TUGAS 1 PEROLEHAN INFORMASI

Batas Pengumpulan: 5 Oktober 2019 Jam 20:00

Capaian pembelajaran:

- Mahasiswa mampu menerapkan dasar text processing dengan menggunakan bahasa pemrograman Perl.

Jenis Tugas: Pemrograman PER

Petunjuk pengerjaan tugas:

1. Buatlah program perl untuk menjawab soal-soal dibawah ini.
2. Pengumpulan tugas terdiri dari:
   - Kode program untuk soal nomor 1 dan nomor 2 dengan format **T1_NPM_Nama.pl**.
   - Laporan jawaban untuk soal nomor 1 dan 2 dibuat ke dalam file pdf dengan format **T1_NPM_Nama.pdf** dan hasil output program disimpan ke dalam file hasil_1.txt dan hasil_2.txt.
   - Simpan kode program dan laporan jawaban ke dalam satu file .zip dengan format: **T1_NPM_Nama.zip**.
3. Berikan komentar yang jelas terhadap variabel dan fungsi modul yang dibuat pada kode program.
4. Jelaskan asumsi jawaban anda pada laporan jawaban untuk memperjelas mengapa hasil tersebut anda dapatkan.
5. Dikumpulkan pada hari 5, Oktober 2019 jam 20.00 WIB di SCELE kelas Information Retrieval.

Soal Nomor 1

a. Hitunglah jumlah kata unik (*vocabulary*) yang terdapat pada keseluruhan korpus-1.txt.
b. Hitunglah jumlah kalimat yang terdapat pada keseluruhan korpus-1.txt.
c. Hitunglah jumlah token berupa angka (termasuk bilangan bulat, desimal, dan romawi) yang terdapat pada keseluruhan korpus-1.txt.
d. Apakah distribusi kata dalam korpus mengikuti distribusi zipf? Jelaskan.

Soal Nomor 2

a. Buatlah aturan untuk mengidentifikasi entitas tanggal dan lokasi sesuai dengan korpus-2.txt (yang merupakan bagian dari korpus-1.txt dokumen nomor DOC-001 sampai DOC-030).

b. Lakukan analisis dan hitung berapa banyak entitas yang bisa diidentifikasi, menggunakan aturan yang telah dibuat sebelumnya, pada korpus-1.txt dokumen nomor DOC-001 sampai DOC-030, dan bandingkanlah dengan korpus-2.txt.

Keterangan Korpus:

- korpus-1.txt berisi 300 Dokumen
    - Setiap dokumen diidentifikasi dengan tag <DOC> … </DOC>
    - Setiap dokumen memiliki nomor dokumen <DOCID> … </DOCID>, judul dokumen <TITLE> … </TITLE>, dan isi dokumen <TEXT> … </TEXT>
- korpus-2.txt berisi 30 Dokumen dari Korpus-1.txt Nomor DOC-001 sampai DOC-030 yang telah memiliki tag identifikasi entitas tanggal dan lokasi.
    - Entitas tanggal diidentifikasi dengan tag <ENT TYPE=”TANGGAL”> … </ENT>
    - Entitas lokasi diidentifikasi dengan tag <ENT TYPE=”LOKASI”> … </ENT>