Cosine Similarity Antar Dokumen
Sebuah Contoh

Lunix96 at {yahoo.com, gmail.com}

Diketahui terdapat 6 dokumen (D1 s.d D6) 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. Beberapa pertandingan sepakbola yang dilakoni persebaya pada masa kampanye Pilkada 2010 Kota surabaya akan ditunda.
- D6. Sepakbola Indonesia memang belum bangkit. Manajemen tim, pertandingan dan tiket perlu ditingkatkan, bukan hanya fokus pada kemenangan tim.

Pertanyaan. Jika terdapat query (Q): “menang pertandingan”, tentukan daftar dokumen yang paling relevan dengan Query tersebut!.

Langkah 1. *Preprocessing* terhadap semua (n= 7) dokumen yang terlibat, yaitu Q, D1, D2, D3, D4, D5 dan D6.

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

Dokumen	**Term yang mewakili dokumen**
Q	Menang tanding
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
D5	tanding sepakbola persebaya kampanye pilkada 2010 kota surabaya tunda
D6	sepakbola indonesia bangkit manajemen tim tanding tiket tingkat fokus menang tim

Langkah 1b. Tentukan bobot untuk setiap term dari 7 dokumen tersebut.

	tf								*idf*	*Wdt = tf.idf*						
Term	**Q**	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	***df***	***log(n/df)***	**Q**	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**
indonesia							1	1	0,845	0	0	0	0	0	0	0,845
percaya			1					1	0,845	0	0	0,845	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	2	0,544	0	0	0	0	0	0,544	0,544
tokoh		1						1	0,845	0	0,845	0	0	0	0	0
politik		1	1					2	0,544	0	0,544	0,544	0	0	0	0
menang	1			2			1	3	0,368	0,368	0	0	0,736	0	0	0,368
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	0	0
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] query Q dengan setiap dokumen yang ada. 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 Q dan 6 dokumen lain. Hasilnya perkalian dari setiap dokumen dengan Q dijumlahkan (sesuai pembilang pada rumus di atas)

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

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

WD5*WDi						Panjang Vektor						
D1	D2	D3	D4	D5	D6	Q	D1	D2	D3	D4	D5	D6
0	0	0	0	0	0	0,000	0	0	0	0	0	0,714
0	0	0	0	0	0	0,000	0	1	0	0	0	0
0	0	0,021353	0,02135	0,02135	0,0214	0,021	0	0	0,021	0,021	0,021	0,021
0	0	0	0	0	0	0,000	0	0	0	0	0,296	0,296
0	0	0	0	0	0	0,000	0,714	0	0	0	0	0
0	0	0	0	0	0	0,000	0,296	0,296	0	0	0	0
0	0	0,270814	0	0	0,1354	0,135	0	0	0,542	0	0	0,135
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
0,000	**0,000**	**0,292**	**0,021**	**0,021**	**0,157**	**0,157**	**7,033**	**6,81**	**6,955**	**3,888**	**4,48**	**4,738**
						0,396	**2,652**	**2,610**	**2,637**	**1,972**	**2,117**	**2,177**

Langkah 2c: Terapkan rumus *cosine similarity*. Hitung kemiripan Q dengan D1, D2 dan seterusnya sampai dengan D6.

Cos (Q, D1) = 0/(0,396*2,652) = **0**

Cos (Q, D4) = 0,021/(0,396*1,972) = **0,0274**

Dan seterusnya.

Cos (Q, D6) = 0,157/(0,396*2,177) = **0,1819**

Hasil perhitungan tersebut diperlihatkan tabel berikut:

D1	D2	D3	D4	D5	D6
0	**0**	**0,27981**	**0,0274**	**0,0255**	**0,1819**

Langkah 3: Urutkan hasil perhitungan kemiripan, diperoleh:

1	2	3	4	5	6
D3	**D6**	**D4**	**D5**	**D1**	**D2**

Benarkah Query relevan dengan dokumen D3? Perhatikan secara seksama. Ternyata? Silakan anda simpulkan sendiri! Jika “menang tanding” yang dimaksud adalah terkait perlombaan partai politik maka D3 memang layak sebagai jawaban tetapi jika yang dimaksudkan adalah pertandingan olah raga, maka D3 harusnya tidak hadir, apalagi pada urutan pertama.

Berapa jumlah total dokumen yang relevan dengan Query tersebut (anggap “menang tanding” pada olahraga)? Terdapat 3 dokumen, yaitu D4, D5, dan D6.

Berapa nilai recall-nya jika keempat dokumen tersebut diserahkan kepada pengguna?

Recall = 3/3 x 100 % = 100%

Bagaimana dengan presisi? Jika dikembalikan hanya 1 dokumen, maka presisi = 0. Jika diberikan kepada pengguna keempat dokumen tersebut, padahal hanya 3 dokumenyang relevan, maka:

Presisi = 3/4 x 100 % = 75 %

Recall dan Presisi ini digunakan sebagai parameter untuk mengetahui bagus dan tidaknya hasil pemrosesan terhadap query yang dilakukan oleh sistem temu balik informasi.

Semoga bermanfaat ☺