Penambangan Data Genap 2023/2024

Assignment 07

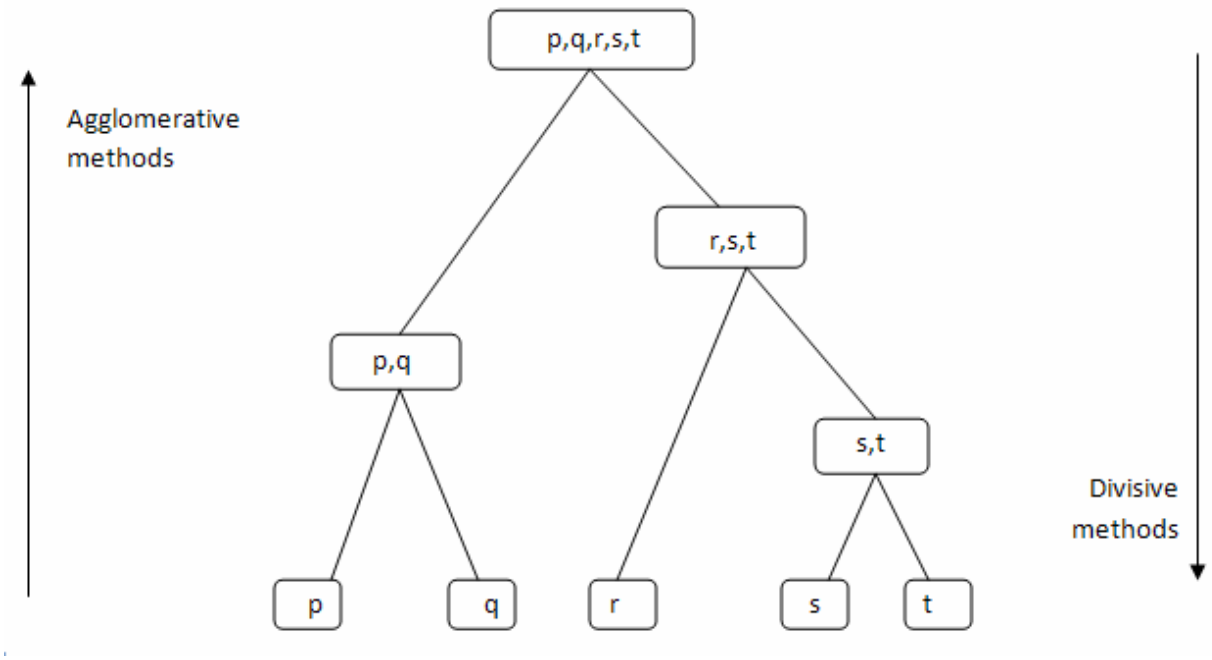
Penambangan Teks

Review Penambangan Teks

Text mining adalah proses mendapatkan informasi atau pengetahuan yang bermanfaat dari teks yang tidak terstruktur. Dalam konteks ini, "teks" dapat merujuk pada dokumen, artikel, tweet, email, atau entitas teks lainnya. Tujuan utama dari text mining adalah untuk mengidentifikasi pola, tren, informasi penting, atau hubungan dalam teks yang besar dan kompleks.

Text clustering adalah teknik dalam text mining yang bertujuan untuk mengelompokkan dokumen teks menjadi kelompok-kelompok yang memiliki kesamaan. Dalam konteks ini, "kesamaan" dapat diukur berdasarkan kata-kata yang digunakan, topik yang dibahas, atau konteks umum lainnya. Salah satu algoritma clustering yang umum digunakan dalam text clustering adalah K-Means, Hierarchical Clustering, DBSCAN, dan masih banyak lagi. Text clustering memiliki banyak aplikasi, termasuk klasifikasi otomatis dokumen, pengelompokkan berita atau artikel berdasarkan topik, dan analisis sosial media untuk mengidentifikasi tren atau kelompok pendapat yang berbeda.

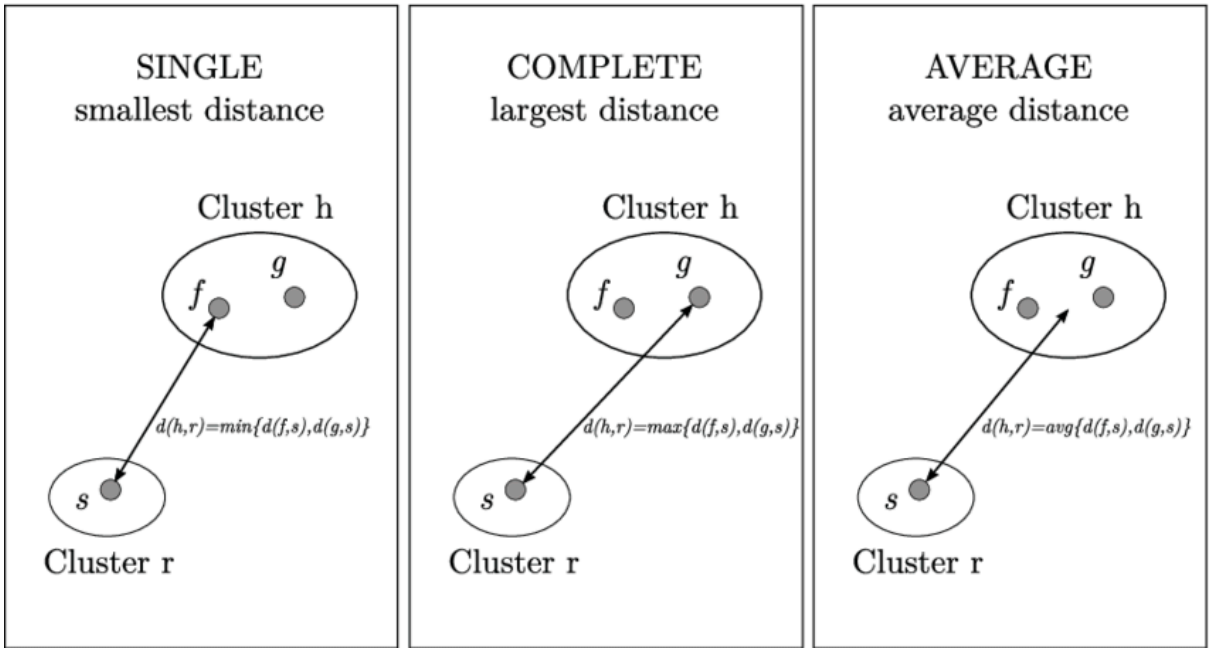
Ada dua jenis Hierarchical Clustering, yaitu Agglomerative dan Divisive. Agglomerative Hierarchical Clustering dimulai dengan setiap dokumen teks dianggap sebagai sebuah kluster tunggal, kemudian pada setiap iterasi, dua kluster yang paling mirip (atau terdekat) digabungkan menjadi satu kluster baru. Divisive Hierarchical Clustering dimulai dengan dimulai dengan satu kluster tunggal yang berisi seluruh dataset, kemudian pada setiap iterasi, kluster yang paling tidak homogen (atau paling berbeda) dibagi menjadi dua kluster baru.



sumber:
https://www.researchgate.net/publication/276332381/figure/fig1/AS:520107005038592@1501014581002/Difference-between-agglomerative-and-divisive-hierarchical-clustering-methods.png

Dalam hierarchical clustering, terdapat beberapa teknik pengelompokan yang biasa digunakan:

- Single linkage (jarak terdekat atau tautan tunggal): Teknik yang menggabungkan kluster-kluster menurut jarak antara anggota-anggota terdekat di antara dua kluster.
- Average linkage (jarak rata-rata atau tautan rata-rata): Teknik yang menggabungkan kluster-kluster menurut jarak rata-rata pasangan anggota masing-masing pada himpunan antara dua kluster.
- Complete linkage (jarak terjauh atau tautan lengkap): Teknik yang menggabungkan kluster-kluster menurut jarak antara anggota-anggota terjauh di antara dua kluster.



sumber:
https://www.researchgate.net/publication/329208978/figure/fig5/AS:755481513562120@1557132237914/Different-linkage-methods-for-hierarchical-clustering.png

Ada beberapa rumus jarak yang biasa digunakan untuk perhitungan jarak, di antaranya adalah:

- Cosine similarity:

$$\text{Cos}(A,B) = \frac{A \cdot B}{||A||\,||B||}$$

A·B is the product (dot) of the vectors A and B

||A|| and ||B|| is the length of the two vectors

The **cosine distance** formula is then: 1 - Cosine Similarity

- Euclidean distance:

$$d(A, B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

Tugas: Melakukan Text Clustering menggunakan Hierarchical Agglomerative Clustering

Pada tugas ini, Anda diberikan beberapa kalimat yang perlu Anda lakukan Text Clustering menggunakan Hierarchical Agglomerative Clustering. Berikut kalimatnya:

1. Masyarakat bersiap menyambut Lebaran dengan penuh harap dan kegembiraan.
2. Indonesia menang 3 gol tanpa balas melawan Vietnam.
3. Korupsi timah merugikan negara Indonesia 271 Triliun
4. Libur lebaran akan dimulai pada tanggal 6 April 2024.
5. Perkembangan teknologi kecerdasan buatan sangat cepat

6. Saya makan nasi goreng untuk buka puasa
7. Indonesia berpeluang lolos ke piala dunia 2026

Soal:

1. Lakukan tahap preprocessing untuk teks berita tersebut secara manual dan jelaskan setiap perubahan yang Anda lakukan dengan lengkap.
2. Lakukan vektorisasi unigram menggunakan algoritma TF-IDF secara manual dan jelaskan setiap langkah yang Anda lakukan dengan lengkap.
3. Lakukan pengelompokkan kalimat menggunakan algoritma Hierarchical Agglomerative Clustering dengan menggunakan teknik pengelompokkan smallest distance dan rumus jarak euclidean distance dan cosine similarity (lakukan 2 kali pengelompokkan, 1 menggunakan euclidean distance dan 1 menggunakan cosine similarity) secara manual dan jelaskan setiap langkah yang Anda lakukan dengan lengkap.
4. Tuliskan hasil akhir yaitu dendrogram yang berhasil dibuat. Anda bisa menggambar menggunakan tulisan tangan kemudian difoto/dipindai untuk hal ini. Pastikan hasil gambar dapat terlihat dengan jelas. Anda cukup menuliskan nomor kalimat untuk mewakili kalimat tersebut dalam dendrogram.

Petunjuk pengerjaan:

- Silakan tulis semua jawaban jelas dan lengkap secara langsung di sebuah file dan kumpulkan dalam format pdf. Anda tidak perlu menuliskan soal, cukup menuliskan jawaban setiap nomor.
- Anda tidak diperbolehkan menggunakan bantuan program apapun dalam proses pengerjaaan setiap soal.
- Dilarang keras menyontek. Plagiarisme tidak ditoleransi dan akan dikenai penalti atau nilai akhir E.
- Pengurangan nilai akibat keterlambatan pengumpulan tugas akan ditentukan berdasarkan jumlah menit keterlambatan Anda dalam mengumpulkan. Misalnya, apabila terlambat 1 menit, nilai akhir akan dikurangi 1 poin, apabila terlambat 10 menit, nilai akhir akan dikurangi 10 poin, dan seterusnya.

Bobot penilaian:

- Soal 1: 20 poin
- Soal 2: 30 poin
- Soal 3: 40 poin
- Soal 4: 10 poin

Pengumpulan tugas:

Kumpulkan berkas dengan format penamaan seperti berikut:

Assignment7_[NPM]_[NamaLengkap].pdf

Contoh:
Assignment7_2006596535_FransiscoWilliamSudianto.pdf