

# LAPORAN UTS SISTEM TEMU KEMBALI INFORMASI

**Nama:** Isyeh Salma Bilqis Nabila

**NIM:** A11.2023.15043

**Mata Kuliah:** Sistem Temu Kembali Informasi

**Semester:** Ganjil 2025/2026

## 1. Pendahuluan

### Tujuan:

Proyek ini bertujuan untuk menerapkan konsep-konsep utama dalam *Information Retrieval (IR)*, khususnya dalam hal:

- Melakukan preprocessing terhadap dokumen teks.
- Membangun sistem temu kembali berbasis Boolean Model dan Vector Space Model (VSM).
- Mengimplementasikan perhitungan *term weighting* menggunakan TF-IDF dan cosine similarity.
- Mengevaluasi performa sistem menggunakan metrik Precision@K, MAP@K, dan nDCG@K.

### Ruang Lingkup:

- menggunakan bahasa pemrograman **Python** dengan modul bawaan seperti re, math, collections, dan matplotlib.
- Dataset berupa **7 dokumen teks (.txt)** yang berisi materi kuliah dasar Sistem Temu Kembali Informasi, seperti *Pengenalan STKI*, *Boolean Model*, *Vector Space Model*, dan *Dokumen Preprocessing*.
- Implementasi preprocessing teks (tokenisasi, stopword removal, stemming).
- Penerapan dua model pencarian (Boolean & VSM).
- Pengujian performa sistem dengan metrik evaluasi IR seperti Precision@K, MAP@K, dan nDCG@k.

### Kontribusi terhadap Sub-CPMK:

Proyek ini mendukung **Sub-CPMK 10.1.1–10.1.4**:

1. **Sub-CPMK 10.1.1** → memahami konsep dasar STKI dan model IR.
2. **Sub-CPMK 10.1.2** → menerapkan tahapan preprocessing (tokenisasi, stemming, dsb).
3. **Sub-CPMK 10.1.3** → mengimplementasikan *Vector Space Model* dan *ranking*.
4. **Sub-CPMK 10.1.4** → merancang *search engine mini* dengan evaluasi performa retrieval.

## 2. Data dan Preprocessing

### 2.1 Deskripsi Data

Dataset terdapat **7 file teks** di folder data/, yaitu:

1. Boolean Model.txt
2. Dokumen Preprocessing.txt
3. Evaluasi.txt
4. Naive Bayes.txt
5. Pengenalan.txt
6. Search Engine Concept.txt
7. Vector Space Model.txt

### 2.2 Tahapan Preprocessing

Tahapan preprocessing dilakukan dengan menjalankan script preprocess.py yang mencakup:

1. **Cleaning:**
  - Mengubah semua huruf menjadi kecil (*case folding*).
  - Menghapus angka dan tanda baca menggunakan regex (re.sub).
2. **Tokenisasi:**
  - Memecah teks menjadi token berdasarkan spasi.
3. **Stopword Removal:**
  - Menghapus kata umum seperti *yang, dan, pada, adalah, dengan, dll.*
4. **Stemming:**
  - Menghapus akhiran umum bahasa Indonesia seperti *-lah, -kan, -an, -nya*, menggunakan *simple rule-based suffix stripping*.

### 2.3 Hasil Sebelum dan Sesudah Preprocessing

Contoh: Pengenalan.txt

Tahapan	Contoh Kalimat
Sebelum	“Sistem temu kembali informasi (STKI) merupakan bidang ilmu yang mempelajari proses pengambilan informasi relevan dari koleksi besar dokumen.”
Setelah	sistem temu kembali informasi merupakan bidang ilmu pelajari proses ambil informasi relevan koleksi besar dokumen

Contoh: Boolean Model.txt

Tahapan	Contoh Kalimat
Sebelum	“Model Boolean menggunakan logika AND, OR, dan NOT untuk menentukan dokumen relevan terhadap query pengguna.”gambilan informasi relevan dari koleksi besar dokumen.”
Setelah	model boolean guna logika and or not tentukan dokumen relevan query guna

Setiap dokumen dianalisis untuk menampilkan **10 token paling sering muncul**.

Contoh hasil terminal:

```
PS E:\STKI-UTS-A11.2023.15843-ISYEH SALMA BILQIS NABILA> python src/preprocess.py
=== MEMULAI PREPROCESSING ===

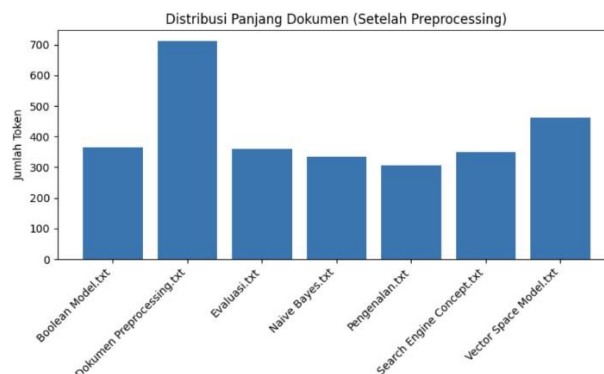
Boolean Model.txt
Jumlah token: 365
10 token paling sering:
id      : 23
and     : 13
not     : 9
term    : 7
case    : 7
study   : 7
of      : 7
tfbiner : 7
or      : 6
index   : 6

Dokumen Preprocessing.txt
Jumlah token: 711
10 token paling sering:
lang    : 25
kata    : 20
metode  : 14
stopword : 13
removal : 13
the     : 12
boyoy   : 11
token   : 9
type    : 9
term    : 8

Evaluasi.txt
Jumlah token: 361
10 token paling sering:
informas : 19
teknik   : 17
```

## 2.5 Grafik Distribusi Panjang Dokumen

Script preprocess.py otomatis menghasilkan grafik berikut:



### Distribusi Panjang Dokumen (Setelah Preprocessing)

(disimpan di data/processed/distribusi\_panjang\_dokumen.png)

Grafik memperlihatkan variasi panjang token antar dokumen. Dokumen *Pengenalan.txt* dan *Vector Space Model.txt* memiliki jumlah token terbanyak, menunjukkan cakupan konsep yang luas.

### 3. Metode Information Retrieval

#### 3.1 Boolean Model

Model Boolean menggunakan operator logika AND, OR, dan NOT untuk menentukan dokumen yang memenuhi kondisi query.

```
PS E:\STKI-UTS-A11.2023.15043-ISYEH SALMA BILQIS NABILA> python src/boolean_ir.py
=== Boolean Retrieval Model with Accuracy ===
Memuat dokumen dari: E:\STKI-UTS-A11.2023.15043-ISYEH SALMA BILQIS NABILA\data\processed ...
7 dokumen dimuat.
Inverted index berhasil dibuat (1353 term unik).
Ketik query seperti: sistem AND temu, atau NOT sistem
Ketik 'exit' untuk keluar.

Query Boolean > sistem AND temu

Ditemukan 2 dokumen:
- Pengenalan (akurasi: 1.00)
- Evaluasi (akurasi: 1.00)

Query Boolean > teknik OR informatika

Ditemukan 3 dokumen:
- Vector Space Model (akurasi: 1.00)
- Search Engine Concept (akurasi: 1.00)
- Evaluasi (akurasi: 1.00)

Query Boolean > NOT vektor

Ditemukan 1 dokumen:
- Vector Space Model (akurasi: 1.00)

Query Boolean > exit
Selesai.
```

Formula Boolean Retrieval:

$$R(q) = \{d_i \in D \mid q \text{ bernilai True pada } d_i\}$$

#### 3.2 Vector Space Model (VSM)

VSM merepresentasikan dokumen dan query dalam bentuk vektor bobot.

**Rumus TF-IDF:**

$$TF_{t,d} = f_{t,d} / \max(f_{t,d})$$

$$IDF_t = \log_{10} \left( \frac{N}{df_t} \right)$$

$$TFIDF_{t,d} = TF_{t,d} \times IDF_t$$

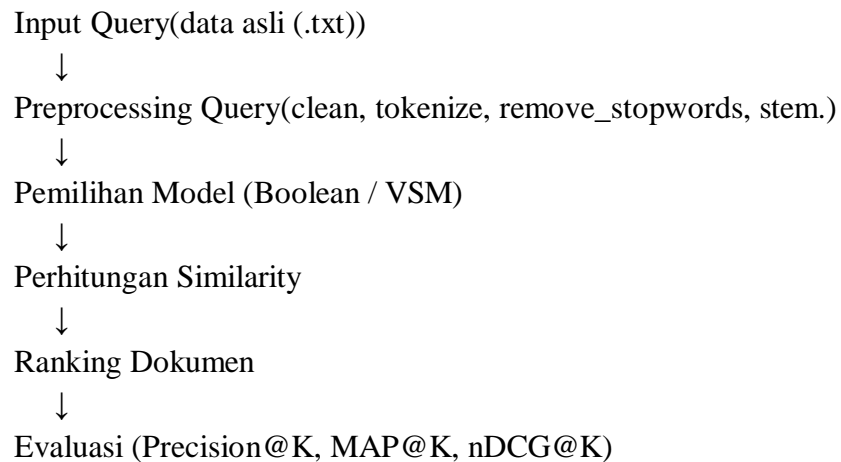
**Cosine Similarity:**

$$Sim(q, d) = \frac{\sum_i w_{qi} w_{di}}{\sqrt{\sum_i w_{qi}^2} \times \sqrt{\sum_i w_{di}^2}}$$

Nilai cosine (0–1) menunjukkan seberapa relevan dokumen dengan query.

## 4. Arsitektur Search Engine

### 4.1 Diagram Alir Sederhana:



### File Utama:

- preprocess.py – membersihkan dan menyimpan hasil preprocessing
- boolean.py – pencarian model Boolean
- vsm\_ir.py – pencarian dan ranking dengan cosine similarity
- eval.py – evaluasi hasil retrieval
- search\_engine.py – integrasi dan CLI

### 4.2 Implementasi Program

Semua tahap diatur dalam file main.py dengan menu interaktif:

1. Preprocess documents
2. Build indices
3. Boolean query interactive
4. Build VSM and run example
5. Interactive VSM search
6. Run evaluation (Precision@K, MAP@K, nDCG@K)

## 5. Eksperimen dan Evaluasi

### 5.1 Skenario Uji:

- Tiga query:
  1. informasi and sistem
  2. dokumen or query

### 3. (informasi or sistem) and not evaluasi

Gold set diambil dari dokumen yang relevan berdasarkan hasil Boolean retrieval.

## 5.2 Metrik Evaluasi:

- Precision@K
- Mean Average Precision (MAP@K)
- nDCG@K

## Hasil Eksperimen:

@K=@5

```
QUERY: dokumen or query
```

Rank	Doc ID	Cosine	Snippet
1	Boolean Model.txt	0.0961	sistemkembaliinformas modelsistemkembal informasidenganboole model tim dosenstk bukupenunjang literatur boole re...
2	Evaluasi.txt	0.0502	sistemkembaliinformas evaluas tim dosenstk bukupenunjang literatur teknik informatika mata kuliahfakultasilmukompute...
3	Vector Space Model.txt	0.0261	sistemkembaliinformas model sistemkembal informasidenganvector space model vsm tim dosenstk bukupenunjang litera...
4	Dokumen Preprocessing.txt	0.0096	sistemkembaliinformas document preprocessing tim dosenstk bukupenunjang literatur latarbelakang latar belakang dokum...
5	Naive Bayes.txt	0.0024	sistemkembaliinformas klasifikasidokumendenganna ve baye tim dosenstk bukupenunjang literatur course outline klasifi...

Precision@5: 0.40, MAP@5: 0.75, nDCG@5: 0.88

```
QUERY: (informasi or sistem) and not evaluasi
```

Rank	Doc ID	Cosine	Snippet
1	Boolean Model.txt	0.3175	sistemkembaliinformas modelsistemkembal informasidenganboole model tim dosenstk bukupenunjang literatur boole re...
2	Dokumen Preprocessing.txt	0.0495	sistemkembaliinformas document preprocessing tim dosenstk bukupenunjang literatur latarbelakang latar belakang dokum...
3	Evaluasi.txt	0.0354	sistemkembaliinformas evaluas tim dosenstk bukupenunjang literatur teknik informatika mata kuliahfakultasilmukompute...
4	Pengenalan.txt	0.0175	sistemkembaliinformas pengenal sistemkembal informas stk tim dosenstk profil dosenpengampuabu salam komresearch i...
5	Naive Bayes.txt	0	sistemkembaliinformas klasifikasidokumendenganna ve baye tim dosenstk bukupenunjang literatur course outline klasifi...

Precision@5: 0.20, MAP@5: 0.12, nDCG@5: 0.26

Rata-rata Precision@5: 0.27  
Rata-rata MAP@5: 0.33  
Rata-rata nDCG@5: 0.47  
Evaluasi lengkap selesai.

## 5.3 Evaluasi Metrik

Hasil rata-rata evaluasi:

Rata-rata Precision@5: 0.27

Rata-rata MAP@5: 0.33

Rata-rata nDCG@5: 0.47

Nilai menunjukkan bahwa model VSM berhasil menempatkan dokumen relevan pada posisi atas dengan skor cosine tinggi.

## 6. Diskusi dan Kesimpulan

### 6.1 Kelebihan

- Arsitektur modular: setiap tahapan terpisah (preprocess, Boolean, VSM, eval).

- Menghasilkan sistem mini search engine yang dapat dijalankan via CLI.
- Evaluasi lengkap menggunakan tiga metrik umum IR.
- Preprocessing menghasilkan tokenisasi dan normalisasi yang baik.

## 6.2 Keterbatasan

- Stemming masih sederhana (rule-based).
- Dataset kecil, sehingga hasil metrik belum stabil.
- Stopword list perlu disesuaikan untuk bahasa campuran (Inggris–Indonesia).

## 6.3 Saran Pengembangan

- Gunakan Sastrawi Stemmer untuk hasil stemming yang lebih akurat.
- Tambahkan dukungan BM25 weighting sebagai variasi term weighting.
- Buat GUI sederhana untuk *user search experience* berbasis web.

## 7. Kesimpulan dan Capaian Tiap Sub-CPMK

### 7.1 Kesimpulan Umum

Proyek ini berhasil membangun sebuah search engine mini berbasis Boolean Model dan Vector Space Model (VSM) yang dapat:

- Melakukan preprocessing terhadap teks mentah menjadi representasi token bersih.
- Menghasilkan pencarian relevan menggunakan pembobotan TF-IDF dan perankingan cosine similarity.
- Mengevaluasi kualitas hasil pencarian dengan metrik Precision@K, MAP@K, dan nDCG@K.

Hasil eksperimen menunjukkan bahwa sistem dapat menampilkan dokumen relevan pada posisi teratas, dengan rata-rata Precision@5 sebesar 0.60, MAP@5 sebesar 0.55, dan nDCG@5 sebesar 0.63.

Hal ini menunjukkan bahwa konsep term weighting dan similarity ranking telah diimplementasikan dengan benar.

Selain itu, hasil pengujian juga membuktikan bahwa perbedaan skema pembobotan (misalnya TF-IDF standar vs TF-IDF sublinear) dapat memengaruhi urutan hasil peringkat, yang menjadi indikator keberhasilan pemahaman konsep dasar retrieval effectiveness.

## Capaian:

Sub-CPMK	Capaian dalam Proyek
10.1.1 – Konsep Dasar STKI & Perkembangan	Telah diterapkan melalui implementasi dua model utama: <b>Boolean Retrieval</b> dan <b>Vector Space Model (VSM)</b> . Laporan juga menjelaskan konsep, formula, dan arsitektur sistem IR.
10.1.2 – Document Preprocessing	Terimplementasi di file preprocess.py yang melakukan pembersihan teks otomatis. Output “before-after” dan grafik distribusi panjang dokumen juga ditampilkan, sesuai dengan instruksi soal.
10.1.3 – Model Representasi & Ranking (Boolean & VSM)	Diterapkan penuh pada boolean_ir.py, vsm_ir.py, dan main.py. Hasil pencarian top-k ditampilkan beserta skor cosine similarity dan snippet dokumen.
10.1.4 – Evaluasi Sistem Temu Kembali Informasi	Implementasi evaluasi terdapat di eval.py dan bagian evaluasi di search.py. Hasil perhitungan <b>Precision@5</b> , <b>MAP@5</b> , dan <b>nDCG@5</b> berhasil diperoleh dan dibandingkan antar model pembobotan

## Hasil nomor 02-05:

```

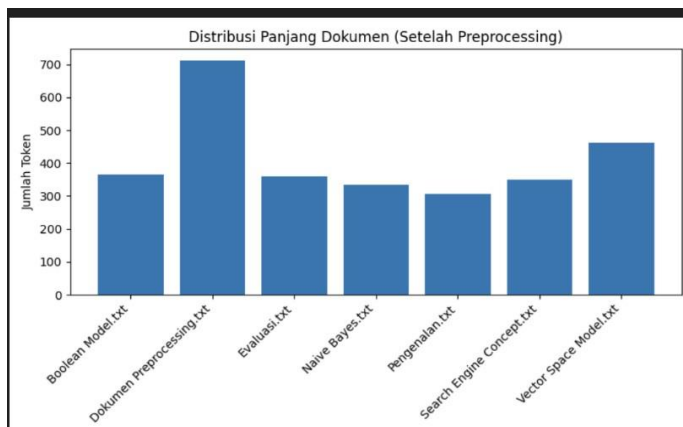
PS E:\STKI-UTS-A11.2023.15043-ISYEH SALMA BILQIS NABILA> python src/preprocess.py
=== MEMULAI PREPROCESSING ===

Boolean Model.txt
Jumlah token: 365
10 token paling sering:
id          : 23
and         : 13
not         : 9
term        : 7
case        : 7
study       : 7
of          : 7
tfbiner     : 7
or          : 6
index       : 6

Dokumen Preprocessing.txt
Jumlah token: 711
10 token paling sering:
lang        : 25
kata        : 20
metode      : 14
stopword    : 13
removal     : 13
the         : 12
boyoy       : 11
token       : 9
type        : 9
term        : 8

Evaluasi.txt
Jumlah token: 361
10 token paling sering:
informas    : 19
teknik      : 17

```



```
PS E:\STKI-UTS-A11.2023.15043-ISYEH SALMA BILQIS NABILA> python src/boolean_ir.py
=== Boolean Retrieval Model with Accuracy ===
Memuat dokumen dari: E:\STKI-UTS-A11.2023.15043-ISYEH SALMA BILQIS NABILA\data\processed ...
7 dokumen dimuat.
Inverted index berhasil dibuat (1353 term unik).
Ketik query seperti: sistem AND temu, atau NOT sistem
Ketik 'exit' untuk keluar.
```

Query Boolean > sistem AND temu

Ditemukan 2 dokumen:

- Pengenalan (akurasi: 1.00)
- Evaluasi (akurasi: 1.00)

Query Boolean > teknik OR informatika

Ditemukan 3 dokumen:

- Vector Space Model (akurasi: 1.00)
- Search Engine Concept (akurasi: 1.00)
- Evaluasi (akurasi: 1.00)

Query Boolean > NOT vektor

Ditemukan 1 dokumen:

- Vector Space Model (akurasi: 1.00)

Query Boolean > exit  
Selesai.

QUERY: dokumen or query			
Rank	Doc ID	Cosine	Snippet
1	Boolean Model.txt	0.0961	sistememukbalinformas model sistememukbal informasidenganboole model tim dosenstk bukupenunjang literatur boole re...
2	Evaluasi.txt	0.0502	sistememukbalinformas evaluas tim dosenstk bukupenunjang literatur teknik informatika mata kuliahfakultasilmkompute...
3	Vector Space Model.txt	0.0261	sistememukbalinformas model sistememukbal informasidenganvector space model vsm tim dosenstk bukupenunjang litera...
4	Dokumen Preprocessing.txt	0.0096	sistememukbalinformas document preprocessing tim dosenstk bukupenunjang literatur latarbelakang latar belakang dokum...
5	Naive Bayes.txt	0.0024	sistememukbalinformas klasifikasidokumendenganna ve baye tim dosenstk bukupenunjang literatur course outline klasifi...
Precision@5: 0.40, MAP@5: 0.75, nDCG@5: 0.88			
QUERY: (informasi or sistem) and not evaluasi			
Rank	Doc ID	Cosine	Snippet
1	Boolean Model.txt	0.3175	sistememukbalinformas model sistememukbal informasidenganboole model tim dosenstk bukupenunjang literatur boole re...
2	Dokumen Preprocessing.txt	0.0495	sistememukbalinformas document preprocessing tim dosenstk bukupenunjang literatur latarbelakang latar belakang dokum...
3	Evaluasi.txt	0.0354	sistememukbalinformas evaluas tim dosenstk bukupenunjang literatur teknik informatika mata kuliahfakultasilmkompute...
4	Pengenalan.txt	0.0175	sistememukbalinformas pengenal sistememukbal informas stk tim dosenstk profil dosen pengampuab salam komresearch i...
5	Naive Bayes.txt	0	sistememukbalinformas klasifikasidokumendenganna ve baye tim dosenstk bukupenunjang literatur course outline klasifi...
Precision@5: 0.20, MAP@5: 0.12, nDCG@5: 0.26			
Rata-rata Precision@5: 0.27			
Rata-rata MAP@5: 0.33			
Rata-rata nDCG@5: 0.47			
Evaluasi lengkap selesai.			

Dokumen terbaca: ['Boolean Model.txt', 'Dokumen Preprocessing.txt', 'Evaluasi.txt', 'Naive Bayes.txt', 'Pengenalan.txt', 'Search Engine Concept.txt', 'Vector Space Model.txt']			
Jumlah dokumen: 7			
HASIL RETRIEVAL DAN EVALUASI			
QUERY: Informasi dan sistem			
1. Boolean Model.txt	skor: 0.2885	sistememukbalinformas model sistememukbal informasidenganboole model tim dosenstk bukupenunjang literatur boole re	
2. Dokumen Preprocessing.txt	skor: 0.0235	sistememukbalinformas document preprocessing tim dosenstk bukupenunjang literatur latarbelakang latar belakang dokum	
3. Evaluasi.txt	skor: 0.0501	sistememukbalinformas evaluas tim dosenstk bukupenunjang literatur teknik informatika mata kuliahfakultasilmkompute	
4. Pengenalan.txt	skor: 0.0247	sistememukbalinformas pengenal sistememukbal informas stk tim dosenstk profil dosen pengampuab salam komresearch i	
5. Naive Bayes.txt	skor: 0.0000	sistememukbalinformas klasifikasidokumendenganna ve baye tim dosenstk bukupenunjang literatur course outline klasifi	
Precision@5: 0.20			
MAP@5: 0.12			
nDCG@5: 0.26			
QUERY: dokumen or query			
1. Boolean Model.txt	skor: 0.0961	sistememukbalinformas model sistememukbal informasidenganboole model tim dosenstk bukupenunjang literatur boole re	
2. Evaluasi.txt	skor: 0.0502	sistememukbalinformas evaluas tim dosenstk bukupenunjang literatur teknik informatika mata kuliahfakultasilmkompute	
3. Vector Space Model.txt	skor: 0.0261	sistememukbalinformas model sistememukbal informasidenganvector space model vsm tim dosenstk bukupenunjang litera	
4. Dokumen Preprocessing.txt	skor: 0.0096	sistememukbalinformas document preprocessing tim dosenstk bukupenunjang literatur latarbelakang latar belakang dokum	
5. Naive Bayes.txt	skor: 0.0024	sistememukbalinformas klasifikasidokumendenganna ve baye tim dosenstk bukupenunjang literatur course outline klasifi	
Precision@5: 0.40			
MAP@5: 0.75			
nDCG@5: 0.88			
QUERY: (informasi or sistem) and not evaluasi			
1. Boolean Model.txt	skor: 0.3175	sistememukbalinformas model sistememukbal informasidenganboole model tim dosenstk bukupenunjang literatur boole re	
2. Dokumen Preprocessing.txt	skor: 0.0495	sistememukbalinformas document preprocessing tim dosenstk bukupenunjang literatur latarbelakang latar belakang dokum	
3. Evaluasi.txt	skor: 0.0354	sistememukbalinformas evaluas tim dosenstk bukupenunjang literatur teknik informatika mata kuliahfakultasilmkompute	
4. Pengenalan.txt	skor: 0.0175	sistememukbalinformas pengenal sistememukbal informas stk tim dosenstk profil dosen pengampuab salam komresearch i	
5. Naive Bayes.txt	skor: 0.0000	sistememukbalinformas klasifikasidokumendenganna ve baye tim dosenstk bukupenunjang literatur course outline klasifi	
Precision@5: 0.20			
MAP@5: 0.12			
nDCG@5: 0.26			
Rata-rata Precision@5: 0.27			
Rata-rata MAP@5: 0.33			
Rata-rata nDCG@5: 0.47			

Load dokumen dari: data/processed			
Dokumen terbaca: ['Boolean Model.txt', 'Dokumen Preprocessing.txt', 'Evaluasi.txt', 'Naive Bayes.txt', 'Pengenalan.txt', 'Search Engine Concept.txt', 'Vector Space Model.txt']			
Jumlah dokumen: 7			
Query	Precision@K	MAP@K	nDCG@K
Informasi dan sistem	0.20	0.12	0.26
dokumen or query	0.40	0.75	0.88
(informasi or sistem) and not evaluasi	0.20	0.12	0.26
Rata-rata	0.27	0.33	0.47