Yohanes Dimas Pratama

A11.2021.13254 - A11.4705

Klasifikasi Dokumen dengan KNN

Data Training:

* D1

Teks: Sepakbola PSIS tahun ini semakin indah.

Class: Olahraga

* D2

Teks: Presiden Indonesia menaikan harga BBM.

Class: Politik

* D3

Teks: Partai politik Indonesia berburu suara.

Class: Politik

* D4

Teks: Timnas Indonesia gagal juara AFC.

Class: Olahraga

Data Testing:

* D5

Teks: Presiden menaikan gaji Timnas Indonesia.

Class: ?

Dokumen teks diatas diklasifikasikan ke dalam dua kelas yaitu Class 1 (Olahraga) = D1&D4 dan Class 2 (Politik) = D2&D3. Menggunakan k-NN, dengan nilai K=3, Tentukan kelas dari D5!

1. Hitung jarak antara D5 dengan setiap dokumen pelatihan menggunakan cosine similarity.

* Vektorisasi teks

D1: [1, 0] = Olahraga

D2: [0, 1] = Politik

D3: [0, 1] = Politik

D4: [1, 0] = Olahraga

D5: [1, 0] = ?

* Hitung cosine similarity

Cosine Similarity(D5, D1) =

Cosine Similarity(D5, D2) =

Cosine Similarity(D5, D3) =

Cosine Similarity(D5, D4) =

* Tentukan kelas untuk D5

D5 lebih mirip dengan D1 dan D4, karena cosine similarity tertinggi dengan keduanya. Oleh karena itu, kita dapat menentukan Class untuk D5: Olahraga.

1. Pilih K dokumen pelatihan dengan jarak terdekat dengan D5. Dalam hal ini, K=3.

D1: Cosine Similarity = 1

D4: Cosine Similarity = 1

D2: Cosine Similarity = 0

Jadi, K dokumen pelatihan dengan jarak terdekat dengan D5 adalah D1, D4, dan D2.

1. Tentukan kelas mayoritas dari K dokumen pelatihan tersebut.

D1: Kelas Olahraga

D4: Kelas Olahraga

D2: Kelas Politik

Kelas mayoritas dari K dokumen pelatihan tersebut adalah "Olahraga", karena dua dari tiga dokumen memiliki kelas Olahraga.

1. Tetapkan kelas mayoritas sebagai kelas dari D5.

Berdasarkan perhitungan sebelumnya, kelas mayoritas dari K dokumen pelatihan terdekat dengan D5 (D1, D4, dan D2) adalah "Olahraga". Oleh karena itu, kita dapat menetapkan kelas "Olahraga" sebagai kelas untuk D5.