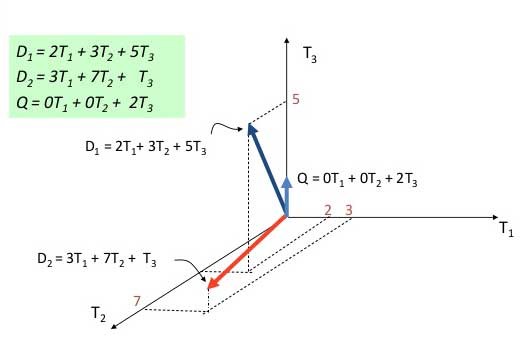
TUGAS PEMEROLEHAN INFORMASI

Pencarian dengan Metode Vektor Space Model (VSM

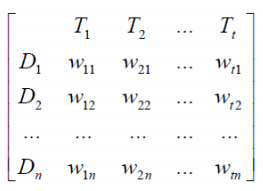
Vector space model adalah suatu model yang digunakan untuk mengukur kemiripan antara suatu dokumen dengan suatu query. Query dan dokumen dianggap sebagai vektor-vektor pada ruang n-dimensi, dimana t adalah jumlah dari seluruh term yang ada dalam leksikon. Leksikon adalah daftar semua term yang ada dalam indeks. Selanjutnya akan dihitung nilai cosinus sudut dari dua vektor, yaitu W dari tiap dokumen dan W dari kata kunci.

Dokumen dalam **Vector Space Model (VSM)** berupa matriks yang berisi bobot seluruh kata pada tiap dokumen. Bobot tersebut menyatakan kepentingan atau kontribusi kata terhadap suatu dokumen dan kumpulan dokumen. Kepentingan suatu kata dalam dokumen dapat dilihat dari frekuensi kemunculannya terhadap dokumen.



Gambar diatas menunjukkan pemodelan dokumen teks di ruang dimensi dimana **(D)** adalah kalimat dokumen sedangkan **(T)** adalah term atau kata.

Misalkan terdapat sejumlah **n** kata yang berbeda sebagai kamus kata (vocabulary) atau indeks kata (terms-index). Kata-kata ini akan membentuk ruang vektor yang memiliki dimensi sebesar **n**. Setiap kata **i** dalam dokumen atau query diberikan bobot sebesar **wi**. Baik dokumen maupun query direpresentasikan sebagai vektor berdimensi **n**.



Untuk mendapatkan nilai jarak atau kemiripan dokumen, dapat menggunakan berbagai macam varian rumus perhitungan jarak diantaranya adalah (1) Cosine, (2) Jaccard, (3) Dice, (3) Euclidean, (4) Manhattan, (5) Minkowski, (6) Mahalanobis, (8) Weighted.

Perhatikan contoh berikut, misalnya:

Dokumen 5 (D5): “ aktivis icw dianiaya aktivis icw, tama satya langkun, dianiaya oleh segerombolan orang yang tak dikenal di kawasan duren tiga, jakarta selatan, kamis dini hari.”

Dokumen 6 (D6) : “ cici tegal dipanggil kpk kpk akan memeriksa pelawak cici tegal dan pesinetron meidiana hutomo terkait dugaan korupsi pengadaan alat rontgen portable di kementerian kesehatan pada 2007.”

Tabel 1. Kata Stoplist Untuk Pencarian Pada D5 Dan D6

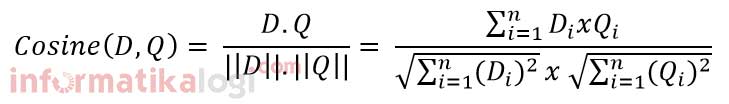


Tabel 2. Kemunculan Kata Pada Dokumen Pencarian Pada D5 Dan D6



**1. Cosine**

**Koefisien Cosine** merupakan metode ukuran kesamaan yang paling popular, yaitu dengan menghitung sudut antara vektor dokumen dengan vektor kueri. Jika vektor adalah satuan panjang, cosinus dari sudut antara mereka hanyalah dot product dari vektor, persamaannya sebagai berikut:



Dimana :  
**D** : vektor D  
**Q** : vektor Q  
**|D|** : besar dari vektor D  
**|Q|** : besar dari vektor Q

Karena berdasarkan cosinus sudut antara dua vektor, maka nilainya berkisar pada 0 sampai dengan 1, dimana 0 menandakan bahwa kedua dokumen tidak mirip sama sekali, dan 1 menandakan bahwa antar dokumen benar-benar identik.