 **DEPARTEMEN INFORMATIKA**

**FAKULTAS TEKNOLOGI INFORMASI DAN KOMUNIKASI**

**INSTITUT TEKNOLOGI SEPULUH NOPEMBER**

**USULAN TUGAS AKHIR**

# IDENTITAS PENGUSUL

**NAMA : Rozana Firdausi**

**NRP : 05111540007001**

**DOSEN WALI : Dr. Agus Zainal Arifin, S.Kom., M.Kom**

**DOSEN PEMBIMBING : 1. Rizky Januar Akbar, S.Kom., M.Eng.  
 2. Anny Yuniarti, S.Kom., M.Comp.Sc**

# JUDUL TUGAS AKHIR

“Rancang Bangun Modul Rekomendasi Dokumen Penelitian Berbasis Web Menggunakan Algoritma *Cosine Similarity*”

# LATAR BELAKANG

SI Kepegawaian ITS (Simpeg) merupakan sistem terintegrasi ITS berbasis web yang dapat diakses secara *online* dengan koneksi internet. Simpeg dapat diakses oleh dosen dan pegawai di ITS. Pada Simpeg terdapat beberapa fitur yang dapat diakses oleh dosen seperti memasukkan dokumen penelitian, memasukkan data diri, melihat capaian kinerja, dan sebagainya. Dokumen penelitian dapat berupa karya ilmiah, jurnal, artikel seminar, proposal penelitian, laporan penelitian, buku, paten, karya seni, dan poster.

Beberapa layanan yang memungkinkan pengguna melakukan pencarian dokumen penelitian yaitu *Google* *Scholar*, *Scopus, DOAJ (Directory of Open Access Journals)* dan *Thompson Reuters*. Sedangkan ITS memiliki aplikasi sejenis yaitu Silacak yang dapat membaca sumber data dari *Google Scholar*, *Scopus, DOAJ,* dan *Thompson Reuters*. Seluruh dokumen penelitian yang dimiliki civitas ITS tersimpan dalam database Silacak.

Penelitian jarang dilakukan sendiri, melainkan dilakukan secara bersama-sama. Satu dokumen penelitian berafiliasi dengan banyak penulis. Simpeg mewajibkan dosen untuk menginputkan dokumen penelitiannya ke Simpeg, tetapi terdapat kasus dimana dosen menginputkan dokumen penelitian tanpa menambahkan anggota peneliti/penulis yang lain. Hal ini menyebabkan duplikasi dokumen penelitian sehingga dibutuhkan suatu modul yang dapat menampilkan rekomendasi untuk mengurangi duplikasi dokumen penelitian.

Sebelumnya terdapat modul rekomendasi yang dibuat namun fokus utamanya berbeda. Modul tersebut merekomendasikan kategori dari dokumen penelitian. Modul tersebut juga belum terintegrasi dengan Simpeg sehingga dibutuhkan modul baru yang dapat mengintegrasikan dokumen penelitian dengan database yang ada yaitu Silacak dan Simpeg. [1]

Maka dari itu, modul rekomendasi dokumen penelitian hadir sebagai solusi agar tidak terjadi duplikasi dokumen penelitian. Modul rekomendasi ini terintegrasi dengan database Silacak dan Simpeg, sehingga modul akan mengecek database Simpeg dan Silacak ketika dibutuhkan. Ketika pengguna menginputkan judul penelitian yang belum pernah ada pada database, modul akan memasukkan dokumen penelitian ke dalam database Simpeg. Begitupun sebaliknya, apabila judul penelitian sudah pernah ada dalam database, modul akan menampilkan rekomendasi dokumen penelitian berdasarkan database Simpeg dan Silacak.

# RUMUSAN MASALAH

Perumusan masalah yang terdapat pada tugas akhir ini, antara lain adalah:

* + - 1. Bagaimana membaca sumber data dokumen penelitian dari masukan dosen, Silacak dan Simpeg?
      2. Bagaimana mendeteksi kemiripan dokumen penelitian menggunakan metode *Cosine Similarity*?
      3. Bagaimana menyajikan rekomendasi dokumen-dokumen penelitian kepada pengguna?

# BATASAN MASALAH

Batasan masalah yang terdapat pada tugas akhir ini, sebagai berikut:

1. Aplikasi pendeteksi kemiripan ini digunakan untuk mendeteksi kemiripan dokumen melalui perangkat web.
2. Aplikasi ini hanya dapat digunakan untuk dokumen teks yang tidak meliputi dokumen yang berisi gambar.
3. Sumber data yang digunakan dalam aplikasi ini adalah dokumen penelitian dari basis data kepegawaian ITS, Silacak.
4. Dokumen yang digunakan untuk perbandingan teks adalah dokumen berbahasa Indonesia dan berbahasa Inggris.
5. *Corpus* yang digunakan sebagai input dalam dokumen adalah judul, penulis, afiliasi dan abstrak.

# TUJUAN PEMBUATAN TUGAS AKHIR

Tujuan dari pembuatan tugas akhir ini adalah:

* + - 1. Membaca sumber data dokumen penelitian dari masukan dosen, Silacak dan Simpeg.
      2. Mendeteksi kemiripan dokumen penelitian menggunakan metode *Cosine Similarity.*
      3. Menyajikan kemiripan antara dokumen-dokumen penelitian kepada pengguna.

# MANFAAT TUGAS AKHIR

Tugas akhir ini diharapkan dapat menjadi modul yang memberikan rekomendasi kepada pengguna untuk menghindari redundansi pada database penelitian. Tugas akhir ini akan menerapkan algoritma *Cosine Similarity*, sehingga tingkat kemiripan suatu dokumen dapat diketahui.

# TINJAUAN PUSTAKA

Dalam mengerjakan tugas akhir ini, terdapat beberapa tinjauan pustaka yang digunakan, yaitu:

1. *Information Retrieval*

*Information Retrieval* pada dasarnya merupakan proses untuk menentukan dokumen dalam koleksi yang harus ditemubalikkan untuk memenuhi keinginan pengguna akan informasi. Suatu sistem temu balik informasi dikatakan ideal jika sistem tersebut dapat menemukan seluruh dokumen yang relevan dan sistem hanya menemukan dokumen yang relevan saja. [2]

1. Algoritma Nazief dan Adriani

Algoritma Nazief & Adriani adalah algoritma untuk mengubah kata berimbuhan menjadi kata dasar dalam Bahasa Indonesia. Kata-kata yang akan dibandingkan diubah terlebih dahulu menjadi kata dasar sehingga memungkinkan hasil perhitungan Cosine Similarity menjadi lebih optimal. [3]

1. Algoritma Porter

Algoritma Porter stemming (atau 'Porter stemmer') adalah proses untuk menghilangkan akhir morfologi dan infleksional yang lazim dari kata-kata dalam bahasa Inggris. Penggunaan utamanya adalah sebagai bagian dari proses normalisasi jangka yang biasanya dilakukan saat mengatur sistem Pengambilan Informasi. [4]

1. *Stopword* Bahasa Indonesia dan Bahasa Inggris

*Stopword* adalah kata umum yang dianggap tidak memiliki makna. Contoh stopword seperti kata ‘ke’, ‘dengan’, ‘jika’ dan lain sebagainya. *Stopword* yang dihilangkan baik yang bahasa Indonesia maupun bahasa Inggris. [5]

1. *Similarity*

Kemiripan dokumen (*document similarity*) dapat digunakan sebagai alat pencarian informasi lain yang sejenis, sehingga dapat mempersingkat waktu. Kemampuan pencarian kemiripan dokumen biasanya diimplementasikan pada sebuah artikel berita dan jurnal. [6]

1. *Cosine Similarity*

*Cosine Similarity* digunakan untuk melakukan perhitungan kesamaan dari dokumen. Terdapat beberapa algoritma lain yang bertujuan sama seperti *Jaccard Similarity* dan *Dice Similarity*. Namun berdasarkan percobaan yang telah dilakukan pada pengujian perbandingan metode *Jaccard* dan *Cosine Similarity* pada aplikasi pencarian terjemah Al-Qur’an dalam Bahasa Indonesia, didapat metode *Cosine* yang nilai kemiripannya tertinggi dari metode *Jaccard* dan *K-Nearest Neighbor* (*KNN*). Hal ini karena metode *Cosine Similarity* mempunyai konsep normalisasi panjang vektordata dengan membandingkan *N-gram* yang sejajar satu sama lain dari 2 pembanding. Sedangkan pada metode *Jaccard* hanya membandingkan isi *N-gram* dengan eksak dan hanya melihat apakah ada suatu *N-gram* tertentu pada pembanding tanpa melihat posisi penulisan yang berbeda. Pada *euclidean distance* yang diterapkan di metode *K-Nearest Neighbor* (*KNN*) tidak mempunyai konsep normalisasi panjang vektor data, sehingga nilai akurasi metode dipengaruhi oleh panjang 2 data pembanding dan harus menentukan nilai dari parameter K (jumlah dari tetangga terdekat). Maka metode *Cosine Similarity* menjadi usulan alternatif untuk mencari kemiripan dari teks *mining*. [7]

# RINGKASAN ISI TUGAS AKHIR

Tugas akhir ini bertujuan untuk membuat aplikasi pendeteksi kemiripan dokumen untuk dosen ITS yang akan mempermudah dosen ITS dalam memasukkan dokumen penelitian yang sudah dibuat. Adapun tahapan yang dilakukan dalam pembuatan aplikasi ini adalah sebagai berikut :

* + - 1. Akuisisi data

Data penelitian diambil dari 2 database yaitu database Silacak dan Simpeg. Database Silacak mengambil dokumen penelitian yang tersedia di *Google* *Scholar*, *Scopus, DOAJ* dan *Thompson Reuters*. Untuk mengakses ketersediaan dokumen penelitian pada 2 database diatas dibutuhkan API (*Application Programming Language*).

* + - 1. *Preprocessing*

Fungsi *preprocessing* pada program ini adalah untuk mendapatkan *keyword* yang akan digunakan sebagai perbandingan dokumen.

* + - * 1. *Case Folding*

*Case folding* adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap *delimiter*. [8]

* + - * 1. *Tokenizing*

*Tokenizing* adalah proses menghilangkan tanda baca pada kalimat sehingga terbentuk kata-kata yang berdiri sendiri.

* + - * 1. *Stopword Removal*

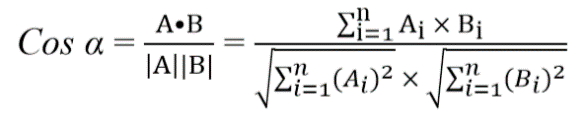
Tahap *stopword removal* adalah tahap pengambilan kata yang penting dari hasil *tokenizing.* Tahap ini menghilangkan kata-kata yang terdapat pada daftar *stopword* Bahasa Inggris dan Bahasa Indonesia.

* + - * 1. *Stemming*

*Stemming* adalah proses mengubah kata menjadi kata dasarnya dengan menghilangkan imbuhan-imbuhan pada kata dalam dokumen. [9] Dalam tahap *stemming* diperlukan suatu algoritma. Algoritma *stemming* untuk Bahasa Indonesia yaitu Nazief dan Adriani. Sedangkan untuk Bahasa Inggris yaitu Porter.

* + - 1. Penerapan Metode *Cosine Similarity*

Cosine Similarity digunakan untuk melakukan perhitungan kesamaan dari dokumen. Rumus yang digunakan oleh Cosine Similarity adalah

 [10]

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|

* + - 1. Menyajikan rekomendasi

Terdapat dua metode dalam menyajikan rekomendasi untuk dosen yang memasukkan dokumen penelitian, diantaranya:

* + - * 1. Aktif

Modul melakukan cek ke database Simpeg dan Silacak. Apabila sudah ada dokumen penelitian dalam database, maka dokumen tersebut direkomendasikan kepada pengguna.



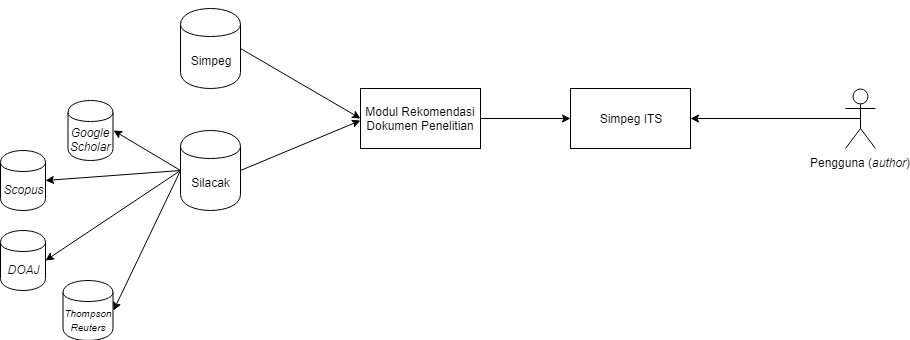
* + - * 1. Pasif

Modul melakukan cek ke database Simpeg dan Silacak. Apabila belum ada dokumen penelitian dalam database, maka pengguna menginputkan *metadata* dokumen penelitiannya ke Simpeg. *Metadata* terdiri dari judul, penulis, afiliasi dan abstrak.

A screenshot of a social media post

Description generated with very high confidence

Diagram arsitektur yang digunakan dalam modul berikut adalah sebagai berikut:



# METODOLOGI

## Penyusunan proposal tugas akhir

Tahap pertama dalam proses pengerjaan tugas akhir ini adalah menyusun proposal tugas akhir. Pada proposal tugas akhir ini diajukan aplikasi pendeteksi kemiripanuntuk dosen ITS Berbasis Web.

## Studi literatur

Pada tahap ini, akan dicari studi literature yang relevan untuk dijadikan referensi dalam pengerjaan tugas akhir. Studi literatur ini didapatkan dari paper yang didapat dari *Google Scholar*. Sumber lain yaitu dokumentasi resmi GitHub untuk mengimplementasikan *text preprocessing* dan metode *Cosine Similarity.* Dan materi pendukung yang berasal dari materi-materi kuliah yang berhubungan dengan metode yang akan digunakan.

## Analisis dan desain perangkat lunak

Aplikasi pendeteksi kemiripan dokumen untuk dosen ITS merupakan aplikasi yang berjalan pada web dengan dukungan aktivitas pengguna dalam kondisi terkoneksi ke internet.

## Implementasi perangkat lunak

Pembangunan aplikasi pendeteksi kemiripan dokumen untuk dosen ITS akan dilakukan dengan menggunakan bahasa pemrograman PHP.

* 1. **Pengujian dan evaluasi**

Pengujian dilakukan kepada beberapa dosen ITS untuk mengetahui tingkat keberhasilan aplikasi sudah berjalan dengan baik.

## Penyusunan buku tugas akhir

Pada tahap ini dilakukan penyusunan laporan yang menjelaskan dasar teori dan metode yang digunakan dalam tugas akhir ini. Pada tahap ini juga disertakan hasil dari implementasi metode dan algoritma yang telah dibuat. Sistematika penulisan buku tugas akhir ini secara garis besar antara lain:

1. Pendahuluan
   1. Latar Belakang
   2. Rumusan Masalah
   3. Batasan Tugas Akhir
   4. Tujuan
   5. Metodologi
   6. Sistematika Penulisan
2. Tinjauan Pustaka
3. Desain dan Implementasi
4. Pengujian dan Evaluasi
5. Kesimpulan dan Saran
6. Daftar Pustaka

# JADWAL KEGIATAN

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Tahapan | 2018 | | | | | | |
| Mei | Juni | Juli | Agustus | September | Oktober | Nopember |
| Penyusunan Proposal |  |  |  |  |  |  |  |
| Studi Literatur |  |  |  |  |  |  |  |
| Perancangan Sistem |  |  |  |  |  |  |  |
| Implementasi |  |  |  |  |  |  |  |
| Pengujian dan Evaluasi |  |  |  |  |  |  |  |
| Penyusunan Buku |  |  |  |  |  |  |  |

# DAFTAR PUSTAKA

|  |  |
| --- | --- |
| [1] | A. Z. Syahran, . D. O. S and A. Saikhu, "Perancangan dan Pembangunan Modul Rekomendasi Section pada Open Journal System (OJS)," *JURNAL TEKNIK POMITS,* 2012. |
| [2] | D. Yogatama, "STUDI PENGGUNAAN STEMMING UNTUK MENINGKATKAN PERFORMANSI SISTEM TEMU BALIK INFORMASI," 10 September 2017. |
| [3] | A. Firdaus, E. Ernawati and V. Arie, "APLIKASI PENDETEKSI KEMIRIPAN PADA DOKUMEN TEKS MENGGUNAKAN ALGORITMA NAZIEF & ADRIANI DAN METODE COSINE SIMILARITY," *UNIB Scholar Repository,* 12 Desember 2013. |
| [4] | M. Porter, "The Porter Stemming Algorithm," [Online]. Available: https://tartarus.org/martin/PorterStemmer/index-old.html. [Accessed 1 Juni 2018]. |
| [5] | G. R. Dantes, . I. . K. R. Arthana and I. . K. Purnamawan, "SISTEM PENDETEKSI PAKAR BERDASARKAN KOLEKSI DOKUMEN ILMIAH UNDIKSHA," p. 3, 2016. |
| [6] | Sugiyamta, "SISTEM DETEKSI KEMIRIPAN DOKUMEN DENGAN ALGORITMA COSINE SIMILARITY DAN SINGLE PASS CLUSTERING," *Dinamika Informatika,* p. 1, 2015. |
| [7] | O. Nurdiana, J. and D. Nursantika, "PERBANDINGAN METODE COSINE SIMILARITY DENGAN METODE JACCARD SIMILARITY PADA APLIKASI PENCARIAN TERJEMAH AL-QUR’AN DALAM BAHASA INDONESIA," *Dinamika Informatika,* p. 4, 2015. |
| [8] | Y. Wibisono and M. L. Khodra, "Clustering Berita Berbahasa Indonesia," *KK Informatika Sekolah Teknik Elektro dan Informatika I T B, jln. Ganesa 10 Bandung,* p. 2, 2005. |
| [9] | G. Karyono and F. S. Utomo, "TEMU BALIK INFORMASI PADA DOKUMEN TEKS BERBAHASA INDONESIA DENGAN METODE VECTOR SPACE RETRIEVAL MODEL," 23 Juni 2012. |
| [10] | R. T. Wahyuni, D. Prastiyanto and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," p. 20, 2017. |