Kisi-kisi UAS
Pengantar Temu Balik Informasi (XI084)
UAS tgl 26 Januari 2016 18:30-20:30

Diketahui sejumlah dokumen sebagai koleksi dengan isi sebagai berikut:

D1: football cricket football
D2: cricket termite grasshopper
D3: football football hockey
D4: football goal
D5: obama romney football
D6: football cricket hockey termite

1. Hitunglah pembobotan TFIDF untuk setiap kata dalam dokumen. Gunakan rumus:

$$w_{i,j} = (1 + \log_{10}(\mathrm{tf}_{i,j})) \cdot \log_{10}(\frac{N}{n_i})$$

$w_{i,j}$	weight assigned to term *i* in document *j*
$\mathrm{tf}_{i,j}$	number of occurrence of term *i* in document *j*
N	number of documents in entire collection
n_i	number of documents with term *i*

Lengkapi tabel di bawah ini sesuai dengan rumus TFIDF di atas:

	TERM FREQUENCIES $\mathrm{tf}_{i,j}$							
Vector	**Football**	**Cricket**	**Termite**	**Grasshopper**	**Hockey**	**Goal**	**Obama**	**Romney**
D1								
D2								
D3								
D4								
D5								
D6								

	TERM WEIGHTS $w_{i,j}$							
Vector	**Football**	**Cricket**	**Termite**	**Grasshopper**	**Hockey**	**Goal**	**Obama**	**Romney**
D1								
D2								
D3								
D4								
D5								
D6								

2. Jika terdapat kueri ***q = football obama***; Urutkan similaritas dokumen pada koleksi soal nomor 1 dengan menggunakan *Cosine Similarity*.

No. 2 $$Cosine(D_i, Q) = \frac{\sum_{j=1}^{t} d_{ij} \cdot q_j}{\sqrt{\sum_{j=1}^{t} {d_{ij}}^2 \cdot \sum_{j=1}^{t} {q_j}^2}}$$

No. 3

Term	Document	R	NR
Term present	$x_i = 1$	p_i	s_i
Term absent	$x_i = 0$	$1 - p_i$	$1 - s_i$

3. Jika diketahui bahwa dokumen D1, D3, D4 dan D6 adalah ***relevan*** pada soal nomor 1. Buatlah tabel kontingensi untuk kueri pada nomor 2. Gunakan Laplace smoothing jika diperlukan.

4. Hitunglah ***retrieval status value*** (RSV) untuk setiap dokumen pada nomor 1 sesuai kontingensi tabel pada nomor 3.

5. Urutkan dokumen-dokumen pada nomor 1 menggunakan ***MLE unigram*** untuk kueri pada nomor 2, dengan nilai $\lambda = ½$.

6. Jika diketahui pengelompokan dokumen sebagai berikut:

Dok	Konten	Topik
D1	football cricket football	Sport
D2	cricket termite grasshopper	Other
D3	football football hockey	Sport
D4	football goal	Sport
D5	obama romney football	Other
D6	football cricket hockey termite	???

Berikan nilai probabilitas setiap kata dalam topik yang tersedia di atas.

Kata	$P(t_i \mid Sport)$	$P(t_i \mid Other)$
Football		
Cricket		
Termite		
Grasshopper		
Hockey		
Goal		
Obama		
Romney		

Sesuai dengan probabilitas yang dihasilkan, apakah D6 termasuk dalam topik Sport atau Other, berikan perhitungannya.

7. Jelaskan arsitektur sistem temu balik Lucene berikut ini sebagaimana pengalaman Anda dalam tugas praktikum, diambil dari:

Białecki, Andrzej, Robert Muir, and Grant Ingersoll. "Apache lucene 4." SIGIR 2012 workshop on open source information retrieval. 2012.

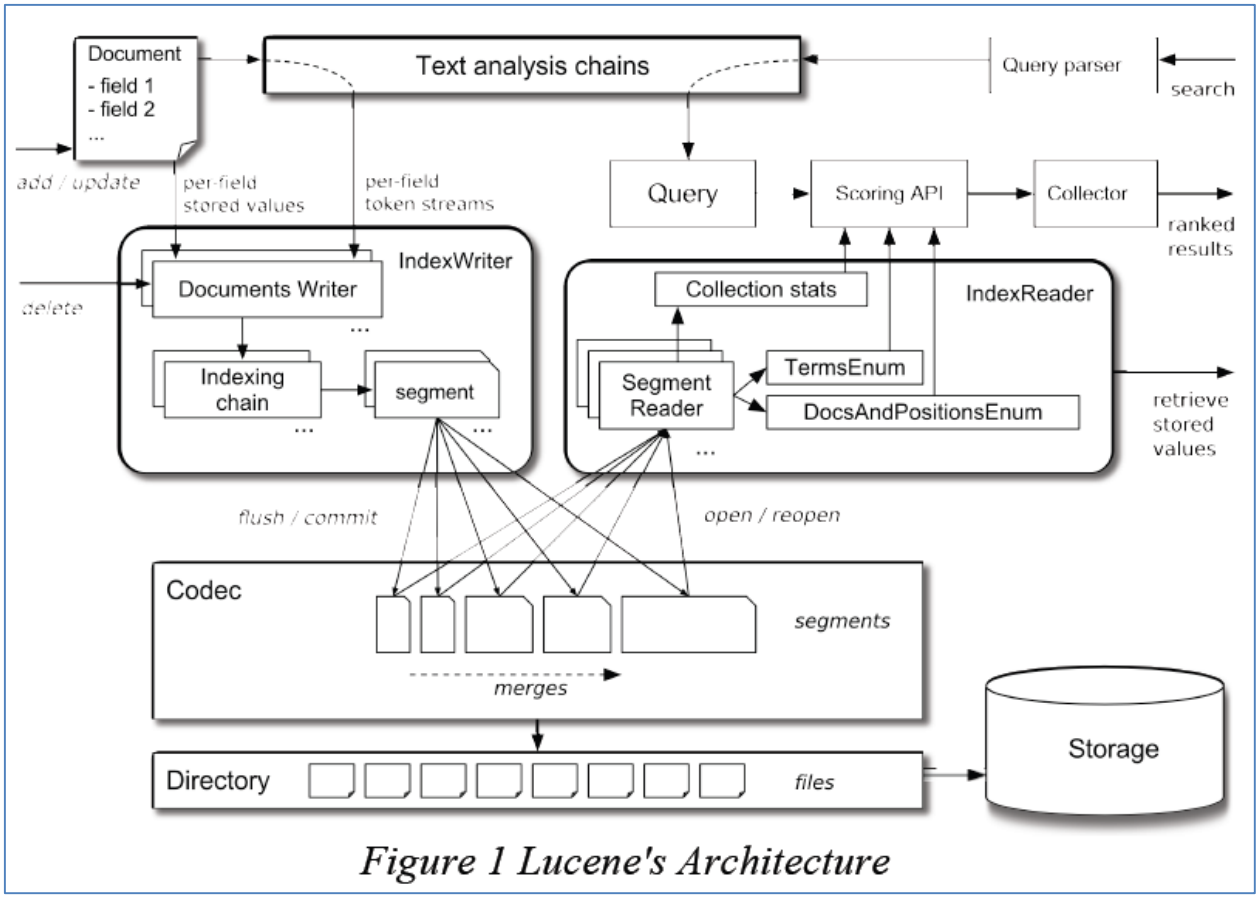

Figure 1 Lucene's Architecture