|  |  |  |
| --- | --- | --- |
| **Nama:**  **(Isi Nama Anda)**  **NIM: (Isi NIM Anda)** | A black background with a black square  Description automatically generated with medium confidence | **MODUL 9**  **Nama Dosen:**  **( )** |
| **Hari/Tanggal:**  **Hari, Tanggal Bulan Tahun** | **Praktikum Pemrosesan Data** | **Nama Asisten Labratorium:**   1. **( )** |

**Similarity Search with Text Embedding**

1. **Teori Singkat**

Ketika membaca artikel menarik, kita mungkin ingin menemukan artikel lain yang serupa dari kumpulan dokumen besar. Mencari secara manual tentu membutuhkan banyak waktu. Oleh karena itu, Artificial Intelligence (AI) dapat digunakan untuk menyelesaikan masalah ini secara efisien. Pendekatan modern menggunakan model berbasis **deep learning** seperti **SciBERT**, yang mampu memahami makna teks ilmiah. SciBERT mengubah setiap dokumen menjadi representasi vektor (disebut embedding), yang merepresentasikan makna teks tersebut secara matematis.

**Similarity Search** dengan **Text Embedding** adalah proses mencari dokumen atau teks yang memiliki kemiripan makna dengan teks tertentu menggunakan representasi vektor (embedding). Pendekatan ini mengandalkan model deep learning yang mengubah teks menjadi vektor berdimensi tinggi, yang disebut text embedding.

A diagram of a text embedding

Description automatically generated

Embedding ini merepresentasikan fitur semantik (makna) dari teks, sehingga memungkinkan pencarian berbasis makna, bukan hanya pencocokan kata secara langsung.

Prosesnya melibatkan langkah-langkah berikut:

1. **Pembentukan Embedding**: Setiap dokumen atau teks dikonversi menjadi embedding menggunakan model bahasa seperti **BERT** atau **SciBERT**.
2. **Perhitungan Kemiripan**: Kemiripan antar embedding dihitung menggunakan metrik seperti **Cosine Similarity** atau **Euclidean Distance**, yang menunjukkan seberapa dekat vektor embedding satu teks dengan yang lain.
3. **Pencarian Cepat**: Dengan bantuan algoritma seperti **k-Nearest Neighbors (k-NN)** atau alat seperti **FAISS**, dokumen-dokumen paling mirip dapat ditemukan secara efisien, bahkan dari kumpulan data besar.

Metode ini digunakan dalam berbagai aplikasi, seperti rekomendasi artikel, pengelompokan dokumen, pencarian informasi, dan analisis data tekstual. Keunggulannya terletak pada kemampuannya memahami makna teks secara kontekstual, meskipun kata-katanya berbeda.

1. **Latihan Pertama**

*Find Similar Documents From Scientific Corpus Using Deep Learning With SciBERT*

* 1. **Import Dataset dari Kaggle**

**A screenshot of a computer program

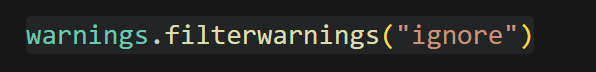
Description automatically generated**

|  |
| --- |
|  |

* 1. **Import Libraries**

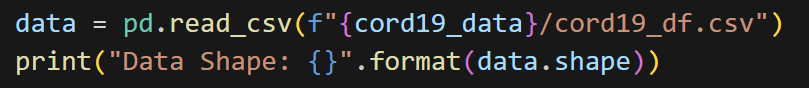
**A screen shot of a computer program

Description automatically generated**

****

|  |
| --- |
|  |

* 1. **Load Data & Memeriksa Missing Value**

****

**A black background with white text

Description automatically generated**

**A computer code with colorful text

Description automatically generated**

|  |
| --- |
|  |

* 1. **Menampilkan Artikel Random (Acak)**

**A screen shot of a computer code

Description automatically generated**

|  |
| --- |
|  |

* 1. **Model SciBERT**

**A screen shot of a computer program

Description automatically generated**

|  |
| --- |
|  |

* 1. **Konversi Teks ke Embedding**

**A computer screen shot of a program code

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

|  |
| --- |
|  |

* 1. **Embedding untuk Dataset**

**A screen shot of a computer

Description automatically generated**

|  |
| --- |
|  |

* 1. **Preprocessing Dataset & Embedding untuk Semua Teks**

**A screen shot of a computer code

Description automatically generated**

**A screen shot of a computer program

Description automatically generated**

**A computer screen shot of text

Description automatically generated**

|  |
| --- |
|  |

* 1. **Fungsi Proses Query**

**A screen shot of a computer code

Description automatically generated**

|  |
| --- |
|  |

* 1. **Fungsi Pencarian Berdasarkan Cosine Similarity**

**A computer screen shot of text

Description automatically generated**

|  |
| --- |
|  |

* 1. **Menguji Fungsi Pencarian Cosine Similarity**

**A screen shot of a computer program

Description automatically generated**

**#Uji 1**

**A computer screen shot of a computer code

Description automatically generated**

**#Uji 2**

**A black screen with blue and yellow text

Description automatically generated**

**#Uji 3**

A screen shot of a computer code

Description automatically generated

**#Uji 4**

A black screen with white text

Description automatically generated

**#Uji 5**

**A black screen with white text

Description automatically generated**

|  |
| --- |
|  |

* 1. **Membuat Indeks FAISS untuk Pencarian Vektor**

**A screen shot of a computer program

Description automatically generated**

|  |
| --- |
|  |

* 1. **Menambahkan Embedding ke Indeks FAISS**

**A screen shot of a computer

Description automatically generated**

|  |
| --- |
|  |

* 1. **Mencari Artikel Menggunakan FAISS**

A computer screen with text

Description automatically generated

**A screenshot of a computer

Description automatically generated**

|  |
| --- |
|  |

* 1. **Menampilkan Hasil Pencarian FAISS**

**A computer screen with text and symbols

Description automatically generated**

|  |
| --- |
|  |

1. **Kesimpulan**
   1. Kita dapat mengetahui… (Tolong Isi lebih dari 3 baris!)
2. **Tugas Kelas**
3. Mengapa data dibersihkan sebelum dikonversi menjadi embedding?
4. Apa perbedaan antara *manual similarity search* dan pendekatan berbasis embedding?

|  |
| --- |
|  |

1. **Cek List (✔)**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Elemen Kompetensi** | **Penyelesaian** | |
| **Selesai** | **Tidak Selesai** |
| **1.** | Latihan Pertama |  |  |
| **2.** | Tugas Kelas |  |  |

1. **Formulir Umpan Balik**

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Elemen Kompetensi** | **Waktu Pengerjaan** | **Kriteria** |
| **1.** | Latihan Pertama | … Menit | … |
| **2.** | Tugas Kelas |  |  |

Keterangan:

1. Menarik
2. Baik
3. Cukup
4. Kurang

Penanggung Jawab Praktikum Kepala Lab. Praktikum

( ) ( )