Contoh Perhitungan Kemiripan Cosinus pada Model Ruang Vektor

Persoalan 1:

Ada 4 dokumen (D1 s.d D4):

D1: dolar naik harga naik penghasilan turun
D2: harga naik harusnya gaji juga naik
D3: Premium tidak terpengaruh dolar
D4: harga laptop naik

Dan ada Query (Q): kenaikan harga

Tugas: hitung ranking kemiripan Q dengan semua dokumen dan urutkan! Gunakan cosine similarity (cosim) dalam model ruang vektor (VSM).

Jawaban:

Ada beberapa langkah yang harus dilakukan untuk mendapatkan dokumen yang paling mirip dengan Query (teranking), yaitu:

1. Lakukan tokenisasi untuk mengambil setiap kata yang ada di dalam dokumen. Kata yang diambil tersebut kemudian dinamakan sebagai token.
2. Setelah tokenisasi, ada banyak perlakukan yang harus diterapkan terhadap token. Ini mencakup penyesuaian singkatan (misal: MU menjadi Manchester United), perubahan token ke huruf kecil, lematisasi dan stemming (biasanya cukup dinamakan stemming), dan stop word removal. Token berubah menjadi Term.
3. Membuat Inverted Index, yaitu suatu daftar (Indeks) yang dengan jelas menunjukkan dimana keberadaan dari suatu term.
4. Dikarenakan beberapa hal (lihat di slide dan di buku IIR), setiap term perlu diberikan bobot tertentu. Skema pembobotan paling umum adalah tf.idf. tf adalah jumlah kemunculan suatu term dalam suatu dokumen (jadi tf dari suatu term t di dalam suatu dokumen dapat berbeda dengan tf term t pada dokumen lain). df adalah jumlah dokumen yang di dalamnya terdapat suatu term t. Artinya, df adalah tergantung pada jumlah dokumen yang dikoleksi. Idf adalah inverse dari df. Nilai idf dari suatu term t didefinisikan oleh ruus berikut:

   $idf_t = \log_2 (N/df_t)$

   Dimana N adalah jumlah total dokumen dalam koleksi (tidak termasuk Query Q).

5. Menghitung panjang vektor dari setiap dokumen. Panjang vektor dari dokumen D1, ditulis |D1| adalah akar dari penjumlahan bobot kuadrat dari setiap term yang hadir di dalam dokumen D1.

6. Menghitung kemiripan Query antara Q dengan setiap dokumen (D1 s.d D4) menggunakan cosine similarity. Rumusan sederhananya adalah

   CoSim (Q, D) = penjumlahan hasil perkalian bobot dari term yang berisan antara Q dan D dibagi perkalian panjang vektor dari Q dan D.

   Rumus dalam bentuk notasi matematis formalnya dapat dilihat pada slide kuliah dan buku IIR (bab 6).

7. Mengurutkan hail perhitungan di atas secara *descending*. Cosim terbesar harus mendapatkan ranking tertinggi.

Kita mulai langkah penyelesaian persoalan di atas secara detail:

1. Tokenisasi
2. Stemming dkk.
3. Indexing

Ketiga langkah di atas harus dilakukan secara urut terhadap dokumen, satu demi satu dokumen.

Terhadap D1, diperoleh token

dolar naik harga naik penghasilan turun

Setelah dilakukan Stemming (dkk.) diperoleh 5 Term

dolar naik harga hasil turun

Kelima term tersebut harus dimasukkan ke dalam Indeks. Bentuk inverted index yang banyak dipakai pada tugas akhir terdiri dari Field Id, Term, DokID, tf dan tf-idf:

ID	Term	DokID	Tf	Tf-idf
1	Dolar	1	1	
2	Naik	1	2	
3	Harga	1	1	
4	Hasil	1	1	
5	turun	1	1	

Ulangi ketiga langkah di atas terhadap dokumen D2, D3 dan D4.

Akhir dari tiga proses di atas terhadap 4 dokumen adalah suatu inverted index berikut:

ID	Term	DokID	Tf	Tf-idf
1	dolar	1	1	
2	naik	1	2	
3	harga	1	1	
4	hasil	1	1	
5	turun	1	1	
6	harga	2	1	
7	naik	2	2	
8	gaji	2	1	
9	premium	3	1	
10	pengaruh	3	1	
11	dolar	3	1	
12	harga	4	1	
13	laptop	4	1	
14	naik	4	1	

4. Pembobotan (weighting)

Pertama, hitung idf untuk setiap term yang ada di dalam indeks (N=4), yaitu dolar, naik, harga, hasil, turun, gaji, premium, pengaruh, laptop.

Idf (dolar) = $\log_2(N/df_{dolar}) = \log_2 (4/2) = \log_2 (2) = 1$
Term dolar hadir dalam 2 dokumen, yaitu D1 dan D3

Idf (naik) = $\log_2(N/df_{naik}) = \log_2 (4/3) = 0,41$
Term naik hadir dalam 3 dokumen, yaitu D1, D2 dan D4.

Idf (harga) = $\log_2(N/df_{harga}) = \log_2 (4/3) = 0,41$
Idf (hasil) = $\log_2(N/df_{hasil}) = \log_2 (4/1) = \log_2 (4) = 2$
Term hasil muncul di dalam 1 dokumen saja, yaitu D1

Idf (turun) = $\log_2(N/df_{turun}) = \log_2 (4/1) = \log_2 (4) = 2$
Idf (gaji) = $\log_2(N/df_{gaji}) = \log_2 (4/1) = \log_2 (4) = 2$
Idf (premium) = $\log_2(N/df_{premium}) = \log_2 (4/1) = \log_2 (4) = 2$
Idf (pengaruh) = $\log_2(N/df_{pengaruh}) = \log_2 (4/1) = \log_2 (4) = 2$
Idf (laptop) = $\log_2(N/df_{laptop}) = \log_2 (4/1) = \log_2 (4) = 2$

Kemudian melakukan perhitungan bobot (tf-idf) terhadap setiap term untuk setiap dokumen yang hadir di dalam indeks:

tf.idf (dolar dalam D1) = $tf_{dolar\ (dalam\ D1)} * idf_{dolar\ (dalam\ koleksi)} = 1 * 1 = 1$
tf.idf (naik dalam D1) = $tf_{naik\ (dalam\ D1)} * idf_{naik\ (dalam\ koleksi)} = 2 * 0,41 = 0,82$
tf.idf (harga dalam D1) = $tf_{harga\ (dalam\ D1)} * idf_{harga\ (dalam\ koleksi)} = 1 * 0,41 = 0,41$
tf.idf (hasil dalam D1) = $tf_{hasil\ (dalam\ D1)} * idf_{hasil\ (dalam\ koleksi)} = 1 * 2 = 2$
tf.idf (turun dalam D1) = $tf_{turun\ (dalam\ D1)} * idf_{turun\ (dalam\ koleksi)} = 1 * 2 = 2$

Silakan hitung tf.idf untuk term-term lain dalam indeks. Hasilnya, digunakan untuk mengupdate filed tf-idf pada tabel indeks di atas. Sehingga indeks sekarang menjadi:

ID	Term	DokID	Tf	Tf-idf
1	dolar	1	1	1
2	naik	1	2	0,82
3	harga	1	1	0,41
4	hasil	1	1	2
5	turun	1	1	2
6	harga	2	1	0,41
7	naik	2	2	0,82
8	gaji	2	1	2
9	premium	3	1	2
10	pengaruh	3	1	2
11	dolar	3	1	1
12	harga	4	1	0,41
13	laptop	4	1	2
14	naik	4	1	0,41

Terlihat jelas, bahwa semakin jarang dokumen mengandung term t maka akan semakin tinggi nilai bobotnya. Semakin sering term t hadir di dalam suatu dokumen, maka semakin tinggi pula bobotnya di dalam dokumen tersebut.

5. Hitung pajang vektor

Panjang vektor dari dokumen D1 adalah akar dari penjumlahan kuadrat dari bobot setiap term di dalam D1 tersebut.

$|D1| = \text{akar } (1^2 + 0{,}82^2 + 0{,}41^2 + 2^2 + 2^2) = 3{,}14$
$|D2| = \text{akar } (0{,}41^2 + 0{,}82^2 + 2^2) = 2{,}2$
$|D3| = \text{akar } (2^2 + 2^2 + 1^2) = 2{,}23$
$|D4| = \text{akar } (0{,}41^2 + 2^2 + 0{,}41^2) = 2{,}1$

6. Hitung cosim

Sebelum dapat menghitung kemiripan cosinus Query dengan setiap dokumen, maka perlu dihitung tf.idf untuk setiap term dalam Query Q. Setelah melewati fase tokenisasi, stemming dan kawan-kawan, diperoleh 2 term dari Query, yaitu naik dan harga.

tf.idf (naik dalam Q) = $tf_{\text{naik (dalam Query)}} * idf_{\text{naik (dalam koleksi)}} = 1 * 0{,}41 = 0{,}41$
tf.idf (harga dalam Q) = $tf_{\text{harga (dalam Query)}} * idf_{\text{harga (dalam koleksi)}} = 1 * 0{,}41 = 0{,}41$

Sehingga panjang vektor dari Query Q adalah

$|Q| = \text{akar}(0{,}41^2 + 0{,}41^2) = 0{,}58$

Sekarang hitung cosim Q dengan keempat dokumen yang ada dalam koleksi. Ingat, hanya lihat term yang beririsan.

Cosim (Q, D1) = penjumlahan hasil perkalian bobot term naik dan harga dibagi perkalian panjang vektor

$\text{Cosim}(Q, D1) = (\text{tf.idf}_{\text{naik dalam Q}} * \text{tf.idf}_{\text{naik dalam D1}}) + (\text{tf.idf}_{\text{harga dalam Q}} * \text{tf.idf}_{\text{harga dalam D1}})/(|Q| * |D1|)$
$\text{Cosim}(Q, D1) = (0{,}41 * 0{,}82 + 0{,}41 * 0{,}41) / (0{,}58 * 3{,}14) = 0{,}28$

Lakukan perhitungan cosim antara Q dengan D2, D3 dan D4.

$\text{Cosim}(Q, D2) = (\text{tf.idf}_{\text{naik dalam Q}} * \text{tf.idf}_{\text{naik dalam D2}}) + (\text{tf.idf}_{\text{harga dalam Q}} * \text{tf.idf}_{\text{harga dalam D2}})/(|Q| * |D2|)$
$\text{Cosim}(Q, D2) = (0{,}41 * 0{,}82 + 0{,}41 * 0{,}41) / (0{,}58 * 2{,}2) = 0{,}4$

$\text{Cosim}(Q, D4) = (\text{tf.idf}_{\text{naik dalam Q}} * \text{tf.idf}_{\text{naik dalam D4}}) + (\text{tf.idf}_{\text{harga dalam Q}} * \text{tf.idf}_{\text{harga dalam D4}})/(|Q| * |D4|)$
$\text{Cosim}(Q, D4) = (0{,}41 * 0{,}41 + 0{,}41 * 0{,}41) / (0{,}58 * 2{,}23) = 0{,}26$

Cosim(Q, D3)? Tidak ada term yang beririsan, sehingga tidak ada kemiripan, alias kemiripan antara Q dena D3 bernilai 0.

7. Ranking

Urutkan secara descending, dari tinggi ke rendah, diperoleh ranking: **D2 - D1 - D4**. Bagaimana dengan D3, apakah dimasukkan ke dalam ranking (ranking 4) dan dikembalikan kepada pengguna (pada sistem IR sungguhan)? TIDAK. Karena kemiripan antara Q dan D3 bernilai 0 maka tidak boleh dikembalikan kepada pengguna, tidak masuk daftar ranking.

MOHON PERIKSA KEMBALI PERHITUNGAN DI ATAS, KESALAHAN HITUNG DAPAT TERJADI. CERMAT DAN RE-CHECK HASIL PERHITUNGAN ANDA.

Persoalan 2:

Suatu koleksi C berisi 3 dokumen berikut:

d1: “berita jawa yang terbaru”
d2: “baru se-jawa di surabaya”
d3: “beritanya tentang kuliah di kampusku”

Ada Query (Q): baru berita terbaru

Berikan daftar dokumen yang paling mirip dengan Query
Gunakan cosine similarity (cosim) dalam model ruang vektor (VSM)

Jawaban Ringkas (detailnya lihat catatan perkuliahan):

Tahap Tokenisasi, Stemming dan Indexing, diperoleh inverted index berikut:

ID	Term	DokID	Tf	Tf-idf
1	berita	1	1	
2	jawa	1	1	
3	baru	1	1	
4	baru	2	1	
5	jawa	2	1	
6	surabaya	2	1	
7	berita	3	1	
8	kuliah	3	1	
9	kampus	3	1	

Hitung idf untuk setiap term yang hadir dalam Indeks, yaitu (N=3):

Idf (berita) $= \log_2 (3/2) = 0,6$
Idf (jawa) $= \log_2 (3/2) = 0,6$
Idf (baru) $= \log_2 (3/2) = 0,6$
Idf (surabaya) $= \log_2 (3/1) = 1,6$
Idf (kuliah) $= \log_2 (3/1) = 1,6$
Idf (kampus) $= \log_2 (3/1) = 1,6$

Hitung tf.idf untuk setiap term yang hadir dalam inverted index:

tf.idf (berita dalam D1) = 1 * 0,6 = 0,6

Silakan hitung tf.idf untuk term yang lain!

Diperoleh indeks berikut:

ID	Term	DokID	Tf	Tf-idf
1	berita	1	1	0,6
2	jawa	1	1	0,6
3	baru	1	1	0,6
4	baru	2	1	0,6
5	jawa	2	1	0,6
6	surabaya	2	1	1,6
7	berita	3	1	0,6
8	kuliah	3	1	1,6
9	kampus	3	1	1,6

Hitung panjang vektor dari setiap dokumen:

$|D1| = \text{akar}\ (0{,}6^2 + 0{,}6^2 + 0{,}6^2) = 1{,}04$
$|D2| = \text{akar}\ (0{,}6^2 + 0{,}6^2 + 1{,}6^2) = 1{,}81$
$|D3| = \text{akar}\ (0{,}6^2 + 1{,}6^2 + 1{,}6^2) = 2{,}34$

Sebelum menghitung kemiripan Q dengan D1 s.d D3, harus dihitung terlebih dahulu tf.idf setiap term dalam Q (2 term baru dan 1 term berita) serta panjang vektornya.

Tf.idf (baru dalam Q) = 2 * 0,6 = 1,2
Tf.idf (berita dalam Q) = 1 * 0,6 = 0,6

Sehingga panjang vektor dari Q:

$|Q| = \text{akar}\ (1{,}2^2 + 0{,}6^2) = 1{,}34$

Kemiripan antara Q dan (D1, D2 dan D3):

$\text{Cosim}(Q, D1) = (\text{tf.idf}_{\text{baru dalam Q}} * \text{tf.idf}_{\text{baru dalam D1}} + \text{tf.idf}_{\text{berita dalam Q}} * \text{tf.idf}_{\text{berita dalam D1}})/(|Q|*|D1|)$
Cosim(Q, D1) = (1,2*0,6 + 0,6 * 0,6) / (1,34 * 1,04) = 0,77
Cosim(Q, D2) = (1,2*0,6) / (1,34 * 1,81) = 0,30 ← hanya term baru
Cosim(Q, D3) = (0,6 * 0,6) / (1,34 * 2,34) = 0,11 ← hanya term berita

Ranking: **D1 – D2 – D3**

MOHON PERIKSA KEMBALI PERHITUNGAN DI ATAS, KESALAHAN HITUNG DAPAT TERJADI. CERMAT DAN RE-CHECK HASIL PERHITUNGAN ANDA.