Indonesia AI
AI for Everyone, AI for Indonesia

#2023JAGOTEKNOLOGIAI

AI Career Bootcamp

Frequency-Based Text Vectorization

Indonesia AI (PT. Teknologi Artifisial Indonesia)

Pembukaan

Guide Book ini memuat beberapa informasi-informasi utama yang akan disampaikan oleh mentor di program AI Career Bootcamp yang bisa dijadikan pegangan para students untuk mempersiapkan diri sebelum sesi Live Class berlangsung.

Deskripsi

Frequency-Based Text Vectorization adalah salah satu teknik yang digunakan dalam NLP untuk mengubah teks menjadi representasi numerik yang dapat diproses oleh algoritma machine learning. Teknik ini berfokus pada frekuensi kemunculan kata-kata dalam teks sebagai dasar untuk pembentukan vektor representasi.

Pada Frequency-Based Text Vectorization, setiap dokumen atau kalimat dalam teks diubah menjadi vektor berdasarkan frekuensi kemunculan kata-kata di dalamnya. Pendekatan yang populer digunakan adalah Term Frequency-Inverse Document Frequency (TF-IDF). Metode ini memberikan bobot pada kata-kata yang sering muncul dalam sebuah dokumen tetapi jarang muncul di seluruh koleksi dokumen. Hal ini membantu dalam mengidentifikasi kata-kata yang lebih signifikan dalam suatu konteks.

Proses Frequency-Based Text Vectorization dimulai dengan tokenisasi, yaitu memisahkan teks menjadi unit-unit yang lebih kecil seperti kata-kata. Selanjutnya, frekuensi kemunculan setiap kata dihitung dalam setiap dokumen atau kalimat. Kemudian, bobot TF-IDF dihitung dengan mengalikan frekuensi kata dengan invers dari frekuensi kata tersebut di seluruh dokumen. Hasil akhirnya adalah vektor numerik yang merepresentasikan dokumen atau kalimat tersebut, dengan setiap dimensi vektor merepresentasikan kata tertentu dan nilainya menggambarkan bobot kata tersebut dalam konteks teks.

Kelebihan dan Kekurangan

Kelebihan Frequency-Based Text Vectorization:

1. Sederhana dan Mudah diimplementasikan

   Teknik ini relatif mudah diimplementasikan dan tidak memerlukan persiapan data yang kompleks. Penghitungan frekuensi kemunculan kata-kata dalam teks dapat dilakukan dengan cepat menggunakan library pemrosesan teks yang tersedia.

2. Menjaga Informasi Frekuensi Kemunculan Kata

   Metode ini mempertahankan informasi tentang frekuensi kemunculan kata dalam dokumen. Kata-kata yang muncul lebih sering akan mendapatkan bobot yang lebih tinggi, dan kata-kata yang jarang muncul akan mendapatkan bobot yang lebih rendah. Hal ini membantu dalam mengidentifikasi kata-kata kunci yang dapat memiliki kontribusi yang signifikan dalam

Kekurangan Frequency-Based Text Vectorization

1. Sensitivitas terhadap Panjang Dokumen

   Metode ini cenderung memberikan bobot yang lebih tinggi pada kata-kata yang muncul lebih sering dalam dokumen yang lebih panjang. Hal ini dapat menyebabkan bias dalam representasi teks, di mana kata-kata yang muncul lebih sering dalam dokumen panjang dianggap lebih penting daripada kata-kata yang muncul lebih sering dalam dokumen yang

2. Tidak Memperhitungkan Konteks dan Urutan Kata

   Frequency-Based Text Vectorization hanya mempertimbangkan frekuensi kemunculan kata-kata dalam teks, tanpa memperhatikan konteks atau urutan kata. Ini dapat mengakibatkan hilangnya informasi penting seperti struktur sintaksis atau hubungan antara kata-kata dalam teks.