**Pembobotan *Vector Space Model* Korpus *Twitter* Tentang *Data Science Text Mining Text Retrieval* Menggunakan *Cosine Smiliarity***

**TUGAS KELOMPOK**

**Disusun Untuk Memenuhi Tugas Mata Kuliah Temu Kembali Informasi**

**Dosen Pengampu : Retnani Latifah, M.Kom**

**Disusun Oleh :**

|  |  |
| --- | --- |
| **MUHAMMAD REZA** | **2019470055** |
| **SELAMET SAPUTRA** | **2019470069** |
| **SYECHAN AHMAD ZIDAN** | **2019470110** |

PROGRAM STUDI TEKNIK INFORMATIKAFAKULTAS TEKNIKUNIVERSITAS MUHAMMADIYAH JAKARTA2022

**DAFTAR ISI**

**BAB 1**

* 1. **Latar belakang masalah**

*Information retrieval* atau pengambilan informasi adalah tugas untuk mengambil informasi yang sesuai atau relevan dari kumpulan korpus yang mewakili permintaan (kueri) (Djenouri et al., 2021).

Ekstrasi fitur atau *Term Frequency Inverse Document Frequency* yaitu perkalian dari term *frequency row (tf raw)* yang dipakai untuk menghitung untuk menghitung jumlah kemunculan kata atau *term* untuk tiap kalimat pada teks. Sedangkan *inverse document frequency* adalah perhitungan untuk menentukan sebuah bobot pada suatu kata dalam suatu teks korpus. Jadi dari hasil nilai ekstrasi fitur tf-idf ini digunakan untunk perhitung similaritas, dan untuk beberapa metode dalam pendekatan statistika (Setyawan et al., 2021).

*Vector Space Model* (VSM). Sebuah model yang digunakan untuk mengukur sebuah kueri antara suatu dokumen dengan suatu kata kunci atau *keyword* (Susanti et al., 2020). *Vector space* adalah geometri berdimensi besar, ruang yang batas-batasnya ditentukan oleh vector. *Vector space model* yang menarik bagi penulis adalah model numerik yang menempatkan teks atau kata dalam sebuah ruang representasi dimensi tinggi. Secara lebih luas masuk akal, kita mungkin mempertimbangkan matriks jangka dokumen, yang pada dasarnya adalah tabel frekuensi kata yang disejajarkan oleh kosakata umum sehingga setiap vektor mewakili distribusi kosa kata ini dalam individu teks (Dobson, 2022). Konsep dasar *vector space model* adalah menghitung jarak vector antara dokumen dengan kata kunci yang dimasukkan kemudian mengurutkan berdasarkan tingkat kedekatannya (Susanti et al., 2020).

BAB 2

- Data ACC

- EDA

- PREPROP

BAB 3

- TFIDF

- VSM