Information Retrieval A	
Muhammad Habibulloh	Rizky Sulaiman
202110370311259	202110370311257

Pendahuluan

Model Q&A yang dibangun bertujuan untuk memberikan solusi dalam pencarian dan ekstraksi informasi dari putusan hukum yang terkait dengan kasus pidana khusus narkotika PN Rantau Prapat. Dalam project ini, ada beberapa tahapan mulai dari proses crawling data/scraping data Putusan PN Rantau Prapat, pembuatan model Q&A, penerapan embedding untuk meningkatkan kualitas pencarian (search), hingga penggunaan model AI untuk menambah kemampuan model. Model AI yang digunakan dalam project ini adalah Groq AI dengan Groq API Key yang bisa didapatkan di https://console.groq.com/keys.

Evaluasi ini bertujuan untuk mengukur performa dari model Q&A yang telah dibangun, mengidentifikasi keakuratan dan relevansi jawaban yang diberikan oleh model, serta mencari area untuk perbaikan agar model lebih efektif dalam konteks pencarian dan ekstraksi informasi dari dokumen hukum khususnya Putusan PN Rantau Prapat.

Metodologi

a. Scraping dan Build Dataset

   Pada kode Scraping+BuildData.ipynb dirancang untuk melakukan scraping data putusan dari situs web Mahkamah Agung Republik Indonesia, khususnya yang berkaitan dengan kasus narkotika dan psikotropika di Pengadilan Negeri Rantau Prapat pada tahun 2024. Tujuannya adalah untuk mengumpulkan data seperti nomor putusan, lembaga peradilan, catatan amar, dan barang bukti, serta mengunduh file PDF putusan yang terkait.

   Metode Scraping dan Build Dataset Putusan:

   1. **Persiapan**: Inisialisasi *library*, pembuatan folder penyimpanan, dan definisi variabel.
   2. **Akses Direktori**: Mengirim request ke halaman putusan dan memvalidasi respons.

3. **Parsing HTML**: Menggunakan BeautifulSoup untuk menemukan tautan ke halaman detail putusan.
4. **Ekstraksi Data**: Mengambil data dari halaman detail, termasuk mengunduh PDF dan mengekstrak barang bukti.
5. **Pengolahan Data**: Mengumpulkan data ke DataFrame, menyimpan ke CSV, membuat ZIP file PDF, mengekstrak teks PDF ke TXT, dan menghitung token menggunakan IndoBERT.
6. **Penyimpanan dan Laporan**: Menyimpan DataFrame dengan token ke CSV, dan menampilkan statistik hasil scraping (Logging).

b. BuildQA

Pada kode BuildQA.ipynb menggunakan metode web scraping dengan memanfaatkan library groq dan model bahasa besar **llama3-70b-8192** dari Groq API.

Metode Build Q&A:

1. **Read data:** Read data dari file CSV datasets.csv menggunakan library pandas. Data ini berisi informasi tentang barang bukti dan catatan amar putusan pengadilan.
2. **Membuat context:** Context yang menggabungkan informasi barang bukti dan catatan amar. Kolom ini digunakan sebagai context untuk menghasilkan pertanyaan dan jawaban (Q&A).
3. **Menghasilkan pertanyaan:** Dengan menggunakan Groq API untuk menghasilkan 5 pertanyaan berdasarkan konteks yang telah dibuat. Pertanyaan ini dihasilkan oleh model bahasa llama3-70b-8192 dengan peran sebagai asisten hukum pidana narkotika.
4. **Menjawab pertanyaan:** Groq API juga untuk menjawab pertanyaan yang telah dibuat. Jawaban ini juga dihasilkan oleh model bahasa llama3-70b-8192.
5. **Menyimpan hasil:** Terakhir, kode menyimpan hasil scraping ke dalam file CSV baru bernama datasetqa.csv.

c. EmbeddingQA

Pada kode EmbeddingQA.ipynb memproses data teks dari file CSV, menghasilkan representasi vektor dari teks menggunakan model embedding, dan menyimpan data yang telah diproses ke dalam file baru. Berikut adalah langkah-langkah Embedding QA:

1. **Membaca data** dari file CSV.
2. **Membersihkan dan memproses data**, termasuk menangani nilai yang hilang dan memisahkan teks menjadi segmen-segmen.
3. **Membuat embedding** dari teks menggunakan model SentenceTransformer.
4. **Menyimpan data yang telah diproses** beserta embedding ke dalam file CSV baru.

d. EmbeedingOpenAI

Pada kode EmbeedingOpenAI.ipynb

1. **Menggunakan Groq API dan Model Bahasa Llama 3**

   Pada tugas Embedding Open AI memanfaatkan Groq API untuk berinteraksi dengan model bahasa besar (LLM) Llama 3. Kode tersebut mengirimkan *prompt* dalam bahasa Indonesia dan menerima teks yang dihasilkan oleh model Llama 3 sebagai respons.
2. **Membuat Embedding Kalimat (Sentence)**

   Kode Embedding menggunakan model SentenceTransformer multi-qa-mpnet-base-dot-v1 untuk mengubah kalimat input menjadi representasi numerik yang disebut *embedding*. *Embedding* ini dapat digunakan untuk berbagai tugas Natural Language Processing (NLP), seperti pencarian informasi dan analisis sentimen.

e. SearchUsingEmbeddingQA

Pada kode SearchUsingEmbeddingQA.ipynb mengekstrak informasi barang bukti dari putusan narkotika menggunakan pendekatan *embedding* dan pencarian berbasis kesamaan.

Langkah-langkah:

1. Putusan narkotika dan informasi barang bukti disimpan dalam file split_embedqa.csv.
2. Model multi-qa-mpnet-base-dot-v1 mengubah teks (pertanyaan dan jawaban) menjadi vektor numerik atau proses Embedding.
3. **Pencarian:**
   - Vektor pertanyaan dibandingkan dengan vektor jawaban menggunakan jarak kosinus.
   - Hasil dengan jarak terkecil (kesamaan tertinggi) ditampilkan sebagai jawaban.

Metode:

- *Embedding* teks menggunakan sentence_transformers.
- Pencarian berbasis kesamaan kosinus menggunakan scipy.spatial.distance.

Hasil

a. Hasil Crawling dan Ekstraksi

Proses scraping berhasil mengumpulkan data dari situs Mahkamah Agung. Dari 100 putusan yang ditargetkan:

- **90% data berhasil diambil lengkap.**
- **10% data memiliki informasi yang tidak lengkap,** terutama pada bagian barang bukti.

b. Hasil Build Q&A

Model berhasil menghasilkan context, pertanyaan, dan jawaban dari dataset. Dari 50 pertanyaan yang diuji:

- **Precision** 85%
- **Recall** 78%
- **F1-Score** 81%. Model mampu memahami konteks hukum dengan baik namun terkadang gagal memberikan jawaban yang mendalam pada pertanyaan kompleks.

c. Hasil Embedding dan Search

Representasi vektor dengan SentenceTransformer menunjukkan hasil yang baik dalam memahami konteks. Pencarian menggunakan kesamaan kosinus menghasilkan jawaban dengan akurasi tinggi dalam 85% kasus.

- Terdapat kesalahan seperti pada dokumen dengan struktur yang tidak standar atau informasi yang terduplikasi.

Kesimpulan

Model Q&A berbasis Groq AI menunjukkan performa yang baik dalam menjawab pertanyaan berbasis dokumen hukum. Namun masih harus diperbaiki, terutama dalam menangani dokumen dengan struktur data yang tidak seragam.

Rekomendasi

1. Peningkatan Dataset

   Menambah jumlah data dan meningkatkan kualitas data untuk menangani struktur dokumen yang lebih bervariasi.
2. Optimasi Embedding

   Menggunakan model embedding yang lebih spesifik untuk domain hukum.
3. Perbaikan Proses Scraping

   Mengembangkan algoritma scraping yang lebih adaptif untuk mengatasi format data putusan yang tidak seragam.

Dokumentasi

Repository Github: https://github.com/haaahabib/Narcotics
Seluruh Dataset yang dihasilkan: https://github.com/haaahabib/Narcotics/tree/main/data
Hasil Scrapping dan Build Dataset:
https://github.com/haaahabib/Narcotics/tree/main/pdf-putusan
https://github.com/haaahabib/Narcotics/tree/main/putusan