A. Ide Penyelesaian

Ide penyelesaian untuk kasus ini dijelaskan dalam tiga tahap sebagai berikut.

1. Persiapan Dokumen

Pada soal dinyatakan bahwa dokumen yang diringkas adalah top 5 dokumen yang diperoleh terhadap masing - masing kueri yang telah ditentukan. Pencarian Top 5 dokumen dilakukan menggunakan program yang telah dibuat sebelumnya pada tugas 2. Hasil keluaran dari program pencarian top 5 dokumen adalah 5 id dokumen untuk amsing - masing kueri. Ada 3 jenis kueri yang diminta yakni, kueri monolingual berbahasa indonesia, kueri bahasa inggris hasil terjemahan mesin penerjemah Google Translate dan kueri bahasa inggris hasil terjemahan kamus bilingual yang disediakan oleh http://dictionary.cambridge.org/dictionary/english-indonesian/.

Daftar kueri untuk masing - masing jenis kueri disimpan dalam file berikut.

- [Original] Queri_Indo.txt merupakan kueri monolingual
- [Trans Ingg GL] Queri_Indo.txt merupakan kueri bahasa Inggris hasil terjemahan dengan mesin penerjemah Google Translate
- [Trans Ingg CAM] Queri_Indo.txt merupakan query bahasa Inggris hasil terjemahan dengan kamus bilingual

Daftar top 5 dokumen untuk masing - masing jenis kueri disimpan dalam file berikut.

- [TOP5_Original] Queri_Indo.txt untuk kueri monolingual berbahasa indonesia
- [TOP5_Trans Ingg GL] Queri_Indo untuk terjemahan dengan mesin penerjemah Google Translate
- [TOP5_Trans Ingg CAM] Queri_Indo.txt untuk terjemahan dengan kamus bilingual http://dictionary.cambridge.org/dictionary/english-indonesian/

2. Algoritma Meringkas

Algoritma yang digunakan adalah Linear Future Combination. Elemen yang digunakan untuk meringkas dokumen adalah sebagai berikut.

a. Thematic Term

Thematic term diterapkan dengan menghitung nilai tf/idf masing - masing kata di kalimat dalam suatu dokumen terhadap keseluruhan dokumen yang ada. Weight tiap kalimat diperoleh dengan menerapkan rumus berikut.

$$\text{Weight}(U) = A/B$$

Keterangan :

U adalah kalimat
A total tf/idf kalimat
B total tf/idf seluruh kalimat di dokumen

b. Add Term

Addterm diterapkan dengan menambahkan Weight(U) apabila kata di U juga terdapat pada kalimat utama, heading, judul atau kueri. Weight tiap kalimat diperoleh dengan menerapkan rumus berikut.

$$Weight(U) = (C/D)*0.1$$

Keterangan :

U adalah kalimat

C adalah total kata di U yang muncul di kalimat utama, heading, judul atau kueri

D adalah total kata di kalimat utama, heading, judul, atau kueri

c. Lokasi

Lokasi diterapkan dengan menambahkan Weight(U) sesuai dengan lokasi kemunculan kalimat. Weight tiap kalimat diperoleh dengan menerapkan rumus berikut.

Weight(U) = E/F

Keterangan :

U adalah kalimat

E adalah lokasi kalimat

F adalah total kalimat di dokumen

Semua Weight untuk masing - masing kalimat di dokumen dijumlahkan. Berdasarkan compression rate yang telah ditentukan pada soal sebesar 10% maka summary dihasilkan dengan menggabungkan n kalimat yang memiliki weight terbesar (dengan n = 10% x total kalimat di dokumen). Hasil ringkasan masing - masing dokumen untuk suatu kueri digabungkan menjadi satu paragraf. Hasil Ringkasan disimpan pada berkas dengan format berikut.

<Kueri></Kueri>

<Ringkasan></Ringakasan>

Hasil ringkasan untuk masing - masing katergori kueri (monolingual, terjemahan mesin, kamus bilingual) disimpan ke 3 file berbeda.

B. Analisa Hasil Ringkasan

Jaccard similarity digunakan untuk menganalisa hasil ringkasan yang diperoleh dari masing - masing kueri. Proses analisa dilakukan dengan menjalankan program AnalisisSimilarity.pl. Program ini akan membandingkan hasil ringkasan yang diperoleh antara kueri monolingual dengan kueri bahasa inggris hasil terjemahan mesin penerjemah google translate dan antara kueri monolingual dengan kueri bahasa inggris hasil terjemahan kamus bilingual yang disediakan oleh dictionary.cambridge.org.

Masukan atau input program ini adalah tiga dokumen hasil ringkasan. Program menghitung nilai jaccard similarity antara hasil ringkasan kueri monolingual dengan kueri bahasa inggris hasil terjemahan. Detail nilai jaccard dapat dilihat pada berkas daftar_nilai_jaccard.txt.

Berdasarkan rata - rata nilai jaccard yang diperoleh disimpulkan bahwa kueri bahasa inggris hasil terjemahan mesin terjemah lebih baik dari pada kamus bilingual. Hasilnya mendekati hasil ringkasan dari kueri monolingual.

Similarity antara Kueri Monolingual dan Mesin Penerjemah = 0.994933616356192
Similarity antara Kueri Monolingual dan Kamus Bilingual = 0.981166399143927

C. Petunjuk Penggunaan Program

Program sudah di set untuk memproses ketiga jenis kueri sekaligus. Output yang dihasilkan juga disimpan dalam 3 berkas yang berbeda sesuai dengan jenis kueri (bahasa indonesia, bahasa inggris hasil terjemahan google translate, dan bahasa inggris hasil terjemahan kmaus cambridge online).

Pengguna hanya perlu menjalankan program pada command promt dengan mengetikkan perintah berikut.
Perl [spasi] T4_1106022654_Gina Andriyani_SourceCode.pl

Program akan memproses 3 berkas berisi Top 5 dokumen untuk masing - masing kueri dan menghasilkan ringkasan ke 3 berkas berbeda pula, yaitu.

- [HasilRingkasan_Original] Queri_Indo.txt untuk hasil ringkasan dari kueri monolingual
- [HasilRingkasan_Trans Ingg GL] Queri_Indo.txt untuk hasil ringkasan dari kueri bahasa inggris hasil terjemahan mesin penerjemah Google Translate
- [HasilRingkasan_Trans Ingg CAM] Queri_Indo.txt untuk hasil ringkasan dari kueri bahasa inggris hasil terjemahan kamus bilingual yang disediakan pada website dictionary.cambridge.org

Informasi tambahan terkait cara mendapatkan top 5 dokumen dijelaskan sebagai berikut.

- Tiga berkas yang berisi Top 5 dokumen merupakan hasil generate program pada tugas 2 (tugas sebelumnya) untuk mengambil 5 Top Dokumen.