Text Mining

Text Clustering

| Outline

1. Intro Text Clustering
2. Jarak antar Cluster
3. Tahapan Clustering
4. Algoritma K-Means
5. Algoritma Agglomerative Hierarchical (single linked)

| #1 Intro Text Clustering

- Cluster: kumpulan objek data
  - Anggota cluster yang sama memiliki kemiripan satu sama lain, tetapi berbeda dengan anggota cluster lain.
- Cluster analysis
  - Menemukan kemiripan data berdasarkan karakteristik dan mengelompokan data yang mirip ke dalam cluster.
  - Mengelompokkan objek menjadi satu kelompok jika objek-objek tsb “mirip”(berkaitan/dekat) dan membuat kelompok yang berbeda jika objek itu “berbeda”
- Unsupervised learning: class tidak ditentukan sebelumnya

| #1 Intro Text Clustering

Problem Statement:

- Dinyatakan:
  - Himpunan dokumen D = {d1, d2, …, dN}
  - Jumlah cluster, K
  - Fungsi objektif untuk evaluasi clustering
    - Fungsi objektif didefinisikan dalam istilah kempiripan atau jarak antar dokumen
- Ingin dihitung persamaan γ = D → {1, …, K} yang menimalkan fungsi objektif atau memastikan tidak ada K cluster yang kosong}

| #1 Intro Text Clustering

Jumlah K:

- Dalam kebanyakan algoritma clustering, pemilihan jumlah cluster atau K sangan ditentukan oleh proses inisialisasi
- Beberapa *rule of thumb*
  - $k \approx \sqrt{\frac{n}{2}}$
  - *(m x n)/t*
    - m = jumlah dokumen
    - n = jumlah term
    - t = jumlah non-zero entri

$$D = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array} \begin{array}{c} \begin{array}{cccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set#cite_note-0

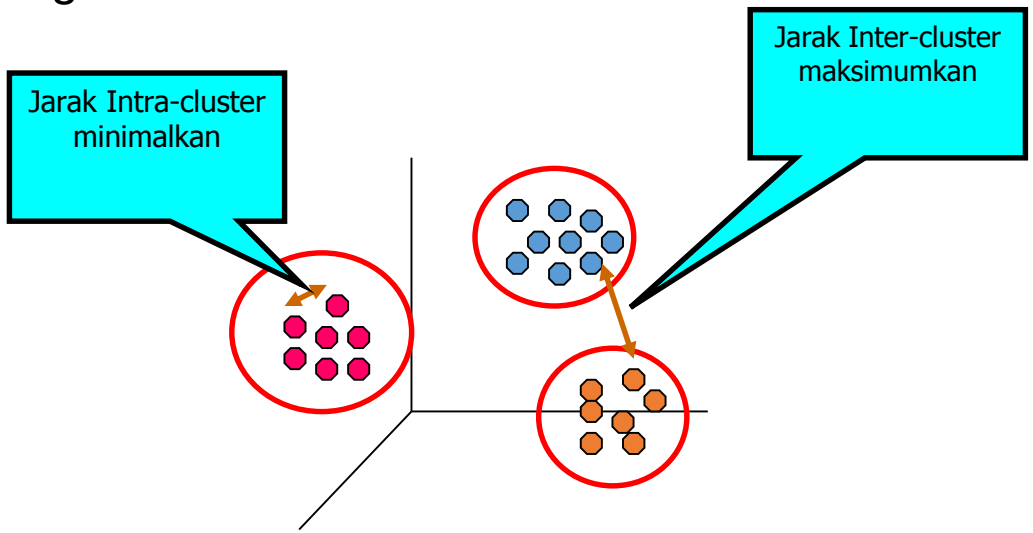
| #1 Intro Text Clustering

Parameter Evaluasi

- Metode yang bagus akan menghasilkan:
  - intra-class similarity yang tinggi (anggota di dalam kelas yang sama mirip)
  - low inter-class similarity (anggota di kelas yang lain, jauh berbeda)
- Kualitas cluster bergantung kepada ukuran kemiripan yang digunakan oleh metode clustering.

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, m_i)^2$$

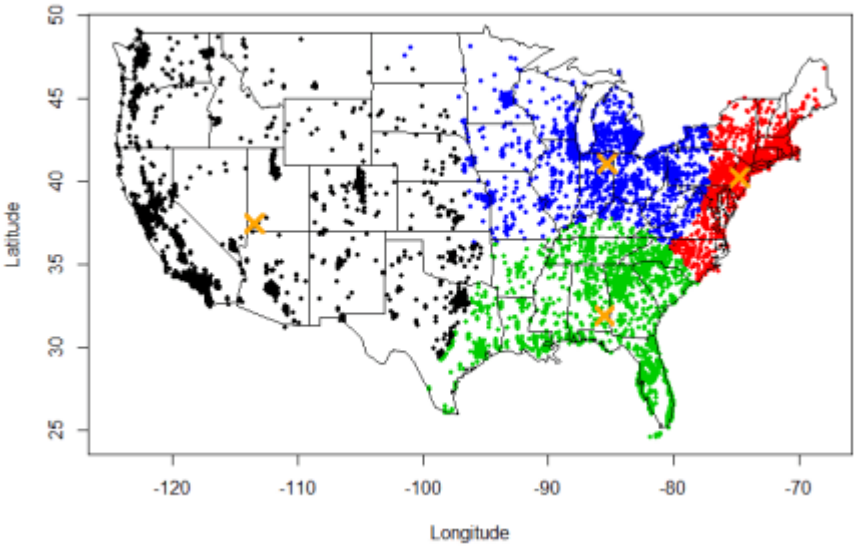
- Kualitas juga ditentukan sejauh mana clustering dapat menemukan pola tersembunyi.



| #1 Intro Text Clustering

Contoh Implementasi :

- Marketing :
  - Membantu marketer untuk mengelompokan kebiasaan pelanggannya
  - Membantu marketer menentukan jenis barang yang laku dijual di daerah tertentu (c. Alfamart)
- Asuransi :
  - Kelompok motor /mobil yang mudah rusak sehingga kemungkinan besar akan terjadi klaim asuransi
- Tata Kota :
  - Daerah yang cocok untuk dibangun perkantoran, perumahan, perbelanjaan, dsb
- Text mining:
  - Pengelompokan dokumen berita, pengelompokan sentiment social media



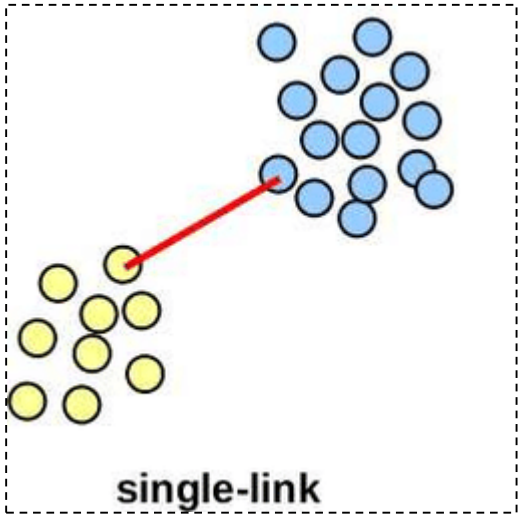
| #1 Intro Text Clustering

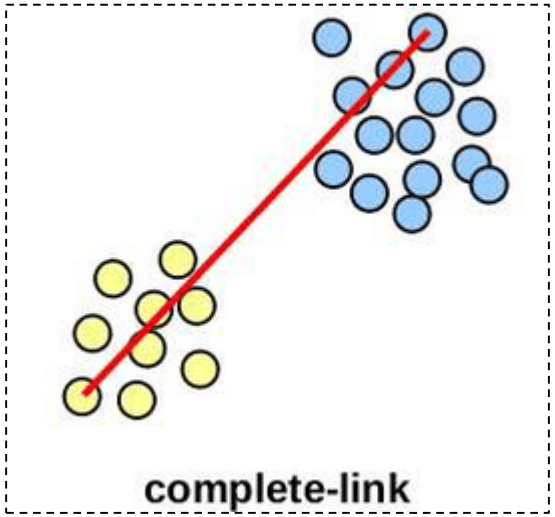
Beberapa metode clustering:

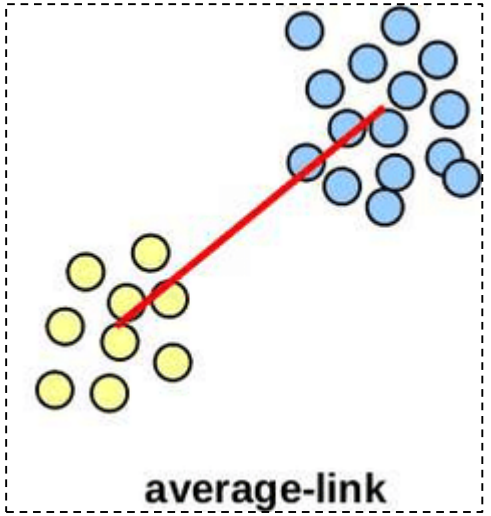
- Partisi :
  - Buat partisi dan evaluasi berdasarkan kriteria tertentu, misalnya meminimalkan sum of square errors
  - Metode: k-means, k-medoids, CLARANS
- Hirarkis:
  - Buat struktur hierarchical menggunakan kriteria tertentu
  - Metode: Diana, Agnes, BIRCH, ROCK, CAMELEON
- Density-based :
  - Berdasarkan connectivity dan density functions
  - Metode: DBSACN, OPTICS, DenClue
- Grid based :
  - data dianggap sebagai benda n dimensi
  - STING,CLIQUE
- Yang lain: Grid-based approach, model-based, frequent pattern-based, user-guided or constraint-based.

| #2 Jarak Antar Cluster

- Single link: jarak terpendek antar elemen di dua cluster $dis(K_i, K_j) = min(t_{ip}, t_{jq})$
- Complete link: jarak terjauh antar elemen di dua cluster, i.e., $dis(K_i, K_j) = max(t_{ip}, t_{jq})$
- Average: rata2 jarak i.e., $dis(K_i, K_j) = avg(t_{ip}, t_{jq})$
- Centroid: jarak antara centroids, i.e., $dis(K_i, K_j) = dis(C_i, C_j)$
- Medoid: jarak antarta medoids, i.e., $dis(K_i, K_j) = dis(M_i, M_j)$
  - Medoid: elemen yang dipilih dan dianggap merupakan titik tengah cluster







#2 Jarak Antar Cluster

Beberapa macam jarak yang dapat digunakan

- Jarak Euclidean

$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- Jarak Manhattan atau Cityblock

$$d(x,y) = \sum_{i=1}^{n}(|x_i - y_i|)$$

- Jarak Minkowski

$$d(x,y) = \|x - y\|_q = (\sum |x - y|^q)^{\frac{1}{q}}$$

- Jarak Mahalanobis

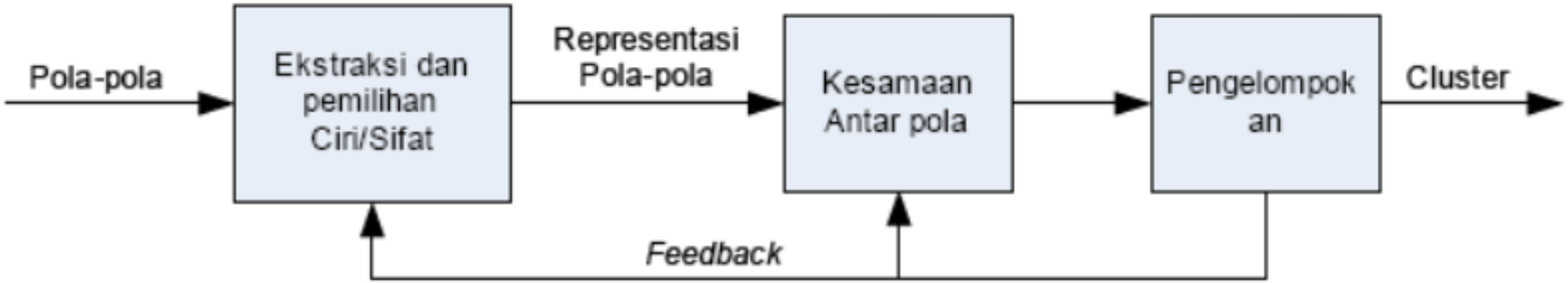
$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1}(\vec{x} - \vec{y})}$$

Jarak yang umum digunakan untuk clustering dokumen adalah cosine distance

$$\boldsymbol{Cosine\ distance = 1 - cosine\ similarity}$$

| #3 Tahapan Clustering

1. Representasi pola : jumlah cluster, tipe cluster, karakteristik sesuai dengan algoritma
2. Pemilihan fitur : proses identifikasi fitur/ciri yang lebih efektif digunakan dalam algoritma clustering, sedangkan extraksi fitur adalah pemakaian satu atau lebih trasofrmasi fitur yang ada sebelumnya untuk mendapatkan fitur yang lebih menonjol
3. Kedekatan pola biasanya diukur dengan fungsi jarak antar dua pasang pola



| #4 K-Means

- Merupakan salah satu algoritma *Partitional clustering*
- Setiap klaster berkaitan dengan sebuah titik pusat klaster (*centroid*)
- Setiap titik dimasukkan ke dalam klaster dengan centroid terdekat.
- Jumlah klaster harus ditentukan sebelumnya

[+]

- Mudah dilakukan saat pengimpelementasian dan di jalankan.
- Waktu yang di butuhkan untuk melakukan pembelajaran relatif lebih cepat.
- Sangat fleksibel, adaptasi yang mudah untuk di lakukan
- Sangat umum penggunaannya.
- Menggunakan prinsip yang sederhana dapat di jelaskan dalam non-statistik.

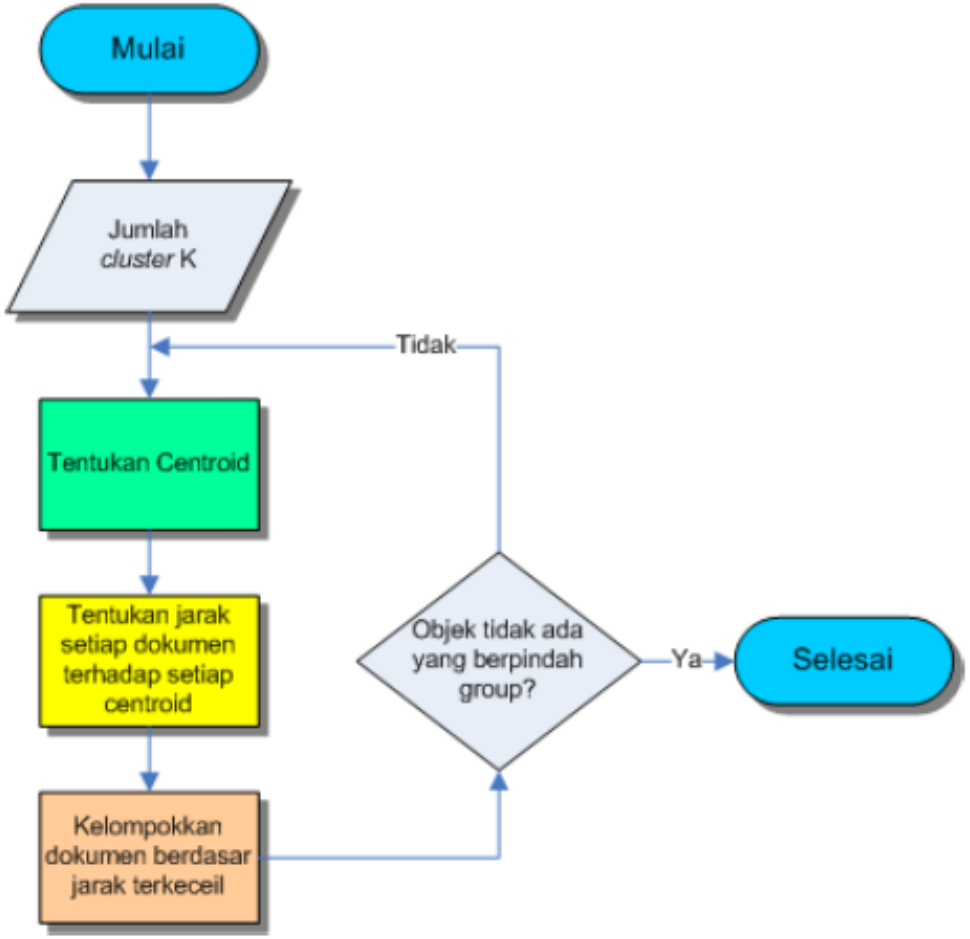
[-]

- K-Means sangat sensitif pada pembangkitan centroid awal secara random
- Memungkinkan suatu cluster tidak mempunyai anggota
- K-means sangat sulit untuk mencapai global optimum
- Tidak dapat menangani outlier

| #4 K-Means

Algorithm 1 Basic K-means Algorithm.

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change



| #4 K-Means (simulasi)

d1

Sekarang saya sedang suka memasak. Masakan kesukaan saya sekarang adalah nasi goreng. Cara memasak nasi goreng adalah nasi digoreng

d2

Ukuran nasi sangatlah kecil, namun saya selalu makan nasi

d3

Nasi berasal dari beras yang ditanam di sawah. Sawah berukuran kecil hanya bisa ditanami sedikit beras

d4

Mobil dan bus dapat mengangkut banyak penumpang. Namun, bus berukuran jauh lebih besar dari mobil, apalagi mobil-mobilan

d5

Bus pada umumnya berukuran besar dan berpenumpang banyak, sehingga bus tidak bisa melewati persawahan

Dokumen yang ada akan dikelompokkan menjadi **2 cluster**

#4 K-Means (simulasi)

Bobot TF-IDF ternormalisasi

TERM	D1	D2	D3	D4	D5
suka	0,51932239	0	0	0	0
masak	0,58961142	0	0	0	0
nasi	0,18713901	0,36624274	0,158012148	0	0
goreng	0,58961142	0	0	0	0
ukur	0	0,28150215	0	0,148958474	0,278625389
makan	0	0,88691758	0	0	0
beras	0	0	0,647708116	0	0
tanam	0	0	0,647708116	0	0
sawah	0	0	0,368755414	0	0,499782802
mobil	0	0	0	0,751874826	0
bus	0	0	0	0,347626715	0,650232416
angkut	0	0	0	0,469317522	0
tumpang	0	0	0	0,267193467	0,499782802

| #4 K-Means (simulasi)

Penentuan Centroid dan hitung jarak

- Misal, D1 dipilih sebagai c_1 (centroid 1) dan D4 dipilih sebagai c_2 (centroid 2)

c_1	0.519	0.590	0.187	0.590	0	0	0	0	0	0	0	0	0
c_2	0	0	0	0	0.149	0	0	0	0	0.752	0.348	0.469	0.267

- Hitung jarak dari masing-masing dokumen ke tiap centroid
  - $d(D1, c_1) = 1 - cosine(D1, c_1) = 1 - 1 = 0$
  - $d(D1, c_2) = 1 - cosine(D1, c_2) = 1 - 0 = 1$

| #4 K-Means (simulasi)

Penentuan keanggotaan cluster

Dokumen	Jarak ke c_1	Jarak ke c_2	Min	Anggota cluster
D1	0	1	0	1
D2	0.931	0.958	0.931	1
D3	0.970	1	0	1
D4	1	0	0	2
D5	1	0.599	0.599	2

#4 K-Means (simulasi)

Hitung ulang centroid

Cluster 1													
D1	0.519	0.590	0.187	0.590	0	0	0	0	0	0	0	0	0
D2	0	0	0.366	0	0.282	0.887	0	0	0	0	0	0	0
D3	0	0	0.158	0	0	0	0.648	0.648	0.369	0	0	0	0
c_1	0.173	0.197	0.237	0.197	0.094	0.296	0.216	0.216	0.123	0	0	0	0

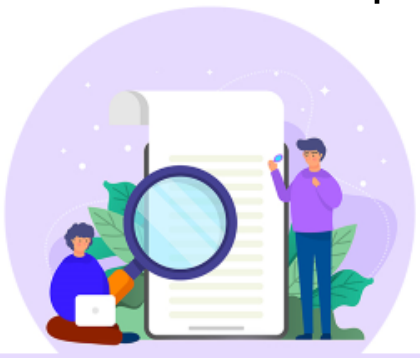
Cluster 2													
D4	0	0	0	0	0.149	0	0	0	0	0.752	0.348	0.469	0.267
D5	0	0	0	0	0.279	0	0	0	0.5	0	0.65	0	0.5
c_2	0	0	0	0	0.214	0	0	0	0.250	0.376	0.499	0.235	0.383

| #4 K-Means (simulasi)

Hitung jarak dan penentuan anggota cluster

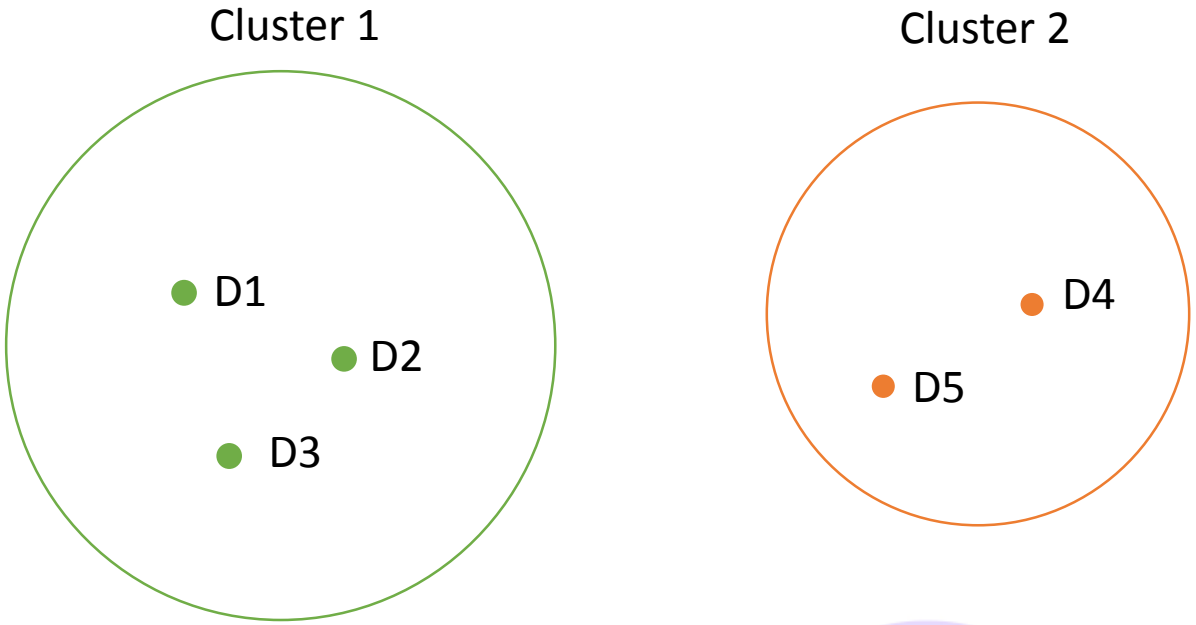
Dokumen	Jarak ke c_1	Jarak ke c_2	Min	Anggota cluster	Anggota cluster sebelumnya
D1	0.634	1	0.634	1	1
D2	0.625	0.940	0.625	1	1
D3	0.638	0.908	0.638	1	1
D4	0.986	0.299	0.299	2	2
D5	0.912	0.299	0.299	2	2

- Keanggotaan cluster tidak berubah antara iterasi sekarang dan iterasi berikutnya, sehingga K-Means telah mencapai keadaan konvergen dan iterasi dihentikan



| #4 K-Means (simulasi)

Hasil akhir clustering (ilustrasi)



| #5 Agglomerative Hierarchical

- Hierarchical Clustering adalah metode analisis kelompok yang berusaha untuk membangun sebuah hirarki kelompok data.
- Strategi pengelompokannya umumnya ada 2 jenis yaitu **Agglomerative (Bottom-Up)** dan **Devisive (Top-Down)**.

- **Algorithm with Single-Linkage (Pseudo Code)**(2)

Input: $\boldsymbol{D}=\{\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_n\}$ // *Set of elements;*

$\boldsymbol{A}$ // $\boldsymbol{n}$*$\boldsymbol{n}$ *proximity or adyacency matrix* $\boldsymbol{A} = [d(i,j)]$ *that showing distance between* $\boldsymbol{x}_i$, $\boldsymbol{x}_j$;

C_r // $\boldsymbol{r}$-*th cluster, with* $1 \le \boldsymbol{r} \le \boldsymbol{n}$; $d\,[C_r, C_s]$ // *Proximity between clusters* C_r *and* C_s;

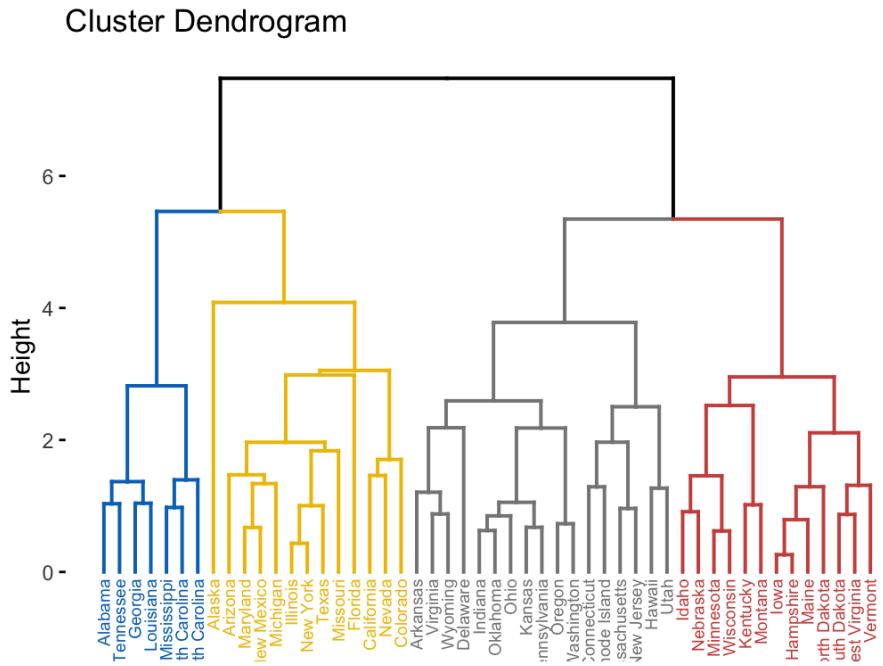
$\boldsymbol{k}$ // *Sequence number, with* $\boldsymbol{k}=0,1,\ldots\boldsymbol{n}-1$; $\boldsymbol{L}(\boldsymbol{k})$ // *Distance-level of the* $\boldsymbol{k}$-*th clustering;*

Output: // *Dendrogram;*

Algorithm:

1. *Begin with* $\boldsymbol{n}$ *clusters, each containing one object and having level* $\boldsymbol{L}(0) = 0$ *and sequence number* $\boldsymbol{k} = 0$.
2. *Find the least dissimilar pair* (C_r, C_s) *in the current clustering, according to*
   $d[C_r, C_s] = \boldsymbol{min}(d[C_i, C_j])$
   where the minimum is over all pairs of clusters (C_i, C_j) *in the current clustering.*
3. *Increment the sequence number :* $\boldsymbol{k} = \boldsymbol{k} + 1$. *Merge clusters* C_r *and* C_s *into a single cluster to form the next clustering* $\boldsymbol{k}$. *Set the level of this clustering to* $\boldsymbol{L}(\boldsymbol{k}) = d[C_r, C_s]$.
4. *Update the proximity matrix,* $\boldsymbol{D}$, *by deleting the rows and columns corresponding to clusters* C_r *and* C_s *and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted* C_{r+s} *and old cluster* C_a *is defined in this way:*
   $d[C_a, C_{r+s}] = \boldsymbol{min}(d[C_a, C_r], d[C_a, C_s])$.
5. *If all objects are in one cluster, stop. Else, go to step 2.*

(2) *Ref. S. C. Johnson (1967): "Hierarchical Clustering Schemes" Psychometrika, 2:241-254*



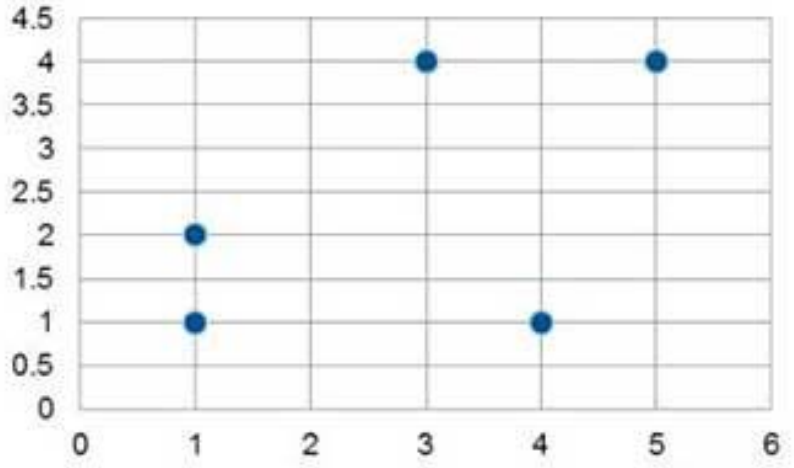
| #5 Agglomerative Hierarchical (single linked)

Langkah Algoritma Agglomerative Hierarchical Clustering :

1. Hitung Matrik Jarak antar data.
2. Gabungkan dua kelompok terdekat berdasarkan parameter kedekatan yang ditentukan.
3. Perbarui Matrik Jarak antar data untuk merepresentasikan kedekatan diantara kelompok baru dan kelompok yang masih tersisa.
4. Ulangi langkah 2 dan 3 higga hanya satu kelompok yang tersisa.

contoh :

Data	Fitur x	Fitur y
1	1	1
2	4	1
3	1	2
4	3	4
5	5	4



| #5 Agglomerative Hierarchical (single linked)

Data	Fitur x	Fitur y
1	1	1
2	4	1
3	1	2
4	3	4
5	5	4

Menghitung Jarak Pada Semua Pasangan dua data
Manhattan:

D_{man} (*Data1*, *Data1*) = |1-1| + |1-1| = 0
D_{man} (*Data1*, *Data2*) = |1-4| + |1-1| = 3
D_{man} (*Data1*, *Data3*) = |1-1| + |1-2| = 1
D_{man} (*Data1*, *Data4*) = |1-3| + |1-4| = 5
D_{man} (*Data1*, *Data5*) = |1-5| + |1-4| = 7
D_{man} (*Data2*, *Data3*) = |4-1| + |1-2| = 4
D_{man} (*Data2*, *Data4*) = |4-3| + |1-4| = 4
D_{man} (*Data2*, *Data5*) = |4-5| + |1-4| = 4
D_{man} (*Data3*, *Data4*) = |1-3| + |2-4| = 4
D_{man} (*Data3*, *Data5*) = |1-5| + |2-4| = 6
D_{man} (*Data4*, *Data5*) = |3-5| + |4-4| = 2

Dman	1	2	3	4	5
1	0	3	1	5	7
2	3	0	4	4	4
3	1	4	0	4	6
4	5	4	4	0	2
5	7	4	6	2	0

| #5 Agglomerative Hierarchical (single linked)

- Dengan memperlakukan data sebagai kelompok, selanjutnya kita pilih jarak dua kelompok yang terkecil.
- $\min(D_{man}) = \min(d_{13}) = 1$
- Terpilih kelompok 1 dan 3, sehingga kedua kelompok ini digabungkan.
- Menghitung jarak antar kelompok (1 dan 3) dengan kelompok lain yang tersisa, yaitu 2, 4 dan 5.
  - $d_{(13)2} = \min \{d_{12}, d_{32}\} = \min \{3,4\} = 3$
  - $d_{(13)4} = \min \{d_{14}, d_{34}\} = \min \{5,4\} = 4$
  - $d_{(13)5} = \min \{d_{15}, d_{35}\} = \min \{7,6\} = 6$
- Dengan menghapus baris-baris dan kolom-kolom matrik jarak yang bersesuaian dengan kelompok 1 dan 3, serta menambahkan baris dan kolom untuk kelompok (13).

↓

Dman	(13)	2	4	5
(13)	0	3	4	6
2	3	0	4	4
4	4	4	0	2
5	6	4	2	0

| #5 Agglomerative Hierarchical (single linked)

Dman	(13)	2	4	5
(13)	0	3	4	6
2	3	0	4	4
4	4	4	0	2
5	6	4	2	0

- Selanjutnya dipilih jarak dua kelompok yang terkecil.
- $\min(D_{man}) = \min(d_{45}) = 2$
- Menghitung jarak antar kelompok (4 dan 5) dengan kelompok lain yang tersisa, yaitu (13) dan 2.
  - $d_{(45)(13)} = \min \{d_{41} , d_{43} , d_{51} , d_{53}\} = \min \{5,4,7,6\} = 4$
  - $d_{(45)2} = \min \{d_{42} , d_{52}\} = \min \{4,4\} = 4$
- Menghapus baris dan kolom matrik yang bersesuaian dengan kelompok 4 dan 5, serta menambahkan baris dan kolom untuk kelompok (45)

Dman	(45)	(13)	2
(45)	0	4	4
(13)	4	0	3
2	4	3	0

| #5 Agglomerative Hierarchical (single linked)

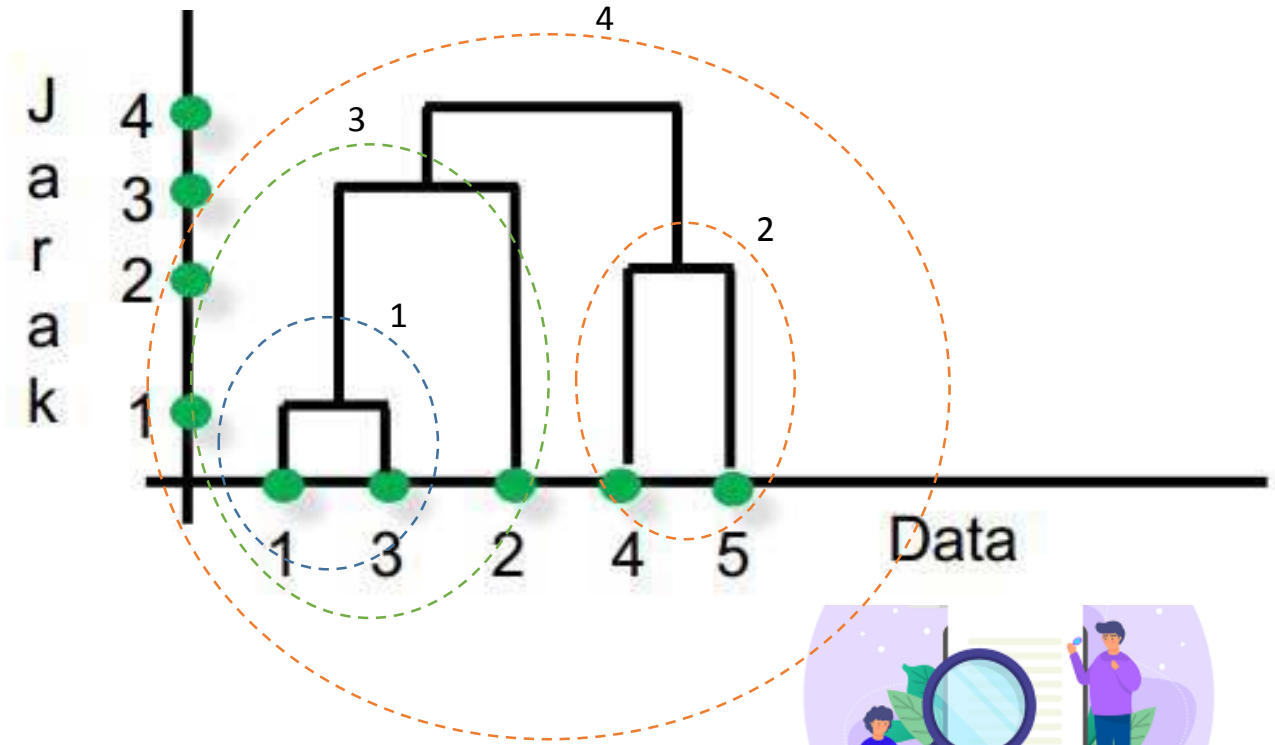
Dman	(45)	(13)	2
(45)	0	4	4
(13)	4	0	3
2	4	3	0

- Selanjutnya dipilih jarak dua kelompok yang terkecil.
- $\min(D_{man}) = \min(d_{(13)2}) = 3$
- Terpilih kelompok (13) dan 2, sehingga kedua kelompok ini digabungkan. (Melanjutkan pengelompokan).
- Menghitung jarak antar kelompok ((13) dan 2) dengan kelompok lain yang tersisa, yaitu (45).
  - $d_{(132)(45)} = \min \{d_{14}, d_{15}, d_{34}, d_{35}, d_{24}, d_{25}\} = \min \{5,7,4,6,4,4\} = 4$
- Menghapus baris dan kolom matrik yang bersesuaian dengan kelompok (13) dan 2, serta menambahkan baris dan kolom untuk kelompok (123).

Dman	(132)	(45)
(132)	0	4
(45)	4	0

| #5 Agglomerative Hierarchical (single linked)

- Jadi kelompok (132) dan (45) digabung untuk menjadi kelompok tunggal dari lima data, yaitu kelompok (13245) dengan jarak terdekat 4.
- Berikut Dendogram Hasil Metode Single Linkage :



TASK

Text PreProcessing with Python + Project (idea)

check our gclassroom

Thank you

</ti>
FST | UINAM
teknik informatika

UNIVERSITAS ISLAM NEGERI
ALAUDDIN
MAKASSAR