

# IMPLEMENTASI METODE AGGLOMERATIVE HIERARCHICAL CLUSTERING UNTUK SISTEM REKOMENDASI FILM

Vanesa Nellie <sup>1)</sup> Viny Christanti Mawardi <sup>2)</sup> Novario Jaya Perdana <sup>3)</sup>

<sup>1)</sup> Teknik Informatika, FTI, Universitas Tarumanagara  
Jl. Letjen S Parman no 1, Jakarta 11440 Indonesia

[vanesa.535190010@stu.untar.ac.id](mailto:vanesa.535190010@stu.untar.ac.id) <sup>1)</sup> [viny@fti.untar.ac.id](mailto:viny@fti.untar.ac.id) <sup>2)</sup> [novariojp@fti.untar.ac.id](mailto:novariojp@fti.untar.ac.id) <sup>3)</sup>

## ABSTRACT

*People can now watch movies on their cellphones or other devices using applications, in addition to watching them on television or in theaters. The user's entered keywords are used as the basis for a system that suggests movies from among the many that have appeared over time. Later, similarity between these keywords and text data, such as movie titles and descriptions, will be assessed. This recommendation system will include preprocessing, and the TF-IDF method will be used to determine the weight value. After the weight values have been determined, the grouping calculations will be performed using agglomerative hierarchical clustering. Previously, the Manhattan Distance method will be used to calculate the distance. After that, the distance that is closest can be determined. The data will be clustered according to the shortest distance once the distance calculation is complete. Following that, the system will display the grouping as a dendrogram. The data used was updated as of the date of scraping, which is November 25, 2022, and contains a total of 2467 data. The Agglomerative Hierarchical Clustering method yielded the best silhouette coefficient value, 0.5025559374455285, forming 20 clusters.*

## Key words

*Preprocessing, Agglomerative Clustering, Manhattan Distance*

## 1. Pendahuluan

Netflix merupakan salah satu aplikasi yang memberikan layanan pada penggunanya dengan menyediakan film/drama/konten terbaru sehingga pengguna tidak akan kehabisan tontonan. Netflix juga menyediakan deskripsi serta genre pada informasi film agar konsumen dapat melihat ringkasan informasi dari film tersebut. Dikarenakan banyaknya judul film yang disediakan, maka akan dibuatkan sistem rekomendasi

untuk menampilkan rekomendasi film sesuai dengan kata kunci yang dicari konsumen. Metode yang akan digunakan untuk mencari film sesuai dengan kemiripan pada kata kunci yang dimasukkan oleh konsumen adalah metode *text mining*. *Text mining* merupakan proses pengolahan data yang bertujuan untuk mendapatkan kata-kata yang berkaitan dengan isi dari suatu kolom data sehingga dapat dilakukan untuk mendeteksi korelasi antar kolom data yang akan dipakai. Pada sistem rekomendasi yang dirancang, kata kunci yang dimasukkan akan dibandingkan dengan deskripsi yang berupa data teks dari masing-masing film. Jika kata kunci tersebut mirip dengan deskripsi film, maka film tersebut akan menjadi rekomendasi.

Salah satu teknik dari metode *text mining* yang dapat diterapkan adalah teknik *clustering*. Teknik *clustering* ini adalah salah satu metode untuk mengelompokkan data menjadi beberapa kluster berdasarkan keterkaitan atau kemiripan suatu data dengan data lainnya.

Proses *text mining* dapat dimanfaatkan sebagai sistem rekomendasi, dimana misalnya pengguna memasukkan suatu kata kunci, dokumen teks akan diolah untuk di kelompokkan menjadi beberapa kelompok yang memiliki jarak terdekat, lalu hasil pengelompokan tersebut dapat dibandingkan kemiripannya dengan kata kunci yang dimasukkan.

Pendekatan dalam mengembangkan metode *clustering* salah satunya adalah pendekatan secara hirarki atau yang biasa disebut sebagai *hierarchical clustering*. *Hierarchical clustering* merupakan teknik pengelompokan untuk membentuk suatu hirarki kelompok data. Metode hirarki ini memiliki 2 jenis strategi pengelompokan, yaitu *divisive (top-down)* dan *agglomerative (bottom-up)*. *Agglomerative* merupakan metode pengelompokan data yang dimulai dari setiap salah satu objek data akan dianggap sebagai satu kluster, lalu akan terus mengelompokkan objek data,

jika memiliki kemiripan akan digabungkan menjadi satu kluster. Sedangkan *divisive* merupakan metode pengelompokan data dimulai dari membuat suatu kluster yang terdiri dari seluruh objek kemudian objek akan dipisahkan ke dalam kluster terpisah jika merupakan objek yang paling berbeda kemiripannya.

Sistem rekomendasi ini nantinya akan menggunakan metode *Agglomerative Hierarchical Clustering* (AHC) untuk membuat kluster dari data-datanya. Kluster ini nantinya akan dibandingkan kemiripan/keterkaitannya berdasarkan dari kata kunci yang diinput oleh pengguna dengan deskripsi dari setiap konten yang ada di data Netflix. Teknik yang diterapkan pada metode AHC ini ada 3, yaitu *single linkage*, *average linkage*, dan *complete linkage*. Keunggulan dari menggunakan metode AHC ini dibandingkan metode *clustering* lainnya adalah metode ini dapat menampilkan hasil dari jarak kedekatan antar data dengan menampilkan dendrogram. Dendrogram merupakan suatu bagan hirarki yang tujuannya untuk menampilkan kemiripan dari setiap data.

Tujuan dari pembuatan sistem rekomendasi konten Netflix ini adalah untuk menghasilkan aplikasi berbasis web yang dapat membantu pengguna untuk mencari konten yang relevan dengan yang dicari serta menerapkan sistem rekomendasi konten Netflix dengan menggunakan metode AHC.

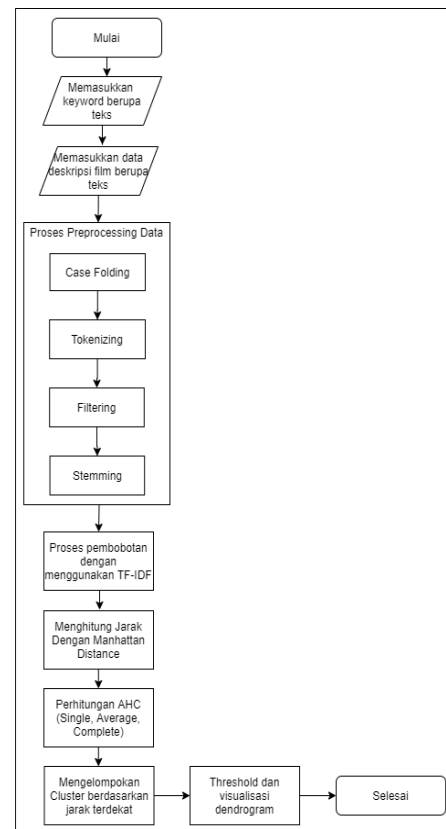
Kegunaan dari perancangan sistem rekomendasi konten Netflix ini adalah pengelolaan data dapat dilakukan secara efisien untuk menghasilkan data yang lebih akurat, pengguna dapat mengetahui konten-konten yang menarik dari Netflix, membantu pengguna untuk menemukan konten yang serupa dengan yang dicari, serta pengguna juga dapat lebih mudah memilih konten aplikasi Netflix yang ingin ditonton berdasarkan dari daftar rekomendasi konten.

## 2. Landasan Teori

Proses pada rancangan ini berawal dari memasukkan data yang berbentuk excel. Lalu data yang berupa teks akan melalui proses perhitungan untuk menghitung nilai bobotnya menggunakan metode TF-IDF. Setelah menghitung nilai bobot, maka akan masuk ke perhitungan matriks jaraknya dengan menggunakan metode *Manhattan Distance*. Lalu, akan dilakukan pengelompokan dengan menggunakan *Agglomerative Hierarchical Clustering*. Jika perhitungan jarak tersebut telah selesai, maka data akan dikelompokkan ke dalam cluster berdasarkan jarak terdekatnya. Setelah itu, matriks jaraknya akan diperbarui dan sistem akan membuat visualisasi pengelompokannya dalam bentuk dendrogram. Pemotongan banyaknya *cluster* juga dapat

di lakukan dengan menentukan nilai ambang batas (*threshold*).

Sistem yang akan dirancang akan berupa suatu aplikasi berbasis website. Aplikasi ini merupakan aplikasi yang dirancang untuk menjadi sistem rekomendasi yang prosesnya akan berawal dari *user* memasukkan suatu keyword pada sistem aplikasi, lalu outputnya adalah beberapa judul dan deskripsi konten Netflix yang mirip dengan *keyword* yang di cari.



Gambar 1. Flowchart perhitungan

### 2.1. Text Preprocessing

Tahapan *preprocessing* adalah proses untuk mengolah data mentah sebelum dilakukan pemrosesan lain. Tahapan *preprocessing* ini dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh system. [1] Pada tahap ini, akan dilakukan tahapan semantik dan sintaktik pada teks yang akan diolah. Tahap pengolahan tersebut berupa:

- Case folding* yaitu pengubahan huruf kapital menjadi huruf kecil. Contohnya adalah mengubah kata dari “HeLLo” menjadi “hello”.
- Tokenization* yaitu proses pemotongan/pemisahan teks menjadi beberapa bagian lebih kecil. Contohnya

adalah memisahkan teks dari “saya ingin makan” menjadi “[‘saya’, ‘ingin’, ‘makan’].

- c. *Stemming* yaitu mengubah bentuk kata menjadi kata dasar/penghapusan imbuhan. Contohnya adalah mengubah kata dari “merespon” menjadi “respon”
- d. *Stopword Removal* yaitu pembuangan kata-kata yang tidak terlalu penting/tidak memiliki makna. Contohnya yaitu dari kalimat “Budi sering membeli barang yang murah dan praktis” dipilih kata pentingnya menjadi “[‘Budi’, ‘sering’, ‘beli’, ‘murah’, ‘praktis’]. [2]

## 2.2. TF-IDF (*Term Frequency – Inverse Document Frequency*)

Dalam proses *text mining*, data yang berupa teks akan diberikan nilai pembobotan. Pembobotan tersebut dapat menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). TF-IDF merupakan suatu metode algoritma dalam menghitung pembobotan setiap teks pada data. Semakin banyak frekuensi kemunculan sebuah kata pada suatu data/dokumen, maka akan semakin besar nilainya. Metode TF-IDF menggabungkan dua teori untuk melakukan perhitungan bobot yaitu perhitungan frekuensi kemunculan suatu kata pada sebuah dokumen yang akan di gunakan dan perhitungan invers frekuensi dokumen yang mengandung kata tersebut. Berikut adalah rumus untuk *Term Frequency* (TF). [3]

$$TF = 0.5 + 0.5 \times \frac{tf}{\max(tf)} \tag{1}$$

Keterangan:

$tf$  = frekuensi *term* ( $t$ ) pada *document* ( $d$ )  
 $\max(tf)$  = jumlah munculnya *term* terbanyak pada dokumen tersebut

Berikut adalah rumus untuk menghitung *Inverse Document Frequency* (IDF).

$$IDF_m = \log\left(\frac{D}{df_m}\right) \tag{2}$$

Keterangan:

$D$  = jumlah dokumen  
 $df_m$  = jumlah dokumen yang mengandung *term* ( $t_m$ )  
 IDF = *inverse document frequency* ( $\log_2\left(\frac{d}{df}\right)$ )

Berikut adalah rumus untuk menghitung *Term Weighting* TF-IDF yang merupakan gabungan dari rumus TF dan IDF.

$$w_{(i,n)} = tf_{(i,n)} \times IDF_{(i,n)} \tag{3}$$

Keterangan:

$i$  = dokumen ke- $i$   
 $n$  = kata ke- $n$  dari kata kunci  
 $w$  = bobot dari dokumen ke- $i$  terhadap kata- $n$

## 2.3. Hierarchical Clustering

Pengelompokan dengan metode hirarki atau dapat disebut sebagai pengelompokan bersarang (*nested*) merupakan teknik pengelompokan yang berdasarkan tingkatan tertentu yang membentuk sebuah konstruksi hirarki seperti struktur pohon. Proses pengelompokan dari metode ini dilakukan secara bertahap dan hasil dari pengelompokan ini akan disajikan dalam bentuk dendrogram. Metode *Hierarchical Clustering* dapat berupa *agglomerative* dan *divisive* yang bergantung pada dekomposisi hirarki tersebut terbentuk secara *bottom-up* (penggabungan) atau *top-down* (pemisahan).

Metode *agglomerative hierarchical clustering* adalah suatu metode yang menggunakan strategi *bottom-up* (penggabungan) dimana biasanya akan dimulai dari setiap objek membentuk cluster sendiri secara iteratif hingga cluster akan menjadi lebih besar dan setiap objek masuk dalam *cluster*.

Dalam merancang suatu sistem rekomendasi, diperlukan penerapan metode *clustering*. Metode *clustering* diperlukan untuk memudahkan sistem untuk melakukan proses pemisahan dokumen ke dalam *cluster* berdasarkan masing-masing kriteria yang sama, sehingga sistem dapat langsung memanggil/menampilkan rekomendasi film yang serupa dengan kata kunci.

Berikut adalah beberapa metode pada *agglomerative hierarchical clustering*.

### a. Single Linkage (Jarak Terdekat)

*Single linkage* adalah perhitungan jarak yang berfokus pada jarak terkecil. Jika terdapat dua objek yang terpisah dan memiliki jarak yang pendek, maka dua objek tersebut akan digabungkan menjadi satu *cluster*.

$$d_{uv} = \min\{d_{uv}\}, d_{uv} \in D \tag{4}$$

Keterangan:

$d$  = kriteria  
 $d_{uv}$  = jarak data U dan V dari masing-masing *cluster*  
 $\epsilon$  = eliminasi

### b. Average Linkage (Jarak Rata-Rata)

*Average linkage* adalah perhitungan jarak yang berfokus pada perhitungan rata-rata jarak dari seluruh objek dalam suatu *cluster* dengan jarak dari seluruh objek pada *cluster* lain.

$$d_{uv} = \frac{1}{u_n \times v_n} \sum d_{uv}, d_{uv} \in D \tag{5}$$

Keterangan:

- d = kriteria
- $d_{uv}$  = jarak data U dan V dari masing-masing cluster
- $\epsilon$  = eliminasi
- $u_n, v_n$  = jumlah data yang berada dalam masing-masing cluster

c. *Complete Linkage* (Jarak Terjauh)

*Complete Linkage* adalah perhitungan jarak yang berfokus pada jarak terjauh. Jika terdapat dua objek yang terpisah dan memiliki jarak yang jauh, maka dua objek tersebut akan digabungkan menjadi satu cluster. [4]

$$d_{uv} = \max\{d_{uv}\}, d_{uv} \in D \tag{6}$$

Keterangan:

- d = kriteria
- $d_{uv}$  = jarak data U dan V dari masing-masing cluster
- $\epsilon$  = eliminasi

2.4. *Manhattan Distance*

Perhitungan matriks jarak pada metode AHC dapat menggunakan rumus *Manhattan Distance* (menjumlahkan hasil selisih dari jarak xi dan yi) ataupun *Euclidean Distance* (perhitungan jarak dua titik dengan konsep Pythagoras). Rumus dari Manhattan Distance adalah sebagai berikut. [5]

$$d = \sum_{i=1}^n |x_i - y_i| \tag{7}$$

Keterangan:

- d = kriteria
- $x_i, y_i$  = Data 1 dan Data 2

2.5. *Silhouette Coefficient*

*Silhouette Coefficient/Score* adalah suatu metode perhitungan evaluasi yang biasanya digunakan untuk menghitung keakuratan dan baiknya suatu cluster. Untuk memperoleh nilai *silhouette*, ada beberapa tahap perhitungan yang perlu dilakukan, perhitungan tersebut adalah sebagai berikut.

1. Menghitung jarak rata-rata data ke-i pada seluruh data di cluster yang sama. Misal diasumsikan data ke-i berada pada cluster A, maka rumus dari a(i) dapat ditulis seperti pada persamaan (8).

$$a(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i, j) \tag{8}$$

Keterangan:

- a(i) = rata-rata jarak objek ke-i dengan objek pada cluster lainnya
- A = banyaknya data pada cluster A
- d(i, j) = jarak antara objek data ke-i dengan j
- j = objek data lain dalam satu cluster A

2. Menghitung jarak rata-rata objek ke-i dengan seluruh objek yang ada di cluster berbeda, kemudian ambil nilai terkecilnya dengan menggunakan rumus pada persamaan nomor (9).

$$d(i, C) = \frac{1}{|C|} \sum_{j \in C} d(i, j) \tag{9}$$

$$b(i) = \min_{C \neq A} d(i, C) \tag{10}$$

Keterangan:

- d(i, C) = jarak rata-rata objek ke-i dengan objek pada cluster lainnya
- C = banyaknya data pada cluster C

3. Setelah menghitung d(i, C), maka ambil nilai jarak yang paling kecil untuk nilai pada b(i).

4. Setelah memperoleh hasil dari nilai a(i) dan b(i), maka hitung nilai *silhouette coefficient* seperti pada persamaan (11).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i) - b(i)\}} \tag{11}$$

Nilai dari *Silhouette Coefficient* berkisar dari -1 hingga 1. Skor 1 menunjukkan bahwa data ke-i tergolong baik. Skor 0 menunjukkan bahwa data ke-i tergolong berada di tengah antara dua cluster. Skor -1 menunjukkan bahwa data ke-i tergolong lemah. [6]

Tabel 1 Struktur Golongan Nilai pada *Silhouette Coefficient*

<i>Silhouette Coefficient</i>	Struktur Golongan
$0.7 < \textit{Silhouette Coefficient} \leq 1$	Kuat
$0.5 < \textit{Silhouette Coefficient} \leq 0.7$	Sedang
$0.25 < \textit{Silhouette Coefficient} \leq 0.5$	Lemah
$\textit{Silhouette Coefficient} \leq 0.25$	Tidak Terstruktur

### 3. Hasil Pengujian

Pengujian dari *clustering* ini adalah menggunakan metode *silhouette coefficient* yang nantinya akan menampilkan hasil penggunaan banyaknya *n\_cluster* dan metode *linkage* terbaik. Jumlah data yang digunakan untuk melakukan pengujian adalah 2468 data. Hasil pengujian tersebut dapat dilihat pada Tabel 2.

Tabel 2 Hasil Pengujian

Linkage	N_Cluster	Silhouette Coefficient
Single	20	0.35243574880973255
	50	0.2323934421848187
	100	0.2127179833399376
	150	0.18871838077108724
Average	20	0.4985112710632082
	50	0.47529099273241865
	100	0.4628487159003454
	150	0.4457014076383949
Complete	20	0.5025559374455285
	50	0.4816997939769783
	100	0.4615103842353544
	150	0.44705346688593844

Hasil pengujian dari Tabel 2, dapat dilihat bahwa semakin banyak *cluster* yang terbentuk maka nilai *silhouette coefficient* nya semakin kecil. Hal ini dapat terjadi karena data yang diproses memiliki struktur yang kompleks atau hanya sedikit memiliki kemiripan antar data lain sehingga tidak dapat masuk ke sejumlah kecil. Pada kasus seperti ini, lebih tepat untuk menggunakan jumlah *cluster* yang lebih kecil. Dapat diketahui juga bahwa *silhouette coefficient* hanya suatu metrik untuk mengevaluasi *cluster* dan memungkinkan bahwa evaluasi menggunakan metode ini tidak selalu menjadi nilai evaluasi yang tepat untuk semua jenis kumpulan data.

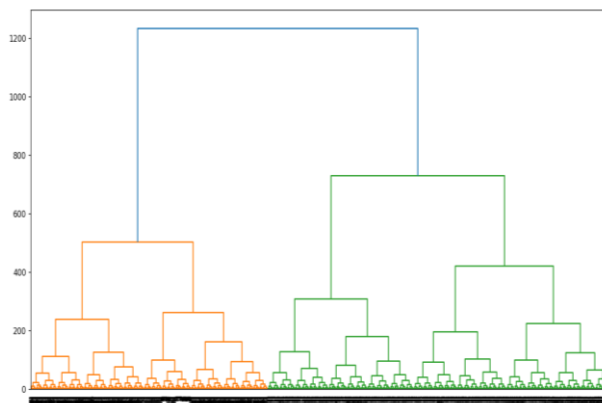
Untuk hasil dari pencarian rekomendasi Film dengan menggunakan kata kunci “Komedi”, dapat dilihat hasilnya pada Gambar 2.

Judul	Deskripsi	Genre	Keccokan
Sebastian Maniscalco: What's Wrong w	Dengan kritiknya yang menggelitik tentang	Komedi Stand-Up, Komedi	Cocok
Sebastian Maniscalco: Why Would You	Di Teater Beacon yang legendaris, stand-up	Komedi Stand-Up, Komedi	Cocok
Shawn Mendes: In Wonder	Dalam dokumenter jujur yang direkam sela	Musik, Konser, Film Dokum	Tidak Cocok
Simon Amstell: Set Free	Simon Amstell, komika yang jujur dan intro	Inggis, Komedi Stand-Up,	Cocok
Stand Out: An LGBTQ+ Celebration	Para legenda komedi LGBTQ+ dan bintang-	Komedi Stand-Up, Komedi	Cocok
Stand Up and Away! with Brian Regan	Dalam serial komedi hibrid ini, Brian Regan	Komedi Stand-Up, Acara T	Cocok
Still LAUGH-IN: The Stars Celebrate	Perayaan acara komedi terobosan "Laugh-	Variety TV, Film AS, Komedi	Cocok
Stutz	Dalam obrolan jujur dengan aktor Jonah Hill	Film Dokumenter, Gaya Hi	Tidak Cocok
That's My Time with David Letterman	David Letterman mengundang beberapa bir	Komedi Stand-Up, Talk Sho	Cocok
The Comedy Lineup	Sekumpulan komedian yang sedang naik da	Komedi Stand-Up, Acara T	Cocok
The Dirt	Biopik blak-blakan adaptasi buku laris Möt	Musik, Film Berdasarkan K	Tidak Cocok
The Hall: Honoring George Carlin, Robin	Komedi spesial: penghormatan Jon Stewart	Komedi Stand-Up, Komedi	Cocok
The Netflix Afterparty: The Best Shows	David Spade, Fortune Feimster, dan London	Variety TV, Film AS, Komedi	Cocok
The Pursuit of Happyness	Setelah diusir, Chris dan putranya menghad	Drama Isu Sosial, Film Berd	Tidak Cocok
The Standups	Para komika baru beraksi di panggung LA se	Komedi Stand-Up, Acara T	Cocok
Tiffany Haddish Presents: They Ready	Dipilih oleh Tiffany Haddish, enam komedia	Komedi Stand-Up, Acara T	Cocok
Tiffany Haddish: Black Mitzvah	Di ulang tahunnya yang ke-40, Tiffany Haddi	Komedi Stand-Up, Komedi	Cocok
Tiffany Haddish: She Ready! From the H	Dalam debut komedinya ini, Tiffany Haddish	Komedi Stand-Up, Komedi	Cocok
Tom Papa: You're Doing Great!	Tom Papa membahas masalah citra tubuh,	Komedi Stand-Up, Komedi	Cocok
Trese After Dark	Para bintang dan kreator berkumpul memb	Memberi Inspirasi	Tidak Cocok
Vir Das: For India	Dari Weda, Vasco da Gama, hingga berbagi	Komedi Stand-Up, Komedi	Cocok
Vir Das: Losing It	Dunia ini penuh dengan masalah, tetapi Vir	Komedi Stand-Up, Komedi	Cocok
Wanda Sykes: Not Normal	Wanda Sykes membahas politik, realita, TV,	Komedi Stand-Up, Komedi	Cocok

Gambar 2. Hasil Clustering dari Kata Kunci “Komedi”

Metode yang diterapkan adalah metode *complete linkage*. Untuk percobaan ini, banyaknya *cluster* yang digunakan adalah 150 agar hasil rekomendasinya tidak terlalu banyak. Hasil data tersebut merupakan data yang satu *cluster* dengan kata kunci yang dimasukkan, sehingga data pada *cluster* tersebut akan ditampilkan sebagai rekomendasinya.

Banyaknya data yang diperoleh dalam *cluster* tersebut adalah sebanyak 104 data. Dari data tersebut terdapat 104 data yang diantaranya diperoleh 95 data yang cocok dengan kata kunci dan 9 data yang tidak cocok dengan kata kunci. Untuk gambar visualisasi *clustering* dapat dilihat pada Gambar 3.



Gambar 3. Visualisasi Dendrogram

### 4. Kesimpulan

Dari hasil pengujian untuk memperoleh rekomendasi film, dapat disimpulkan bahwa:

1. Penerapan metode *Agglomerative Hierarchical Clustering* berhasil dilakukan dengan jumlah *cluster* terbaik adalah 20 dimana nilai rata-rata dari

*Silhouette Coefficient* nya adalah 0.5025559374455285.

2. *Complete Linkage* merupakan metode terbaik untuk melakukan pengelompokan data Netflix ini.
3. Dari data yang diproses, semakin banyak *cluster* yang terbentuk maka nilai *silhouette coefficient* nya semakin kecil. Ini memungkinkan terjadi dikarenakan data yang diproses memiliki sedikit kemiripan datanya dengan data lain.
4. Pengujian sistem rekomendasi dengan menerapkan metode *Agglomerative Hierarchical Clustering* pada website telah berjalan dengan baik.

## REFERENSI

- [1] Cindy Winata; Viny Christanti Mawardi; dan Janson Hendryli, *Movilizer Application with Genre and Rating Classification Using NW-KNN Method*, 3<sup>rd</sup> Asia Pacific Information Technology Conference, h. 33-38, Januari 2021.
- [2] Devin Abipraya; Viny Christanti Mawardi; dan Novario Jaya Perdana, *Implementasi Opinion Mining Untuk Provider Internet Menggunakan Metode Naïve Bayes*, *Jurnal Ilmu Komputer dan Sistem Informasi*, Vol.9, No.2, 2021.
- [3] Online Edition 2009 Cambridge, *Scoring, Term Weighting, and the vector Space Model*, doi:10.1017/CBO9780511809071.007, ISBN 978-0-511-80907-1, h. 128.
- [4] Jiawei Han; Micheline Kamber; dan Jian Pei, *Data Mining Concepts and Technique*, Waltham: Morgan Kaufmann Publications, h. 457-459, 2012.
- [5] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square", *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, Vol. 4, Nomor 1, h. 20-24, Januari 2019.
- [6] Edy Susanto; Viny Christanti Mawardi; dan Manatap Dolok Lauro, *Aplikasi Clustering Berita dengan Metode K-Means dan Peringkat Berita dengan Metode Maximum Marginal Relevance*, *Jurnal Ilmu Komputer dan Sistem Informasi*, Vol.9, No.1, h.62-68, 2021.

**Vanesa Nellie**, seorang mahasiswa pada program studi Fakultas Teknologi Informasi di Universitas Tarumanagara.

**Viny Christanti Mawardi**, memperoleh gelar S.Kom dari Universitas Tarumanagara tahun 2004 dan M.Kom dari Universitas Indonesia tahun 2008. Saat ini aktif sebagai dosen tetap Program Studi Teknik Informatika, Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta.

**Novario Jaya Perdana**, memperoleh gelar S.Kom dari ITS tahun 2011 dan M.T. dari Universitas Indonesia tahun 2016. Saat ini aktif sebagai dosen tetap Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta.