e-*ISSN 2548-6438*
p-ISSN 2614-7173

PENERAPAN TEKS MINING DAN *COSINE SIMILARITY* UNTUK MENENTUKAN KESAMAAN DOKUMEN SKRIPSI

APPLICATION OF TEXT MINING AND COSINE SIMILARITY TO DETERMINE THE SIMILARITY OF THESIS DOCUMENTS

Asriyani Arsad, Mustamin Hamid, M Santosa
Fakultas Teknik, Program Studi Teknik Informatika
Universitas Muhammadiyah Maluku Utara
Email : asriyaniarsad@gmail.com

Abstrak

Digitalisasi dokumentasi-dokumen penelitian termasuk skripsi memudahkan para mahasiswa untuk mendapatkan sumber referensi sebagai bahan untuk melakukan penelitian atau pembuatan skripsi. Sumber referensi yang sangat banyak yang seharusnya menjadi bahan penelitian lebih lanjut kadangkala oleh sebagian mahasiswa yang malas hal ini bukan menjadi referensi akan tetapi sebagai bahan copy paste yang menyebabkan rusaknya tujuan dari skripsi yaitu mahasiswa bisa mengembangkan ilmu pengetahuan yang diperoleh selama proses perkuliahan. Praktek plagiarisme yang merugikan pihak lain dan sangat menuntungkan pihak lain harus di cegah, dibatasi bahkan bisa dihilangkan. Untuk mengatasi hal tersebut pada penelitian ini di implementasikan algoritma cosine similarty untuk mencari kemiripan dokument skripsi pada perpustakaan prodi Teknik Informatika Universitas Muhammadiyah Maluku Utara. Dari implementasi sistem yang dilakukan Pengukuran kesamaan menggunakan metode *Cosine Similarity*, dengan membandingkan 30 judul skripsi lama dan 5 judul skripsi baru. Hasil dari sistem ini menunjukkan nilai Cosine Similarity (ranger nilai cosine similarity adalah 0 sampai 1) terbesar pada setiap sampel judul skripsi baru, yaitu 0,1 (10%), 0,49 (49%), 0,4 (40%), 0,16 (16%).

Kata Kunci: Skripsi, Kesamaan, Cosine Similarity

Abstract

Digitization of research documents, including theses, makes it easier for students to obtain reference sources as material for conducting research or writing a thesis. There are so many reference sources that should be used as material for further research, sometimes by some lazy students this is not used as a reference but as copy-paste material which causes damage to the aim of the thesis, namely that students can develop the knowledge gained during the lecture process. The practice of plagiarism which is detrimental to other parties and greatly disadvantages other parties must be prevented, limited, and even eliminated. To overcome this, in this research, the cosine similarity algorithm was implemented to find similarities in thesis documents in the Informatics Engineering study program library, Muhammadiyah University, North Maluku. From the implementation of the system, similarity measurements were carried out using the Cosine Similarity method, by comparing 30 old thesis titles and 5 new thesis titles. The results of this system show the largest Cosine Similarity value (range value of cosine similarity is 0 to 1) for each sample of new thesis titles, namely 0.1 (10%), 0.49 (49%), 0.4 (40%), 0.16 (16%).

Keywords: Thesis, Similarity, Cosine Similarity

PENDAHULUAN

Skripsi adalah momen yang sangat meyakinkan untuk siswa tahun tertentu. Salah satu syarat untuk memperoleh gelar sarjana baik di Perguruan Tinggi Negeri (PTN) maupun Perguruan Tinggi Swasta (PTS) adalah mahasiswa harus menulis skripsi sesuai dengan pokok bahasan penelitiannya. (Khazari *et al*.,2017)

Tingkat kelulusan mahasiswa yang berdampak pada jumlah skripsi yang bertambah, dapat mengakibatkan kemungkinan munculnya kesamaan serta kemiripan pada setiap dokumen skripsi. Hal ini dapat menimbulkan tindakan plagiarisme. Oleh karena itu, harus di buat sistem kemiripan document skripsi agar mempermudah dalam proses pencarian kesamaan dokumen – dokumen skripsi tersebut.

Teknik Informatika Universitas Muhammadiyah Maluku Utara memiliki data dokumen - dokumen skripsi yang masih belum di buatkan sistem kemiripan dokumen, sehingga untuk menilai kesamaan antar dokumen – dokumen skripsi masih harus dilakukan pencarian secara manual. Hal ini tentu saja tidak selalu akurat dan dapat terjadi kesalahan, yang dapat menghambat mahasiswa dalam penyusunan skripsi dan kinerja dosen pembimbing menjadi tidak efisien dan memakan waktu

Berdasarkan penelitian yang dilakukan oleh Ardi Sanjaya dan Sempu Dwi, "Uji Kemiripan Kalimat Menggunakan Fungsi Terbilang Pada Pre-Processing Dan Cosine Similarity Dalam Bahasa Indonesia" dan di terbitkan pada tahun 2022. Eksplorasi dimulai dengan melihat dan memperhatikan upaya penanganan rutin untuk mengevaluasi kedekatan antar kalimat. Kalimat yang digunakan pada tahap ini adalah kalimat bahasa Indonesia dengan angka dan tanda baca yang menambahkan unsur yang lebih kompleks. 12 dari 13 data pengujian pada pengujian pertama dan kedua memperoleh hasil sebesar 92,30 persen.

Penelitian berikutnya oleh Daniel Oktodeli berjudul "Implementasi Natural Language Processing (NLP) dan Algoritma Cosine Similarity dalam Penilaian Ujian Esai Otomatis" tahun 2022. Penelitian ini menerapkan Natural Language Processing (NLP) dalam mengolah data jawaban mahasiswa. NLP untuk memproses teks dan memeriksa kesamaan dokumen. Pengecekan kemiripan dokumen digunakan dalam penelitian untuk mengetahui tingkat kemiripan antara kunci jawaban soal dengan data jawaban siswa. Algoritma yang digunakan untuk memeriksa kesamaan dokumen adalah Cosine Similarity. Hasil pengujian memiliki jawaban dengan tingkat kemiripan 90,58%

LANDASAN TEORI

Pengertian Skripsi

Skripsi adalah suatu karya ilmiah yang ditulis dalam bentuk presentasi tertulis hasil penelitian yang membahas suatu permasalahan faktual dengan menggunakan kaidah-kaidah ilmu pengetahuan yang berlaku pada jurusan yang sedang ditempuh. Skripsi merupakan suatu karya ilmiah yang harus memenuhi standar penulisan ilmiah seperti menggunakan bahasa yang baku dan efisien, mengutip kutipan, dan menarik kesimpulan berdasarkan penalaran yang logis. Skripsi juga merupakan laporan tentang sesuatu yang telah dilakukan (penelitian), yang merupakan suatu karya empiris. (Tim Penyusun dkk. 2014)

Pengertian *Text Mining*

Proses mendapatkan data yang berguna dari gudang database yang tepat disebut data mining. Selain itu, data mining dapat dipahami sebagai proses memperoleh informasi baru dari sejumlah besar data. bantuan dalam pengambilan keputusan. Penambangan data terkadang disebut sebagai knowledge discovery (Prasetyo, 2012)

Bagian dari penambangan data (*data mining*) adalah penambangan teks (*text mining*). Klasifikasi dokumen tekstual, dimana dokumen akan diklasifikasikan menurut topik dokumennya, biasanya merupakan penerapan fungsi text mining. Kata-kata dalam sebuah artikel dapat digunakan untuk menentukan jenis kategori artikel dengan bantuan text mining. Oleh karena itu, text mining dapat dengan cepat membantu dalam pengelompokan dokumen. (Ginting dkk. 2021)

Pengertian *Text Pre-processing*

Text preprocessing adalah proses mengubah data tekstual yang tidak terstruktur menjadi data terstruktur untuk disimpan dalam database (Langgeni dkk. 2010). Seperangkat indeks istilah yang dapat mewakili dokumen adalah tujuan dari p*reprocessing*. Ada beberapa bagian pada bagian preprocessing teks, antara lain:

1) Tokenisasi: Tokenisasi adalah proses pemisahan kata dari kumpulan data dan menentukan struktur masing-masing untuk memotong kalimat menjadi kata. Bagian yang dibuang dapat berupa angka, gambar, dan aksentuasi. Pemrosesan teks tidak terpengaruh oleh hal ini. (Fadli,2018)
2) Filtering: Pada tahap ini, kata-kata penting dihilangkan dari hasil tokenisasi. Hasil penguraian dokumen teks kemudian disaring untuk menghilangkan kata-kata yang "tidak relevan" dengan membandingkannya dengan stoplist, yang berisi kumpulan kata-kata "tidak relevan".
   Sistem pemisahnya dapat menggunakan perhitungan stoplist (menghilangkan kata-kata yang kurang penting) atau wordlist (menjaga kata-kata penting). Stoplist/stopwords adalah kata-kata tidak jelas yang dapat dibuang dengan pendekatan kumpulan kata. (Setiawan,dkk. 2016)
3) Stemming: Stemming adalah proses menghilangkan imbuhan pada suatu kalimat untuk menemukan bentuk kata dasarnya. Stemming adalah siklus yang mengubah kata-kata dalam laporan menjadi kata dasar dengan menggunakan prinsip-prinsip tertentu.
   Stemming dilakukan untuk menyamakan bentuk kata. Tujuan dari proses stemming adalah menghilangkan imbuhan baik yang berupa prefiks, sufiks maupun konfiks pada setiap kata. (Yuniar,dkk. 2022)

Pengertian *Term Weighting*

Setelah tahap preprocessing laporan yang menghasilkan bermacam-macam istilah atau kata, selanjutnya pada tahap tersebut dilakukan tahap pembobotan istilah yang nantinya akan diberikan bobot atau nilai dimana bobot tersebut menunjukkan pentingnya suatu istilah terhadap catatan. Menghitung bobot setiap istilah yang dicari dalam setiap laporan diharapkan dapat menentukan aksesibilitas dan kemiripan suatu istilah dalam catatan. (Pausta dkk. 2013) Semakin sering suatu istilah muncul dalam daftar koleksi, semakin tinggi nilai atau bobot istilah tersebut. Setelah tahap

pembobotan selesai, kita lanjutkan ke sistem pengelompokan. Sedangkan untuk pembobotan, teknik yang digunakan dalam pembobotan adalah strategi Tf-Idf. (Yudiarta dkk. 2018).

Pengertian *Cosine Similarity*

Metode Cosine Similarity merupakan suatu alat untuk menentukan derajat kemiripan antara dua hal. Ukuran kesamaan ruang vektor umumnya digunakan sebagai dasar perhitungan metode ini. Menggunakan kata kunci dari suatu dokumen sebagai ukuran, metode kesamaan kosinus ini menentukan seberapa mirip dua objek, seperti D1 dan D2, yang dinyatakan dalam dua vektor (Nurdiana, dkk. 2016)

Cosine Similarity merupakan perhitungan dalam text mining, kemampuan yang dimiliki untuk mengelompokkan suatu laporan atau teks. Ide di balik kesamaan kosinus adalah untuk menormalkan panjang vektor dengan membandingkan dokumen A dan B.

$$Cos\ \alpha = \frac{A \bullet B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

Keterangan :
A= Vektor A, yang akan dibandingkan kemiripannya
B = Vektor B, yang akan dibandingkan kemiripannya
A ● B = dot product antara vektor A dan vektor B
|A| = Panjang vektor A
|B| = Panjang vektor B
|A||B| = cross product antara |A| dan |B| (Riyani, dkk. 2019)

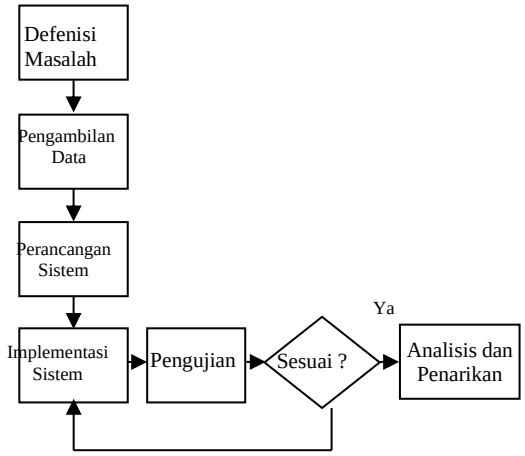
Pengertian Python

Python adalah bahasa pemrograman yang menekankan pada keterbacaan kode pada filosofi perancangannya. Python diakui sebagai bahasa yang menggabungkan kapasitas,dengan tanda baca kode yang sangat jelas, dan dilengkapi dengan kegunaan perpustakaan standar yang sangat besar dan lengkap. Selain itu, Python memiliki komunitas dukungan yang besar. (Syahrudin, dkk. 2018). Beberapa keunggulan bahasa Python adalah Python dapat digunakan di server untuk membuat aplikasi web, Python membaca dan mengubah dokumen, dapat digunakan untuk menangani informasi yang sangat besar dan melakukan sains yang kompleks.

Motivasi menggunakan dan mempelajari Python adalah karena Python dapat menangani berbagai tahapan (Windows, Macintosh, Linux, Raspberry Pi, dan sebagainya), Python memiliki tanda baca dasar seperti bahasa Inggris, Python memiliki struktur kalimat yang memungkinkan para insinyur untuk membuat program dengan baris yang lebih sedikit. Daripada beberapa dialek pemrograman lainnya, Python berjalan pada kerangka penerjemah, kode penting dapat berupa dieksekusi saat dibuat. Ini menyiratkan bahwa pembuatan prototipe bisa sangat cepat, Python dapat ditangani dengan cara prosedural, cara item terletak atau cara praktis, dan Python memiliki banyak perpustakaan (Alfian. 2020)

METODE PENELITIAN

Metode penelitian yang digunakan dalam sistem ini terdiri dari beberapa tahap, yaitu definisi masalah khusus untuk menentukan penelitian yang akan dieksplorasi, pengumpulan informasi dan rencana kerangka pelaksanaan sehingga siklus selanjutnya adalah investigasi dan pengambilan keputusan.



Gambar 1. Kerangka metode penelitian (Arsad Asriyani,2024)

Identifikasi Masalah

Pada program studi teknik informatika Universitas Muhammadiyah Maluku Utara tahap ini untuk mengkaji permasalahan yang akan diangkat oleh sistem, menentukan hal-hal penting sebagai dasar penyelesaian masalah melalui analisis kebutuhan, serta merancang dan mengimplementasikan sistem kesamaan judul skripsi.

Tahap defenisi masalah yaitu dilakukan analisis dan melihat documen skripsi program studi teknik informatika semakin banyak, sehinga akan mempersulit dosen pembimbing dan mahasiswa yang menghadapi tugas akhir (skripsi) terhadap kemiripan judul skripsi

Pengambilan Data

Data judul skripsi yang digunakan dalam penelitian ini adalah arsip dokumen skripsi program studi teknik informatika yang diperoleh dari perpustakan program studi teknik informatika, berjumlah 35 judul skripsi dimana 30 data judul skripsi yang di input dalam database dan 5 judul yang dijadikan judul baru untuk pengujian sistem kemiripan menggunakan metode cosine similarty

Kebutuhan Perangkat Keras

a. Laptop Lenovo
b. Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz 2.60GHz
c. RAM 8 GB
d. Keyboard
e. Monitor

Kebutuhan Perangkat Lunak

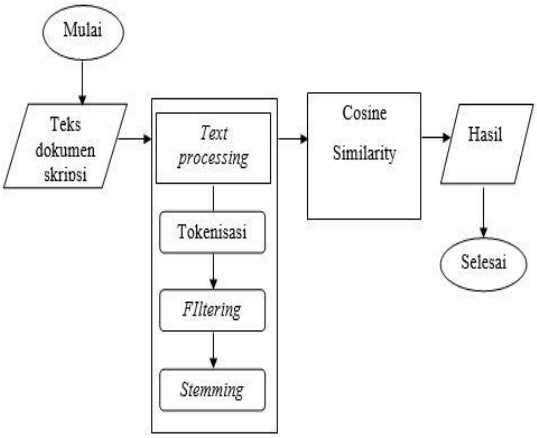
a. Sistem Operasi menggunakan Windows 10
b. Bahasa program *Python*
c. Browser

Sistem Yang Berjalan

Pada dasarnya telah diterapkan sistem untuk pengelolaan dokumen-dokumen skripsi yang ada di perpustakaan Prodi Teknik Informatika ini, namun sistem yang ada masih manual menggunakan aplikasi *Microsoft Excel*, dan hanya sebatas pendataan berdasarkan judul dari dokumen skripsi yang ada. Hal ini tentu saja belum sepenuhnya merangkum isi dari suatu dokumen skripsi

Sistem Yang Diusulkan

Kemiripan dilakukan dengan metode cosine similarty. Sebelum mencari kemiripan, dokumen terlebih dahulu di input, data yang di ambil hanya bagian judul dari dokumen skripsi. Selanjutnya data tersebut akan diproses hingga mendapatkan hasil akhir berupa kumpulan data kata yang dapat mewakili masing-masing dokumen skripsi. Hasil dari proses tersebut akan di cari kemiripan menggunakan metode *cosine similarty*.



Gambar 2. Sistem yang diusulkan (Arsad Asriyani,2024)

Pada gambar menjelaskan dari awal bagian input data teks dokumen skripsi hingga hasil akhir berupa kumpulan data. Teks dokumen skripsi yang dimasukkan ke dalam sistem ini akan melalui *Text Preprocessing* yang memiliki tahapan-tahapan seperti Tokenisasi, *Filtering* dan *Stemming.* Selanjutnya dilakukan pencarian dengan metode *cosine similarty*

IMPLEMENTASI SISTEM

Halaman Sistem

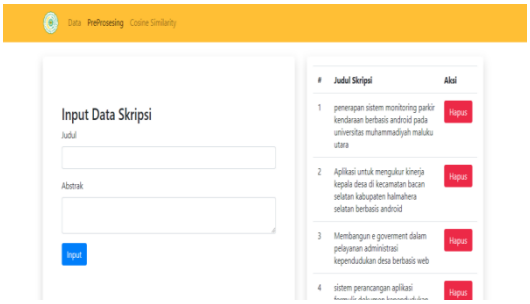
1. Tampilan Halaman Beranda

Halaman beranda adalah tampilan halaman saat sistem pertama kali dibuka

Gambar 3. halaman beranda

2. Tampilan Halaman Data

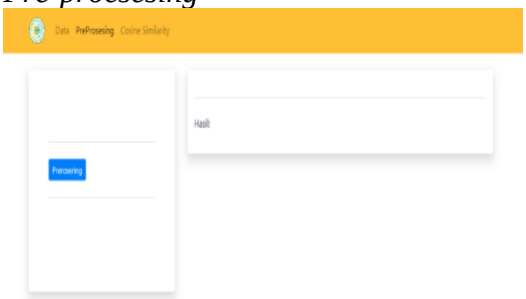
Tampilan data terdapat form inputan judul skripsi dan menampilkan judl skripsi yang telah diinput



Gambar 4. Halaman data

3. Tampilan Halaman *Pre-processing*

Halaman *Preprocessing* adalah tampilan yang nantinya melakukan preprosesing data judul skripsi dan menampilkan hasil *Pre-procsesing*



Gambar 5. Halaman Pre-processing

4. Tampilan Halaman Pengecekan Kemiripan

Pada halaman ini dilakukan proses pengecekan kemiripan judul skripsi lama dalam database dan judul skrpsi baru yang nantinya akan diinput. Pada halaman ini juga ditampilkan hasil kemiripan judul skripsi

Gambar 6. Halaman pengecekan kemiripan judul

Data yang Digunakan

Data yang digunakan adalah data judul skripsi program studi teknik informatika universitas muhammadiyah

maluku utara pada tahun 2018 – 2020 yang berjumlah 35 judul dimana 30 judul di input dalam database sistem dan 5 judul dijadikan judul ujicoba kemiripan menggunakan metode cosine similarty.

Tabel 1. Data judul skripsi

NO	JUDUL SKRIPSI
1	peningkatkan layanan internet di kantor loka montor spektrum frekuensi radio ternate dengan perbandingan metode queue burst dan token bucket
2	sistem pakar penyakit mata dengan menggunakan metode forward chaining berbasis android
3	penerapan sistem monitoring parkir kendaraan berbasis android pada universitas muhammadiyah maluku utara
4	Aplikasi untuk mengukur kinerja kepala desa di kecamatan bacan selatan kabupaten halmahera selatan berbasis android
5	Membangun e-goverment untuk administrasi kependudukan desa berbasis web
6	sistem perancangan aplikasi formulir dokumen kependudukan dan pencatatan sipil berbasis android
7	Recovery dan analisis barang bukti digital yang dihapus pada usb flash drive
8	media bembelajaran huruf abjad braille untuk siswa tunanetral slb negri sasa tingkat sdlh berbasis mikrokontroler
9	aplikasi penangan dini bencana kebakaran dengan rute gps photo tagging berbasis android
10	sistem pengolahan data biro camaru universitas muhammadiyah maluku utara
11	sistem monitoring absensi pegawai berbasis website terintegrasi finger print
12	sistem informasi bahan laboratorium karantina tumbuhan berbasis android pada balai karantina pertanian kelas 2 ternate
13	aplikasi penilaian laporan pendidikan kurikulum 2013 berbasis website pada sd islamiyah 4 kota ternate
14	sistem informasi sekolah berbasis website dengan fitur e lerning dan vidio converence pada sma negri tidore kepulauan
15	sistem informasi jadwal guru berbasis android pada madrasa aliyah negri kota ternate
16	sistem pendukun keputusan penerimaan bonus karyawan pada toko dengan menggunakan metode simple assitive meighting
17	desain dan pengembangan aplikasi elektronik loan eloan berbasis android untuk pengujian kredit pada pt bank tabungan negara persero tbk kantor cabang ternate
18	membangun sistem lelang online produk perkebunan di desa berbasis web studi kasus desa amasing kali kabupaten halmahera selatan
19	mengenal dan menebak surat pendek dalam Al-qur'an berbasis android
20	sistem informasi pendataan guru honorer tingkat Sd dsn Smp berbasis android pada kantor dinas pendidikan kota ternate kepulauan
21	membangun aplikasi permintaan data cuaca meteorologi berbasis android, informatika, universitas muhammadiyah maluku utara
22	Sistem pakar mendiagnosa penyakit vertigo dengan metode backward chaining berbasis web
23	sistem pendukung keputusan menentukan pekerjaan bagi alumni teknik informatika dengan metode weighted
24	sistem pendukung keputusan kelayakan pemberi pinjaman menggunakan metode fuzzy logic pada pt mitra dana top finance
25	sistem pendukung keputusan penentuan tingkat kelulusan mahasiswa prodi teknik informatika universitas muhammadiyah maluku utara dengan metode fuzzy mamdani
26	Membangun website promosi penjualan ikan tuna loin pada pt.ud raul
27	klasifikasi pengaduan mayarakat manggunakan algoritma TF_ODF dan cosine similarty
28	aplikasi mobile notivikasi promosi ukm di ummu berbasis android
29	penentuan jenis malaria dengan menggunakan metode forward chaining dan naive bayes berbasis mobile
30	penerapan data mining menggunakan algoritma naive bayes untuk melakukan klasifikasi kelulusan pada mahasiswa tingkat akhir

Text Pre-processing

Filtering/Stopward

a. Teks judul sebelum filtering/ stopword

Tabel 2. Data sebelum filtering

NO	Sebelum Filtering/stopword
1	Recovery ~~dan~~ analisis barang bukti digital ~~yang~~ dihapus ~~pada~~ usb flash drive
2	sistem pendukun keputusan penerimaan bonus karyawan ~~pada~~ toko ~~dengan~~ menggunakan metode simple assitive meighting
3	desain ~~dan~~ pengembangan aplikasi elektronik loan eloan berbasis android ~~untuk~~ pengujian kredit ~~pada~~ pt bank tabungan negara persero tbk kantor cabang ternate
4	membangun sistem lelang online produk perkebunan ~~di~~ desa berbasis web studi kasus desa amasing kali kabupaten halmahera selatan
5	sistem informasi pendataan guru honorer tingkat Sd dsn Smp berbasis android ~~pada~~ kantor dinas pendidikan kota ternate kepulauan
6	sistem pendukung keputusan menentukan pekerjaan ~~bagi~~ alumni teknik informatika ~~dengan~~ metode weighted

b. Teks judul setelah filtering/stopword

Tabel 3. Hasil filtering/stopword

NO	Sesudah Filtering/stopword
1	Recovery analisis barang bukti digital dihapus usb flash drive
2	sistem pendukun keputusan penerimaan bonus karyawan toko menggunakan metode simple assitive meighting
3	desain pengembangan aplikasi elektronik loan eloan berbasis android pengujian kredit pt bank tabungan negara persero tbk kantor cabang ternate
4	membangun sistem lelang online produk perkebunan desa berbasis web studi kasus desa amasing kali kabupaten halmahera selatan

5	sistem informasi pendataan guru honorer tingkat Sd dsn Smp berbasis android kantor dinas pendidikan kota ternate kepulauan
6	sistem pendukung keputusan menentukan pekerjaan alumni teknik informatika metode weighted

Tokenisasi

Tahapan ini dilakukan pemisahan per-kata, yang nantinya dihitung berapa jumlah kata yang keluar dari setiap kata

Tabel 4. Hasil tokenisasi

Hasil Tokenisasi
'alumni', 'amasing', 'analisis', 'android', 'aplikasi', 'assitive', 'bank', 'barang', 'berbasis', 'bonus', 'bukti', 'cabang', 'desa', 'desain', 'digital', 'dihapus', 'dinas', 'drive', 'dsn', 'elektronik', 'eloan', 'flash', 'guru', 'halmahera', 'honorer', 'informasi', 'informatika', 'kabupaten', 'kali', 'kantor', 'karyawan', 'kasus', 'kepulauan', 'keputusan', 'kota', 'kredit', 'lelang', 'loan', 'meighting', 'membangun', 'menentukan', 'menggunakan', 'metode', 'negara', 'online', 'pekerjaan', 'pendataan', 'pendidikan', 'pendukun', 'pendukung', 'penerimaan', 'pengembangan', 'pengujian', 'perkebunan', 'persero', 'produk', 'pt', 'recovery', 'sd', 'selatan', 'simple', 'sistem', 'smp', 'studi', 'tabungan', 'tbk', 'teknik', 'ternate', 'tingkat', 'toko', 'usb', 'web', 'weighted'

Stemming

Tahap ini dari text hasil tokenisasi dilakukan proses perubahan bentuk kata menjadi kata dasar

Tabel 5. Hasil stemming

Hasil Stemming
'alumni', 'amasing', 'analisis', 'android', 'aplikasi', 'assitive', 'bank', 'barang', 'bas', 'bonus', 'bukti', 'cabang', 'desa', 'desain', 'digital', 'hapus', 'dinas', 'drive', 'dsn', 'elektronik', 'elo', 'flash', 'guru', 'halmahera', 'honorer', 'informasi', 'informatika', 'kabupaten', 'kali', 'kantor', 'karyawan', 'kasus', 'pulau', 'putus', 'kota', 'kredit', 'lelang', 'loan', 'meighting', 'bangun', 'tentu', 'guna', 'metode', 'negara', 'online', 'kerja', 'data', 'didik', 'dukun', 'dukung', 'terima', 'kembang', 'uji', 'kebun', 'persero', 'produk', 'pt', 'recovery', 'sd', 'selatan', 'simple', 'sistem', 'smp', 'studi', 'tabung', 'tbk', 'teknik', 'ternate', 'tingkat', 'toko', 'usb', 'web', 'weighted'

Term Weighting
TF-IDF

Setelah dilakukan tahapan filtering/stopword, tokenizasi, stemming. tahap selanjutnya dilakukan pembobotan kata menggunakan metode TF-IDF

Tabel 6. Hasil proses TF-IDF

Hasil Pembobotan Tf-Idf
0. 0. 0.33333333 0. 0. 0. 0. 0.33333333 0. 0. 0.33333333 0. 0. 0. 0.33333333 0.33333333 0. 0.33333333 0. 0. 0. 0.33333333 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. 0.33333333 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.33333333 0. 0.
0. 0. 0. 0. 0. 0.3057546 0. 0. 0. 0.3057546 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.3057546 0. 0. 0.25072314 0. 0. 0. 0.0.3057546 0. 0. 0.3057546 0.25072314 0. 0. 0. 0. 0. 0.3057546 0. 0.3057546 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.3057546 0.18139171 0. 0. 0. 0. 0.0. 0. 0.3057546 0. 0. 0.
0. 0.23681493 0. 0. 0. 0. 0. 0. 0.16394991 0. 0. 0. 0.47362985 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.23681493 0. 0. 0. 0.23681493 0.23681493 0. 0. 0.23681493 0. 0. 0. 0. 0.23681493 0. 0. 0.23681493 0. 0. 0. 0. 0.23681493 0. 0. 0. 0. 0. 0. 0. 0. 0.23681493 0. 0.23681493 0. 0. 0. 0.23681493 0. 0.14049262 0. 0.23681493 0. 0. 0. 0. 0. 0. 0. 0.23681493 0.
0. 0. 0. 0.212804 0. 0. 0. 0. 0.17966376 0. 0. 0. 0. 0. 0. 0. 0.25951255 0.0.25951255 0. 0. 0. 0.25951255 0.0.25951255 0.25951255 0. 0. 0. 0.212804 0. 0. 0.25951255 0. 0.25951255 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.25951255 0.25951255 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.25951255 0. 0. 0.15395819 0.25951255 0. 0. 0. 0. 0.212804 0.25951255 0. 0. 0. 0.
0.33909405 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. 0.33909405 0. 0. 0. 0. 0. 0. 0.27806198 0. 0. 0. 0. 0.0. 0.33909405 0. 0.27806198 0. 0. 0.33909405 0. 0. 0. 0.33909405 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.20117065 0 0. 0. 0. 0.33909405 0. 0. 0. 0. 0. 0.33909405

HASIL DANPEMBAHASAN

1. Judul skripsi yang di uji

Terdapat 5 judul skripsi yang digunakan dalam pengujian pengecekan kemiripan, yaitu dapat dilihat pada tabel berikut

Tabel 7. Judul skripsi yang diuji

	Judul skripsi yang di uji
1	analisis user experience terhadap website progrezcloud dengan metode usability testing
2	sistem pakar diagnosa penyakit tanaman cengkeh dengan menggunakan metode forward shaing berbasis web
3	rancang bangun sistem informasi audit mutu internal universitas muhammadiyah maluku utara berbasis web
4	klasifikasi pengaduan masyarakat menggunakan algoritma term frequency inverse document frequency dan cosine similarty (studi kasus BMKG ternate)
5	sistem monitoring status gizi untuk menegah stunting pada bayi berbasis android pada

	puskesmas kelurahan tomalou

2. Hasil Pengecekan kemiripan 5 judul ujicoba terhadap judul lain dalam database sistem, dapat dilihat pada tabel dibawah ini

Keterangan :
q1 = analisis user experience terhadap website progrezcloud dengan metode usability testing
q2= sistem pakar diagnosa penyakit tanaman cengkeh dengan menggunakan metode forward shaing berbasis web
q3= rancang bangun sistem informasi audit mutu internal universitas muhammadiyah maluku utara berbasis web
q4= klasifikasi pengaduan masyarakat menggunakan algoritma term frequency inverse document frequency dan cosine similarty (studi kasus BMKG ternate)
q5= sistem monitoring status gizi untuk menegah stunting pada bayi berbasis android pada puskesmas kelurahan tomalou

Tabel 8. Hasil kemiripan

JUDUL SKRIPSI DALAM DATABASE	Hasil Pengecekan Kemiripan Judul Skripsi				
	q1	q2	q3	q4	q5
peningkatkan layanan internet di kantor loka montor spektrum frekuensi radio ternate dengan perbandingan metode queue burst dan token bucket	0,03	0,03	0,0	0,03	0,0
sistem pakar penyakit mata dengan menggunakan metode forward chaining berbasis android	0,05	0,5	0,05	0,04	0,08
penerapan sistem monitoring parkir kendaraan berbasis android pada universitas muhammadiyah maluku utara	0,0	0,0	0,0	0,0	0,16
Aplikasi untuk mengukur kinerja kepala desa di kecamatan bacan selatan kabupaten halmahera selatan berbasis android	0,0	0,01	0,0	0,0	0,0
Membangun e-goverment untuk administrasi kependudukan desa berbasis web	0,0	0,1	0,17	0,0	0,02
sistem perancangan aplikasi formulir dokumen kependudukan dan pencatatan sipil berbasis android	0,0	0,05	0,04	0,0	0,07
Recovery dan analisis barang bukti digital yang dihapus pada usb flash drive	0,1			0,0	
media bembelajaran huruf abjad braille untuk siswa tunanetral slb negri sasa tingkat sdlh berbasis mikrokontroler	0,0	0,01	0,01	0,0	0,01
aplikasi penangan dini bencana kebakaran dengan rute gps photo tagging berbasis android	0,0	0,02	0,02	0,0	0,04
sistem pengolahan data biro camaru universitas muhammadiyah maluku utara	0,0	0,03	0,34	0,0	0,03
sistem monitoring absensi pegawai berbasis website terintegrasi finger print	0,09	0,05	0,04	0,0	0,13
sistem informasi bahan laboratorium karantina tumbuhan berbasis android pada balai karantina pertanian kelas 2 ternate	0,0	0,04	0,09	0,03	0,06
aplikasi penilaian laporan pendidikan kurikulum 2013 berbasis website pada sd islamiyah 4 kota ternate	0,07	0,02	0,02	0,03	0,01
sistem informasi sekolah berbasis website dengan fitur e lerning dan vidio converence pada sma negri tidore kepulauan	0,07	0,04	0,09	0,0	0,03
sistem informasi jadwal guru berbasis android pada madrasa aliyah negri kota ternate	0,0	0,05	0,11	0,04	0,07
sistem pendukun keputusan penerimaan bonus karyawan pada toko dengan menggunakan metode simple assitive meighting	0,0	0,1	0,02	0,03	0,02
mengenal dan menebak surat pendek dalam Al-qur'an berbasis android	0,0	0,02	0,02	0,0	0,05
sistem informasi pendataan guru	0,0	0,04	0,09	0,03	0,06

honorer tingkat Sd dsn Smp berbasis android pada kantor dinas pendidikan kota ternate kepulauan					
membangun aplikasi permintaan data cuaca meteorologi berbasis android, informatika, universitas muhammadiyah maluku utara	0,0	0,02	0,35	0,0	0,05
sistem pendukung keputusan menentukan pekerjaan bagi alumni teknik informatika dengan metode weighted	0,0	0,07	0,03	0,0	0,02
sistem pendukung keputusan kelayakan pemberi pinjaman menggunakan metode fuzzy logic pada pt mitra dana top finance	0,0	0,09	0,02	0,03	0,02
Membangun website promosi penjualan ikan tuna loin pada pt.ud raul	0,0	0,08	0,07	0,0	0,01
klasifikasi pengaduan mayarakat manggunakan algoritma TF_IDF dan cosine similarty	0,0	0,05	0,0	0,4	0,0
aplikasi mobile notivikasi promosi ukm di ummu berbasis android	0,0	0,02	0,02	0,0	0,05
penentuan jenis malaria dengan menggunakan metode forward chaining dan naive bayes berbasis mobile	0,0	0,2	0,02	0,03	0,02
penerapan data mining menggunakan algoritma naive bayes untuk melakukan klasifikasi kelulusan pada mahasiswa tingkat akhir	0,0	0,0	0,0	0,14	0,0
Hasil Kemiripan (nilai terbesar)	0,1	0,5	0,35	0,4	0,16

KESIMPULAN

Tahapan pemrosesan di mulai dengan *text preprocessing* yaitu tokenisasi, *filtering*, dan *stemming*. Selanjutnya tahapan pembobotan kata atau *term weighting*, setelah setiap kata memilliki bobot maka dilakukan proses pengecekan dengan metode *cosine similarity*.

Dari hasil pengujian yang telah dilakukan terhadap 5 judul skripsi yaitu pada judul pertama kemiripanya 1%, judul kedua kemiripanya 49%, judul ketiga kemiripanya 36%, judul keempat kemiripanya 11% dan judul kelima kemiripanya 16%, ditarik kesimpulan bahwa sistem yang dibangun menggunakan metode *cosine similarity* dapat melakukan pengecekan kemiripan judul skripsi untuk menghindari plagiarisme terhadap pengajuan judul skripsi baru program studi teknik informatika universitas muhammadiyah maluku utara.

DAFTAR PUSTAKA

Sihombing, D. O. (2022). *Implementasi Natural Language Processing (NLP) dan Algoritma Cosine Similarity dalam Penilaian Ujian Esai Otomatis*. Jurnal Sistem Komputer dan Informatika (JSON), 4(2), 396-406.

E. Prasetyo, *Data Mining –Konsep Dan Aplikasi Manggunakan Maltab*. Yogyakarta : ANDI , 2012.

Fadlil, A. (2018). *Aplikasi Sitem Temu Kembali Angket Mahasiswa Menggunakan Application of Information Retrieval for Opinion Student.* Jurnal Teknologi Informasi Dan Ilmu Komputer, 6(1), 33-40. https://doi.org/10.25126/jtiik.201961184

Khazari, A. S., Marisa, F., & Wijaya, I. D. (2017). *Sistem Rekomendasi Penentuan Judul Skripsi Menggunakan Algoritma Decision*

Tree. Jurnal Teknologi dan Manajemen Informatika, 3(1)

Langgeni, Baizal & Firdaus.. 2010. *Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection. Yogyakarta : Seminar Nasional Informatika*

Maarif, A. L. F. I. A. N. (2020). Buku Ajar *Pemrograman Lanjut Bahasa Pemrograman Python.* Universitas Ahmad Dahlan Yogtakarta.

Naf'an, Muhammad Zidny; Burhanuddin, Auliya; Riyani, Ade. *Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen.* Jurnal Linguistik Komputasional, 2019, 2.1: 23-27.

Nurdiana, O., Jumadi, J., & Nursantika, D. (2016). *Perbandingan metode Cosine Similarity dengan metode Jaccard Similarity pada aplikasi pencarian terjemah Al-Qurâ€™ an dalam Bahasa Indonesia*. Jurnal Online Informatika, 1(1), 59-63.

Pausta Yugianus, Harry Soekotjo Dachlan, dan Rini Nur Hasanah, 2013 "*Pengembangan Sistem Penelusuran Katalog Perpustakaan Dengan Metode Rocchio Relevance Feedback*", EECCIS Vol. 7, No. 1, Juni 2013

PENYUSUN, Tim, et al. *Pedoman penulisan skripsi. Universitas Negeri Surabaya*, 2014.

Sanjaya, A., & Sasongko, S. D. (2022). *Uji Kemiripan Kalimat Menggunakan Fungsi Terbilang Pada Pre-Processing Dan Cosine Similarity Dalam Bahasa Indonesia*. J. Ilm. Nero, 7(2), 95-104.

Sari, H., Ginting, G. L., Zebua, T., & Mesran, M. (2021). *Penerapan Algoritma Text Mining dan TF-IDF Untuk Pengelompokan Topik Skripsi Pada Aplikasi Repository STMIK Budi Darma.* TIN: Terapan Informatika Nusantara, 2(7), 414-432.

Setiawan, A., Astuti, I. F., & Kridalaksana, A. H. (2016). *Klasifikasi dan pencarian buku referensi akademik menggunakan metode naïve bayes classifier (nbc)(studi kasus: perpustakaan daerah provinsi kalimantan timur)*. Informatika Mulawarman: Jurnal Ilmiah Ilmu Komputer, 10(1), 1-10.

Syahrudin, A. N., & Kurniawan, T. (2018). *Input dan output pada bahasa pemrograman python*. Jurnal Dasar Pemograman Python STMIK, 20, 1-7.

Yuniar, E., Utsalinah, D. S., & Wahyuningsih, D. (2022). *Implementasi Scrapping Data Untuk Sentiment Analysis Pengguna Dompet Digital dengan Menggunakan Algoritma Machine Learning*. Jurnal Janitra Informatika dan Sistem Informasi, 2(1), 35-42.