem Pendukung Keputusan Evaluasi Kinerja Karyawan Menggunakan Metode Promethee Studi Kasus PT. Manulife Indonesia	2019/2020
7	T3114260	Mariati R Akim	Memprediksi Penentuan Besarnya Pinjaman di Koperasi simpan Pinjam Menggunakan Algoritma Naïve Bayes	2019/2020
8	T3115035	Isman Bahutala	Rancang Bangun Prototype Monitoring Sampah Berbasis Arduino Uno	2019/2020
9	T3113215	Alham Djoli	Clustering Komoditi Unggulan Daerah Provinsi	2019/2020

No	Nim	Nama_Mhs	Judul	Tahun
			Gorontalo Menggunakan Algoritma K-Means	
10	T3115084	Mohammad Syarif G. Ishak	Deteksi Warna Kulit Berbasis Analisis HSV Color Model Untuk Klasifikasi Konten Pornografi	2019/2020
11	T3115016	Mahtias Sutiyo	Pengenalan Wajah Menggunakan Metode ANN Dengan Fitur Ekstraksi Local Binary Pattern Histogram	2019/2020
12	T3115398	Ridwan Nandipinta Dali	Prediksi Jumlah Pencari Kerja Menurut Jenis Kelamin dengan Metode Regresi Linear	2019/2020
13	T3114137	Idrak P. Dotu	Penerapan Data Mining Untuk Menentukan Potensi Hujan Harian Dengan Menggunakan Algoritma K-Nearest Neighbor (KNN)	2019/2020
14	T3112159	Sandri Due	Implementasi Data Mining Dalam Memprediksi Jumlah Pasien Menggunakan Metode Regresi Linier Sederhana	2019/2020
15	T3114267	Riston Ruhban	Prediksi Penggunaan Listrik Menggunakan Metode Linier Regresi Berganda	2019/2020
16	T3115149	Idris Haras	Klasifikasi Calon Penerima Sertifikat Redistribusi Tanah Pertanian Menggunakan Metode Naïve Bayes	2019/2020
17	T3114155	Moh. Abdullah Ishak	Klasifikasi Pegawai Berdasarkan Kinerja Dengan Menggunakan Algoritma Naïve Bayes	2019/2020

No	Nim	Nama_Mhs	Judul	Tahun
18	T3114249	Rizal Pontoiyo	Prediksi Anggaran Pendapatan dan Belanja Daerah Menggunakan Algoritma C.45	2019/2020
19	T3114227	Rama Iswandi Marada	Klasifikasi Penyakit Gastritis Menggunakan Metode Naïve Bayes	2019/2020
20	T3114253	Ajan	Prediksi Harga Ecer Ikan Kakap Merah Menggunakan Metode K-NN	2019/2020
21	T3115394	Husni Mubarak	Implementasi Metode Algoritma KNN Untuk Memprediksi Status Stunting Pada Balita	2019/2020
22	T3115146	Rendra Haswandi Puyo	Analisa dan Perancangan Sistem Keamanan Jaringan Internet Menggunakan Teknik Filtering (Studi kasus : Universitas Ichsan Gorontalo)	2019/2020
23	T3114090	Yolanda Djafar	Model K-Nearest Neighbor Yang efisien Untuk Prediksi Penyakit Jantung	2019/2020
24	T3115110	Paksa Gwijangge	Prediksi Tingkat Kelulusan Siswa SMK Negeri 4 Gorontalo Menggunakan Metode K-Nearest Neighbor	2019/2020
25	T3115005	Adrian Salawali	Aplikasi Data Mining Untuk Clustering Tingkat Kepadatan Penduduk di Kota Gorontalo Menggunakan Algoritma K-Means	2019/2020
26	T3115190	Rahmawati H. Kuntuamas	Penerapan Algoritma K-Means Clustering Untuk Data pelanggaran Lalu Lintas Kota Gorontalo	2019/2020

No	Nim	Nama_Mhs	Judul	Tahun
27	T3116021	Dana Pratama Jambo	Pengenalan Ekspresi Wajah Secara Real Time Menggunakan Klasifikasi K- NN (K-Nearest Neighbor)	2020/2021
28	T3116307	Ikbal Iskandar	Prototype Kandang Ayam cerdas Berbasis Mikrokontroller Arduino Uno	2020/2021
29	T3116123	Rifai Gobel	Prototype Deteksi Suhu Tubuh Menggunakan Sensor GY-906MLXESF Berbasis Internet Of Things	2020/2021
30	T3116002	Moh. Yasir	Prototype Pendeteksi Kebocoran Gas dan Api Menggunakan Mikrokotroller Arduino dan Modul Wifi berbasis Android	2020/2021
31	T3116058	Karim Kamali	Prediksi Jumlah Kasus Illegal Fishing Menggunakan Regresi Linear Berganda	2020/2021
32	T3116147	Sabrin Umonti	Analisis Interferensi Jaringan Wireless dan Kualitas Kinerja Hotspot Universitas Ichsan Gorontalo	2020/2021
33	T3116154	Rizki Malahedi	Game Simulasi Pembelajaran Budidaya Tanaman Jagung Berbasis Real Time Strategi	2020/2021
34	T3116101	Mohammad Rezky B. Pakaya	Pengenalan Tingkat Kematangan Buah Mangga Menggunakan Metode KNN	2020/2021
35	T3116006	Ramdani Karel	Deteksi Mutu Jambu Kristal Berdasarkan Fitur Warna dan Tekstur Menggunakan Metode GLCM	2020/2021
36	T3116072	Maman Kurniawan Abdul	Prediksi Jumlah Produksi Konveksi Jeans Dengan Metode Regresi Linier	2020/2021

No	Nim	Nama_Mhs	Judul	Tahun
37	T3115188	Sitti Nur Aisyah Lamama	Kamus Bahasa Mongondow- Indonesia Menggunakan Algoritma Algoritma Knuth Morris Pratt	2020/2021
38	T3114212	Ismail S. Manto	Implementasi Algoritma Brute Force Untuk Pencarian Judul Buku Berbasis Android	2020/2021
39	T3116052	Alfin Bimbing	Integrasi Color Moments Dan GLCM Untuk Identifikasi Penyakit Buah Kakao Menggunakan KNN	2020/2021
40	T3115252	Supriadi AS Suruh	Peramalan Produksi Tandan Buah Segar (TBS) Kelapa Sawit Menggunakan Metode Seasonal Arima	2020/2021
41	T3115004	Meri Ramadani Lasulika	Metode Regresi Linier Sederhana Untuk Prediksi Harga Komoditi Jagung	2020/2021
42	T3114270	Apris Umar	Penerapan Metode Profile Mathcing Pada Sistem Pendukung Keputusan Penilaian Pegawai Yang Akan Melanjutkan pendidikan Magister (Studi Kasus di Kantor BMKG Gorontalo)	2020/2021
43	T3114178	Vikran Djakatara	Penerapan Algoritma Naïve Bayes Untuk Memprediksi Kelayakan Pelanggan dalam Promosi Pemasaran Produk	2020/2021
44	T3114153	Muhamad Sukran Tuki	Prediksi Jumlah Produksi SIM Menggunakan Metode Least Square	2020/2021
45	T3115142	Alpin Adjim	Penerapan Algoritma K- Means Clustering Untuk Pengelompokan Potensi Produksi Hasil Pertanian di	2020/2021

No	Nim	Nama_Mhs	Judul	Tahun
			Kabupaten Bolaang Mongondow Selatan	
46	T3115167	Asdiranto A. Marisang	Implementasi Algoritma Turbo Boyer Moore Pada Aplikasi Kamus Bahasa Banggai Berbasis Android	2020/2021
47	T3116088	Paisal Aboya	Aplikasi Bank Darah Berbasis Mobile di PMI Kota Gorontalo	2020/2021
48	T3115111	Zulkifli Ibrahim	Mendeteksi Serangan DDOS (Distributed Denial Of Service) Menggunakan DDOS Deflate	2020/2021
49	T3115106	Zulfikar Nurkholis Yasin	Face Detection Berbasis Raspberry	2020/2021
50	T3116068	Iman Rizali	Prototype Pintu Portal Kompleks Perumahan Berbasis Mobile	2020/2021
51	T3116098	Imran Hamzah	Analisa & Perancangan Sistem Keamanan Jaringan Terhadap Serangan DHCP Rogue Pada Router Mikrotik	2020/2021
52	T3117074	Zulkifli Katili	Prediksi Penjualan Bright Gas 5,5 Kg Menggunakan Metode Least Square	2021/2022
53	T3117090	Siti Maryam Gaib	Analisis Efektivitas Pembelajaran Daring Google Classroom Menggunakan Metode Naive Bayes	2021/2022
54	T3117132	Riton Pajiu	Prediksi Penjualan Mie Basah Menggunakan Metode Least Square	2021/2022

No	Nim	Nama_Mhs	Judul	Tahun
55	T3117046	I Kadek Suardika	Prediksi Luas Serangan Hama Tanaman jagung Menggunakan Metode Fuzzy Time Series	2021/2022
56	T3116070	Sela Marsela Ismail	Identifikasi Kematangan Buah Pepaya Menggunakan Metode LPB dan K-NN	2021/2022
57	T3116029	Dimaswara Hendry Putra	Identifikasi Kematangan Buah Jambu Biji Dengan Image Processing Menggunakan Metode KNN dan Fitur Ekstraksi GLCM	2021/2022
58	T3116067	Mohamad Rizal Kasim	Pengenalan Kematangan Buah Nanas Menggunakan Metode Gray Level Co-Occurrence Matrix (GLCM) dan K-Nearest Neighbor	2021/2022
59	T3117075	Zulfikri Halid Salehe	Text Mining Untuk Membedakan Berita Hoax Menggunakan Naive Bayes	2021/2022
60	T3117039	Andre Suharto	Prediksi Tingkat Inflasi Menggunakan Metode Regresi Linear Berganda di Provinsi Gorontalo	2021/2022
61	T3116356	Tahir A. Mohune	Alat Pendeteksi Tingkat Kelayakan Daging Ayam RAS Berbasis Arduino Uno	2021/2022
62	T3117043	Sri Wahyuni Sahi	Prediksi Jumlah Pemakaian Air Bersih Menggunakan Algoritma Least Square	2021/2022
63	T3117081	Ermita	Prediksi Jumlah Pemakaian Bahan Baku Laundry Habis Pakai Menggunakan Metode Least Square	2021/2022

No	Nim	Nama_Mhs	Judul	Tahun
64	T3117092	Moh. Riskal S. Madilau	Penerapan Metode K-Nearest Neighbor Untuk Prediksi Penjualan Peralatan Depot Air terlaris	2021/2022
65	T3117080	Fadly Rizky Setiawan	Pengukuran Ketinggian Gelombang Laut Menggunakan Arduino Berbasis IOT dengan Platform ThingSpeak	2021/2022
66	T3117051	Iswan Harun	Prediksi Produksi Tahu Menggunakan Least Square	2021/2022
67	T3116019	Ican Fajriansyah	Perancangan Sistem Kontrol Kipas Angin Menggunakan Sensor Suara Berbasis Mikrokontroller Nodemcu	2021/2022
68	T3116093	Muhammad Alfaridzi Jafar	Rancang Bangun Lampu Taman Menggunakan Arduino Uno dan Sensor Gerak	2021/2022
69	T3115218	Sigit Ramadhan Muda	Pengenalan Alat Musik Tradisional Gorontalo (Polopalo) Berbasis Android	2021/2022
70	T3117142	Suprpto Daapala	Implementasi Metode LEAST SQUARE Untuk Prediksi Produksi Minyak Crude Palm Oil (CPO)	2021/2022
71	T3116037	Apriliyanto Abd. Rahman	Analisa dan Perbandingan Kinerja Protokol Video Streaming Menggunakan Media Wired dan Wireless di Jaringan Local	2021/2022
72	T3116139	Rapin Igrisa	Analisa Kebutuhan BandWidth dan Kualitas Kecepatan Jaringan WIFI UNISAN Pada Game Online	2021/2022

No	Nim	Nama_Mhs	Judul	Tahun
73	T3114147	Zainal J. Kuo	Sistem Pendukung Keputusan Kenaikan Jabatan Menggunakan Metode SAW	2021/2022
74	T3117135	Alfian Fazrul Sadu	prediksi persediaan oli yamalube dengan metode regresi linier sederhana pada PT. Hasjrat Abadi Gorontalo	2021/2022
75	T3116164	Dhea Ningrum S. Bilantu	Prediksi Persediaan Blangko Pada pembuatan Kartu Identitas Anak (KIA) di Kabupaten Gorontalo Menggunakan Metode Linear Regresi	2021/2022
....				
320	T3114230	Abdul Wahid Talib	Sistem Pendukung Keputusan Seleksi Penerimaan Beasiswa Menggunakan Metode Naive Bayes	2021/2022

No	Nama	Judul
1	A. Wantoro	Sistem Pendukung Keputusan Pemberian Kredit Mobil Dengan Fuzzy Tsukamoto (Studi Kasus: Pt Clipan Finance)
2	S.A Wardana	Algoritma Machine Learning untuk penentuan Model Prediksi Produksi Telur Ayam Petelur di Sumatera
3	G.Guntoro	Prediksi Jumlah Kendaraan di Provinsi Riau Menggunakan Metode Backpropagation
4	GN Ayuni	Penerapan Metode Regresi Linear Untuk Prediksi Penjualan Properti pada PT XYZ
5	IW Saputro	Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa
...	...	...
32	T Susim	Pengolahan Citra untuk Pengenalan Wajah (Face Recognition) Menggunakan OpenCV



DAFTAR RIWAYAT HIDUP

Nama : Sitti Meautiah Devi Amelia Jan
Tempat, Tanggal Lahir : Manado, 11 Juni 1998
Pekerjaan : Mahasiswa
Email : semuthstyles@gmail.com

Riwayat Pendidikan:

1. Tahun 2004, menyelesaikan Pendidikan di Taman Kanak-Kanak Shandy Putra Telkom.
2. Tahun 2010, menyelesaikan Pendidikan di Sekolah Dasar Negeri 30 Kota Gorontalo.
3. Tahun 2013, menyelesaikan Pendidikan di Sekolah Menengah Pertama Negeri 1 Gorontalo.
4. Tahun 2016, menyelesaikan Pendidikan di Sekolah Menengah Atas Negeri 3 Manado.
5. Tahun 2019, diterima menjadi mahasiswa di Perguruan Tinggi Swasta Universitas Ichsan Gorontalo.

Surat Penelitian



**KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS ICHSAN GORONTALO
LEMBAGA PENELITIAN**

Kampus Unisan Gorontalo Lt.3 - Jln. Achmad Nadjamuddin No. 17 Kota Gorontalo
Telp: (0435) 8724466, 829975 E-Mail: lembagapenelitian@unisan.ac.id

Nomor : 4319/PIP/LEMLIT-UNISAN/GTO/X/2022

Lampiran : -

Hal : Permohonan Izin Penelitian

Kepada Yth,

Dekan Fakultas Ilmu Komputer Universitas Ichsan Gorontalo

di,-

Tempat

Yang bertanda tangan di bawah ini :

Nama : Dr. Rahmisyari, ST.,SE.,MM

NIDN : 0929117202

Jabatan : Ketua Lembaga Penelitian

Meminta kesediannya untuk memberikan izin pengambilan data dalam rangka penyusunan **Proposal / Skripsi**, kepada :

Nama Mahasiswa : Sitti Meautiah Devi Amelia Jan

NIM : T3119082

Fakultas : Fakultas Ilmu Komputer

Program Studi : Teknik Informatika

Lokasi Penelitian : FAKULTAS ILMU KOMPUTER UNIVERSITAS ICHSAN GORONTALO

Judul Penelitian : PENERAPAN ALGORITMA COSINE SIMILARITY UNTUK DETEKSI TINGKAT KEMIRIPAN JUDUL TUGAS AKHIR MAHASISWA TEKNIK INFORMATIKA

Atas kebijakan dan kerja samanya diucapkan banyak terima kasih.

Gorontalo, 12 Oktober 2022
Ketua


Dr. Rahmisyari, ST.,SE.,MM
NIDN 0929117202

Surat Keterangan Penelitian



KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS ICHSAN GORONTALO
FAKULTAS ILMU KOMPUTER
SURAT KEPUTUSAN MENDIKNAS RI NOMOR 84/D/O/2001
Jl. Achmad Nadjamuddin No. 17 Telp (0435) 829975 Fax (0435) 829976 Gorontalo

SURAT KETERANGAN PENELITIAN

Nomor : 10/FIKOM-UIG/SKP/1/2023

Yang bertanda tangan dibawah ini :

Nama : Irvan Abraham Salih, M.Kom
NIDN : 0928028101
Jabatan : Dekan Fakultas Ilmu Komputer

Dengan ini Menerangkan bahwa :

Nama Mahasiswa : Sim Meautiah Devi Amelia Jan
NIM : T3119082
Program Studi : Teknik Informatika

Bahwa yang bersangkutan benar-benar telah melakukan penelitian tentang "Integrasi TF-IDF dan Algoritma Cosine Similarity Untuk Deteksi Tingkat Kemiripan Judul Penelitian Teknik Informatika (studi kasus Fakultas Ilmu Komputer Universitas Ichsan Gorontalo) " Guna untuk menyelesaikan Studi pada Program Studi Teknik Informatika Fakultas Ilmu Komputer, dan bersangkutan telah menyelesaikan penelitian tersebut pada TGL 10 Januari 2023 sesuai dengan waktu yang telah di tentukan.

Demikian Surat Keterangan ini dibuat dan digunakan untuk seperlunya.



Gorontalo, 10 Januari 2023

Dekan

Irvan Abraham Salih, M.Kom
NIDN. 0928028101

Surat Keterangan Bebas Plagiasi



KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS ICHSAN GORONTALO
FAKULTAS ILMU KOMPUTER
SURAT KEPUTUSAN MENDIKNAS RI NOMOR 84/D/O/2001
Jl. Achmad Najamuddin No. 17 Telp. (0435) 829975 Fax (0435) 829976 Gorontalo

SURAT REKOMENDASI BEBAS PLAGIASI
No. 119/FIKOM-UIG/R/IV/2023

Yang bertanda tangan di bawah ini :

Nama : Irvan Abraham Salihi, M.Kom
NIDN : 0928028101
Jabatan : Dekan Fakultas Ilmu Komputer

Dengan ini menerangkan bahwa :

Nama Mahasiswa : Sitti Meautiah devi Amelia Jan
NIM : T3119082
Program Studi : Teknik Informatika (S1)
Fakultas : Fakultas Ilmu Komputer
Judul Skripsi : Integrasi TF-IDF Dan Algoritma Cosine Similarity Untuk
Deteksi Tingkat Kemiripan Judul Penelitian Teknik
Informatika

Sesuai hasil pengecekan tingkat kemiripan skripsi melalui aplikasi **Turnitin** untuk judul skripsi di atas diperoleh hasil *Similarity* sebesar **9%**, berdasarkan Peraturan Rektor No. 32 Tahun 2019 tentang Pendeteksian Plagiat pada Setiap Karya Ilmiah di Lingkungan Universitas Ichsan Gorontalo dan persyaratan pemberian surat rekomendasi verifikasi calon wisudawan dari LLDIKTI Wil. XVI, bahwa batas kemiripan skripsi maksimal 30%, untuk itu skripsi tersebut di atas dinyatakan **BEBAS PLAGIASI** dan layak untuk diujikan.

Demikian surat rekomendasi ini dibuat untuk digunakan sebagaimana mestinya.

Mengetahui
Dekan,

Irvan Abraham Salihi, M.Kom
NIDN. 0928028101

Gorontalo, 3 April 2023
Tim Verifikasi,


Zulfrianto Y. Lamasigi, M.Kom
NIDN. 0914089101

Tersampir :
Hasil Pengecekan Turnitin

Surat Keterangan Bebas Pustaka



KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS ICHSAN GORONTALO
FAKULTAS ILMU KOMPUTER
UPT. PERPUSTAKAAN FAKULTAS
SK. MENDIKNAS RI NO. 84/D/0/2001

Jl. Achmad Nadjamuddin No.17 Telp(0435) 829975 Fax. (0435) 829976 Gorontalo

SURAT KETERANGAN BEBAS PUSTAKA

No : 012/Perpustakaan-Fikom/III/2023

Perpustakaan Fakultas Ilmu Komputer (FIKOM) Universitas Ichsan Gorontalo dengan ini menerangkan bahwa :

Nama Anggota : Sitti Meautiah D. A Jan
No. Induk : T3119082
No. Anggota : M202318

Terhitung mulai hari, tanggal : Senin, 13 Maret 2023, dinyatakan telah bebas pinjam buku dan koleksi perpustakaan lainnya.

Demikian keterangan ini di buat untuk di pergunakan sebagaimana mestinya.



Gorontalo, 13 Maret 2023

Mengetahui,
Kepala Perpustakaan

Apriyanto Alhamad, M.Kom

NIDN : 0924048601