Laporan Tugas Besar 2

Aljabar dan Geometri IF2123
Kelompok 46

Rhapsodya Piedro Asmorobangun (13519084),

Benidictus Galih Mahar Putra (13519159),

Made Kharisma Jagaddhita (13519176)

Kelas Mahasiswa (K-4), Jurusan Teknik Informatika

Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung,

Jl. Ganesha no. 10 Bandung, Indonesia, 40132

BAB I - DESKRIPSI MASALAH

Temu-balik informasi (information retrieval) merupakan proses menemukan kembali (retrieval) informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis. Biasanya, sistem temu balik informasi ini digunakan untuk mencari informasi pada informasi yang tidak terstruktur, seperti laman web atau dokumen. Ide utama dari sistem temu balik informasi adalah mengubah search query menjadi ruang vektor.

Setiap dokumen maupun query dinyatakan sebagai vektor w = (w1, w2,..., wn) di dalam Rn, dimana nilai wi dapat menyatakan jumlah kemunculan kata tersebut dalam dokumen (term frequency). Penentuan dokumen mana yang relevan dengan search query dipandang sebagai pengukuran kesamaan (similarity measure) antara query dengan dokumen. Semakin sama suatu vektor dokumen dengan vektor query, semakin relevan dokumen tersebut dengan query. Kesamaan tersebut dapat diukur dengan cosine similarity dengan rumus:

$$sim(\mathbf{Q}, \mathbf{D}) = \cos\theta = \frac{\mathbf{Q} \cdot \mathbf{D}}{\|\mathbf{Q}\|\|\mathbf{D}\|}$$

BAB II – DASAR TEORI

I. Information Retrieval (IR)

Information retrieval (IR) atau sistem temu balik informasi digunakan untuk menemukan kembali informasi-informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis. Salah satu aplikasi umum dari sistem temu kembali informasi adalah search-engine atau mesin pencarian yang terdapat pada jaringan internet. Pengguna dapat mencari halaman-halaman Web yang dibutuhkannya melalui mesin tersebut.

Ukuran efektivitas pencarian ditentukan oleh precision dan recall. Precision adalah rasio jumlah dokumen relevan yang ditemukan dengan total jumlah dokumen yang ditemukan oleh search-engine. Precision mengindikasikan kualitas himpunan jawaban, tetapi tidak memandang total jumlah dokumen yang relevan dalam kumpulan dokumen.

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{documents retrieved}\}|}{|\{\text{documents retrieved}\}|}$$

Recall adalah rasio jumlah dokumen relevan yang ditemukan kembali dengan total jumlah dokumen dalam kumpulan dokumen yang dianggap relevan.

$$\text{Recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{documents retrieved}\}|}{|\{\text{relevant documents}\}|}$$

Model ruang vektor dan model probabilistik adalah model yang menggunakan pembobotan kata dan perangkingan dokumen. Hasil retrieval yang didapat dari model-model ini adalah dokumen terangking yang dianggap paling relevan terhadap query.Terdapat beberapa cara atau metode dalam melakukan pembobotan kata pada metode TF-IDF, yaitu melalui skema pembobotan query dan dokumen.

Dalam model ruang vektor, dokumen dan query direpresentasikan sebagai vektor dalam dalam ruang vektor yang disusun dalam indeks term, kemudian dimodelkan dengan persamaan geometri. Sedangkan model probabilistik membuat asumsi-asumsi distribusi term dalam dokumen relevan dan tidak relevan dalam orde estimasi kemungkinan relevansi suatu dokumen terhadap suatu query.

II. Vektor

Vektor spasial atau vektor Euclidean; biasa disebut vektor dalam matematika dan fisika adalah objek geometri yang memiliki besar dan arah. Vektor dilambangkan dengan tanda panah (→). Besar vektor proporsional dengan panjang panah dan arahnya bertepatan dengan arah panah. Vektor dapat melambangkan perpindahan dari titik A ke B. Vektor sering ditandai sebagai $\overrightarrow{AB}$.

Untuk mencari panjang sebuah vektor dalam ruang euklidian tiga dimensi, dapat digunakan cara berikut:

$$\|\mathbf{a}\| = \sqrt{{a_1}^2 + {a_2}^2 + {a_3}^2}$$

yang merupakan konsekuensi dari Teorema Pythagoras karena vektor dasar e1, e2, e3 merupakan vektor-vektor satuan ortogonal.

Perkalian vektor adalah operasi perkalian dengan dua operand (objek yang dikalikan) berupa vektor. Tetapi hasil operasi ini tidak selalu adalah vektor. Terdapat tiga macam perkalian vektor, yaitu produk skalar atau perkalian titik (bahasa Inggris: dot product atau scalar product, perkalian silang (bahasa Inggris: cross product atau vector product atau directed area product) dan perkalian langsung (bahasa Inggris: direct product).

Produk skalar (atau "perkalian titik") dua buah vektor akan menghasilkan sebuah skalar. Jenis perkalian ini bersifat komutatif.

$$\vec{A} \cdot \vec{B} = (a_x\hat{i} + a_y\hat{j} + a_z\hat{k}) \cdot (b_x\hat{i} + b_y\hat{j} + b_z\hat{k})$$
$$= a_x b_x + a_y b_y + a_z b_z$$

Hasil suatu perkalian silang dua buah vektor adalah juga sebuah vektor. Perkalian silang bersifat tidak komutatif.

$$\vec{A} \times \vec{B} = (a_x\hat{i} + a_y\hat{j} + a_z\hat{k}) \times (b_x\hat{i} + b_y\hat{j} + b_z\hat{k})$$
$$= (a_y b_z - a_z b_y)\hat{i} + (a_z b_x - a_x b_z)\hat{j} + (a_x b_y - a_y b_x)\hat{k}$$

III. *Cosine Similarity*

Kesamaan (sim) antara dua vektor Q = (q1 , q2 , …, qn) dan D = (d1 , d2 , …, dn) diukur dengan rumus cosine similarity yang merupakan bagian dari rumus perkalian titik (dot product) dua buah vektor:

$$sim(\mathbf{Q}, \mathbf{D}) = \cos\theta = \frac{\mathbf{Q} \cdot \mathbf{D}}{\|\mathbf{Q}\|\|\mathbf{D}\|}$$

Jika cos θ = 1, berarti θ = 0, vektor Q dan D berimpit.

BAB III – IMPLEMENTASI PROGRAM

lowStemStopSplit(strings) - Mereturn kata-kata yang ada di strings setelah dilowercase, distem, dan dihilangkan stopwordsnya

getFiles() - Mereturn nama file dan semua kalimat pada semua file .txt yang ada di directory files dalam bentuk array

getKamusData(query) - Mereturn kamus data berdasarkan file dan query

getFileSum(kalimatFile) - Mereturn jumlah kata dalam suatu file

countTerm(kalimat, query) - Mereturn sebuah list dengan perhitungan kemunculan kata berdasarkan kamus data

getFileTerms(kalimatFile, query) - Mereturn sebuah list dengan perhitungan kemunculan kata dari file berdasarkan kamus data

getSimilarity(QTerms, DTerms) - Mereturn nilai similarity dari terms query dan terms satu dokumen

getAllSim(queryTerms, fileTerms) - Mereturn nilai similarity dari terms query dan terms semua dokumen

getTableValue(queryTerms, fileTerms) - Mereturn nilai-nilai dalam tabel

getKalimatPertama(kalimatFile) - Mereturn kalimat pertama dari sebuah file.

BAB IV – IMPLEMENTASI PROGRAM

8. **artikel 10.txt**

Jumlah kata: 169

Tingkat kemiripan: 0.0%

Kalimat pertama dari dokumen: Sandiaga Uno (Foto: dok

9. **artikel 11.txt**

Jumlah kata: 383

Tingkat kemiripan: 0.0%

Kalimat pertama dari dokumen: REPUBLIKA

10. **artikel 13.txt**

Jumlah kata: 190

Tingkat kemiripan: 0.0%

Kalimat pertama dari dokumen: JAKARTA – Gubernur DKI Jakarta Anies Baswedan terus kritis terhadap lahan reklamasi di Teluk Jakarta

11. **artikel 14.txt**

Jumlah kata: 251

Tingkat kemiripan: 0.0%

Kalimat pertama dari dokumen: Jakarta, CNN Indonesia -- Kepergian Cristiano Ronaldo mulai memberikan efek negatif bagi Real Madrid

12. **artikel 3.txt**

Jumlah kata: 153

Tingkat kemiripan: 0.0%

Kalimat pertama dari dokumen: REPUBLIKA

13. **artikel 5.txt**

Jumlah kata: 448

Tingkat kemiripan: 0.0%

Kalimat pertama dari dokumen: REPUBLIKA

14. **artikel 7.txt**

Jumlah kata: 315

Tingkat kemiripan: 0.0%

Kalimat pertama dari dokumen: REPUBLIKA

15. **artikel 9.txt**

Jumlah kata: 180

Tingkat kemiripan: 0.0%

Kalimat pertama dari dokumen: Jakarta - Total sudah sembilan pelaku penganiayaan pria dengan keterbelakangan mental, Ali Achmad Firmansyah atau Iyan (20), yang ditangkap

Term	Query	artikel 15.txt	artikel 1.txt	artikel 2.txt	artikel 12.txt	artikel 8.txt	artikel 6.txt	artikel 4.txt	artikel 10.txt	artikel 11.txt	artikel 13.txt	artikel 14.txt	artikel 3.txt	artikel 5.txt	artikel 7.txt	artikel 9.txt
hari	1	4	2	2	1	1	1	1	0	0	0	0	0	0	0	0

Perihal

BAB VI – KESIMPULAN, SARAN, REFLEKSI

Hasil yang telah dicapai kelompok kami dalam tubes ini cukup memuaskan. Uji kasus yang kami lakukan sudah menghasilkan output yang diharapkan, walau loading time untuk menampilkan hasil pencarian kami masih membutuhkan waktu.

Hal yang dapat kami kembangkan dari program ini tentu adalah web scraping, yang belum sempat kami tambahkan. Perbandingan antara dua dokumen juga merupakan sebuah fitur yang ingin kami implementasikan.

Refleksi dari Piedro: Saya sangat kecewa dengan kinerja saya dalam tugas besar kali ini. Kontribusi saya sangat sedikit, seperti dilihat dari history commit di Github. Saya hanya membantu sedikit di backend, dan sama sekali tidak membantu menyelesaikan front end. Padahal bagian front end adalah bagian paling susah dari tugas besar kali ini. Dan saya juga kecewa karena saya bahkan tidak mengerjakan pekerjaan saya dengan benar, yaitu laporan ini yang masih belum selesai. Mohon maaf karena kelalaian saya ini.

Refleksi dari Kharisma: Menjadi deadliner itu sangat menyenangkan.

DAFTAR REFERENSI

1. Sistem Temu Balik Informasi. (2019, Juni 6). Di Wikipedia, Ensiklopedia Bebas. Diakses pada November 16, 2020, dari https://id.wikipedia.org/wiki/Sistem_temu_balik_informasi

2. Vektor Euklidean. (2020, Oktober 11). Di Wikipedia, Ensiklopedia Bebas. Diakses pada November 16, 2020, dari https://id.wikipedia.org/wiki/Vektor_Euklidean

3. Perkalian Vektor. (2020, Agustus 15). Di Wikipedia, Ensiklopedia Bebas. Diakses pada November 16, 2020, dari https://id.wikipedia.org/wiki/Perkalian_vektor

4. Slide bahan kuliah IF2123 tahun 2020

5. "Dataset: Artikel" oleh Feryandi Nurdiantorois dilisensikan di bawah Creative Commons Attribution-ShareAlike 4.0 International License.