Laporan UTS Sistem Temu Kembali Informasi (STKI)

Disusun oleh:

Aditya Rendy Setyawan – A11.2023.15189

Teknik Informatika

Fakultas Ilmu Komputer

Universitas Dian Nuswantoro

Semarang

2025

1. Pendahuluan

1.1 Tujuan

Tujuan proyek ini adalah membangun sistem *Sistem Temu Kembali Informasi (Information Retrieval System)* sederhana yang dapat melakukan pencarian dokumen berbasis teks menggunakan dua pendekatan utama, yaitu Boolean Retrieval Model dan Vector Space Model (VSM).
Proyek ini juga bertujuan untuk memahami proses *preprocessing*, pembuatan *index*, serta evaluasi performa sistem pencarian menggunakan metrik standar seperti Precision, Recall, F1-score, dan MAP.

1.2 Ruang Lingkup

Proyek ini mencakup:

1. Pemrosesan awal dokumen teks (.txt) yang berisi Rencana Pembelajaran Semester (RPS).
2. Implementasi model Boolean IR dan VSM (TF-IDF + Cosine Similarity).
3. Evaluasi hasil pencarian menggunakan metrik evaluasi IR.
4. Pembuatan antarmuka CLI sederhana (search engine mini) untuk interaksi pengguna.

1.3 Keterkaitan dengan Sub-CPMK

Sub-CPMK	Deskripsi	Capaian Proyek
10.1.1	Menjelaskan konsep dasar STKI dan arsitektur IR klasik	Soal 01
10.1.2	Melakukan preprocessing teks	Soal 02
10.1.3	Membangun model Boolean Retrieval	Soal 03
10.1.4	Menerapkan Vector Space Model (TF-IDF, cosine)	Soal 04
10.1.5	Mengevaluasi sistem IR menggunakan metrik performa	Soal 05

2. Data dan Preprocessing

2.1 Sumber Data

Dataset terdiri dari lima dokumen RPS mata kuliah:

1. RPS Kriptografi
2. RPS Sistem Informasi
3. RPS Sistem Temu Kembali Informasi
4. RPS Sistem Terdistribusi
5. RPS Manajemen Proyek Teknologi Informasi

Semua file disimpan dalam folder data/ dengan format .txt.

2.2 Tahapan Preprocessing

Preprocessing dilakukan dengan tahapan berikut:

1. Case Folding: mengubah seluruh huruf menjadi huruf kecil.
2. Tokenisasi: memecah teks menjadi token berdasarkan spasi dan tanda baca.
3. Stopword Removal: menghapus kata umum seperti “yang”, “dan”, “atau”.
4. Stemming: mengubah kata ke bentuk dasarnya menggunakan library *Sastrawi*.

Hasil preprocessing disimpan di folder data_processed/ dalam bentuk:

CLEAN_RPS Kriptografi.txt
CLEAN_RPS Sistem Informasi.txt

2.3 Contoh Before & After

Sebelum	**Sesudah**
“Mahasiswa mampu memahami konsep dasar sistem informasi dan implementasinya.”	“mahasiswa mampu paham konsep dasar sistem informasi implementasi”

3. Metode Information Retrieval

3.1 Boolean Retrieval

Boolean model menggunakan operasi logika untuk pencarian:

— AND: dokumen harus mengandung semua term.
— OR: dokumen mengandung salah satu term.
— NOT: dokumen yang tidak mengandung term tertentu.

Contoh query:

"kriptografi AND keamanan"

Sistem membangun inverted index, yaitu struktur data yang memetakan setiap term ke daftar dokumen yang memuatnya.

3.2 Vector Space Model (VSM)

VSM merepresentasikan setiap dokumen dan query sebagai vektor dalam ruang term. Relevansi dihitung dengan **cosine similarity** antara vektor query dan dokumen.

Rumus TF-IDF

- Term Frequency (TF):

  TF = 1+log10(ft,d)

- Inverse Document Frequency (IDF):

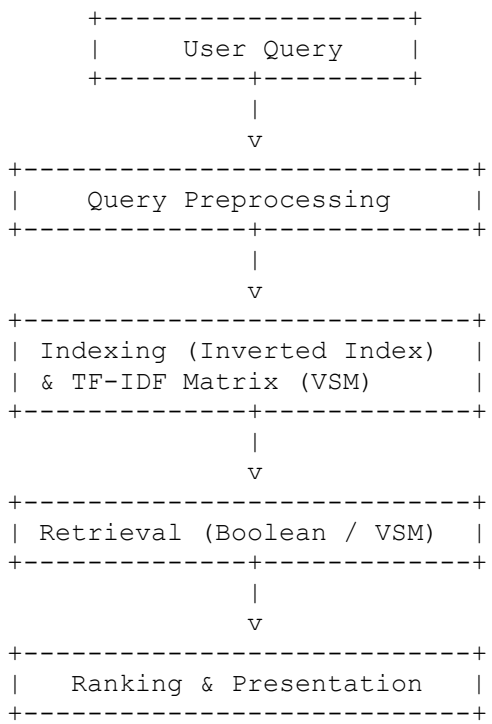
  IDF = log10(N/dft)

- TF-IDF Weight:

$w_{t,d} = TF \times IDF$

4. Arsitektur Search Engine

Arsitektur sistem mengikuti alur klasik mesin pencari sederhana:



5. Eksperimen dan Evaluasi

5.1 Skenario Uji

Tiga query digunakan untuk menguji performa sistem:

1. manajemen proyek teknologi
2. sistem terdistribusi
3. algoritma enkripsi rsa

5.2 Hasil Boolean Retrieval

Model Boolean diuji dengan operator `AND`, `OR`, dan `NOT`. Contoh:

```
QUERY: "kriptografi AND keamanan"
> Retrieved Docs: ['D4']
> Precision=1.0, Recall=1.0, F1=1.0
```

5.3 Hasil Vector Space Model (TF-IDF & Cosine)

Contoh hasil top-5 ranking dari `vsm_ir.py`:

Query	Gold Set	Precision@5	AP
manajemen proyek teknologi	D5, D1	0.40	0.42
sistem terdistribusi	D3	0.20	1.00
algoritma enkripsi rsa	D4	0.20	0.50
Mean Average Precision (MAP)	—	—	**0.6389**

5.4 Analisis

- Query 2 (sistem terdistribusi) berhasil sempurna karena term sangat spesifik.
- Query 1 dan 3 menghasilkan nilai AP < 1 karena term juga muncul di beberapa dokumen lain.
- MAP 0.6389 menunjukkan sistem sudah cukup relevan untuk dataset kecil.

6. Diskusi

Kelebihan:

- Menggunakan dua model IR (Boolean dan VSM) dalam satu sistem.
- Implementasi preprocessing lengkap (stopword, stemming).
- Evaluasi otomatis menggunakan metrik akademik (MAP, Precision, Recall).

Keterbatasan:

- Dataset kecil (5 dokumen) sehingga distribusi term kurang representatif.
- Belum mendukung query kompleks dengan tanda kurung atau operator campuran.
- Model VSM belum dioptimasi untuk *query expansion* atau *synonym handling*.

Saran Pengembangan:

- Menambahkan fitur *web scraping* untuk menambah dataset.
- Mengintegrasikan *ranking learning* atau *semantic search (BERT)*.
- Mengembangkan antarmuka web sederhana menggunakan Flask/Streamlit.

7. Kesimpulan

Proyek ini berhasil merealisasikan sistem temu kembali informasi sederhana yang mencakup seluruh proses utama: preprocessing, indexing, retrieval, ranking, dan evaluasi. Keluaran setiap tahap sesuai dengan Sub-CPMK:

- Soal 01: Pemahaman konsep dan arsitektur IR.
- Soal 02: Preprocessing teks menggunakan Sastrawi.
- Soal 03: Boolean model dengan inverted index.
- Soal 04: VSM dengan TF-IDF dan cosine similarity.
- Soal 05: Evaluasi performa sistem IR dengan metrik MAP dan nDCG.

Dengan rata-rata performa sistem (MAP = 0.6389), sistem ini menunjukkan hasil yang relevan dan sesuai teori IR klasik.