TDAI SESI 10

MUHAMMAD FAJRUL ASLIM - 20200804031

Jelaskan perbedaan fitur bag-of-words dan TF-IDF.

JAWAB

Bag of Words

Bag of Words (BoW) merupakan salah satu metode paling sederhana dalam mengubah data teks menjadi vektor yang dapat dipahami oleh komputer. Metode ini sejatinya hanya menghitung frekuensi kemunculan kata pada seluruh dokumen.

Ukuran korpus Bag of Words mengikuti jumlah kata unik dari seluruh dokumen. Artinya, jika nantinya terdapat berbagai kata unik baru maka ukuran korpus juga akan semakin membesar. Tentunya hal ini akan berpengaruh pada komputasi yang dibutuhkan pada saat kita melatih model machine learning.

Bag of Words menghilangkan konteks kalimat akibat tidak memperhatikan urutan kata.

TF-IDF

TF-IDF biasa digunakan ketika kita ingin mengubah data teks menjadi vektor namun dengan memperhatikan apakah sebuah kata tersebut cukup informatif atau tidak. Mudahnya, TF-IDF membuat kata yang sering muncul memiliki nilai yang cenderung kecil, sedangkan untuk kata yang semakin jarang muncul akan memiliki nilai yang cenderung besar. Kata yang sering muncul disebut juga Stopwords biasanya dianggap kurang penting, salah satu contohnya adalah kata hubung (yang, di, akan, dengan, dll).

Term Frequency (TF) menghitung frekuensi jumlah kemunculan kata pada sebuah dokumen. Karena panjang dari setiap dokumen bisa berbeda-beda, maka umumnya nilai TF ini dibagi dengan panjang dokumen (jumlah seluruh kata pada dokumen).

TF-IDF sejatinya berdasar pada Bag of Words (BoW), sehingga TF-IDF pun tidak bisa menangkap posisi teks dan semantiknya.

TF-IDF hanya berguna sebagai fitur di level leksikal.

Berikan ringkasan artikel berikut ini

https://www.datacamp.com/community/tutorials/simplifying-sentiment-analysispython

JAWAB

Analisis sentimen adalah topik penting di bidang NLP. Ini dengan mudah menjadi salah satu topik terpanas di lapangan karena relevansinya dan banyaknya masalah bisnis yang dipecahkan dan telah mampu dijawabnya. Dalam tutorial ini, Anda akan membahas topik yang tidak terlalu sederhana ini dengan cara yang sederhana. Anda akan menguraikan semua matematika kecil di baliknya, dan Anda akan mempelajarinya. Anda juga akan membuat pengklasifikasi sentimen sederhana di akhir tutorial ini. Secara khusus, Anda akan mencakup:

- Memahami analisis sentimen dari perspektif praktisi
- Merumuskan pernyataan masalah analisis sentiment
- Klasifikasi Naive Bayes untuk analisis sentiment
- Studi kasus dengan Python
- Bagaimana analisis sentimen memengaruhi beberapa alasan bisnis
- Bacaan lebih lanjut tentang topik ini

Pada dasarnya, analisis sentimen atau klasifikasi sentimen termasuk dalam kategori luas tugas klasifikasi teks di mana Anda diberikan frasa, atau daftar frasa dan pengklasifikasi Anda seharusnya memberi tahu apakah sentimen di balik itu positif, negatif atau netral. Terkadang, atribut ketiga tidak diambil untuk menjaganya agar tetap menjadi masalah klasifikasi biner. Dalam tugas baru-baru ini, sentimen seperti "agak positif" dan "agak negatif" juga sedang dipertimbangkan.

Sebelum memahami pernyataan masalah tugas klasifikasi sentimen, Anda harus memiliki gagasan yang jelas tentang masalah klasifikasi teks umum. Mari kita definisikan secara formal masalah tugas klasifikasi teks umum.

Mengapa analisis sentimen begitu penting?

Analisis sentimen memecahkan sejumlah masalah bisnis yang sebenarnya:

- Ini membantu untuk memprediksi perilaku pelanggan untuk produk tertentu.
- Ini dapat membantu untuk menguji kemampuan beradaptasi suatu produk.
- Mengotomatiskan tugas laporan preferensi pelanggan.
- Ini dapat dengan mudah mengotomatiskan proses menentukan seberapa baik film berjalan dengan menganalisis sentimen di balik ulasan film dari sejumlah platform.
- Dan masih banyak lagi!