JUDUL

“Perancangan Metode Hybrid untuk Deteksi Similarity Index Dokumen Multi-Format Menggunakan Modul Ekstraksi, Tokenisasi, dan Pendekatan Dual Source”

“Perancangan Metode Hybrid untuk Deteksi Similarity Index Dokumen Multi-Format Menggunakan Modul Ekstraksi, Tokenisasi, dan Pendekatan Dual Source Berbasis Website”

Konsep Sistem Deteksi Plagiarisme

Mengimplementasikan sistem dasar untuk mendeteksi plagiarisme pada dokumen (PDF dan DOCX) dengan membandingkannya terhadap dua sumber utama:

1. **Dokumen Lokal:** Sekumpulan dokumen (dataset) yang sudah diindeks dan disimpan secara lokal.
2. **Sumber Internet:** Menggunakan Google Search untuk memeriksa keberadaan kalimat tertentu secara online.

Alur Kerja Utama:

1. **Persiapan Lingkungan:** Mengatur koneksi Google Drive dan membuat struktur folder yang diperlukan untuk aplikasi, database, dan laporan.
2. **Ekstraksi Dokumen:** Mengimplementasikan fungsi untuk membaca dan mengekstrak teks dari file PDF dan DOCX. Teks dibersihkan dari karakter non-ASCII. Dokumen PDF dikonversi sementara ke DOCX untuk proses selanjutnya.
3. **Pengindeksan Dokumen Lokal:** Memindai folder dataset, mengekstrak teks dari setiap dokumen yang didukung (PDF/DOCX), dan menyimpan teks serta metadata (nama file, path, ukuran, tanggal indeks, dll.) ke dalam folder database dan file database.csv. Setiap dokumen diindeks ke dalam subfolder berdasarkan tipe file-nya (misalnya, database/pdf, database/docx).
4. **Tokenisasi Kalimat:** Dokumen yang akan diperiksa dipecah menjadi kalimat-kalimat individual menggunakan NLTK.
5. **Similarity Checking (Pengecekan Kemiripan):** Untuk setiap kalimat dari dokumen yang diperiksa:
    - **Prioritas Google Search:** Kalimat dicek terlebih dahulu di Google Search (exact phrase dan broader search). Jika ditemukan, kalimat tersebut ditandai sebagai terindikasi plagiarisme dari sumber internet.
    - **Pengecekan Lokal:** Jika kalimat tidak ditemukan di Google, barulah dilakukan pengecekan terhadap semua dokumen lokal yang sudah diindeks. Metode yang digunakan adalah:

- **Exact Match:** Memeriksa apakah kalimat ada persis di dokumen lokal.
- **Word Overlap:** Menghitung persentase kata yang sama antara kalimat dan dokumen lokal jika tidak ada exact match. Kalimat ditandai terindikasi plagiarisme lokal jika skor kemiripan melebihi threshold tertentu (default 75%).

6. **Agregasi Hasil:** Hasil pengecekan dari semua kalimat diagregasi untuk menghitung persentase kemiripan secara keseluruhan (Internet vs Lokal) dan per sumber spesifik.
7. **Highlighting Dokumen:** Dokumen asli yang diperiksa di-highlight pada bagian kalimat yang terindikasi plagiarisme. Setiap sumber (lokal atau internet) diberi warna highlight yang unik untuk memudahkan identifikasi.
8. **Pembuatan Laporan:** Laporan deteksi plagiarisme dibuat dalam format DOCX. Laporan ini berisi ringkasan persentase kemiripan (Internet dan Lokal), breakdown kemiripan per sumber (dengan indikator warna highlight), dan daftar kalimat-kalimat yang terindikasi plagiarisme per sumber. Laporan ini kemudian ditambahkan ke akhir dokumen yang sudah di-highlight.
9. **Cleanup:** File-file temporary yang dibuat selama proses (misalnya, konversi PDF ke DOCX) dihapus. Laporan akhir diunduh untuk pengguna.

Sistem ini memprioritaskan pengecekan online (Google Search) untuk efisiensi, dan baru beralih ke database lokal jika tidak ada kecocokan di internet. Hasil disajikan dalam laporan baca dengan visualisasi highlight.

METODE:
Prototpe / Rapid Application Development (RAD)

HASIL DETEKSI PLAGIARISME (SISINDO)

Tanggal Pemeriksaan: 26-11-2025 12:53

Kalimat Terdeteksi	Sumber	Skor