

**Seri bahan kuliah Algeo #12**

# Aplikasi *Dot Product* pada Sistem Temu-balik Informasi (*Information Retrieval System*)

Bahan kuliah IF2123 Aljabar Linier dan Geometri

Oleh: Rinaldi Munir

**Program Studi Teknik Informatika  
STEI-ITB**

# Temu-balik Informasi

- **Temu-balik informasi (*information retrieval*)**: menemukan kembali (*retrieval*) informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis.

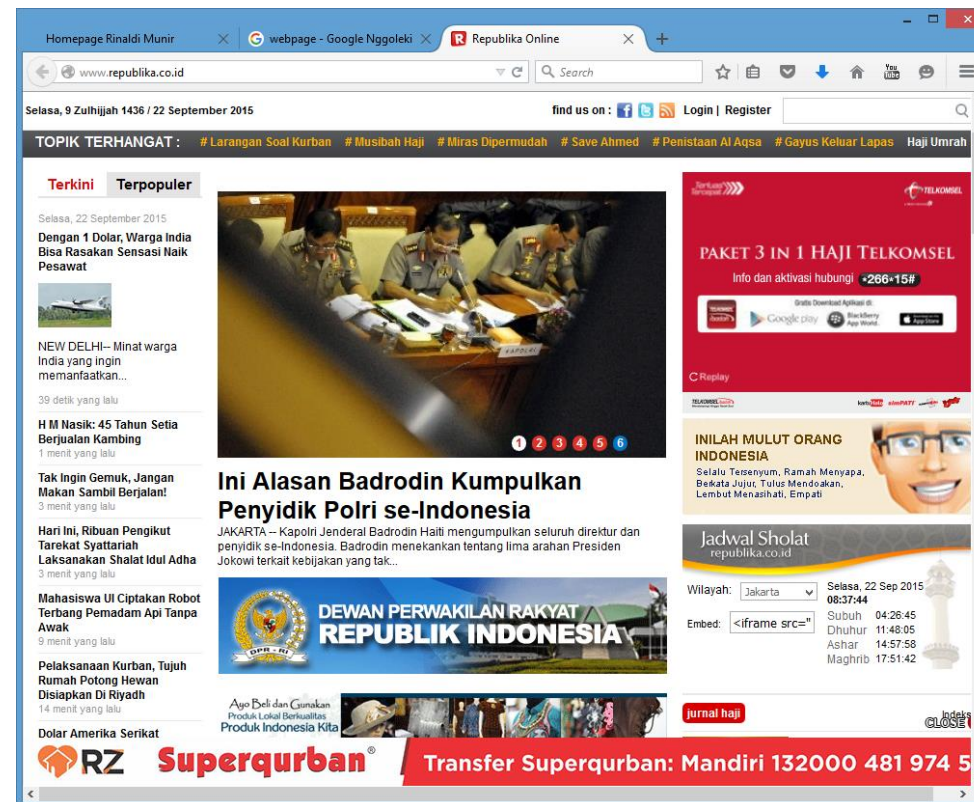


Sumber gambar: <https://sites.google.com/site/berbagiinformasidanekspresi/arsip/pengantar-temu-kembali-informasi-information-retrieval>

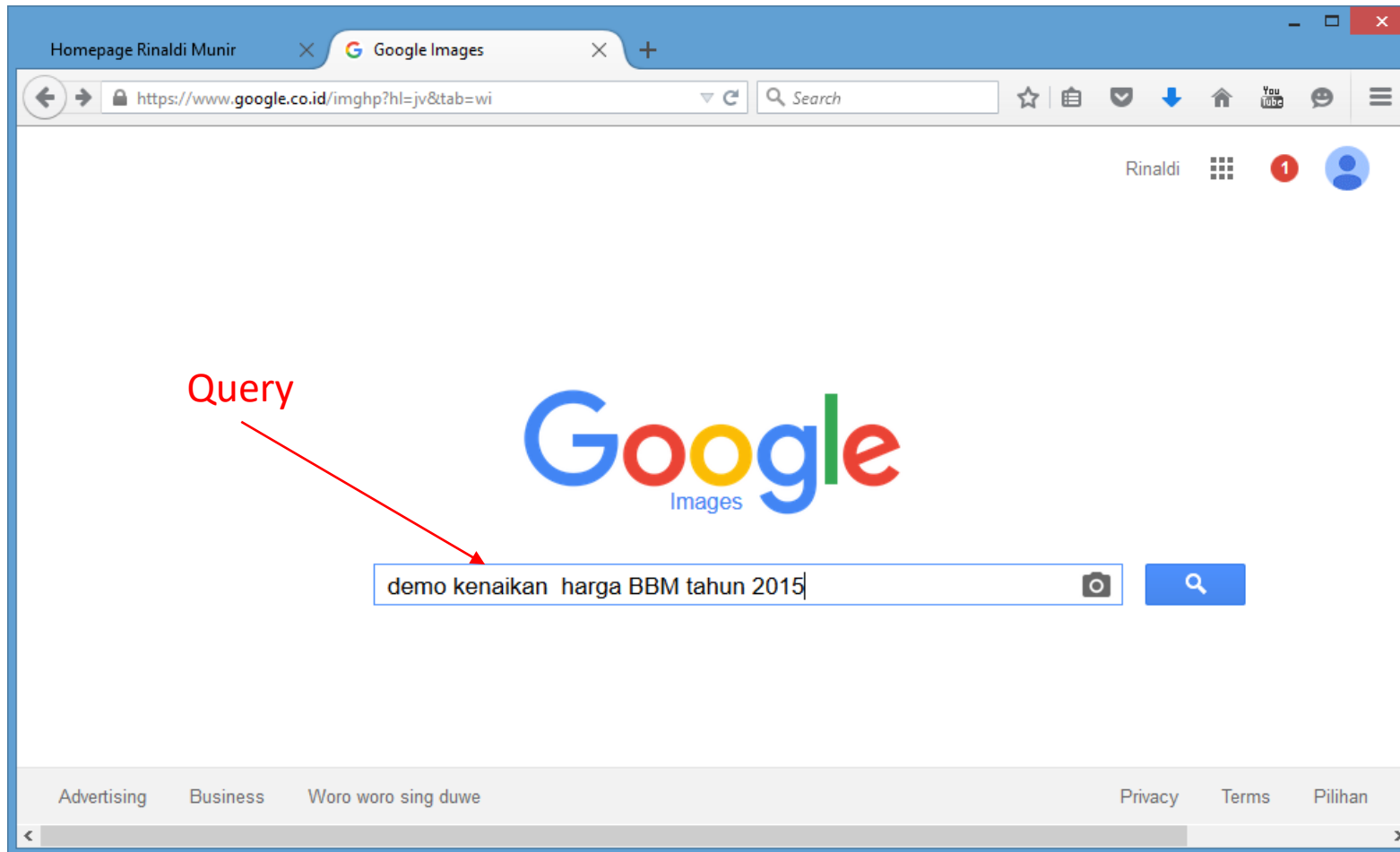
- IR tidak sama dengan pencarian di dalam basisdata (*database*)
- IR umumnya digunakan pada pencarian informasi yang isinya tidak terstruktur
- Informasi terstruktur: tabel-tabel di dalam basisdata (*database*)

Tabel mahasiswa						
NO	NAMA	NIM	JENIS KELAMIN	Umur	Tahun Lahir	Asal
1	Yusuf R	10018149	L	18	1992	Jogja
2	Lukman Reza	10018148	L	18	1992	Sulawesi
3	Aril	10018154	L	18	1992	Sumatra
4	Kifli	10018156	L	18	1992	Jogja
5	Khairuddin	10018151	L	18	1992	Papua
6	Angga	10018181	L	18	1992	Wonosobo
7	Nely	10018170	P	18	1992	Jogja
8	Reza	10018129	L	18	1992	Jogja
9	Ana	10017213	P	20	1990	Jogja
10	Nina	10012312	P	19	1991	Jogja

- Informasi tak-terstruktur:
  - dokumen (isinya bergantung pembuatnya)
  - laman web (*webpage*)



- Aplikasi IR: *search engine*



Hasil pencarian:

The screenshot shows a Google search results page for the query "demo kenaikan harga BBM tahun 2015". The browser's address bar shows the URL "https://www.google.co.id/search?q=demo+kenaikan++harga+BBM+tahun+2015". The search bar contains the same query. Below the search bar, there are tabs for "Web", "Gambar - gambar", "Videos", and "More". The search results are displayed below, starting with "About 663000 results (0.37 seconds)".

The first result is titled "Kenaikan Harga BBM Disambut Aksi Demo - Nasional" from nasional.sindonews.com. The snippet mentions that on Saturday (28/3/2015) at 00:00, a demonstration was held in response to the increase in the price of gasoline (BBM), with some companies at the beginning of the year considering a wage increase.

The second result is titled "Kenaikan harga BBM - Nasional - SINDOnews" from nasional.sindonews.com. The snippet mentions that on Saturday (28/3/2015) at 00:40 WIB, the price of gasoline (BBM) increased. It also mentions a decrease in the price of BBM being considered as a gift at the beginning of the year.

Below the text results, there is a section titled "Images for demo kenaikan harga BBM tahun 2015" with a "Lapurna gambar" link. This section contains four small images: a hand holding a pen over a document, a gas pump nozzle, a banner with the text "SUBSIDY UNTUK RAKYAT TURUNKAN HARGA BAHAN POKOK & LAIN", and a group of people at night.

Below the images, there is a link "More images for demo kenaikan harga BBM tahun 2015".

The third result is titled "Mahasiswa Makassar Demo Tolak Kenaikan Harga BBM ..." from news.okezone.com. The snippet mentions that on Thursday (2 April 2015) at 15:33 WIB, there was a demonstration in Makassar against the increase in the price of gasoline (BBM) by about 100 students from the Makassar Student Union (Himpunan Mahasiswa Makassar) in response to the DPR's approval of the 2016 Budget, Education, and Social Security (Pagu Anggaran Sesneg, Seskab dan KSP Tahun 2016) on Tuesday (22 March 2015).

Homepage Rinal... Kotak Masu... (3) Rinaldi ... Zimbra: Inb... 'Kalau PKS P... Catatanku | I... demo k...

https://www.google.co.id/?gws\_rd=ssl#q=demo+kenaikan+harga+BBM+tahun Search

**Kenaikan Harga BBM | Tempo.co**  
[www.tempo.co/topik/.../2546/Kenaikan-Harga-BBM](http://www.tempo.co/topik/.../2546/Kenaikan-Harga-BBM) ▼ Terjemahke koco iki  
Kenaikan harga BBM juga bisa ditetapkan sebelum April 2015. Sesuai dengan aturan, bisa .... Meleset dari Target. Tekanan inflasi mulai mereda tahun depan. .... PPP Jawa Tengah Dukung Demo Tolak Harga BBM Naik. Menurut Fraksi PPP ...

**Pemerintah Tutup Mata Dengan Harga BBM Picu Kenaikan ...**  
[www.rmol.co/.../2015/.../Pemerintah-Tutup-Mata-D...](http://www.rmol.co/.../2015/.../Pemerintah-Tutup-Mata-D...) ▼ Terjemahke koco iki  
Apr 5, 2015 - Sedang Konsolidasi, Ribuan Buruh Akan Gelar Demo ... "Ribuan buruh akan turun pada Kamis 9 April 2015. ... Selain menolak kenaikan harga BBM, aksi tersebut juga menuntut kenaikan upah 32 persen, ... seperti kenaikan harga BBM yang sesuai harga pasar, upah murah, dan naiknya lima tahun sekali.

**harga bbm naik - Liputan6.com**  
[www.liputan6.com/tag/harga-bbm-naik](http://www.liputan6.com/tag/harga-bbm-naik) ▼ Terjemahke koco iki  
Berita tag harga bbm naik - Harga Pertamina naik Rp 150 per liter menjadi Rp 9.450 per liter. ... Jokowi Minta Relawan Dukung Kenaikan Harga BBM ... Perubahan terhitung mulai 15 Mei 2015 pukul 00.00 WIB untuk wilayah Jakarta dan Jawa ...

**BBM Naik, Mahasiswa Demo Aksi Bakar Ban - Tribun ...**  
[pekanbaru.tribunnews.com](http://pekanbaru.tribunnews.com) > Nasional ▼ Terjemahke koco iki  
Mar 28, 2015 - Kenaikan harga BBM semakin menyengsarakan rakyat miskin. ... rakyat justru malah sebaliknya," ujar Arman, Koordinator Umum Aksi, Sabtu (28/3/2015). Kenaikan ... Jual Vespa exel Tahun 1996 lokasi Kepulauan Riau.

**Go**oooooooooooo**gle** >  
1 2 3 4 5 6 7 8 9 10 terusane

Help Send feedback Privacy Terms

# IR dengan Model Ruang Vektor

- Salah satu model IR adalah **model ruang vektor**
- Model ini menggunakan teori di dalam aljabar vector
- Misalkan terdapat  $n$  kata berbeda sebagai kamus kata (*vocabulary*) atau indeks kata (*term index*).
- Kata-kata tersebut membentuk ruang vektor berdimensi  $n$
- Setiap dokumen maupun *query* dinyatakan sebagai vektor  $\mathbf{w} = (w_1, w_2, \dots, w_n)$  di dalam  $\mathbf{R}^n$ .
- $w_i$  = bobot setiap kata  $i$  di dalam *query* atau dokumen
- Nilai  $w_i$  dapat menyatakan jumlah kemunculan kata tersebut dalam dokumen (*term frequency*)



**Contoh:** Misalkan terdapat tiga buah kata ( $T_1$ ,  $T_2$ , dan  $T_3$ ), dua buah dokumen ( $D_1$  dan  $D_2$ ) serta sebuah *query*  $Q$ . Masing-masing dinyatakan sebagai vector:

$$\mathbf{D}_1 = (2, 3, 5), \quad \mathbf{D}_2 = (3, 7, 1), \quad \mathbf{Q} = (0, 0, 2)$$

$\mathbf{D}_1 = (2, 3, 5)$  artinya dokumen  $D_1$  mengandung 2 buah kata  $T_1$ , 3 buah kata  $T_2$ , dan 5 buah kata  $T_3$ .

$\mathbf{D}_2 = (3, 7, 1)$  artinya dokumen  $D_2$  mengandung 3 buah kata  $T_1$ , 7 buah kata  $T_2$ , dan satu buah kata  $T_3$ .

$\mathbf{Q} = (0, 0, 2)$  artinya *query*  $Q$  hanya mengandung 2 buah kata  $T_3$ .

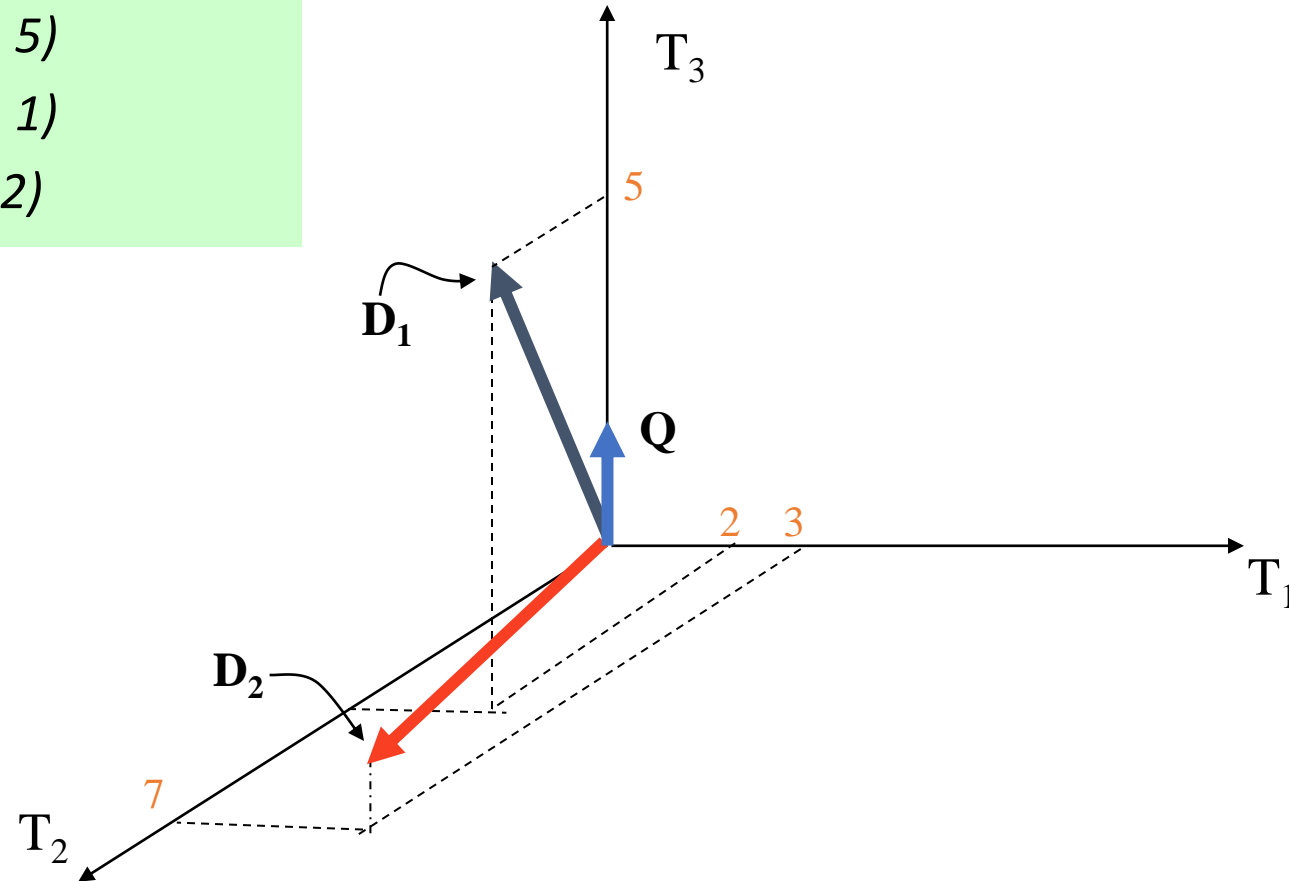
# Representasi grafik vektor

Contoh:

$$\mathbf{D}_1 = (2, 3, 5)$$

$$\mathbf{D}_2 = (3, 7, 1)$$

$$\mathbf{Q} = (0, 0, 2)$$



- Penentuan dokumen mana yang relevan dengan *query* dipandang sebagai pengukuran kesamaan (*similarity measure*) antara *query* dengan dokumen.
- Semakin sama suatu vektor dokumen dengan vektor *query*, semakin relevan dokumen tersebut dengan *query*.
- Kesamaan (*sim*) antara dua vektor  $\mathbf{Q} = (q_1, q_2, \dots, q_n)$  dan  $\mathbf{D} = (d_1, d_2, \dots, d_n)$  diukur dengan rumus *cosinus similarity* yang merupakan bagian dari rumus perkalian titik (*dot product*) dua buah vektor:

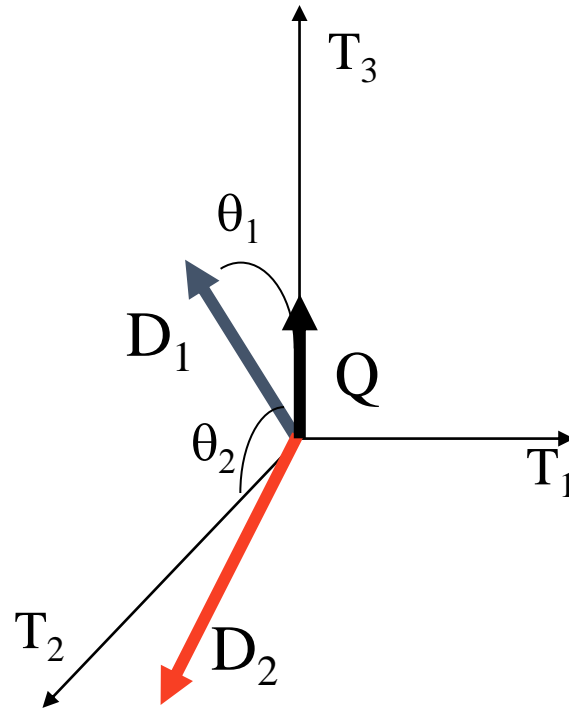
$$\mathbf{Q} \cdot \mathbf{D} = \|\mathbf{Q}\| \|\mathbf{D}\| \cos \theta$$



$$\text{sim}(\mathbf{Q}, \mathbf{D}) = \cos \theta = \frac{\mathbf{Q} \cdot \mathbf{D}}{\|\mathbf{Q}\| \|\mathbf{D}\|}$$

dengan  $\mathbf{Q} \cdot \mathbf{D}$  adalah perkalian titik yang didefinisikan sebagai

$$\mathbf{Q} \cdot \mathbf{D} = q_1 d_1 + q_2 d_2 + \dots + q_n d_n$$



$$\text{sim}(\mathbf{Q}, \mathbf{D}) = \cos \theta = \frac{\mathbf{Q} \cdot \mathbf{D}}{\|\mathbf{Q}\| \|\mathbf{D}\|}$$

- Jika  $\cos \theta = 1$ , berarti  $\theta = 0$ , vektor  $\mathbf{Q}$  dan  $\mathbf{D}$  berimpit, yang berarti dokumen  $D$  sesuai dengan *query*  $Q$ .
- Jadi, nilai *cosinus* yang besar (mendekati 1) mengindikasikan bahwa dokumen cenderung sesuai dengan *query*.

- Setiap dokumen di dalam koleksi dokumen dihitung kesamaannya dengan *query* dengan rumus cosinus di atas.
- Selanjutnya hasil perhitungan *di-ranking* berdasarkan nilai cosinus dari besar ke kecil sebagai proses pemilihan dokumen yang yang “dekat” dengan *query*.
- *Pe-ranking-an* tersebut menyatakan dokumen yang paling relevan hingga yang kurang relevan dengan *query*.
- Nilai cosinus yang besar menyatakan dokumen yang relevan, nilai cosinus yang kecil menyatakan dokumen yang kurang relevan dengan *query*.

- Pada contoh di atas:

$$\mathbf{Q} \cdot \mathbf{D}_1 = (2)(0) + (3)(0) + (5)(2) = 10$$

$$\mathbf{Q} \cdot \mathbf{D}_2 = (3)(0) + (7)(0) + (1)(2) = 2$$

$$\|\mathbf{Q}\| = \sqrt{0^2 + 0^2 + 2^2} = \sqrt{4} = 2$$

$$\|\mathbf{D}_1\| = \sqrt{2^2 + 3^2 + 5^2} = \sqrt{4 + 9 + 25} = \sqrt{38}$$

$$\|\mathbf{D}_2\| = \sqrt{3^2 + 7^2 + 1^2} = \sqrt{9 + 49 + 1} = \sqrt{59}$$

$$\text{sim}(Q, D_1) = \cos \theta_1 = \frac{\mathbf{Q}_1 \cdot \mathbf{D}_1}{\|\mathbf{Q}\| \|\mathbf{D}_1\|} = \frac{10}{2\sqrt{38}} = 0.81$$

$$\text{sim}(Q, D_2) = \cos \theta_2 = \frac{\mathbf{Q}_1 \cdot \mathbf{D}_2}{\|\mathbf{Q}\| \|\mathbf{D}_2\|} = \frac{2}{2\sqrt{59}} = 0.13$$

Karena  $0.81 > 0.13$ , maka dokumen  $D_1$  lebih sesuai dengan query  $Q$  dibandingkan dengan dokumen  $Q_2$

$$\text{sim}(\mathbf{Q}, \mathbf{D}) = \cos \theta = \frac{\mathbf{Q} \cdot \mathbf{D}}{\|\mathbf{Q}\| \|\mathbf{D}\|}$$

$$\mathbf{Q} \cdot \mathbf{D} = q_1 d_1 + q_2 d_2 + \dots + q_n d_n$$

# Latihan (Kuis 2020)

**Google** misalkan menggunakan metode ruang vektor untuk me-ranking website-website berdasarkan keyword yang dimasukkan. Misalkan ada masukan keyword dari pengguna sbb : *“red big car”* dan ada 3 website yang isinya sebagai berikut :

<https://www.algeo.com> : *“Dian wear a red blouse in the house”*

<https://www.aljabargeometri.com> : *“Big Edi ride a red big car in the road”*

<https://www.aljabarlinear.com> : *“Dian ride a very big big red car in the road”*

- a. Carilah similaritas antara keyword yang dimasukkan oleh user tersebut dengan ketiga website tersebut.
- b. Lakukan perangkingan website tersebut.

## Jawaban:

a) Misalkan vector query dilambangkan dengan Q, vektor website <https://www.algeo.com> dilambangkan dengan vector D1, vektor <https://www.aljabargeometri.com> dilambangkan dengan vektor D2, dan vektor <https://www.aljabarlinear.com> dilambangkan dengan D3.

Term	Vektor Query Q	Vektor D1	Vektor D2	Vektor D3
a	0	1	1	1
big	1	0	2	2
blouse	0	1	0	0
car	1	0	1	1
Dian	0	1	0	1
Edi	0	0	1	0
house	0	1	0	0
in	0	1	1	1
red	1	1	1	0
ride	0	0	1	1
road	0	0	1	1
the	0	1	1	1
very	0	0	0	1
wear	0	1	0	0



Vektor-vektor:

$$Q = (0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0)$$

$$D1 = (1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1)$$

$$D2 = (1, 2, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0)$$

$$D3 = (1, 2, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0)$$

- Panjang vektor query  $Q = |Q| = \sqrt{1^2 + 1^2 + 1^2} = 1,73$
- Panjang vektor dokumen  $D1 = |D1| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = 2,83$
- Panjang vector dokumen  $D2 = |D2| = \sqrt{1^2 + 2^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = 3,46$
- Panjang vektor dokumen  $D3 = |D3| = \sqrt{1^2 + 2^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = 3,46$
- Similaritas query Q dengan dokumen  $D1 = \mathbf{Q \cdot D1} / |Q| |D1| = ((1*0)+(1*0)+(1*1)) / (1,73*2,83) = 0,20$
- Similaritas query Q dengan dokumen  $D1 = \mathbf{Q \cdot D2} / |Q| |D2| = ((1*2)+(1*1)+(1*1)) / (1,73*3,46) = 0.67$
- Similaritas query Q dengan dokumen  $D1 = \mathbf{Q \cdot D3} / |Q| |D3| = ((1*2)+(1*1)+(1*1)) / (1,73*3,46) = 0.67$

Alternatif jawaban versi lain yang lebih ringkas:

- Similaritas (Q, D1) =  $\frac{\mathbf{Q} \cdot \mathbf{D1}}{|\mathbf{Q}| |\mathbf{D1}|} = 1/\sqrt{8}$
- Similaritas(Q, D2) =  $\frac{\mathbf{Q} \cdot \mathbf{D2}}{|\mathbf{Q}| |\mathbf{D2}|} = 4/\sqrt{12}$
- Similaritas(Q, D3) =  $\frac{\mathbf{Q} \cdot \mathbf{D3}}{|\mathbf{Q}| |\mathbf{D3}|} = 4/\sqrt{12}$

b) Karena similaritas dokumen D2 dan dokumen D3 sama maka ada 2 kemungkinan ranking yaitu ranking:

- |       |      |       |
|-------|------|-------|
| 1) D2 | atau | 1) D3 |
| 2) D3 |      | 2) D2 |
| 3) D1 |      | 3) D1 |

- Untuk mendalami lebih lanjut tentang model-model lain di dalam Sistem Temu-balik Informasi, maka anda dapat mengambil mata kuliah pilihan **IF4042 Sistem Temu Balik Informasi** di Semester 7.

## Referensi

1. Prof. Dik Lee, *Vector Space Retrieval Models*, Univ. of Science and Tech, Hong Kong.
2. Hendra Bunyamin, *Information Retrieval System dengan Metode Latent Semantic Indexing*, Tesis S2 Informatika ITB, 2005.