Text Mining Exercises 4

March 19, 2014

1. Terdapat sebuah koleksi dokumen yang sudah di-indeks. Berikut ini adalah tabel dengan *terms* (t) yang ada dalam dokumen tersebut beserta dengan persentase dokumen yang mengandung *term* tersebut. Contohnya: 10% dari dokumen yang terdapat pada koleksi dokumen memiliki *term computer*.

Term	% dok
computer	10%
software	10%
bugs	5%
code	2%
developer	2%
programmers	2%

Berikut ini adalah tiga dokumen yang ada pada koleksi dokumen:

D1 = "programmers build computer software"
D2 = "most software has bugs, but good software has less bugs than bad software"
D3 = "some bugs can be found only by executing the software, not by examining the source code"

Hitunglah *tf-idf weights* dari setiap term di setiap dokumen. Anda bisa menggunakan tabel berikut ini untuk membantu:

$term$	$\%dok$	f	tf	$idf = log\left(\frac{100}{\%dok}\right)$	$tf \times idf$
computer	10				
software	10				
bugs	5				
code	2				
developer	2				
programmers	2				

2. Terdapat sebuah query Q =computer software programmers. Hitunglah kesamaan (similarity) antara Q dengan dokumen-dokumen pada nomor (1). Setelah itu tentukan rankingnya.

3. Terdapat sebuah koleksi berisi 100 dokumen dengan ID 1...100. Dokumen yang dianggap relevan adalah dokumen dengan ID 1...20. Terdapat dua sistem IR dimana saat seorang user memasukkan sebuah query hasilnya adalah sebagai berikut:
   S1 = [1, 2, 21, 22, 3, 23, 25, 4, 28, 5, 29, 30, 6, 7, 31, 32, 33, 40, 41, 42, 8, 43, 44, 9, 45, 10, 50, 51, 11, 52, 53, 54, 12, 60, 62, 13, 63, 64, 14, 15, 16, 70, 78, 80, 17, 81, 82, 83, 85, 18, 90, 19, 91, 92, 20, 93, 94, 95, 96, 98]
   S2 = [25, 26, 1, 27, 28, 2, 3, 29, 30, 4, 35, 36, 5, 37, 6, 7, 8, 38, 9, 40, 10, 42, 11, 45, 46, 12, 48, 50, 51, 13, 60, 61, 64, 14, 70, 72, 15, 78, 79, 90]

   Untuk setiap sistem, hitunglah recall, precision dan F-measure (dengan$\alpha = \frac{1}{2}, \alpha = \frac{1}{4}, \alpha = \frac{3}{4}$) untuk query ini.