Klasifikasi Teks Menggunakan k-NN
sebuah contoh

Lunix96 at {yahoo.com, gmail.com}

Diketahui terdapat 8 dokumen (D1 s.d D8) sebagai berikut:

- D1. Tokoh politik dari berbagai partai mengadakan rapat untuk membahas koalisi baru menjelang pemilu 2014 dan beberapa pilkada 2012 dan 2013.
- D2. Partai politik sudah tidak dapat dipercaya. Sebagian besar partai mengutamakan kepentingan partai daripada kebutuhan rakyat
- D3. Partai demokrat memenangkan pemilu 2009 karena figur SBY. Partai Golkar berusaha menang pada 2012. Pertandingan 2 partai ini akan seru.
- D4. Pertandingan pertama antara Persema dan Persebaya diadakan di Malang. Ini akan menguntungkan tuan rumah
- D5. Sebagian besar wasit di Indonesia sulit dipercaya. Beberapa pertandingan sepakbola sering tidak adil. Tim nasional perlu pembenahan Total.
- D6. Suap menyuap sudah lazim di negeri Ini. Pemilu ada suap. Pilkada juga suap. Mungkin pula saat Pilpres.
- D7. Beberapa pertandingan sepakbola yang dilakoni persebaya pada masa kampanye Pilkada 2010 Kota surabaya akan ditunda.
- D8. Sepakbola Indonesia memang belum bangkit. Manajemen tim, pertandingan dan tiket perlu ditingkatkan, bukan hanya fokus pada kemenangan tim.

Jika dokumen-dokumen teks tersebut dikelompokkan (*classification*) ke dalam dua kelas, C1 (Politik) dan C2 (Olahraga), menggunakan kecerdasan manusia, misalnya tiap kelas hanya boleh beranggotakan 3 dokumen, maka kita dapat memperoleh hasil sebagai berikut:

C1 akan beranggotakan D1, D2 dan D3

C2 akan beranggotakan D4, D7 dan D8

Pada k-NN, fase ini dinamakan fase manual atau training. Kita memilih beberapa dokumen contoh (*sample*) dan mengelompokkannya secara manual ke dalam kelas-kelas yang telah didefinisikan.

Pertanyaan. Menggunakan k-NN, tentukan kelas dari dokumen D5!

Langkah 1. *Preprocessing* terhadap semua (terdapat 7) dokumen yang terlibat, yaitu D5, D1, D2, D3, D4, D7 dan D8.

Langkah 1a: Lakukan tokenisasi, *stop words removal* dan *stemming*. Hasilnya diperlihatkan pada tabel berikut:

Dokumen	**Term yang mewakili dokumen**
D5	besar wasit indonesia sulit percaya tanding sepakbola adil tim nasional benah total
D1	tokoh politik partai rapat bahas koalisi baru jelang pemilu 2014 pilkada 2012 2013
D2	partai politik percaya besar partai utama penting partai butuh rakyat
D3	partai demokrat menang pemilu 2009 figur sby partai golkar usaha menang 2012 tanding partai seru
D4	tanding pertama persema persebaya malang untung rumah
D7	tanding sepakbola persebaya kampanye pilkada 2010 kota surabaya tunda
D8	sepakbola indonesia bangkit manajemen tim tanding tiket tingkat fokus menang tim

Langkah 1b. Tentukan bobot untuk setiap term dari 7 dokumen yang terlibat. Total dokumen ada 8. Dokumen yang telah terklasifikasi ada 6 dan yang akan diklasifikasikan (D5) sehingga total yang terlibat adalah 7. Dokuman D6 tidak dilibatkan, belum terklasifikasi dan dapat dijadikan obyek pada klasifikasi berikutnya.

	tf								idf	Wdt = tf.idf						
Term	D5	D1	D2	D3	D4	D7	D8	df	log(n/df)	D5	D1	D2	D3	D4	D7	D8
besar	1							1	0,845	0,845	0	0	0	0	0	0
wasit	1							1	0,845	0,845	0	0	0	0	0	0
indonesia	1						1	2	0,544	0,544	0	0	0	0	0	0,544
sulit	1							1	0,845	0,845	0	0	0	0	0	0
percaya	1		1					2	0,544	0,544	0	0,544	0	0	0	0
tanding	1			1	1	1	1	5	0,146	0,146	0	0	0,146	0,146	0,146	0
sepakbola	1					1	1	3	0,368	0,368	0	0	0	0	0,368	0,368
tokoh		1						1	0,845	0,000	0,845	0	0	0	0	0
adil	1							1	0,845	0,845	0	0	0	0	0	0
politik		1	1					2	0,544	0,000	0,544	0,544	0	0	0	0
benah	1							1	0,845	0,845	0	0	0	0	0	0
menang				2			1	3	0,368	0,000	0	0	0,736	0	0	0,368
tim	1						2	2	0,544	0,544	0	0	0	0	0	1,088
nasional	1							1	0,845	0,845	0	0	0	0	0	0
total	1							1	0,845	0,845	0	0	0	0	0	0
partai		1	3	3				3	0,368	0	0,368	1,104	1,104	0	0	0
rapat		1						1	0,845	0	0,845	0	0	0	0	0
bahas		1						1	0,845	0	0,845	0	0	0	0	0
koalisi		1						1	0,845	0	0,845	0	0	0	0	0
baru		1						1	0,845	0	0,845	0	0	0	0	0
jelang		1						1	0,845	0	0,845	0	0	0	0	0
pemilu		1		1				2	0,544	0	0,544	0	0,544	0	0	0
2014		1						1	0,845	0	0,845	0	0	0	0	0
pilkada		1				1		2	0,544	0	0,544	0	0	0	0,544	0
2012		1		1				2	0,544	0	0,544	0	0,544	0	0	0
2013		1						1	0,845	0	0,845	0	0	0	0	0
percaya			1					1	0,845	0	0	0,845	0	0	0	0
2009			1	1				2	0,544	0	0	0,544	0,544	0	0	0
besar			1					1	0,845	0	0	0,845	0	0	0	0
utama			1					1	0,845	0	0	0,845	0	0	0	0
penting			1					1	0,845	0	0	0,845	0	0	0	0
butuh			1					1	0,845	0	0	0,845	0	0	0	0
rakyat			1					1	0,845	0	0	0,845	0	0	0	0
demokrat				1				1	0,845	0	0	0	0,845	0	0	0
figur				1				1	0,845	0	0	0	0,845	0	0	0
sby				1				1	0,845	0	0	0	0,845	0,000	0,000	0,000
golkar				1				1	0,845	0	0	0	0,845	0	0	0
usaha				1				1	0,845	0	0	0	0,845	0	0	0
seru				1				1	0,845	0	0	0	0,845	0	0	0
pertama					1			1	0,845	0	0	0	0	0,845	0	0
persema					1			1	0,845	0	0	0	0	0,845	0	0
persebaya					1	1		2	0,544	0	0	0	0	0,544	0,544	0
malang					1			1	0,845	0	0	0	0	0,845	0	0
untung					1			1	0,845	0	0	0	0	0,845	0	0
rumah					1			1	0,845	0	0	0	0	0,845	0	0
kampanye						1		1	0,845	0	0	0	0	0	0,845	0
2010						1		1	0,845	0	0	0	0	0	0,845	0
kota						1		1	0,845	0	0	0	0	0	0,845	0
surabaya						1		1	0,845	0	0	0	0	0	0,845	0
tunda						1		1	0,845	0	0	0	0	0	0,845	0
bangkit							1	1	0,845	0	0	0	0	0	0	0,845
manajemen							1	1	0,845	0	0	0	0	0	0	0,845
tiket							1	1	0,845	0	0	0	0	0	0	0,845
tingkat							1	1	0,845	0	0	0	0	0	0	0,845
fokus							1	1	0,845	0	0	0	0	0	0	0,845

Langkah 2: Hitung kemiripan vektor dokumen D5 dengan setiap dokumen yang telah terklasifikasi (D1, D2, D3, D4, D7 dan D8). Kemiripan antar dokumen dapat menggunakan *cosine similarity*. Rumusnya adalah sebagai berikut:

$$\cos(\Theta_{ij}) = \frac{\sum_k (d_{ik} d_{jk})}{\sqrt{\sum_k d_{ik}^2} \sqrt{\sum_k d_{jk}^2}}$$

Langkah 2a: Hitung hasil perkalian skalar antara D5 dan 6 dokumen yang telah terklasifikasi. Hasilnya perkalian dari setiap dokumen dengan D5 dijumlahkan (sesuai pembilang pada rumus di atas)

Langkah 2b: Hitung panjang setiap dokumen, termasuk D5. Caranya, kuadratkan bobot setiap term dalam setiap dokumen, jumlahkan nilai kuadrat tersebut dan kemudian akarkan.

Sisi kiri dari tabel berikut ini mewakili langkah 2a dan sisi kanan memperlihatkan langkah 2b.

WD5*WDi						Panjang Vektor						
D1	D2	D3	D4	D7	D8	D5	D1	D2	D3	D4	D7	D8
0	0	0	0	0	0	0,714	0	0	0	0	0	0
0	0	0	0	0	0	0,714	0	0	0	0	0	0
0	0	0	0	0	0,296	0,296	0	0	0	0	0	0,296
0	0	0	0	0	0	0,714	0	0	0	0	0	0
0	0,296	0	0	0	0	0,296	0	0,296	0	0	0	0
0	0	0,02135	0,0214	0,0214	0,0214	0,021	0	0	0,021	0,021	0,021	0,021
0	0	0	0	0,1354	0,1354	0,135	0	0	0	0	0,135	0,135
0	0	0	0	0	0	0,000	0,714	0	0	0	0	0
0	0	0	0	0	0	0,714	0	0	0	0	0	0
0	0	0	0	0	0	0,000	0,296	0,296	0	0	0	0
0	0	0	0	0	0	0,714	0	0	0	0	0	0
0	0	0	0	0	0	0,000	0	0	0,542	0	0	0,135
0	0	0	0	0	0,592	0,296	0	0	0	0	0	1,184
0	0	0	0	0	0	0,714	0	0	0	0	0	0
0	0	0	0	0	0	0,714	0	0	0	0	0	0
0	0	0	0	0	0	0	0,135	1,219	1,219	0	0	0
0	0	0	0	0	0	0	0,714	0	0	0	0	0
0	0	0	0	0	0	0	0,714	0	0	0	0	0
0	0	0	0	0	0	0	0,714	0	0	0	0	0
0	0	0	0	0	0	0	0,714	0	0	0	0	0
0	0	0	0	0	0	0	0,714	0	0	0	0	0
0	0	0	0	0	0	0	0,296	0	0,296	0	0	0
0	0	0	0	0	0	0	0,714	0	0,000	0	0	0
0	0	0	0	0	0	0	0,296	0	0	0	0,296	0,000
0	0	0	0	0	0	0	0,296	0	0,296	0	0	0
0	0	0	0	0	0	0	0,714	0	0	0	0	0
0	0	0	0	0	0	0	0	0,714	0	0	0	0
0	0	0	0	0	0	0	0	0,296	0,296	0	0	0
0	0	0	0	0	0	0	0	0,714	0	0	0	0
0	0	0	0	0	0	0	0	0,714	0	0	0	0
0	0	0	0	0	0	0	0	0,714	0	0	0	0
0	0	0	0	0	0	0	0	0,714	0	0	0	0
0	0	0	0	0	0	0	0	0,714	0	0	0	0
0	0	0	0	0	0	0	0	0	0,714	0	0	0
0	0	0	0	0	0	0	0	0	0,714	0	0	0
0	0	0	0	0	0	0	0	0	0,714	0	0	0
0	0	0	0	0	0	0	0	0	0,714	0	0	0
0	0	0	0	0	0	0	0	0	0,714	0	0	0
0	0	0	0	0	0	0	0	0	0,714	0	0	0
0	0	0	0	0	0	0	0	0	0	0,714	0	0
0	0	0	0	0	0	0	0	0	0	0,714	0	0
0	0	0	0	0	0	0	0	0	0	0,296	0,296	0
0	0	0	0	0	0	0	0	0	0	0,714	0	0
0	0	0	0	0	0	0	0	0	0	0,714	0	0
0	0	0	0	0	0	0	0	0	0	0,714	0	0
0	0	0	0	0	0	0	0	0	0	0	0,714	0
0	0	0	0	0	0	0	0	0	0	0	0,714	0
0	0	0	0	0	0	0	0	0	0	0	0,714	0
0	0	0	0	0	0	0	0	0	0	0	0,714	0
0	0	0	0	0	0	0	0	0	0	0	0,714	0
0	0	0	0	0	0	0	0	0	0	0	0	0,714
0	0	0	0	0	0	0	0	0	0	0	0	0,714
0	0	0	0	0	0	0	0	0	0	0	0	0,714
0	0	0	0	0	0	0	0	0	0	0	0	0,714
0	0	0	0	0	0	0	0	0	0	0	0	0,714
0	**0,3**	**0,02**	**0,02**	**0,16**	**1,04**	**6,044**	**7,033**	**6,392**	**6,955**	**3,888**	**4,32**	**5,343**
						2,458	**2,652**	**2,528**	**2,637**	**1,972**	**2,078**	**2,312**

Langkah 2c: Terapkan rumus *cosine similarity*. Hitung kemiripan D5 dengan D1, D2 dan seterusnya sampai dengan D8.

Cos (D5, D1) = 0/(2,458*2,652) = **0,000**

Cos (D5, D2) = 0,3/(2,458*2,528) = **0,048**

Dan seterusnya.

Cos (D5, D8) = 1,04/(2,458*2,312) = **0,184**

Hasil perhitungan tersebut diperlihatkan tabel berikut:

D1	D2	D3	D4	D7	D8
0,000	0,048	0,003	0,004	0,031	0,184

Langkah 3: Urutkan hasil perhitungan kemiripan, diperoleh:

1	2	3	4	5	6
D8	D2	D7	D4	D3	D1

Langkah 4: Ambil sebanyak k (k=4) yang paling tinggi tingkat kemiripannya dengan D5 dan tentukan kelas dari D5. Hasilnya:

D8	D2	D7	D4

Dokumen D5 terklasifikasi ke dalam kelas? Pilih kelas yang paling banyak kemunculannya! Apakah C1? Atau C2? Ternyata, untuk k=4, C1 diwakili hanya oleh 1 dokumen yaitu D2, sedangkan C2, diwakili 3 dokumen, yaitu D8, D7 dan D4.

Kemanakah D5 berlabuh?

D5 terklasifikasi ke kelas C2 (Olahraga).

Kasus Khusus:

Bagaimana jika nilai cosim di atas seperti ini:

D1	D2	D3	D4	D7	D8
0,02	0,04	0,003	0,004	0,03	0,184

Jika di ambil 4 (k = 4) dokumen paling dekat dengan D5, diperoleh D8, D2, D1, D7. Kelas C1 dan C2, masing-masing diwakili oleh 2 dokumen. Ke kelas manakah D5 terklasifikasi? Pada kasus demikian, ada beberapa solusi yang dapat ditempuh, yaitu:

1. Kurangi atau tambahkan k (sebesar 1).

Jika k=3, maka D5 masuk ke kelas C2, diwakili oleh dokumen D8 dan D7.
Jika k=5, maka D5 masuk ke kelas C2, diwakili oleh dokumen D8, D7 dan D4.
Jika k=1, maka D5 masuk ke kelas C2, diwakili oleh hanya dokumen D8.

2. K tetap 4, tidak berubah. Jumlahkan tingkat kemiripan dari setiap dokumen untuk kelas yang sama. Diperoleh:
   - Nilai C1 = nilai kemiripan (D5, D1) + nilai kemiripan (D5, D2) = **0,035 + 0,04**
   - Nilai C2 = nilai kemiripan (D5, D8) + nilai kemiripan (D5, D7) = **0,184 + 0,03**

   Nilai C2 lebih besar. D5 harus masuk C2

Kesimpulan:

So, D5 masuk ke dalam C1 atau C2? Jika melihat isi dari D5 maka kita dapat memutuskan bahwa isinya terkait erat dengan olahraga dan harusnya masuk ke dalam C2 (Olahraga), bukan C1, meskipun D5 juga mengandung term-term yang berhubungan dengan Politik (C1).

Pada banyak kasus, jika pemilihan awal (k=4) tidak memberikan solusi klasifikasi, maka dilakukan pengurangan atau penambahan k (sebesar satu). Pada pendekatan ini (sebagaimana di atas), D5 terklasifikasi ke dalam C2 (Olahraga). Pada banyak penelitian, nilai k adalah 3, 4 atau 5, dan terbukti memberikan hasil yang lebih baik.