LAPORAN UTS STKI

Implementasi Boolean Retrieval dan Vector Space Model untuk Pencarian Review Restoran

Oleh:

Layla Thias Rahmawati (A11.2022.14132)

Teknik Informatika

Universitas Dian Nuswantoro

2025/2026

DAFTAR ISI

1. Pendahuluan

1.1 Tujuan

Tujuan dari proyek ini adalah membangun sebuah sistem pencarian sederhana yang dapat menemukan review restoran yang relevan berdasarkan kata kunci yang diberikan pengguna. Untuk mencapai hal tersebut, proyek ini mencakup proses membersihkan teks, membuat indeks dokumen, menerapkan dua metode pencarian (Boolean Retrieval dan Vector Space Model), melakukan perhitungan kemiripan menggunakan TF-IDF, serta menampilkan hasil pencarian dalam bentuk daftar dokumen yang paling cocok. Selain itu, proyek ini juga bertujuan memberikan pemahaman praktis mengenai cara kerja sistem Temu Kembali Informasi (IR) serta bagaimana evaluasi dilakukan untuk mengukur kualitas hasil pencarian.

1.2 Ruang Lingkup

Ruang lingkup proyek ini dibatasi pada pemrosesan sepuluh dokumen review restoran dalam format teks sederhana. Dokumen-dokumen tersebut diolah menggunakan teknik dasar pemrosesan teks seperti case folding, tokenizing, stopword removal, dan stemming. Sistem pencarian yang dibangun hanya memanfaatkan dua pendekatan klasik IR, yaitu Boolean Retrieval dan Vector Space Model, tanpa menggunakan model pencarian modern atau teknik pemahaman makna mendalam. Sistem juga berjalan secara lokal di lingkungan Python, sehingga tidak melibatkan data dari web maupun integrasi antarmuka pengguna yang kompleks.

1.3 Kontribusi Proyek terhadap Sub-CPMK

Proyek ini berkontribusi langsung terhadap capaian pembelajaran mata kuliah STKI karena tidak hanya mencakup pemahaman teori IR, tetapi juga penerapan proses preprocessing, pembuatan indeks, implementasi metode pencarian, hingga evaluasi performa sistem. Melalui proyek ini, dijelaskan bagaimana query dicocokkan dengan dokumen, bagaimana bobot relevansi dihitung, bagaimana sistem mengurutkan dokumen, serta bagaimana kualitas hasil pencarian dinilai menggunakan precision,

recall, dan metrik lainnya. Dengan demikian, proyek ini memberikan pengalaman praktis yang menggambarkan keseluruhan alur kerja IR sesuai Sub-CPMK yang ditetapkan dalam mata kuliah.

2. Data & Preprocessing

2.1 Data

Data yang digunakan dalam proyek ini terdiri dari sepuluh dokumen teks berisi review pelanggan terhadap sebuah restoran. Setiap dokumen berisi komentar singkat tentang pengalaman makan, yang umumnya mencakup aspek rasa makanan, harga, pelayanan, suasana, hingga kepuasan secara keseluruhan. Dokumen-dokumen ini dipilih karena bentuknya yang tidak terstruktur dan bergaya bahasa natural, sehingga cocok digunakan untuk proyek Sistem Temu Kembali Informasi. Data sengaja dibiarkan apa adanya sesuai teks asli pengguna, agar proses preprocessing dapat memperlihatkan perubahan yang terjadi sebelum dokumen digunakan dalam model IR.

2.2 Tahapan Preprocessing

Sebelum dokumen dapat diproses oleh sistem IR, setiap review harus melalui tahapan preprocessing agar teks menjadi lebih bersih, konsisten, dan mudah dianalisis. Proses ini dimulai dengan case folding untuk mengubah semua huruf menjadi huruf kecil, dilanjutkan dengan pembersihan karakter seperti angka, simbol, dan tanda baca. Kemudian teks dipecah menjadi token kata melalui proses tokenizing, dan kata-kata umum yang tidak membawa makna penting dihapus melalui stopword removal. Langkah terakhir adalah stemming menggunakan library Sastrawi, yang berguna untuk mengubah kata menjadi bentuk dasar, sehingga kata seperti “pelayanannya” dapat dipersingkat menjadi “layan” dan dianggap sama dengan kata-kata lain yang berakar sama.

2.3 Contoh Before dan After

Untuk menunjukkan perubahan yang terjadi selama preprocessing, berikut contoh transformasi dari salah satu review. Sebelum diproses, teks biasanya masih mengandung huruf kapital, tanda baca, dan berbagai bentuk kata turunan seperti “pelayanannya sangat ramah, makanannya enak banget!”. Setelah melalui serangkaian tahapan preprocessing, teks tersebut menjadi lebih sederhana dan bersih, misalnya berubah menjadi “layan sangat ramah makan enak banget”. Contoh ini menunjukkan bagaimana preprocessing membantu menstandarkan teks, mengurangi noise, dan mempersiapkan dokumen agar bisa diproses lebih efektif oleh Boolean Retrieval maupun Vector Space Model.

3. Metode Information Retrieval (IR)

3.1 Boolean Retrieval

Boolean Retrieval adalah metode pencarian yang bekerja dengan mencocokkan kata kunci secara langsung terhadap isi dokumen berdasarkan logika AND, OR, dan NOT. Pada metode ini, dokumen hanya akan dianggap relevan apabila memenuhi kondisi Boolean yang diminta oleh pengguna. Misalnya, untuk query “enak AND ramah”, sistem hanya akan mengambil dokumen yang mengandung kedua kata tersebut sekaligus. Model ini memiliki kelebihan karena sederhana dan hasilnya sangat jelas baik dokumen cocok atau tidak cocok. Namun, kekurangannya adalah model ini tidak memberikan peringkat relevansi, sehingga tidak ada informasi tambahan seperti seberapa dekat isi dokumen dengan maksud pengguna. Meskipun sederhana, Boolean Retrieval sangat sesuai untuk memahami fondasi awal pencarian dokumen dalam IR.

3.2 Vector Space Model (VSM)

Vector Space Model bekerja dengan mengubah dokumen dan query pengguna menjadi representasi berbentuk vektor angka berdasarkan perhitungan TF-IDF. Dengan pendekatan ini, sistem tidak hanya melihat apakah kata tertentu muncul,

tetapi juga seberapa penting kata tersebut dalam sebuah dokumen dibandingkan dokumen lainnya. Setelah semua dokumen direpresentasikan dalam bentuk vektor, sistem menghitung tingkat kemiripan antara dokumen dan query menggunakan cosine similarity, yaitu ukuran yang menunjukkan seberapa dekat arah dua vektor. Semakin tinggi nilai similarity, semakin relevan dokumen tersebut terhadap query pengguna. Berbeda dengan Boolean Retrieval, VSM memungkinkan dokumen yang tidak mengandung semua kata secara langsung tetap dianggap relevan jika memiliki konteks yang mirip, sehingga memberikan hasil yang lebih fleksibel dan mendekati pencarian modern.

3.3 Rumus yang Digunakan

Beberapa perhitungan penting digunakan dalam VSM, yaitu Term Frequency (TF) untuk menghitung seberapa sering sebuah kata muncul dalam dokumen, Document Frequency (DF) untuk mengetahui jumlah dokumen yang mengandung kata tersebut, dan Inverse Document Frequency (IDF) yang memberi bobot lebih tinggi pada kata-kata yang jarang muncul di seluruh dokumen. Gabungan TF dan IDF disebut TF-IDF, yaitu nilai yang menunjukkan tingkat kepentingan sebuah kata dalam dokumen tertentu. Setelah dokumen dan query direpresentasikan sebagai vektor TF-IDF, sistem menggunakan cosine similarity untuk mengukur kemiripan dengan menghitung sudut antara dua vektor. Semakin kecil sudutnya, semakin tinggi kemiripannya, dan semakin relevan dokumen tersebut terhadap kata kunci yang dicari pengguna.

4. Arsitektur Search Engine

Arsitektur sistem pencarian yang dibangun pada proyek ini mengikuti alur kerja dasar sebuah search engine klasik, namun dalam bentuk yang jauh lebih sederhana. Proses dimulai dari kumpulan dokumen review restoran yang disimpan dalam folder data. Dokumen-dokumen tersebut kemudian melalui preprocessing agar bersih, konsisten, dan siap diolah. Setelah preprocessing selesai, dokumen dimasukkan ke dua jenis indeks: inverted index untuk Boolean Retrieval dan indeks TF-IDF untuk Vector Space Model. Ketika pengguna memasukkan sebuah query,

query tersebut juga diproses terlebih dahulu agar formatnya sesuai dengan dokumen yang telah dipreproses. Sistem kemudian meneruskan query ke dua jalur pencarian, yaitu jalur Boolean dan jalur VSM. Pada jalur Boolean, sistem mencocokkan kata kunci secara langsung untuk menentukan dokumen mana yang memenuhi kondisi AND, OR, atau NOT. Sementara pada jalur VSM, sistem menghitung tingkat kemiripan antara query dan dokumen menggunakan cosine similarity dan menghasilkan ranking dokumen yang paling relevan. Setelah kedua jalur selesai, hasil pencarian ditampilkan sebagai output agar pengguna dapat melihat dokumen mana saja yang paling sesuai dengan kata kunci yang dimasukkan.

5. Eksperimen & Evaluasi

5.1 Skenario Eksperimen

Eksperimen dilakukan untuk menguji bagaimana sistem pencarian bekerja ketika pengguna memasukkan kata kunci tertentu. Dalam pengujian ini, query yang digunakan adalah “enak ramah”, yang mewakili dua aspek penting dalam review restoran—kualitas makanan dan keramahan layanan. Query ini dipilih karena sering muncul dalam ulasan dan relevan untuk melihat kemampuan sistem dalam menemukan dokumen yang memenuhi dua makna sekaligus. Sistem menjalankan pencarian dengan dua metode, yaitu Boolean Retrieval dan Vector Space Model, untuk melihat perbedaan hasil antara pencarian berbasis kecocokan kata secara langsung dan pencarian berbasis kemiripan konteks.

5.2 Hasil Pencarian (Boolean Retrieval)

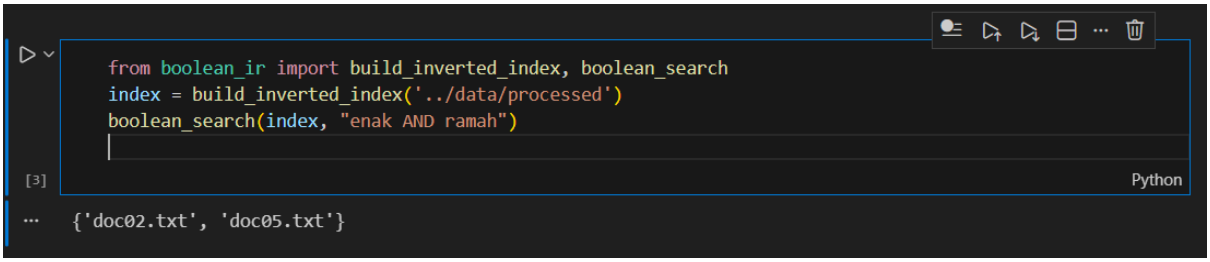
Pada metode Boolean Retrieval, sistem hanya mengembalikan dokumen yang benar-benar mengandung kedua kata dalam query, yaitu “enak” dan “ramah”. Hasil pencarian menunjukkan bahwa hanya sebagian dokumen yang memenuhi kondisi tersebut, sehingga sistem menghasilkan daftar dokumen yang relatif sedikit namun sangat tepat sasaran. Keunggulan Boolean terlihat dari kepastian hasilnya, di mana setiap dokumen yang muncul sudah pasti mengandung semua

kata yang dicari. Namun, metode ini tidak memberikan informasi tambahan seperti dokumen mana yang paling relevan atau memiliki konteks yang lebih kuat.

5.3 Hasil Pencarian (Vector Space Model / VSM)

Berbeda dengan Boolean, metode VSM menampilkan daftar dokumen dalam bentuk ranking berdasarkan tingkat kemiripan antara isi dokumen dan query “enak ramah”. Dari hasil perhitungan cosine similarity, dokumen yang mengandung kedua kata sekaligus cenderung mendapat nilai similarity yang tinggi, namun dokumen lain yang memiliki kata-kata terkait atau konteks mirip juga dapat muncul dalam daftar, meskipun dengan skor yang lebih rendah. Dengan cara ini, VSM memberikan gambaran yang lebih halus dan fleksibel mengenai relevansi dokumen, tidak hanya berdasarkan kecocokan kata, tetapi juga berdasarkan kemiripan makna secara keseluruhan.

Contoh hasil:



5.4 Evaluasi Performa Sistem

Evaluasi dilakukan dengan membandingkan hasil pencarian sistem dengan gold standard atau daftar dokumen yang dianggap relevan berdasarkan penilaian manual. Metrik yang digunakan meliputi precision, recall, Average Precision (AP), dan Mean Average Precision (MAP). Precision mengukur seberapa banyak dokumen yang ditampilkan sistem benar-benar sesuai, sementara recall mengukur seberapa banyak dokumen relevan berhasil ditemukan. AP digunakan untuk menghitung rata-rata nilai precision setiap kali sistem menemukan dokumen relevan pada ranking, dan MAP memberikan nilai rata-rata dari seluruh query yang diuji. Dari evaluasi ini dapat dilihat bagaimana kualitas hasil pencarian dari

kedua metode, serta mana yang bekerja lebih baik dalam konteks dataset review restoran.

6. Diskusi

6.1 Kelebihan Sistem

Sistem pencarian yang dibangun pada proyek ini memiliki beberapa kelebihan yang membuatnya mudah dipahami dan cukup efektif untuk dataset kecil. Boolean Retrieval memberikan hasil yang sangat tegas dan akurat karena hanya menampilkan dokumen yang benar-benar mengandung kata kunci tertentu, sehingga cocok digunakan ketika pengguna membutuhkan kecocokan yang pasti. Di sisi lain, Vector Space Model memberikan hasil yang lebih kaya karena mampu mengurutkan dokumen berdasarkan tingkat kemiripannya, sehingga pengguna bisa melihat dokumen mana yang paling relevan meskipun tidak mengandung kata yang sama persis. Preprocessing menggunakan library Sastrawi juga meningkatkan kualitas teks yang diproses, terutama dalam Bahasa Indonesia yang kaya bentuk kata.

6.2 Keterbatasan Sistem

Namun, sistem ini juga memiliki beberapa keterbatasan yang perlu diperhatikan. Jumlah dokumen yang digunakan relatif sedikit, sehingga hasil pencarian dan evaluasi mungkin belum cukup representatif untuk skala besar. Boolean Retrieval terasa terlalu kaku karena hanya bergantung pada kecocokan kata secara literal dan tidak mempertimbangkan konteks. Sementara itu, VSM masih memiliki kelemahan dalam memahami hubungan makna antara kata, sehingga sinonim seperti “lezat” dan “enak” dianggap berbeda dan tidak dihitung sebagai kemiripan. Sistem juga belum menerapkan metode IR yang lebih modern seperti BM25 atau embedding berbasis semantik, yang sebenarnya dapat meningkatkan kualitas pencarian secara signifikan.

6.3 Saran Pengembangan

Untuk pengembangan selanjutnya, sistem ini sangat berpotensi diperluas menjadi lebih canggih dan praktis. Dataset dapat diperbanyak agar hasil ranking menjadi lebih stabil dan evaluasi lebih akurat. Dari sisi metode, sistem dapat ditingkatkan menggunakan BM25 atau word embeddings seperti Word2Vec, FastText, atau BERT untuk memahami sinonim dan konteks yang lebih luas. Fitur tambahan seperti query expansion, penanganan frasa, atau pemisahan sentiment positif dan negatif juga dapat meningkatkan pengalaman pengguna. Terakhir, sistem ini dapat diintegrasikan ke dalam antarmuka web sederhana agar lebih mudah digunakan dan mendekati bentuk sebuah search engine nyata.