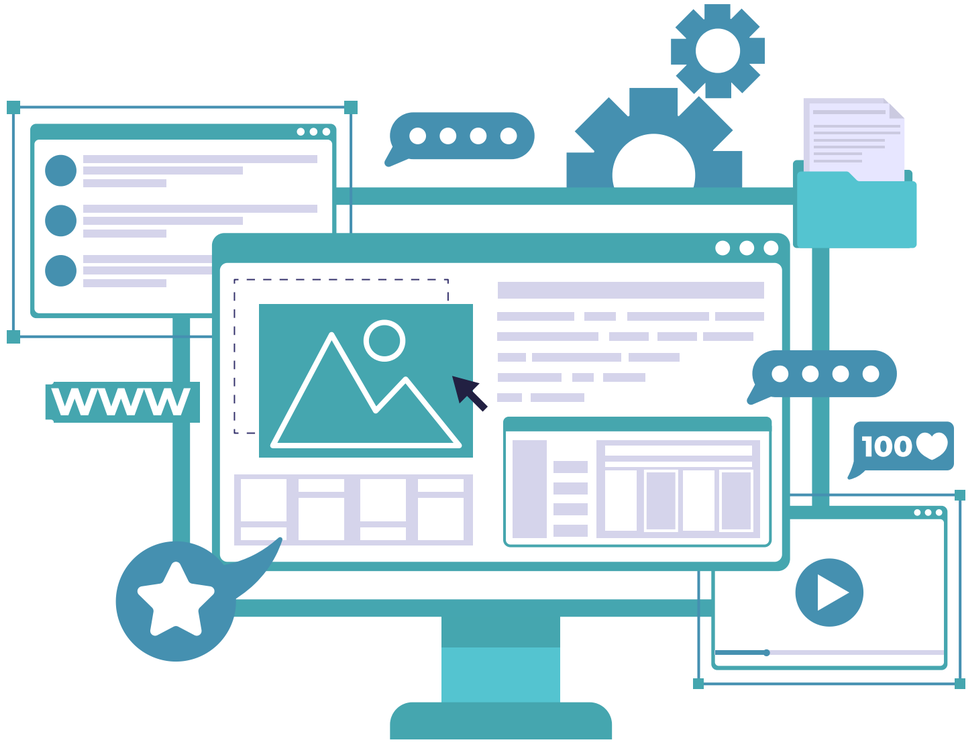
Poltek Harber
POLITEKNIK HARAPAN BERSAMA

BIG DATA - RPL

ANALISIS SENTIMEN BERBASIS SIMILARITY DENGAN COSINE SIMILARITY DI PYTHON

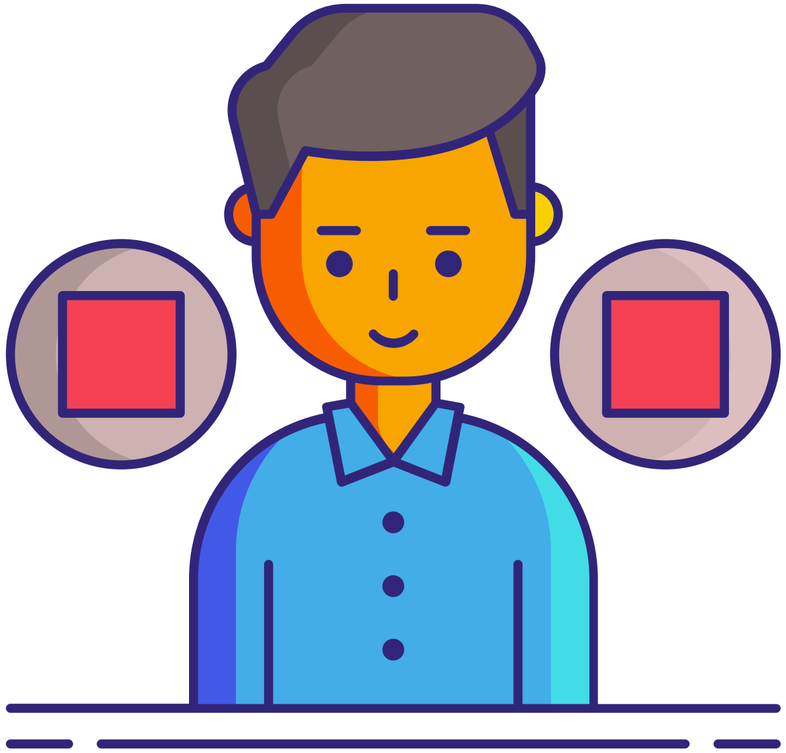


Poltek Harber
POLITEKNIK HARAPAN BERSAMA

TUJUAN PEMBELAJARAN

- Memahami metode cosine similarity untuk analisis sentimen.
- Mengetahui cara menggunakan TF-IDF untuk merepresentasikan teks.
- Mengimplementasikan model similarity untuk prediksi sentimen di Python.
- Menyimpan dan memuat model untuk prediksi teks.

Poltek Harber
POLITEKNIK HARAPAN BERSAMA

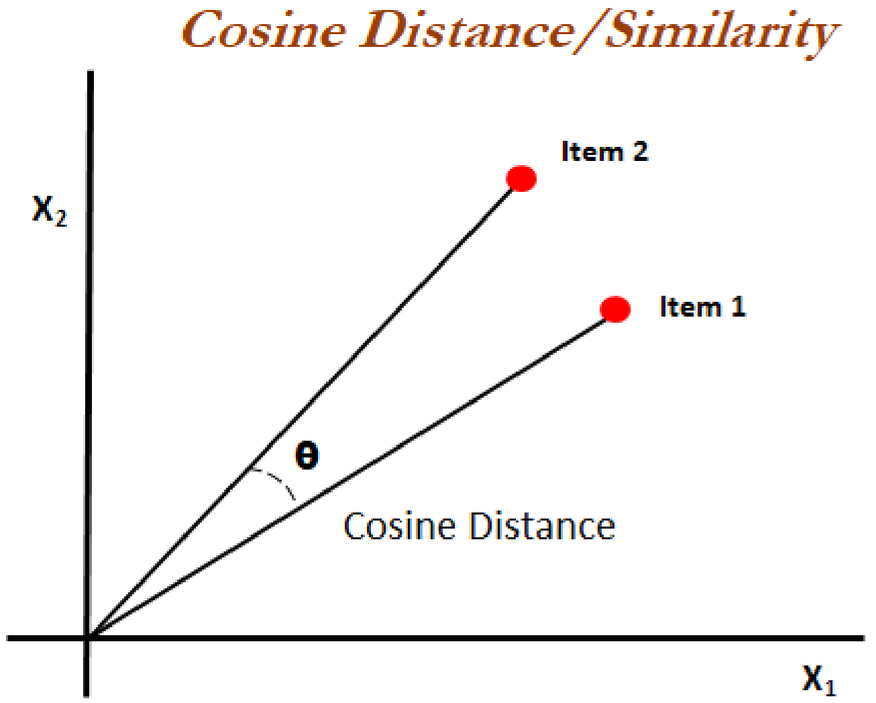


APA ITU COSINE SIMILARITY?

Cosine Similarity adalah cara untuk mengukur seberapa mirip dua teks satu sama lain. Semakin mirip teks tersebut, semakin besar nilai similarity-nya. Nilainya berkisar antara 0 (tidak mirip sama sekali) hingga 1 (sangat mirip).

KONSEP SIMILARITY DAN COSINE SIMILARITY

- Similarity: Metode untuk mengukur kemiripan antara dua teks berdasarkan representasi matematis.
- Cosine Similarity: Mengukur sudut kosinus antara dua vektor teks, memberikan nilai antara -1 (sangat berbeda) dan 1 (sangat mirip).



Poltek Harber
POLITEKNIK HARAPAN BERSAMA

BAGAIMANA CARA KERJANYA?

1. Teks diubah menjadi angka: Teks tidak bisa langsung dihitung menggunakan Cosine Similarity, jadi kita perlu mengubahnya menjadi bentuk angka terlebih dahulu. Salah satu cara yang digunakan adalah dengan teknik yang disebut TF-IDF (Term Frequency - Inverse Document Frequency). Teknik ini memberi bobot pada kata-kata yang penting dalam sebuah teks.

Term Frequency (TF) mengukur seberapa sering kata muncul dalam sebuah teks.
Inverse Document Frequency (IDF) mengukur seberapa jarang kata muncul di seluruh dokumen (semakin jarang, semakin penting kata tersebut).

Membandingkan dua teks: Setelah teks diubah menjadi vektor angka, kita bisa mengukur kemiripannya menggunakan Cosine Similarity. Jika dua teks mirip, nilai similarity-nya akan lebih dekat ke 1. Jika tidak mirip, nilai similarity-nya akan lebih mendekati 0.

Poltek Harber
POLITEKNIK HARAPAN BERSAMA

KELEBIHAN COSINE SIMILARITY:

1. Mudah digunakan untuk teks yang tidak terlalu panjang.
2. Tidak perlu pelatihan model yang rumit.

KEKURANGAN COSINE SIMILARITY:

1. Hanya menghitung kemiripan berdasarkan kata-kata, tidak memperhitungkan makna sebenarnya.
2. Kurang efektif jika teks sangat berbeda meskipun memiliki sentimen yang sama.

Poltek Harber
POLITEKNIK HARAPAN BERSAMA

CONTOH SEDERHANA:

Bayangkan Anda memiliki dua tweet:
"Saya sangat suka acara ini, sangat inspiratif!"
"Acara ini sangat keren dan menginspirasi!"

Jika kita menghitung Cosine Similarity antara keduanya, kita akan mendapatkan nilai yang tinggi (misalnya 0.9), karena kedua tweet tersebut hampir memiliki kata-kata yang sama dan menyampaikan sentimen yang sama.

Namun, jika Anda membandingkan tweet:
"Saya sangat suka acara ini, sangat inspiratif!"
"Saya sedang makan siang."

Cosine Similarity akan memberikan nilai rendah (misalnya 0.2), karena kata-katanya sangat berbeda.

Poltek Harber
POLITEKNIK HARAPAN BERSAMA

KESIMPULAN:

Cosine Similarity membantu kita untuk memahami seberapa mirip sebuah teks dengan teks lainnya, yang bisa sangat berguna dalam analisis sentimen untuk memprediksi apakah suatu tweet atau komentar bersifat positif atau negatif. Dengan menggunakan teknik seperti TF-IDF, kita dapat mengubah teks menjadi angka dan membandingkannya dengan cara yang lebih efisien.