**QUIZ**

Nama : M. Nizar Rahman

NIM : 16/395745/PA/17321

MATA KULIAH : SISTEM TEMU KEMBALI INFORMASI

HARI, TANGGAL : RABU, 08 MEI 2019

WAKTU : 13.30 – 16.00

DOSEN : DRS. EDI WINARKO, MSC., PHD

SIFAT : BUKU TERBUKA

1. Jika terdapat 4 dokumen seperti berikut

**D1: "The sky is blue",  
D2: "The sun is bright",  
D3: "The sun in the sky is bright",  
D4: "We can see the shining sun, the bright sun”**

Hitung jarak antara Query “sun on the sky” dengan masing-masing dokumen menggunakan (a) Jaccard coefficient, (b) tf-matching score

a. Jaccard coefficient

D1. JC(Q,D1) = 1/6 = 16.67%

D2. JC(Q,D2) = 1/6 = 16.67%

D3. JC(Q,D3) = 3/7 = 42.85%

D4. JC(Q,D4) = 2/9 = 22.22%

b. tf-matching score

D1. tf-score(Q,D1) = ((1+log 1) + (1+log 1)) = 1 + 1 = 2

D2. tf-score(Q,D2) = ((1+log 1) + (1+log 1)) = 1 + 1 = 2

D3. tf-score(Q,D3) = ((1+log 1) + (1 + log 2) + (1 + log 1)) = 1 + 1.3 + 1 = 3.3

D4. tf-score(Q,D4) = ((1 + log 2) + (1 + log 2)) = 1.3 + 1.3 = 2.6

1. Hitung kesamaan antara query "*information retrieval system*" dengan dokumen "*information system, information retrieval, and operating system*" dengan cara mengisi nilai-nilai dari kolom yang kosong pada tabel di bawah ini. Asumsikan jumlah dokumen N = 1,000,000 dan term 'and' sebagai stop word. Berapakah nilai akhir kesamaan antara query dengan dokumen?

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| word | query | | | | | document | | | qi . di |
| tf | wf | df | idf | qi = wf-idf | tf | wf | di = normalized wf |
| information  retrieval  operating  system | 1  1  0  1 | 1  1  0  1 | 10000  20000  50000  10000 | 2  1.7  1.3  2 | -1  -0.7  -1.3  -1 | 2  1  1  2 | 1.3  1  1  1.3 | 0.56  0.43  0.43  0.56 | -0.56  -0.301  -0.559  -0.56 |

Final similarity score : -0.56 -0.301 -0.559 -0.56 = -1.98

cos(query,document) = -1.98 / (2.044)\*(0.998) = -0.97

1. Dapatkah nilai TF-IDF dari suatu kata (term) dalam dokumen bernilai lebih dari 1? Jelaskan jawaban anda.

Tidak, karena nilai 1 merupakan maksimum rasio kemunculan term dalam dokumen

1. Kapan nilai TF-IDF dari suatu kata dalam dokumen bernilai 0 (nol)?

Ketika kata tersebut tidak ada dalam dokumen manapun

4. Hitung *edit distance* dari kata "CELAK" dan "BETAH" menggunakan algoritma Levenshtein. Tunjukan matrix hasil perhitungannya dan urut-urutan operasi edit untuk mengubah kata “CELAK” menjadi “BETAH”

Jawab :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | B | E | T | A | H |
|  | 0 | 1 | 2 | 3 | 4 | 5 |
| C | 1 | 1 | 2 | 3 | 4 | 5 |
| E | 2 | 2 | 1 | 2 | 3 | 4 |
| L | 3 | 3 | 2 | 2 | 3 | 4 |
| A | 4 | 4 | 3 | 3 | 2 | 3 |
| K | 5 | 5 | 4 | 4 | 3 | 3 |

Edit Distance = 3

Langkah :

CELAK awal

CELAH ubah K menjadi H

CELAH A tetap

CETAH ubah L menjadi T

CETAH E tetap

BETAH ubah C menjadi B