LAPORAN

IMPLEMENTASI SISTEM TEMU KEMBALI INFORMASI MENGGUNAKAN BOOLEAN MODEL DAN VECTOR SPACE MODEL PADA KORPUS PARFUM HMNS

Disusun oleh:

Alrijal Nur Ilham

A12.2022.14113

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS DIAN NUSWANTORO

SEMARANG

2025

DAFTAR ISI

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Sistem Temu Kembali Informasi (STKI) merupakan salah satu cabang penting dalam ilmu komputer yang mempelajari bagaimana suatu sistem mampu menemukan, menyeleksi, dan menyajikan informasi yang relevan dari sekumpulan dokumen berdasarkan kata kunci atau kebutuhan informasi pengguna. Dalam era digital saat ini, volume data dan dokumen yang terus meningkat menuntut adanya mekanisme pencarian yang cepat, tepat, dan efisien. Teknologi yang digunakan pada mesin pencari seperti Google Search, fitur pencarian pada perpustakaan digital, hingga search engine dalam marketplace modern adalah contoh nyata penerapan konsep STKI. Sistem-sistem tersebut bekerja untuk mengolah teks dalam jumlah besar dan menampilkan hasil pencarian yang paling sesuai dengan maksud pengguna.

Dalam konteks pembelajaran mata kuliah STKI, mahasiswa tidak hanya dituntut memahami teori-teori dasar mengenai pemrosesan teks dan pencarian informasi, tetapi juga mampu mengimplementasikan model pencarian sederhana secara mandiri. Oleh karena itu, pada Ujian Tengah Semester (UTS) ini, penulis membangun sebuah sistem temu kembali informasi dengan memanfaatkan dua pendekatan dasar yang umum digunakan, yaitu Boolean Retrieval Model dan Vector Space Model (VSM). Model Boolean mengandalkan operasi logika seperti AND, OR, dan NOT untuk menentukan apakah sebuah dokumen memenuhi kriteria pencarian yang diberikan pengguna. Pendekatan ini bersifat biner—dokumen dianggap relevan atau tidak, tanpa memberikan peringkat. Sebaliknya, Vector Space Model bekerja dengan cara menghitung bobot istilah menggunakan TF-IDF serta mengukur tingkat kesamaan antara dokumen dan query melalui cosine similarity, sehingga mampu memberikan hasil pencarian yang terurut berdasarkan tingkat relevansi.

Pemilihan tema korpus berupa deskripsi parfum HMNS didasarkan pada karakteristik teksnya yang kaya akan kata-kata yang menggambarkan aroma, nuansa, intensitas, hingga komposisi parfum. Variasi istilah seperti citrus, woody, fresh, warm, floral, dan musk membuat korpus ini sangat cocok dijadikan objek studi dalam pemrosesan teks. Selain itu, deskripsi parfum umumnya memiliki struktur kalimat yang singkat namun informatif, sehingga memudahkan proses preprocessing sekaligus memberikan contoh yang jelas mengenai bagaimana sistem IR bekerja dalam mengenali dan mengurutkan informasi berbasis teks.

1.2 Tujuan

Berdasarkan latar belakang yang telah dijelaskan, maka dapat diidentifikasi tujuan proyek sebagai berikut:

1. Mengimplementasikan dua model IR dasar, yaitu Boolean Retrieval Model dan Vector Space Model (VSM), dalam bentuk program Python yang dapat berjalan pada korpus teks buatan sendiri.
2. Memahami dan menerapkan tahap-tahap preprocessing teks seperti case folding, tokenizing, stopword removal, dan stemming untuk menghasilkan representasi dokumen yang siap diproses oleh model IR.
3. Membangun inverted index dan perhitungan TF-IDF secara mandiri tanpa library crawling atau dataset eksternal.
4. Menguji kinerja sistem menggunakan metrik evaluasi Information Retrieval, seperti precision, recall, dan F1-score.
5. Menyediakan antarmuka pencarian sederhana (CLI) yang memungkinkan pengguna melakukan pencarian Boolean dan VSM terhadap korpus parfum.

1.3 Ruang Lingkup

Ruang lingkup proyek ini mencakup pembangunan sistem temu kembali informasi sederhana dengan menggunakan korpus berisi sepuluh deskripsi parfum HMNS yang disusun secara manual. Proses yang dilakukan meliputi preprocessing teks seperti case folding, tokenisasi, penghapusan stopword, dan stemming. Sistem ini hanya menerapkan dua model dasar, yaitu Boolean Retrieval Model dan Vector

Space Model (VSM) berbasis TF-IDF dan cosine similarity. Evaluasi dilakukan menggunakan precision, recall, dan F1-score, baik secara manual maupun otomatis melalui test case. Implementasi dibatasi pada antarmuka command-line tanpa pengembangan GUI atau web, serta tidak mendukung struktur query kompleks seperti penggunaan tanda kurung. Dengan batasan tersebut, proyek difokuskan sebagai implementasi dasar STKI dari pengolahan dokumen hingga penilaian performa pencarian.

1.4 Kontribusi Proyek vs SUB-CPMK

Proyek ini berkontribusi dalam memberikan pemahaman praktis mengenai bagaimana konsep dasar Sistem Temu Kembali Informasi diterapkan dalam sebuah sistem pencarian sederhana. Melalui implementasi preprocessing teks, model Boolean Retrieval, dan Vector Space Model berbasis TF-IDF, proyek ini membantu menunjukkan bagaimana dokumen direpresentasikan, diindeks, dan dicocokkan dengan query pengguna. Selain itu, proyek ini memberikan pemahaman tentang evaluasi kinerja sistem melalui perhitungan precision, recall, dan F1-score, sehingga mahasiswa dapat menilai efektivitas metode pencarian yang digunakan. Integrasi keseluruhan proses ke dalam aplikas CLI yang interaktif juga menjadi kontribusi penting dalam memperkuat kemampuan pengembangan perangkat lunak sederhana untuk pemrosesan teks. Dengan demikian, proyek ini mendukung pencapaian kompetensi mahasiswa dalam memahami konsep IR, mengimplementasikan metode pencarian, melakukan evaluasi performa, serta membangun sistem temu kembali informasi secara end-to-end.

BAB 2
LANDASAN TEORI

2.1 Sistem Temu Kembali Informasi

Sistem Temu Kembali Informasi (STKI) adalah suatu sistem yang dirancang untuk membantu pengguna menemukan dokumen-dokumen yang relevan berdasarkan query atau kata kunci tertentu. Berbeda dengan sistem tanya-jawab yang memberikan jawaban langsung, STKI bertugas menyeleksi dan mengembalikan dokumen yang *mengandung* informasi yang dicari. STKI banyak digunakan pada mesin pencari, perpustakaan digital, maupun sistem pencarian produk, sehingga mampu memproses dan menyaring informasi dalam jumlah besar secara cepat dan efisien.

2.2 Preprocessing Text

Preprocessing merupakan tahap awal yang berfungsi membersihkan dan menyiapkan teks sebelum masuk ke proses indexing atau pemodelan IR. Tahapan yang digunakan meliputi:

1. **Case folding**

   Mengubah seluruh huruf menjadi huruf kecil agar perhitungan kata konsisten.

2. **Tokenizing**

   Memecah kalimat menjadi unit kata sehingga dapat diproses secara individual.

3. **Stopword removal**

   Menghapus kata-kata umum yang tidak memberikan makna signifikan, seperti “dan”, “yang”, atau “di”.

4. **Stemming**

   Mengubah kata menjadi bentuk dasarnya menggunakan algoritma Sastrawi, sehingga kata seperti *wangi*, *wanginya*, dan *mengharumkan* direduksi ke bentuk dasar yang sama.

2.3 Boolean Retrieval Model

Boolean Retrieval Model merupakan pendekatan pencarian klasik yang bekerja menggunakan logika Boolean. Model ini menentukan relevansi dokumen berdasarkan keberadaan atau ketiadaan kata tertentu dalam dokumen.

Operator yang digunakan meliputi:

1. **AND** → kedua kata harus terdapat dalam dokumen.
2. **OR** → salah satu kata muncul sudah dianggap relevan.
3. **NOT** → mengecualikan dokumen yang mengandung kata tertentu.

Model ini menghasilkan keluaran berupa himpunan dokumen relevan tanpa perankingan. Artinya, semua dokumen yang memenuhi kondisi query dianggap sama relevannya.

2.4 Vector Space Model (VSM)

Usaha Vector Space Model merepresentasikan dokumen dan query sebagai vektor dalam ruang berdimensi kata. Model ini memungkinkan dokumen diranking berdasarkan tingkat kemiripan dengan query.

TF-IDF (Term Frequency – Inverse Document Frequency)

Digunakan untuk memberikan bobot pada setiap kata.

1. **TF (Term Frequency)**

   Jumlah kemunculan kata dalam dokumen.

2. **IDF (Inverse Document Frequency)**

   Mengukur seberapa penting sebuah kata secara global:

$$IDF = \log\left(\frac{N}{df}\right)$$

TF-IDF

Produk antara TF dan IDF, menunjukkan pentingnya kata dalam dokumen sekaligus membedakan kata umum dan kata khusus.

Cosine Similarity

Mengukur tingkat kemiripan antara dokumen dan query:

$$\cos(\theta) = \frac{A \cdot B}{||A|| \times ||B||}$$

Nilai cosine berada pada rentang 0–1, di mana nilai mendekati 1 menunjukkan dokumen memiliki kemiripan yang tinggi dengan query.

2.5 Evaluasi Sistem IR

A Evaluasi dilakukan untuk mengetahui sejauh mana sistem mampu mengembalikan dokumen yang relevan terhadap query pengguna. Metrik yang digunakan meliputi:

1. **Precision**
   Rasio dokumen relevan dari seluruh dokumen yang diambil sistem.
2. **Recall**
   Rasio dokumen relevan yang berhasil ditemukan dibandingkan seluruh dokumen relevan yang ada.
3. **F1-score**
   Nilai harmonisasi antara precision dan recall, memberikan evaluasi yang lebih seimbang.

Ketiga metrik ini membantu menilai efektivitas model Boolean maupun VSM dalam menampilkan hasil pencarian yang sesuai kebutuhan.

BAB 3
IMPLEMENTASI DAN HASIL

3.1 Implementasi Program

Struktur Folder:

```
stki-uts-A11.2022.14113-Alrijal Nur Ilham/
│
├── data/
│   ├── raw/
│   └── processed/
│
├── src/
│   ├── preprocess.py
│   ├── boolean_ir.py
│   ├── vsm_ir.py
│   ├── eval.py
│   ├── analyze_results.py
│   └── search.py
│
├── app/
│   └── main.py
│
├── notebooks/
│   └── UTS_STKI_A11.2022.14113.ipynb
│
├── reports/
│   ├── laporan.pdf
│   └── readme.md
│
└── requirements.txt
```

Program utama dijalankan via python search.py:

```
=== SISTEM PENCARIAN PARFUM HMNS ===
1. Preprocessing Dokumen
2. Pencarian Boolean Retrieval
3. Pencarian Vector Space Model (TF-IDF)
4. Evaluasi Sistem (Precision, Recall, F1)
5. Analisis Hasil Evaluasi (Analyze Results)
0. Keluar

Pilih menu [0-5]:
```

3.2 Hasil Boolean Retrieval

Menjalankan boolean retrieval dari menu search.py dengan mengisi querry “citrus”.

```
MODE: BOOLEAN RETRIEVAL
Inverted Index terbentuk (10 dokumen)

Masukkan query (gunakan AND/OR/NOT, ketik 'exit' untuk kembali): citrus
Dokumen cocok: ['doc3', 'doc8']

Masukkan query (gunakan AND/OR/NOT, ketik 'exit' untuk kembali): exit
```

3.3 Hasil Vector Space Model

Menjalankan vector space model (VSM) dari menu search.py dengan mengisi querry “floral citrus”.

```
MODE: VECTOR SPACE MODEL (TF-IDF)
 10 dokumen dimuat, 111 kosakata unik.

Masukkan query pencarian (ketik 'exit' untuk kembali): floral citrus

Hasil Ranking (Cosine Similarity):
doc6       → skor: 0.1712
doc8       → skor: 0.1687
doc7       → skor: 0.1597
doc3       → skor: 0.1544
----------------------------------------

Masukkan query pencarian (ketik 'exit' untuk kembali):
```

3.4 Hasil Evaluasi

Menjalankan analisis evaluasi dari menu search.py.

```
MODE: ANALISIS HASIL EVALUASI OTOMATIS
Menjalankan evaluasi untuk beberapa query parfum...

Query                          |  Precision |     Recall |   F1-score
-------------------------------------------------------------------------
vanilla floral aroma           |       0.67 |       1.00 |       0.80
woody amber warm               |       0.50 |       0.67 |       0.57
fresh citrus mint              |       1.00 |       1.00 |       1.00
musk jasmine floral            |       0.75 |       1.00 |       0.86
lavender sandalwood calm       |       1.00 |       1.00 |       1.00
-------------------------------------------------------------------------
Rata-rata                      |       0.78 |       0.93 |       0.85
```

3.5 Analisis

Recall sebesar 0.93 menunjukkan bahwa sistem berhasil menemukan seluruh dokumen yang relevan berdasarkan query yang diberikan. Artinya, tidak ada satu pun dokumen relevan yang terlewat oleh model. Namun, nilai precision yang berada pada angka 0.78 mengindikasikan bahwa masih terdapat sebagian dokumen yang tidak relevan namun ikut ditampilkan sebagai hasil pencarian. Kondisi ini menunjukkan bahwa meskipun sistem mampu menangkap seluruh dokumen relevan, masih terjadi *over-retrieval* atau pengambilan dokumen berlebih.

Secara keseluruhan, hasil tersebut menggambarkan bahwa Vector Space Model (VSM) bekerja dengan cukup baik pada korpus kecil yang berisi deskripsi parfum. Hal ini karena korpus parfum memiliki pola kosakata yang relatif stabil, seperti istilah aroma *citrus*, *floral*, *woody*, dan *fresh*, sehingga perhitungan bobot TF-IDF dapat membedakan dokumen secara lebih akurat. Dengan konsistensi istilah tersebut, model mampu melakukan pengukuran kemiripan secara efektif meskipun masih terdapat beberapa kasus kemiripan semu yang membuat beberapa dokumen tidak relevan ikut terambil.

BAB 4
PENUTUP

4.1 Kesimpulan

A Berdasarkan pembangunan dan pengujian sistem temu kembali informasi pada korpus deskripsi parfum HMNS, dapat disimpulkan bahwa model Boolean Retrieval dan Vector Space Model (VSM) mampu diterapkan secara efektif pada korpus berukuran kecil. Boolean Retrieval bekerja baik untuk pencarian berbasis kondisi logika sederhana, namun tidak memberikan perankingan sehingga hasilnya kurang fleksibel untuk query yang lebih kompleks. Sebaliknya, VSM dengan bobot TF-IDF mampu menghasilkan daftar dokumen yang terurut berdasarkan tingkat kemiripan, sehingga memberikan hasil pencarian yang lebih informatif.

Hasil evaluasi menunjukkan bahwa nilai recall mencapai 0.93, menandakan bahwa seluruh dokumen relevan berhasil ditemukan oleh sistem. Namun, precision yang berada pada angka 0.78 mengindikasikan adanya dokumen tidak relevan yang masih ikut terambil. Hal ini menunjukkan bahwa model sudah cukup baik dalam menangkap relevansi secara umum, tetapi masih perlu peningkatan dalam mengurangi pengambilan dokumen yang kurang sesuai. Secara keseluruhan, proyek ini berhasil menggambarkan bagaimana proses preprocessing, indexing, dan retrieval bekerja dalam sistem IR dasar.

4.2 Saran

Untuk pengembangan lebih lanjut, beberapa perbaikan dapat dilakukan agar sistem memiliki performa lebih optimal. Pertama, korpus dapat diperluas agar model memiliki lebih banyak variasi kata dan konteks, sehingga perhitungan TF-IDF menjadi lebih representatif. Kedua, penggunaan stopword list yang lebih komprehensif atau penyesuaian daftar stopword khusus domain parfum dapat membantu meningkatkan akurasi pencarian. Ketiga, penerapan teknik *query expansion* atau *synonym mapping* dapat mengurangi kesalahan pencarian akibat perbedaan pilihan kata pengguna.