Studi Kasus #1

Matakuliah Teks dan Web Mining
Semester Genap Tahun Ajaran 2013/2014
Dosen: Budi Susanto, S.Kom., M.T. (dan Agata Filiana, S.Kom., M.Sc.)

Tim Proyek

Proyek studi kasus #1 dikerjakan secara berkelompok dengan anggota berjumlah 4 mahasiswa. Anggota tiap kelompok diserahkan kepada mahasiswa peserta matakuliah Teks dan Web Mining. Setiap kelompok harus menunjuk satu orang sebagai ketua. Ketua mengirimkan daftar anggota kelompoknya ke budsus@ti.ukdw.ac.id paling lambat 18 Maret 2014. Pemilihan topik dokumen untuk setiap kelompok akan diumumkan di kelas pada tanggal 19 Maret 2014. Setiap kelompok harus mempresentasikan hasil pekerjaannya pada periode tanggal 7, 8, dan 10 April 2014. Penentuan jadwal presentasi akan diundi pada tanggal 20 Maret 2014 dan akan diumumkan melalui e-Class. **Laporan proyek sudah harus diterima pada tanggal 6 April 2014 pukul 23:59 WIB (Lihat bagian laporan bawah)**. Nilai akan berkurang 10 poin setiap hari keterlambatannya. Masing-masing anggota kelompok akan mendapat nilai maksimum 15 untuk laporan dan 10 untuk presentasi. Masing-masing penilaian akan menggunakan rubrik penilaian yang sudah disampaikan di Silabus matakuliah ini.

Platform

Berikut platform minimal yang harus digunakan:

- Sistem operasi dapat Windows atau Linux (Linux lebih disarankan)
- Lucene Core 3.6.2
    - Kebutuhan minimum untuk lucene, silahkan baca feil SYSTEM_REQUIREMENTS.txt dari paket yang tersedia.

Tugas

Untuk setiap kelompok

- Laporan proyek ditulis menggunakan bahasa Indonesia yang baik dan benar.
- Setiap kelompok fokus pada topik tertentu untuk kumpulan dokumen yang dikumpulkan.
- Setiap anggota kelompok WAJIB mengumpulkan 10 dokumen untuk topik yang telah ditentukan dalam kelompok. Dokumen diambil dari situs berita http://www.antaranews.com/.
- Dokumen-dokumen yang terkumpul diindeks menggunakan mesin pencari Lucene. Anda dapat menggunakan antarmuka yang tersedia di Lucene untuk memberikan query.
- Setiap anggota kelompok harus memberikan **satu** tugas pencarian.
- Setiap tugas pencarian, Anda harus membentuk dua *query string*:
    - sebuah ekspresi Boolean (Hasil pencarian adalah dokumen-dokumen yang memenuhi ekspresi tersebut. Dokumen tidak diurutkan).

    - sebuah daftar *terms*, yang disebut sebagai *vector model query* (Hasil adalah daftar dokumen yang dirangking berdasar relevansinya dengan query).
- Kelompok mengevaluasi relevansi semua dokumen dibandingkan dengan setiap tugas pencarian. Untuk setiap tugas pencarian harus terdapat dua evaluasi independen.
- Jalankan query menggunakan Lucene.
- Hitung *recall* dan *precision* untuk semua hasil. Khusus untuk hasil dari daftar terurutkan, gambarkan kurva recall-precision (gunakan rata-rata hasil dari semua tugas pencarian!)

Laporan

Setiap kelompok menuliskan sebuah laporan yang ditulis dengan menggunakan Google Docs. Setiap laporan harus di share (*editable*) ke budsus@ti.ukdw.ac.id. Setiap laporan yang telah di-share, akan diambil versi terakhir pada tanggal **7 April 2014 mulai pukul 00:00 WIB**. Bagian-bagian yang harus ada dalam setiap laporan:

- Deskripsi tentang koleksi dokumen, jumlah dokumen, topik, dan jumlah kata keseluruhan dan jumlah kata per dokumen (termasuk rata-rata per dokumen).
- Tugas pencarian dan bentuk querynya.
- Pengalaman dari evaluasi relevansi.
- Presentasi dari hasil pencarian (*retrieval*) (misalnya sebuah grafik recall-precision untuk rata-rata hasil dari *vector model queries*, rata-rata *precision-recall* untuk query *Boolean*).

oOo