SKRIPSI

PEMBOBOTAN TF-IDF DAN METODE COSINE SIMILARITY UNTUK DETEKSI KEMIRIPAN DALAM PENGAJUAN TOPIK SKRIPSI DI UTDI

SATRIA DWI HARTANTO

NIM : 195410229

PROGRAM STUDI INFORMATIKA

PROGRAM SARJANA

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS TEKNOLOGI DIGITAL INDONESIA

YOGYAKARTA

2023

SKRIPSI

PEMBOBOTAN TF-IDF DAN METODE COSINE SIMILARITY UNTUK DETEKSI KEMIRIPAN DALAM PENGAJUAN TOPIK SKRIPSI DI UTDI

Diajukan sebagai salah satu syarat untuk menyelesaikan studi

Program Sarjana

Program Studi Informatika

Fakultas Teknologi Informasi

Universitas Teknologi Digital Indonesia

Yogyakarta

Disusun Oleh

SATRIA DWI HARTANTO

NIM : 195410229

PROGRAM STUDI INFORMATIKA

PROGRAM SARJANA

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS TEKNOLOGI DIGITAL INDONESIA

YOGYAKARTA

2023

PERNYATAAN KEASLIAN SKRIPSI

Dengan ini saya menyatakan bahwa naskah skripsi ini belum pernah diajukan untuk memperoleh gelar Sarjana Komputer di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara sah diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 17 Juli 2023

Satria Dwi Hartanto

NIM : 195410229

HALAMAN PERSEMBAHAN

Karya ilmiah ini saya persembahkan kepada :

Bapak Supartono S.Pd., dan Ibu Sulistyaningsih beserta seluruh keluarga yang senantiasa mendo'akan saya.

Teman-teman dan orang terdekat saya yang selalu memberi nasihat dan semangat agar saya diberi kesehatan, kemudahan, dan kelancaran serta menyelesaikan karya tulis ini.

MOTTO

“Kenangan terbaik di dunia adalah sholat.” - Gush Baha

“Allah mempunyai cara tak terbatas, dalam menolong hambaNya” – Mbah Nun

KATA PENGANTAR

Puji syukur kehadirat Allah SWT yang telah memberikan rahmat-Nya, sehingga saya dapat menyelesaikan karya tulis ini. Dalam penulisan ini saya mendapat dukungan dan bantuan dari berbagai pihak, maka pada kesempatan kali ini saya ucapkan banyak terima kasih kepada :

1. Bapak Ir. Totok Suprawoto, M.M., M.T, selaku kepala Universitas Teknologi Digital Indonesia Yogyakarta.
2. Bapak Pius Dian Anggoro, S, Si., M.Cs. yang sudah membimbing saya dalam membuat karya tulis ini.
3. Bapak dan Ibu dosen UTDI yang sudah mengajar dan memberikan ilmunya baik di dalam maupun di luar kampus.
4. Orang tua dan keluarga yang telah mendoakan dan memberikan dukungan baik secara moral maupun materi.
5. Teman dan orang terdekat yang selalu mendo'akan kebaikan untuk saya.

Disadari bahwa dalam penyusunan karya tulis ini masih terdapat kekurangan. Oleh karena itu, kritik dan saran yang membangun dari segala pihak sangat dibutuhkan. Semoga karya tulis ini bermanfaat dan dapat memberi inspirasi terhadap pembaca.

Yogyakarta, ...…

Satria Dwi Hartanto

DAFTAR ISI

Hal

HALAMAN COVER i
HALAMAN JUDUL ii
HALAMAN PERSETUJUAN iii
HALAMAN PENGESAHAN iv
PERNYATAAN KEASLIAN SKRIPSI v
HALAMAN PERSEMBAHAN vi
MOTTO vii
KATA PENGANTAR viii
DAFTAR ISI ix
DAFTAR GAMBAR xii
DAFTAR TABEL xiv
DAFTAR LAMPIRAN xv
INTISARI xvi
ABSTRACT xvii
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 2
1.3 Ruang Lingkup 3
1.4 Tujuan Penelitian 4
1.5 Manfaat Penelitian 4
1.6 Sistematika Penulisan 5
BAB II LANDASAN TEORI 7
2.1 Tinjauan Pustaka 7
2.2 Text Mining 12
2.2.1 Text Prepocessing 13
2.2.2 Processing 14
2.3 TF-IDF 14
2.4 Cosine Similarity 16
BAB III METODE PENELITIAN 18

3.1 Metode Pengumpulan Data 18
3.1.1 Metode Observasi 18
3.1.2 Metode Wawancara 18
3.1.3 Kepustakaan 19
3.2 Analisis Kebutuhan 19
3.2.1 Analisis Kebutuhan Input 19
3.2.2 Analisis Kebutuhan Proses 20
3.2.3 Analisis Kebutuhan Output 20
3.2.4 Kebutuhan Hardware 21
3.2.5 Kebutuhan Software 21
3.3 Perancangan Sistem 21
3.3.1 Perancangan Flowchart Sistem 21
3.3.2 Perancangan DFD *(*Data *Flow Diagram)* 23
3.3.3 Perancangan Basis Data 24
3.4 Perancangan Interface 28
3.4.1 Halaman Data Pengujian 28
3.4.2 Halaman Data Pengujian Hasil 28
3.4.3 Halaman Pengujian Perbandingan 29
BAB IV IMPLEMANTASI DAN PEMBAHASAN 30
4.1 Implementasi Sistem 30
4.2 Pembahasan Basis Data 34
4.2.1 Tabel Pengguna 34
4.2.2 Tabel Master 35
4.2.3 Tabel Master Stemming 35
4.2.4 Tabel Pengujian 36
4.2.5 Tabel Pengujian Hasil 37
4.3 Pembuatan Kode Program 38
4.3.1 Koneksi Basis Data 38
4.3.2 Import Data 39
4.3.3 Ambil Dataset 40
4.3.4 Text Pre Processing 41

4.3.5 Stemming Data 42
4.3.6 Proses TF-IDF 43
4.3.7 Cosine Similarity 44
4.3.8 Menampilkan Hasil Perbandingan 45
4.4 Pembahasan Antarmuka Program 46
4.4.1 Halaman Login 46
4.4.2 Halaman Home 46
4.4.3 Halaman Menu Pengguna 47
4.4.3 Halaman Menu Master Data Skripsi 49
4.4.4 Halaman Menu Pengujian 51
4.5 Pengujian Metode 56
4.5.1 Data Pengujian 56
4.5.2 Pengujian Perbandingan Perhitungan 59
4.5.3 Pengujian Hasil Kemiripan 60
4.5.4 Pengujian Perbandingan Stemming dan Non Stemming 62
BAB V KESIMPULAN 64
5.1 Kesimpulan 64
5.2 Saran 65
DAFTAR PUSTAKA 66
LAMPIRAN 67

DAFTAR GAMBAR

Hal

Gambar 3.1 Flowchart Sistem 22
Gambar 3.2 Diagram Konteks 23
Gambar 3.3 DFD Level 1 23
Gambar 3.4 Rancangan Relasi Antartabel 24
Gambar 3.5 Halaman Data Pengujian 28
Gambar 3.6 Halaman Data Pengujian Hasil 29
Gambar 3.7 Halaman Pengujian Perbandingan 29
Gambar 4.1 Spesifikasi Web Server 30
Gambar 4.2 Database Web Server 31
Gambar 4.3 Struktur Folder Project 32
Gambar 4.4 Struktur Data Vendor 33
Gambar 4.5 Isi File Library 33
Gambar 4.6 Implementasi Basis Data 34
Gambar 4.7 Tampilan Tabel Pengguna 35
Gambar 4.8 Tampilan Tabel Master 35
Gambar 4.9 Tampilan Tabel Master Stemming 36
Gambar 4.10 Tampilan Tabel Pengujian 37
Gambar 4.11 Tampilan Tabel Pengujian Hasil 38
Gambar 4.12 Tampilan Kode Program Koneksi Basis Data 39
Gambar 4.13 Tampilan Kode Program Import Data 40
Gambar 4.14 Tampilan Kode Program Ambil Dataset 41
Gambar 4.15 Tampilan Kode Program Text Pre Processing 42
Gambar 4.16 Tampilan Kode Program Stemming Data 42
Gambar 4.17 Tampilan Kode Program Proses TF-IDF 43
Gambar 4.18 Tampilan Kode Program Cosine Similarity 44
Gambar 4.19 Tampilan Kode Program Menampilkan Hasil Perbandingan 45
Gambar 4.20 Tampilan Halaman Login 46
Gambar 4.21 Tampilan Halaman Dashboard 47
Gambar 4.22 Tampilan Halaman Tabel Data Pengguna 48

Gambar 4.23 Tampilan Halaman Tambah Pengguna 48
Gambar 4.24 Tampilan Halaman Tabel Data Master Skripsi 49
Gambar 4.25 Tampilan Halaman Import Data 50
Gambar 4.26 Tampilan Halaman Olah Data 50
Gambar 4.27 Tampilan Halaman Detail Data Skripsi 51
Gambar 4.28 Tampilan Halaman Input Data 52
Gambar 4.29 Tampilan Halaman Hasil Detail Deteksi Admin 53
Gambar 4.30 Tampilan Halaman Hasil Detail Deteksi Mahasiswa 54
Gambar 4.31 Tampilan Halaman Tabel Hasil Kemiripan 55
Gambar 4.32 Tampilan Halaman Tabel Hasil Perbandingan 56
Gambar 4.33 Tampilan Tabel Hasil Perhitungan Manual 60
Gambar 4. 34 Tampilan Tabel Hasil Perhitungan Sistem 60
Gambar 4.35 Tampilan Tabel Hasil Pengujian Kemiripan 61
Gambar 4.36 Tampilan Tabel Hasil Pengujian Stemming dan Non Stemming ... 62

DAFTAR TABEL

Hal

Tabel 2.1 Kajian Penelitian Terdahulu..... 9
Tabel 3.1 Tabel Pengguna..... 25
Tabel 3.2 Tabel Master 25
Tabel 3.3 Tabel Master Stemming..... 26
Tabel 3.4 Tabel Pengujian..... 26
Tabel 3.5 Tabel Pengujian Hasil 27
Tabel 4.1 Dataset Pengujian..... 57

DAFTAR LAMPIRAN

Hal

Lampiran 1 Perhitungan Manual 67
Lampiran 2 Kriteria Kelulusan Ujian Pendadaran 75
Lampiran 3 Catatan Pendadaran 76
Lampiran 4 Keputusan Hasil Ujian Pendadaran 76
Lampiran 5 Bimbingan Via Email 77

INTISARI

Dalam perkembangan teknologi sudah mempermudah mahasiswa dalam proses pencarian tema atau judul skripsi. Hal ini bisa mempunyai kelebihan dan kekurangan masing - masing. Selain itu hal ini dapat dijadikan referensi yang baik dengan banyaknya sumber referensi, sisi lainnya dapat memicu tingkat kemiripan judul yang beragam yang memicu plagiat yang tinggi sehingga menyebabkan kurang beragamnya topik atau judul penelitian.

Proses pendeteksian kemiripan judul yang dilakukan oleh Universitas Teknologi Digital Indonesia masih sebatas membandingkan judul skripsi yang diajukan oleh mahasiswa oleh dosen pembimbing masing - masing. Hal ini akan berdampak kurang lengkapnya sumber informasi judul yang sudah di ajukan dalam basis data yang tersedia di Universitas Teknologi Digital Indonesia. Penelitian ini dilakukan untuk menguji tingkat kemiripan judul proposal skripsi yang akan diajukan dengan dibandingkan dengan judul yang sudah ada di Universitas Teknologi Digital Indonesia.

Tahapan penelitian yang digunakan yaitu *preprocessing* (terdiri dari case folding, tokenizing, filtering, dan stemming), perhitungan pembobotan TF-IDF, dan perhitungan nilai kemiripan menggunakan cosine similarity. Penelitian ini diimplementasikan dalam bentuk web dengan bahasa pemrograman PHP dan basis data Mysql dan tambahan library sastrawi untuk proses preprocesing.

Penelitian menunjukkan bahwa stemming mampu menghilangkan sebagian besar kesalahaan yang disebabkan oleh kata imbuhan: (*ber, me, pe, ter, an, kan, di*). Skenario penelitian dengan *stemming* menghasilkan nilai kemiripan rata-rata lebih tinggi 1.104% daripada tanpa *stemming*. Berdasarkan hasil percobaan cosine similarity dan pembobotan TF-IDF mampu menghasilkan nilai kemiripan dari masing-masing teks pembanding.

Kata kunci : *Cosine Similarity, Mysql, Preprocessing, PHP, Stemming, TF-IDF.*

ABSTRACT

The advancement of technology has significantly facilitated students in the process of searching for thesis themes or titles. While this has numerous advantages, it also comes with its own set of drawbacks. On one hand, it provides a wealth of reference sources, enabling students to explore diverse topics. On the other hand, it can lead to a higher incidence of plagiarism due to the proliferation of similar titles, resulting in a lack of diversity in research topics.

The University of Indonesia's Digital Technology department currently employs a title comparison approach, wherein submitted thesis titles are compared manually by respective faculty advisors. However, this method may lead to incomplete title information in the available university database. Therefore, this research aims to assess the similarity level between proposed thesis titles and existing ones in the University of Indonesia's database.

The research methodology involves several stages, including preprocessing (comprising case folding, tokenizing, filtering, and stemming), TF-IDF weighting, and computing similarity values using cosine similarity. The implementation is carried out as a web-based application using PHP programming language and a MySQL database, with additional support from the Sastrawi library for preprocessing tasks.

The study demonstrates that stemming effectively removes a significant portion of errors caused by affixes (e.g., ber, me, pe, ter, an, kan, di).The research scenarios demonstrate that stemming yields an average similarity score 1.104% higher than without stemming. Based on the cosine similarity experiments and TF-IDF weighting, the study successfully generates similarity scores for each comparative text.

Keywords : *Cosine Similarity, MySQL, Preprocessing, PHP, Stemming, TF-IDF.*