

APLIKASI CLUSTERING FILM BERDASARKAN SINOPSISNYA



LAPORAN

**Disusun untuk Memenuhi Tugas Kelompok
pada Mata Kuliah Pengolahan Bahasa Alami Semester VIII
yang diampu oleh Ibu Sukmawati Nur Endah, S.Si., M.Kom.**

Disusun oleh:

Linggar Maretva Cendani (24060117120031)

Fetty Krisnaeni (24060117120032)

**DEPARTEMEN ILMU KOMPUTER/INFORMATIKA
FAKULTAS SAINS DAN MATEMATIKA
UNIVERSITAS DIPONEGORO
SEMARANG**

2021

ABSTRAK

Dunia perfilman memiliki kisah perjalanan yang cukup panjang, mulai dari film bisu yang tidak berwarna hingga saat ini telah menjadi film yang kaya akan efek dan dapat dengan mudah ditemukan di dunia hiburan. Hingga saat ini sudah banyak film-film berkualitas dengan berbagai macam genre yang dapat disaksikan oleh masyarakat. Genre merupakan salah satu preferensi penonton terhadap film berdasarkan atribut filmnya. Sebagian besar orang akan menonton film apabila mereka suka dengan genre yang diminati oleh mereka. Oleh karena itu pengelompokkan film sangat diperlukan bagi orang-orang yang ingin mendapatkan informasi mengenai film-film yang berada dalam lingkup genre yang sama. Pada penelitian ini akan dibangun Aplikasi Clustering Film Berdasarkan Sinopsisnya yang merupakan aplikasi text clustering untuk mengelompokkan film berdasarkan sinopsisnya. Data judul film dan sinopsis yang akan digunakan pada penelitian ini diambil dari website <https://www.themoviedb.org/> yaitu berupa judul film dan sinopsisnya. Aplikasi ini dibuat dalam bentuk web dengan menggunakan bahasa pemrograman Python dan HTML.

Kata kunci: Text Clustering, Clustering Film, Python, KMeans.

KATA PENGANTAR

Dengan menyebut nama Allah SWT yang Maha Pengasih lagi Maha Penyayang, atas karunia dan rahmat-Nya penulis dapat menyelesaikan pembuatan Aplikasi Teks Clustering dan menyusun laporan yang berjudul “Aplikasi Clustering Film Berdasarkan Sinopsisnya” dengan baik. Laporan ini disusun untuk memenuhi tugas kelompok pada mata kuliah Pengolahan Bahasa Alami di Departemen Ilmu Komputer, Universitas Diponegoro.

Ucapan terima kasih penulis sampaikan kepada Ibu Sukmawati Nur Endah, S.Si., M.Kom. selaku dosen pengampu mata kuliah Pengolahan Bahasa Alami yang senantiasa membimbing penulis. Terima kasih juga penulis sampaikan kepada segenap pihak yang telah memberikan dukungan dan bantuan sehingga penulis dapat menyelesaikan laporan ini tepat pada waktunya.

Penulis menyadari bahwa dalam laporan ini masih banyak kekurangan baik dari segi materi ataupun dalam penyajiannya. Kritik dan saran sangat penulis harapkan untuk perbaikan pada penulisan ilmiah yang akan datang. Penulis berharap laporan ini dapat bermanfaat bagi pembaca pada umumnya dan penulis sendiri pada khususnya.

Semarang, 10 Juni 2021

Penulis

DAFTAR ISI

ABSTRAK	ii
KATA PENGANTAR	iii
DAFTAR ISI	iv
DAFTAR GAMBAR	vi
BAB I PENDAHULUAN	1
1.1. Latar Belakang.....	1
1.2. Rumusan Masalah.....	2
1.3. Sistematika Penulisan	2
BAB II DASAR TEORI.....	3
2.1. Clustering	3
2.2. K-Means Clustering	4
BAB III METODOLOGI	6
3.1. Pengumpulan Data.....	6
3.2. Preprocessing Data	6
3.3. K-Means Model Training.....	7
3.4. Data Labelling	7
3.5. Data Preview for each Cluster	8
3.6. Save All Data from each Cluster	8
3.7. Feature Names of each Cluster.....	8
3.8. Predict Sentences	8
3.9. Evaluation.....	8
3.10. Data Visualization	9
BAB IV HASIL DAN PEMBAHASAN	10
4.1. Tampilan Antar Muka.....	10

4.1.1.	Tampilan Antar Muka Halaman K-Means Model Training	10
4.1.2.	Tampilan Antar Muka Halaman Clustering Result	10
4.1.3.	Tampilan Antar Muka Halaman Feature Names	13
4.1.4.	Tampilan Antar Muka Halaman Data Per Cluster.....	15
4.1.5.	Tampilan Antar Muka Halaman Cluster Prediction	20
4.1.6.	Tampilan Antar Muka Halaman Cluster Prediction (By Title)	22
4.2.	Hasil Penelitian.....	25
4.3.	Analisa Hasil	25
BAB V PENUTUP		27
5.1.	Kesimpulan.....	27
DAFTAR PUSTAKA		vii
LAMPIRAN		viii

DAFTAR GAMBAR

Gambar 2. 1 Visualisasi Algoritma Dasar K-Means Clustering	4
Gambar 3. 1 Data Judul Film dan Sinopsisnya dari website https://www.themoviedb.org/	6
Gambar 3. 2 Data Judul Film dan Sinopsis Setelah dilakukan Case Folding	7
Gambar 3. 3 Data Judul Film dan Sinopsis Setelah dilakukan Proses Data Labelling	8
Gambar 4. 1 Antar Muka Halaman K-Means Model Training	10
Gambar 4. 2 Antar Muka Halaman Clustering Result: Movie Synopsis Labelled	11
Gambar 4. 3 Antar Muka Halaman Clustering Result: Evaluasi Silhouette	12
Gambar 4. 4 Antar Muka Halaman Clustering Result: Evaluasi Elbow Method	12
Gambar 4. 5 Antar Muka Halaman Clustering Result: Visualisasi 2 Dimensi	12
Gambar 4. 6 Antar Muka Halaman Clustering Result: Visualisasi 3 Dimensi	13
Gambar 4. 7 Antar Muka Halaman Feature Names Tiap Cluster	13
Gambar 4. 8 Antar Muka Halaman Feature Names: WordCloud Feature Names	14
Gambar 4. 9 Antar Muka Halaman Data Cluster: 0 dan 1	15
Gambar 4. 10 Antar Muka Halaman Data Cluster: 2 dan 3	16
Gambar 4. 11 Antar Muka Halaman Data Cluster: 4 dan 5	17
Gambar 4. 12 Antar Muka Halaman Data Cluster: 6 dan 7	18
Gambar 4. 13 Antar Muka Halaman Data Cluster: 8 dan 9	19
Gambar 4. 14 Antar Muka Halaman Data Cluster: 10 dan 11	20
Gambar 4. 15 Antar Muka Halaman Cluster Prediction	21
Gambar 4. 16 Antar Muka Halaman Cluster Prediction: Related Movies	22
Gambar 4. 17 Antar Muka Halaman Cluster Prediction (By Title)	23
Gambar 4. 18 Antar Muka Halaman Cluster Prediction (By Title): Related Movies	24

BAB I

PENDAHULUAN

Pada bab ini akan membahas mengenai latar belakang, rumusan masalah, dan sistematika penulisan yang dibuat dalam proses pembuatan Aplikasi Clustering Film Berdasarkan Sinopsisnya.

1.1. Latar Belakang

Dunia perfilman memiliki kisah perjalanan yang cukup panjang, mulai dari film bisu yang tidak berwarna hingga saat ini telah menjadi film yang kaya akan efek dan dapat dengan mudah ditemukan di dunia hiburan. Seiring berjalannya waktu hingga sekitar tahun 1980/1990-an film Indonesia semakin meningkat. Namun peningkatan tersebut juga diikuti dengan masuknya film luar baik Hollywood ataupun Bollywood yang akhirnya mendominasi perfilman negeri. Masyarakat Indonesia cenderung menyukai film luar karena memiliki karakteristik tersendiri, alur cerita yang menarik, dan mindset bahwa menonton film luar lebih membuat mereka keren. Selain itu, kreatifitas dalam pembuatan film luar jauh lebih bagus, berbeda dengan film Indonesia yang bermain aman dengan hanya memproduksi film dengan genre tertentu yang akan banyak ditonton masyarakat.

Hingga saat ini sudah banyak film-film berkualitas dengan berbagai macam genre yang dapat disaksikan oleh masyarakat. Genre merupakan salah satu preferensi penonton terhadap film berdasarkan atribut filmnya. Sebagian besar orang akan menonton film apabila mereka suka dengan genre yang diminati oleh mereka. Genre film dapat kita ketahui dari judul maupun alur cerita yang ada pada film. Kita juga dapat mengetahui genre suatu film berdasarkan sinopsis film yang kita baca.

Pengelompokkan film sangat diperlukan bagi orang-orang yang ingin mendapatkan informasi mengenai film-film yang berada pada satu jenis genre. Selain itu pengelompokkan film akan bermanfaat bagi orang yang hanya menyukai suatu genre film tertentu sehingga mereka dapat menonton film dengan konteks cerita yang disuakinya. Contohnya seperti seseorang yang hanya menyukai film dengan genre horror akan terbantu, karena mereka akan mendapatkan informasi kumpulan film bergenre horror.

Dalam hal ini penulis melakukan penelitian mengenai text clustering yang akan mengelompokkan film berdasarkan sinopsisnya. Data yang akan digunakan pada penelitian ini diambil dari website <https://www.themoviedb.org/> yaitu berupa judul film dan

sinopsisnya. Hasil penelitian berupa aplikasi text clustering untuk menentukan genre film berdasarkan sinopsisnya yang dibuat dalam bentuk web.

1.2. Rumusan Masalah

Berdasarkan uraian permasalahan pada latar belakang di atas, maka penulis merumuskan masalah sebagai berikut:

1. Merancang aplikasi text clustering untuk menentukan genre film berdasarkan sinopsisnya.
2. Merancang desain antar muka aplikasi text clustering untuk menentukan genre film berdasarkan sinopsisnya.

1.3. Sistematika Penulisan

Untuk memberikan gambaran yang urut dan jelas mengenai pembahasan penyusunan Aplikasi Clustering Film Berdasarkan Sinopsisnya maka dibuatlah sistematika penulisan sebagai berikut:

BAB I PENDAHULUAN

Bab ini membahas mengenai latar belakang, rumusan masalah, dan sistematika penulisan yang dibuat dalam proses pembuatan Aplikasi Clustering Berdasarkan Sinopsisnya.

BAB II DASAR TEORI

Bab ini menjelaskan dasar-dasar teori yang mendukung dalam pembuatan Aplikasi Clustering Film Berdasarkan Sinopsisnya.

BAB III METODOLOGI

Bab ini akan membahas metodologi penelitian yang digunakan dalam pembuatan Aplikasi Clustering Film Berdasarkan Sinopsisnya.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjabarkan tampilan antar muka, hasil penelitian, dan analisa hasil dalam pembuatan Aplikasi Clustering Film Berdasarkan Sinopsisnya sesuai dengan metodologi yang digunakan.

BAB V PENUTUP

Bab ini merupakan penutup yang berisi kesimpulan dari bab-bab yang dibahas sebelumnya.

BAB II

DASAR TEORI

Bab ini menjelaskan dasar-dasar teori yang mendukung dalam pembuatan Aplikasi Clustering Film Berdasarkan Sinopsisnya. Dasar teori yang digunakan dalam penelitian ini adalah sebagai berikut:

2.1. Clustering

Clustering merupakan suatu proses mengelompokkan suatu objek ke dalam kelompok-kelompok objek yang sejenis. Bentuk data yang paling umum digunakan dalam clustering yaitu data unsupervised learning. Dalam proses pengelompokannya algoritma clustering akan membagi suatu objek menjadi subset objek, di mana setiap subset berisi objek-objek yang dianggap sejenis.

Clustering dapat digunakan untuk mengorganisasikan dokumen yang diperoleh. Beberapa alasan perlu adanya pengelompokan dokumen adalah sebagai berikut:

- a. Untuk analisa keseluruhan korpus.
- b. Untuk visualisasi koleksi dokumen dan topiknya.
- c. Untuk memperbaiki recall pada hasil pelacakan.
- d. Untuk navigasi yang lebih baik dari hasil pelacakan.

Dalam proses clustering terdapat beberapa istilah yang perlu dipahami, yaitu cluster, objek, dan properti. Cluster merupakan kelompok yang dihasilkan dari proses clustering. Objek merupakan sesuatu yang ditempatkan pada kelompok/cluster. Sedangkan properti merupakan cara yang digunakan untuk merepresentasikan hasil cluster tersebut.

Menurut Nadia Nejdah, et al (2009), pada dasarnya dalam setiap algoritma Feature Selection pasti memiliki tiga produk atau hasil akhir yang menyangkut tentang text clustering, yaitu:

1. Kemampuan memitigasi permasalahan yang disebabkan oleh tingginya dimensionalitas dan sparsitas terkait penilaian kesamaan dokumen.
2. Kemampuan mengurangi biaya komputasi dari algoritma pengelompokan mengenai waktu pemrosesan dan ruang memori.
3. Kemampuan menyediakan seperangkat istilah yang ringkas yang dapat mencirikan dan membedakan klaster yang terbentuk.

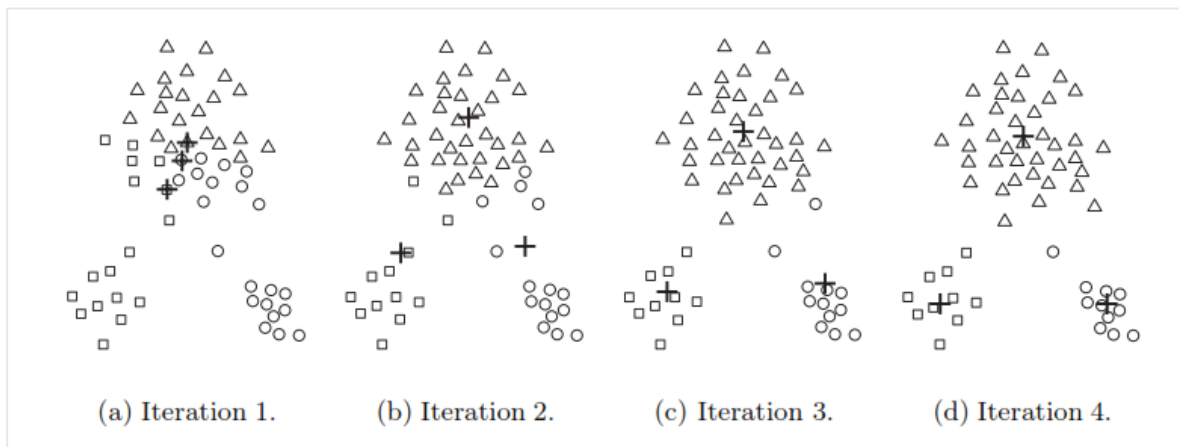
2.2. K-Means Clustering

Pada konsepnya proses K-Means Clustering akan mengelompokkan n objek ke dalam K cluster berdasarkan nilai atribut yang dimiliki oleh objek tersebut. K merupakan jumlah cluster yang berupa bilangan integer positif. K-Means Clustering termasuk jenis hard clustering, di mana satu objek hanya dapat menjadi anggota dari satu cluster secara eksklusif.

Berikut merupakan algoritma dasar K-Means Clustering:

- 1: Pilih suatu set objek sebagai inisial centroid dari cluster.
- 2: **Repeat**
- 3: Tempatkan setiap objek pada cluster terdekat.
- 4: Hitung kembali centroid sebagai center of mass (average) dari anggota-anggotanya.
- 5: **Until** Titik centroid tidak berubah lagi.

Algoritma dasar K-Means Clustering dapat lebih jelas dan mudah dipahami seperti yang disajikan pada gambar berikut:



Gambar 2. 1 Visualisasi Algoritma Dasar K-Means Clustering

Untuk menetapkan titik ke centroid terdekat, kita memerlukan ukuran kedekatan yang mengkuantifikasi gagasan "terdekat" untuk data spesifik yang sedang dipertimbangkan. Beberapa cara yang digunakan untuk menetapkan titik ke centroid terdekat yaitu dapat menggunakan Euclidean Distance, Minkowski Distance, dan Manhattan Distance. Penjelasan sebagai berikut:

1. Euclidean Distance

Euclidean distance merupakan ukuran jarak yang paling sering digunakan untuk data numerik. Jarak Euclidean antara 2 titik atau tuple, $\mathbf{x} = (x_1, x_2, \dots, x_d)$ dan $\mathbf{y} = (y_1, y_2, \dots, y_n)$ pada ruang dimensi d adalah

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

2. Manhattan Distance

Manhattan distance dihitung berdasarkan jumlah jarak dari semua atribut. Jarak Manhattan antara 2 titik atau tuple, $\mathbf{x} = (x_1, x_2, \dots, x_d)$ dan $\mathbf{y} = (y_1, y_2, \dots, y_n)$ pada ruang dimensi d adalah

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

3. Minkowski Distance

Minkowski distance merupakan generalisasi dari jarak Euclidean dan Manhattan. Jarak Manhattan antara 2 titik atau tuple, $\mathbf{x} = (x_1, x_2, \dots, x_d)$ dan $\mathbf{y} = (y_1, y_2, \dots, y_n)$ pada ruang dimensi d adalah

$$d_{min}(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^d |x_i - y_i|^p \right)^{\frac{1}{p}}, \quad p \geq 1$$

Jika $p=1$ maka Manhattan distance

Jika $p=2$ maka Euclidean distance

BAB III

METODOLOGI

Bab ini akan membahas metodologi penelitian yang digunakan dalam pembuatan Aplikasi Clustering Film Berdasarkan Sinopsisnya. Penjelasan mengenai metodologi penelitian yang digunakan pada penelitian ini adalah sebagai berikut:

3.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sinopsis film yang didapatkan dari website <https://www.themoviedb.org/>. Pada website tersebut terdapat film-film dengan berbagai macam genre yang disajikan bersama dengan sinopsis masing-masing film tersebut. Sinopsis film pada website tersebut disajikan dalam Bahasa Inggris.

Pada bagian pengumpulan data ini dilakukan proses pengambilan data genre film serta melakukan pengambilan data judul dan sinopsis film pada website <https://www.themoviedb.org/>. Genre film yang ada pada website tersebut berjumlah sebanyak 19 genre, diantaranya adalah Drama, Crime, Comedy, Action, Thriller, Documentary, Adventure, Science Fiction, Animation, Family, Romance, Mystery, Horror, Fantasy, War, Music, History, Western, dan TV Movie. Sedangkan data judul film dan sinopsis yang diambil dari website <https://www.themoviedb.org/> berjumlah 8.457 film. Berikut merupakan tampilan 5 data teratas judul film dan sinopsis yang didapatkan dari website <https://www.themoviedb.org/>:

	title	synopsis
0	Four Rooms	It's Ted the Bellhop's first night on the job....
1	Judgment Night	While racing to a boxing match, Frank, Mike, J...
2	Life in Loops (A Megacities RMX)	Timo Novotny labels his new project an experim...
3	Star Wars	Princess Leia is captured and held hostage by ...
4	Finding Nemo	Nemo, an adventurous young clownfish, is unexp...

Gambar 3. 1 Data Judul Film dan Sinopsisnya dari website <https://www.themoviedb.org/>

3.2. Preprocessing Data

Preprocessing data merupakan proses yang dilakukan untuk mendapatkan data yang bersih agar dapat digunakan secara maksimal. Pada proses preprocessing data ini dilakukan

dengan menghapus semua missing value yang ada pada data judul film dan sinopsis, kemudian dilakukan case folding, dan dilanjutkan dengan melakukan TF-IDF training.

Menghapus missing value pada data film dan sinopsis dilakukan karena data tersebut tidak memenuhi kualifikasi atau mungkin datanya null sehingga hasilnya tidak valid dan pastinya tidak akan digunakan dalam penelitian. Seperti pada penelitian ini data awal yang didapatkan dari website <https://www.themoviedb.org/> berjumlah 8.457 film, namun karena terdapat beberapa data yang perlu dihapus maka dilakukan proses delete missing value sehingga data yang digunakan untuk melakukan penelitian ini sebanyak 8.214 film.

Case folding dilakukan untuk mengubah semua huruf dalam sinopsis film tersebut menjadi lowercase, kemudian dilanjutkan dengan melakukan TF-IDF training. Berikut merupakan tampilan data judul film dan sinopsis setelah dilakukan proses case folding:

	title	synopsis
0	Four Rooms	it's ted the bellhop's first night on the job....
1	Judgment Night	while racing to a boxing match, frank, mike, j...
2	Life in Loops (A Megacities RMX)	timo novotny labels his new project an experim...
3	Star Wars	princess leia is captured and held hostage by ...
4	Finding Nemo	nemo, an adventurous young clownfish, is unexp...

Gambar 3. 2 Data Judul Film dan Sinopsis Setelah dilakukan Case Folding

3.3. K-Means Model Training

Setelah melewati proses preprocessing data, kemudian data film dan sinopsis yang berjumlah 8.214 akan dilakukan training model K-Means. Proses training model K-Means ini menggunakan jumlah cluster sebanyak 19 cluster dan akan dilakukan iterasi sebanyak 500 kali iterasi. Pada proses ini data 8.214 judul film dan sinopsis akan dikelompokkan ke dalam 19 cluster berdasarkan genre yang sesuai.

3.4. Data Labelling

Proses data labelling dilakukan untuk memberikan label ke semua data film dan sinopsis yang berjumlah 8.214 data. Label yang dimaksud yaitu berupa angka 1 sampai 19 yang menunjukkan cluster ke-1 sampai cluster ke-19. Berikut merupakan tampilan data judul film dan sinopsis setelah dilakukan proses data labelling:

	title	synopsis	label
0	Four Rooms	it's ted the bellhop's first night on the job....	3
1	Judgment Night	while racing to a boxing match, frank, mike, j...	3
2	Life in Loops (A Megacities RMX)	timo novotny labels his new project an experim...	9
3	Star Wars	princess leia is captured and held hostage by ...	3
4	Finding Nemo	nemo, an adventurous young clownfish, is unexp...	13

Gambar 3. 3 Data Judul Film dan Sinopsis Setelah dilakukan Proses Data Labelling

3.5. Data Preview for each Cluster

Proses ini dilakukan untuk menampilkan seluruh data-data judul film beserta sinopsisnya sesuai dengan hasil pengelompokkan clusternya. Setiap cluster akan memberikan informasi film apa saja yang termasuk dalam cluster tersebut diikuti dengan sinopsis filmnya.

3.6. Save All Data from each Cluster

Proses ini dilakukan untuk menyimpan semua data dari masing-masing cluster. Data judul film dan sinopsisnya akan disimpan dalam format file csv.

3.7. Feature Names of each Cluster

Pada proses ini ditampilkan nama-nama fitur untuk setiap cluster. Fitur tersebut merupakan kata-kata yang paling relevan untuk setiap cluster yang menunjukkan karakteristik cluster. Kata-kata untuk nama pada tiap cluster tersebut berasal dari sinopsis film yang termasuk pada cluster tersebut.

3.8. Predict Sentences

Proses predict sentence ini dilakukan untuk memprediksi genre film berdasarkan kalimat atau kata-kata yang diinputkan oleh pengguna pada aplikasi. Kalimat dan kata-kata yang diinputkan dapat berupa sinopsis maupun judul film, kemudian sistem akan menampilkan termasuk ke dalam genre apakah film tersebut. Selain itu, aplikasi juga akan menampilkan judul film yang relevan dengan film tersebut berdasarkan genre yang sama.

3.9. Evaluation

Proses evaluasi cluster dilakukan dengan menggunakan Elbow Method (SSE) dan Silhouette Score. Elbow method pada dasarnya digunakan untuk menentukan berapa banyak jumlah cluster yang paling optimal untuk suatu data clustering. Hal ini ditentukan

berdasarkan perubahan nilai SSE yang paling signifikan (perubahan yang paling banyak) jika dilihat dari grafiknya.

Sedangkan evaluasi menggunakan Silhoutte Score dilakukan untuk mengetahui kualitas hasil clustering. Kualitas yang dimaksud yaitu apabila titik-titik di suatu cluster memiliki kemiripan yang tinggi dan jarak/perbedaan antar cluster semakin tinggi maka hasil clustering tersebut akan semakin bagus. Nilai hasil evaluasi silhoutte berada diantara -1 dan 1. Jika nilainya mendekati -1 maka hasil evaluasi silhoutte ini semakin jelek dan clustering yang dihasilkan juga semakin jelek. Tetapi sebaliknya apabila nilainya mendekati 1 maka hasil evaluasi silhoutte ini semakin bagus dan hasil akhir clustering yang dihasilkan akan semakin bagus juga.

3.10. Data Visualization

Data visualization untuk hasil clustering judul film dan sinopsisnya dilakukan dalam bentuk visualisasi 2 dimensi dan 3 dimensi. Masing-masing bentuk visualisasi akan dilakukan dengan PCA Dimensionality Reduction dan t-SNE Dimensionality Reduction, baik untuk visualisasi 2 dimensi maupun 3 dimensi.

BAB IV

HASIL DAN PEMBAHASAN

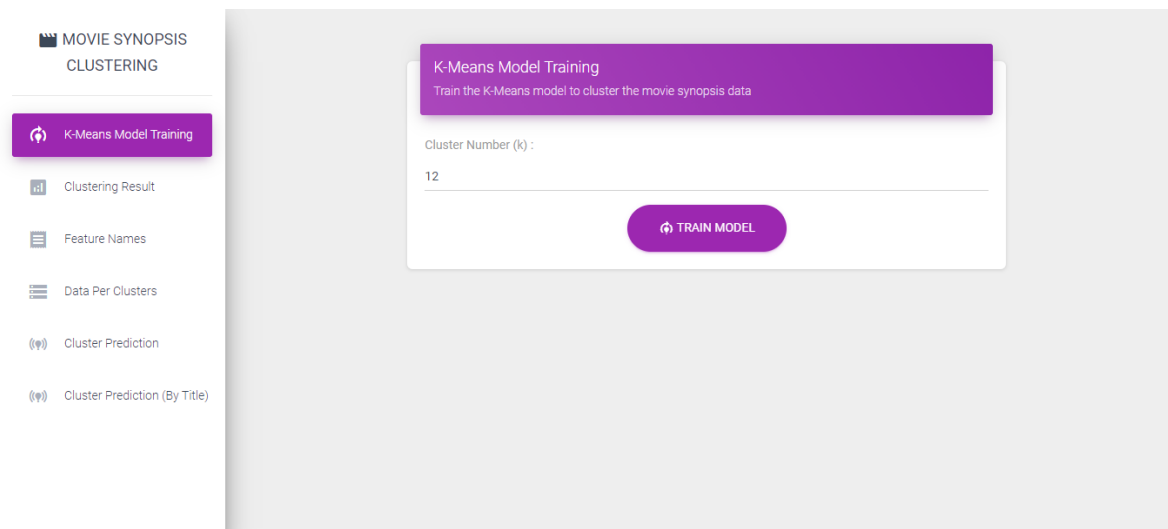
Bab ini menjabarkan tampilan antar muka, hasil penelitian, dan analisa hasil dalam pembuatan Aplikasi Clustering Film Berdasarkan Sinopsisnya sesuai dengan metodologi yang digunakan.

4.1. Tampilan Antar Muka

Berikut merupakan tampilan antar muka untuk setiap halaman pada Aplikasi Clustering Film Berdasarkan Sinopsisnya:

4.1.1. Tampilan Antar Muka Halaman K-Means Model Training

Berikut merupakan tampilan antar muka halaman awal pada Aplikasi Clustering Film Berdasarkan Sinopsisnya. Pada halaman ini pengguna dapat melakukan training model K-Means dengan memasukkan jumlah cluster yang diinginkan. Pada tampilan berikut sebagai contoh pengguna memasukkan jumlah cluster yaitu 12.



Ganbar 4. 1 Antar Muka Halaman K-Means Model Training

4.1.2. Tampilan Antar Muka Halaman Clustering Result

Pada halaman clustering result akan menampilkan lima data hasil clustering yang sudah dilabeli dengan nomor clusternya. Data hasil clustering yang sudah dilabeli dapat diunduh dengan format file .csv melalui tombol untuk mendownload yang sudah disediakan. Pada halaman ini juga disediakan tombol untuk mendownload model K-Means yang telah ditraining dengan format file .sav.

Halaman clustering result juga menampilkan evaluasi clustering. Evaluasi yang ditampilkan yaitu Silhouette Score dan Elbow Method. Kemudian visualisasi data yang digunakan akan ditampilkan menggunakan reduksi fitur dengan PCA dalam bentuk visualisasi 2 dimensi, dan 3 fitur untuk visualisasi 3 dimensi. Berikut merupakan tampilan antar muka halaman clustering result:

MOVIE SYNOPSIS CLUSTERING

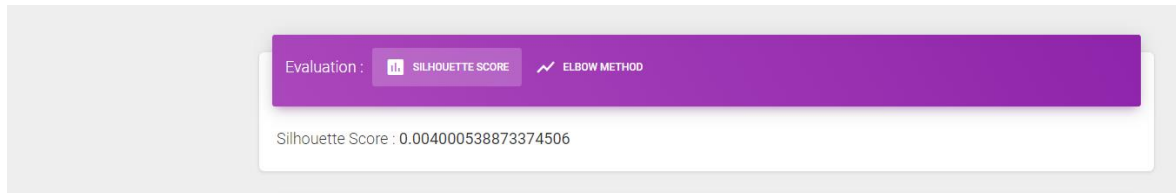
- K-Means Model Training
- Clustering Result**
- Feature Names
- Data Per Clusters
- Cluster Prediction
- Cluster Prediction (By Title)

Movie Synopsis Labeled
Preview of labeled data based on clustering result (5 Data)

ID	Title	Synopsis	Cluster
0	Four Rooms	It's Ted the Bellhop's first night on the job...and the hotel's very unusual guests are about to place him in some outrageous predicaments. It seems that this evening's room service is serving up one unbelievable happening after another.	3
1	Judgment Night	While racing to a boxing match, Frank, Mike, John and Rey get more than they bargained for. A wrong turn lands them directly in the path of Fallon, a vicious, wise-cracking drug lord. After accidentally witnessing Fallon murder a disloyal henchman, the four become his unwilling prey in a savage game of cat & mouse as they are mercilessly stalked through the urban jungle in this taut suspense drama	3
2	Life in Loops (A Megacities RMX)	Timo Novotny labels his new project an experimental music documentary film, in a remix of the celebrated film Megacities (1997), a visually refined essay on the hidden faces of several world "megacities" by leading Austrian documentarist Michael Glawogger. Novotny complements 30 % of material taken straight from the film (and re-edited) with 70 % as yet unseen footage in which he blends original shots unused by Glawogger with his own sequences (shot by Megacities cameraman Wolfgang Thaler) from Tokyo. Alongside the Japanese metropolis, Life in Loops takes us right into the atmosphere of Mexico City, New York, Moscow and Bombay. This electrifying combination of fascinating film images and an equally compelling soundtrack from Sofa Surfers sets us off on a stunning audiovisual adventure across the continents. The film also makes an original contribution to the discussion on new trends in documentary filmmaking. Written by KARLOVY VARY IFF 2006	5
3	Star Wars	Princess Leia is captured and held hostage by the evil Imperial forces in their effort to take over the galactic Empire. Venturesome Luke Skywalker and dashing captain Han Solo team together with the loveable robot duo R2-D2 and C-3PO to rescue the beautiful princess and restore peace and justice in the Empire.	3
4	Finding Nemo	Nemo, an adventurous young clownfish, is unexpectedly taken from his Great Barrier Reef home to a dentist's office aquarium. It's up to his worrisome father Marlin and a friendly but forgetful fish Dory to bring Nemo home -- meeting vegetarian sharks, surfer dude turtles, hypnotic jellyfish, hungry seagulls, and more along the way.	7

[DOWNLOAD K-MEANS MODEL](#) | [DOWNLOAD LABELED DATA](#)

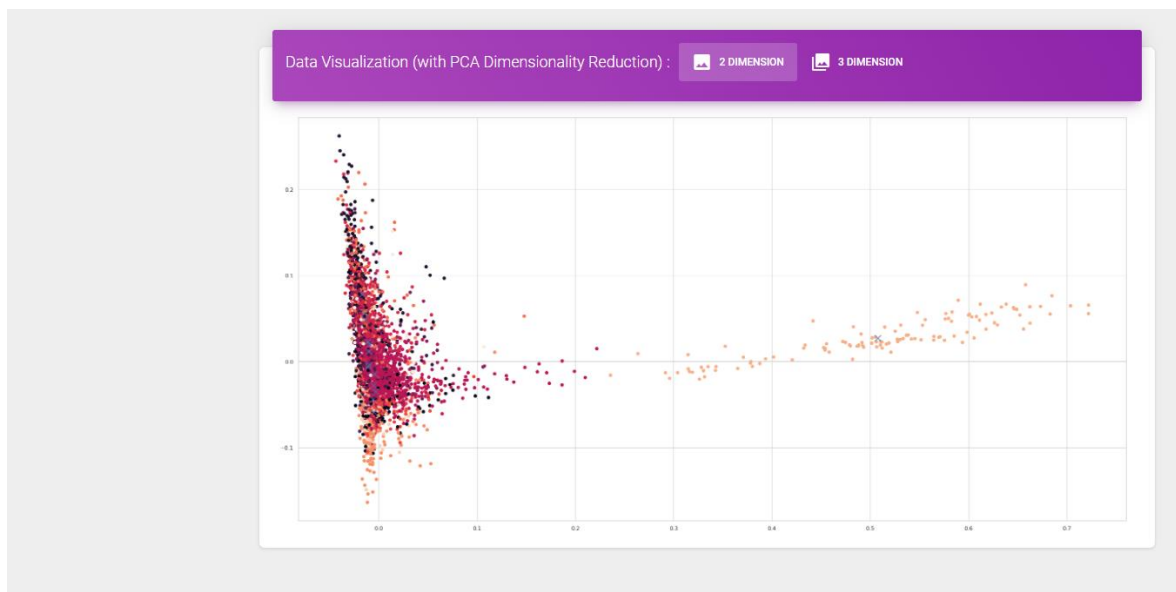
Ganbar 4. 2 Antar Muka Halaman Clustering Result: Movie Synopsis Labelled



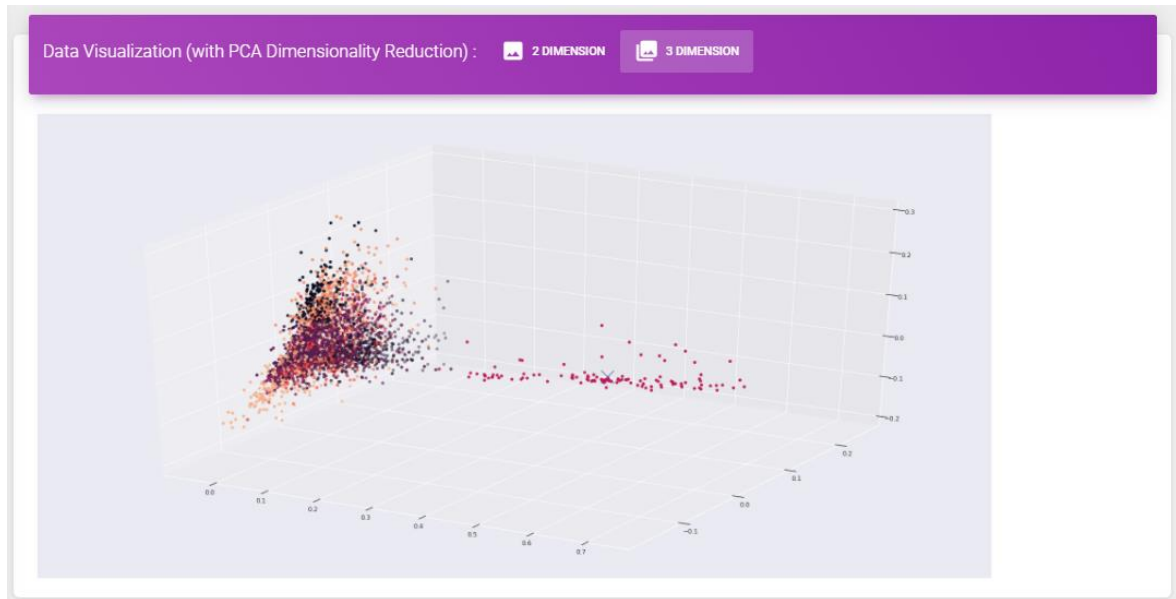
Ganbar 4. 3 Antar Muka Halaman Clustering Result: Evaluasi Silhouette



Ganbar 4. 4 Antar Muka Halaman Clustering Result: Evaluasi Elbow Method



Ganbar 4. 5 Antar Muka Halaman Clustering Result: Visualisasi 2 Dimensi



Ganbar 4. 6 Antar Muka Halaman Clustering Result: Visualisasi 3 Dimensi

4.1.3. Tampilan Antar Muka Halaman Feature Names

Halaman feature names merupakan halaman yang menampilkan 10 fitur kata paling relevan untuk tiap cluster. Pada halaman ini akan ditampilkan juga WordCloud dari fitur-fiitur pada tiap cluster. Berikut merupakan tampilan antar muka halaman feature names:

MOVIE SYNOPSIS

CLUSTERING

K-Means Model Training

Clustering Result

Feature Names

Data Per Clusters

Cluster Prediction

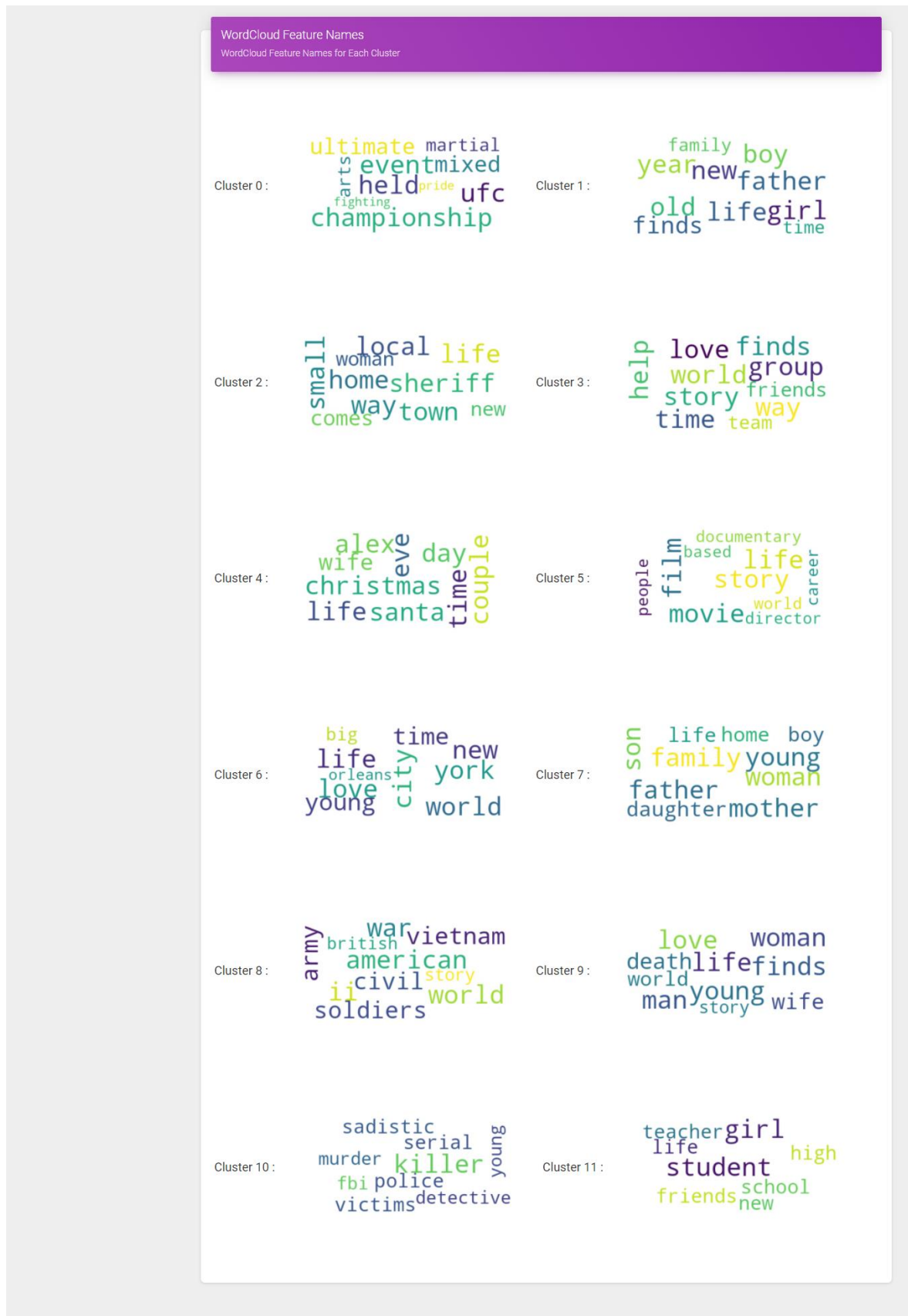
Cluster Prediction (By Title)

Feature Names for Each Cluster

Preview top 10 features that represent each cluster

Cluster 0 :
event, ufc, championship, held, ultimate, mixed, martial, arts, fighting, pride,
Cluster 1 :
old, year, life, boy, new, father, girl, finds, family, time,
Cluster 2 :
town, small, sheriff, local, life, way, home, comes, woman, new,
Cluster 3 :
world, love, story, way, time, group, finds, help, friends, team,
Cluster 4 :
just, christmas, santa, couple, day, alex, life, eve, time, wife,
Cluster 5 :
film, life, story, movie, documentary, world, based, director, people, career,
Cluster 6 :
new, york, city, life, world, love, young, time, big, orleans,
Cluster 7 :
young, family, father, mother, woman, son, daughter, boy, life, home,
Cluster 8 :
war, world, ii, american, vietnam, soldiers, civil, army, story, british,
Cluster 9 :
man, young, woman, love, life, finds, death, wife, world, story,
Cluster 10 :
killer, serial, victims, police, sadistic, detective, murder, killers, fbi, young,
Cluster 11 :
school, high, student, teacher, students, friends, new, girl, girls, life,

Ganbar 4. 7 Antar Muka Halaman Feature Names Tiap Cluster



Ganbar 4. 8 Antar Muka Halaman Featue Names: WordCloud Featue Names

4.1.4. Tampilan Antar Muka Halaman Data Per Cluster

Pada halaman ini akan ditampilkan tabel preview 5 data untuk tiap cluster. Data yang ditampilkan pada masing-masing cluster berupa data id film, judul film, serta sinopsisnya. Pada halaman ini juga disediakan tombol untuk download full data per cluster dengan format .csv. Berikut merupakan tampilan antar muka halaman data per cluster:

MOVIE SYNOPSIS
CLUSTERING

K-Means Model Training

Clustering Result

Feature Names

Data Per Clusters

Cluster Prediction

Cluster Prediction (By Title)

Cluster 0
Preview of cluster 0 data (5 Data)

ID	Title	Synopsis
4792	UFC 88: Breakthrough	UFC 88: Breakthrough was a mixed martial arts event held by the Ultimate Fighting Championship (UFC) on September 6, 2008 at the Philips Arena in Atlanta, Georgia. The event was headlined by a light heavyweight bout between Chuck Liddell and Rashad Evans.
4793	UFC 87: Seek and Destroy	UFC 87: Seek and Destroy was a mixed martial arts event held by the Ultimate Fighting Championship on August 9, 2008, at the Target Center in Minneapolis, Minnesota. The card was headlined by a welterweight championship bout between champion Georges St. Pierre and challenger Jon Fitch.
4794	UFC 86: Jackson vs. Griffin	UFC 86: Jackson vs. Griffin was a mixed martial arts event held by the Ultimate Fighting Championship (UFC) on July 5, 2008, at the Mandalay Bay Events Center in Las Vegas, Nevada. The title bout between Quinton "Rampage" Jackson and Forrest Griffin, coaches on The Ultimate Fighter: Team Rampage vs. Team Forrest, was for Jackson's UFC Light Heavyweight Championship.
4822	UFC 85: Bedlam	UFC 85: Bedlam was a mixed martial arts event held by the Ultimate Fighting Championship (UFC), on June 7, 2008 at The O2 arena in London, England. Former two-time UFC® welterweight champion MATT HUGHES begins his run for an unprecedented third title against a young gun eager to take his place near the top of the 170-pound ladder in Brazilian slugger THIAGO "PITBULL" ALVES.
4823	UFC 84: Ill Will	UFC 84: Ill Will was a mixed martial arts event held by the Ultimate Fighting Championship (UFC) on May 24, 2008, at the MGM Grand Garden Arena in Las Vegas, Nevada. The card featured the return of Sean Sherk, who was suspended and stripped of his UFC lightweight title after he tested for steroids at UFC 73. He faced G.J. Penn, who had since won the vacated title.

DOWNLOAD FULL CLUSTER 0.CSV

Cluster 1
Preview of cluster 1 data (5 Data)

ID	Title	Synopsis
31	Billy Elliot	Set against the background of the 1984 Miners' Strike, 11-year-old Billy Elliot stumbles out of the boxing ring and onto the ballet floor. He faces many trials and triumphs as he strives to conquer his family's set ways, inner conflict, and standing on his toes.
43	Dirty Dancing	Expecting the usual tedium that accompanies a summer in the Catskills with her family, 17-year-old Frances "Baby" Houseman is surprised to find herself stepping into the shoes of a professional hooper—and unexpectedly falling in love.
54	Léon: The Professional	Léon, the top hit man in New York, has earned a rep as an effective "cleaner". But when his next-door neighbors are wiped out by a loose-cannon DEA agent, he becomes the unwilling custodian of 12-year-old Mathilda. Before long, Mathilda's thoughts turn to revenge, and she considers following in Léon's footsteps.
88	Ocean's Twelve	Danny Ocean reunites with his old flame and the rest of his merry band of thieves in carrying out three huge heists in Rome, Paris and Amsterdam – but a Europol agent is hot on their heels.
148	About a Boy	Will Freeman is a good-looking, smooth-talking bachelor whose primary goal in life is avoiding any kind of responsibility. But when he invents an imaginary son in order to meet attractive single moms, Will gets a hilarious lesson about life from a bright, but hopelessly geeky 12-year-old named Marcus. Now, as Will struggles to teach Marcus the art of being cool, Marcus teaches Will that you're never too old to grow up.

DOWNLOAD FULL CLUSTER 1.CSV

Ganbar 4. 9 Antar Muka Halaman Data Cluster: 0 dan 1

Cluster 2
Preview of cluster 2 data (5 Data)

ID	Title	Synopsis
19	The Simpsons Movie	After Homer accidentally pollutes the town's water supply, Springfield is encased in a gigantic dome by the EPA and the Simpsons are declared fugitives.
25	8 Mile	The setting is Detroit in 1995. The city is divided by 8 Mile, a road that splits the town in half along racial lines. A young white rapper, Jimmy "B-Rabbit" Smith Jr. summons strength within himself to cross over these arbitrary boundaries to fulfill his dream of success in hip hop. With his pal Future and the three one third in place, all he has to do is not choke.
47	Anatomy of a Murder	Semi-retired Michigan lawyer Paul Biegler takes the case of Army Lt. Manion, who murdered a local innkeeper after his wife claimed that he raped her. Over the course of an extensive trial, Biegler parries with District Attorney Lodwick and out-of-town prosecutor Claude Dancer to set his client free, but his case rests on the victim's mysterious business partner, who's hiding a dark secret.
53	Lock, Stock and Two Smoking Barrels	A card shark and his unwillingly-enlisted friends need to make a lot of cash quick after losing a sketchy poker match. To do this they decide to pull a heist on a small-time gang who happen to be operating out of the flat next door.
74	Groundhog Day	A narcissistic TV weatherman, along with his attractive-but-distant producer, and his mawkish cameraman, is sent to report on Groundhog Day in the small town of Punxsutawney, where he finds himself repeating the same day over and over.

Download Full Cluster 2.CSV

Cluster 3
Preview of cluster 3 data (5 Data)

ID	Title	Synopsis
0	Four Rooms	It's Ted the Bellhop's first night on the job...and the hotel's very unusual guests are about to place him in some outrageous predicaments. It seems that this evening's room service is serving up one unbelievable happening after another.
1	Judgment Night	While racing to a boxing match, Frank, Mike, John and Rey get more than they bargained for. A wrong turn lands them directly in the path of Fallon, a vicious, wise-cracking drug lord. After accidentally witnessing Fallon murder a disloyal henchman, the four become his unwilling prey in a savage game of cat & mouse as they are mercilessly stalked through the urban jungle in this taut suspense drama
3	Star Wars	Princess Leia is captured and held hostage by the evil Imperial forces in their effort to take over the galactic Empire. Venturesome Luke Skywalker and dashing captain Han Solo team together with the loveable robot duo R2-D2 and C-3PO to rescue the beautiful princess and restore peace and justice in the Empire.
13	Pirates of the Caribbean: The Curse of the Black Pearl	Jack Sparrow, a freewheeling 18th-century pirate, quarrels with a rival pirate bent on pillaging Port Royal. When the governor's daughter is kidnapped, Sparrow decides to help the girl's love save her.
14	Kill Bill: Vol. 1	An assassin is shot by her ruthless employer, Bill, and other members of their assassination circle – but she lives to plot her vengeance.

Download Full Cluster 3.CSV

Ganbar 4. 10 Antar Muka Halaman Data Cluster: 2 dan 3

Cluster 4		
Preview of cluster 4 data (5 Data)		
ID	Title	Synopsis
8	Dancer in the Dark	Selma, a Czech immigrant on the verge of blindness, struggles to make ends meet for herself and her son, who has inherited the same genetic disorder and will suffer the same fate without an expensive operation. When life gets too difficult, Selma learns to cope through her love of musicals, escaping life's troubles - even if just for a moment - by dreaming up little numbers to the rhythmic beats of her surroundings.
41	Raiders of the Lost Ark	When Dr. Indiana Jones – the tweed-suited professor who just happens to be a celebrated archaeologist – is hired by the government to locate the legendary Ark of the Covenant, he finds himself up against the entire Nazi regime.
106	A Clockwork Orange	In a near-future Britain, young Alexander DeLarge and his pals get their kicks beating and raping anyone they please. When not destroying the lives of others, Alex swoons to the music of Beethoven. The state, eager to crack down on juvenile crime, gives an incarcerated Alex the option to undergo an invasive procedure that'll rob him of all personal agency. In a time when conscience is a commodity, can Alex change his tune?
126	Saw II	When a new murder victim is discovered with all the signs of Jigsaw's hand, Detective Eric Matthews begins a full investigation and apprehends Jigsaw with little effort. But for Jigsaw, getting caught is just another part of his plan. Eight more of his victims are already fighting for their lives and now it's time for Matthews to join the game.
146	High Fidelity	When record store owner Rob Gordon gets dumped by his girlfriend, Laura, because he hasn't changed since they met, he revisits his top five breakups of all time in an attempt to figure out what went wrong. As Rob seeks out his former lovers to find out why they left, he keeps up his efforts to win Laura back.
Download Full Cluster 4.csv		
Cluster 5		
Preview of cluster 5 data (5 Data)		
ID	Title	Synopsis
2	Life in Loops (A Megacities RMX)	Timo Novotny labels his new project an experimental music documentary film, in a remix of the celebrated film Megacities (1997), a visually refined essay on the hidden faces of several world "megacities" by leading Austrian documentarist Michael Glawogger. Novotny complements 30 % of material taken straight from the film (and re-edited) with 70 % as yet unseen footage in which he blends original shots unused by Glawogger with his own sequences (shot by Megacities cameraman Wolfgang Thaler) from Tokyo. Alongside the Japanese metropolis, Life in Loops takes us right into the atmosphere of Mexico City, New York, Moscow and Bombay. This electrifying combination of fascinating film images and an equally compelling soundtrack from Sofa Surfers sets us off on a stunning audiovisual adventure across the continents. The film also makes an original contribution to the discussion on new trends in documentary filmmaking. Written by KARLOVY VARY IFF 2006
6	American Beauty	Lester Burnham, a depressed suburban father in a mid-life crisis, decides to turn his hectic life around after developing an infatuation with his daughter's attractive friend.
12	The Endless Summer	Bruce Brown's The Endless Summer is one of the first and most influential surf movies of all time. The film documents American surfers Mike Hynson and Robert August as they travel the world during California's winter (which, back in 1965 was off-season for surfing) in search of the perfect wave and ultimately, an endless summer.
48	Kunstgriff	Kunstgriff is a brilliantly filmed black and white short film. Andre F. Nebe gives proof of his storytelling abilities in this 6 minute film.
61	The Big Lebowski	Jeffrey 'The Dude' Lebowski, a Los Angeles slacker who only wants to bowl and drink White Russians, is mistaken for another Jeffrey Lebowski, a wheelchair-bound millionaire, and finds himself dragged into a strange series of events involving nihilists, adult film producers, ferrets, errant toes, and large sums of money.
Download Full Cluster 5.csv		

Ganbar 4. 11 Antar Muka Halaman Data Cluster: 4 dan 5

Cluster 6
Preview of cluster 6 data (5 Data)

ID	Title	Synopsis
55	Taxi Driver	A mentally unstable Vietnam War veteran works as a night-time taxi driver in New York City where the perceived decadence and sleaze feed his urge for violent action, attempting to save a preadolescent prostitute in the process.
89	Breakfast at Tiffany's	Holly Golightly is an eccentric New York City playgirl determined to marry a Brazilian millionaire. But when young writer Paul Varjak moves into her apartment building, her past threatens to get in their way.
114	Braveheart	Enraged at the slaughter of Murron, his new bride and childhood love, Scottish warrior William Wallace slays a platoon of the local English lord's soldiers. This leads the village to revolt and, eventually, the entire country to rise up against English rule.
125	Saw III	Jigsaw has disappeared. Along with his new apprentice Amanda, the puppet-master behind the cruel, intricate games that have terrified a community and baffled police has once again eluded capture and vanished. While city detective scramble to locate him, Doctor Lynn Denlon and Jeff Reinhart are unaware that they are about to become the latest pawns on his vicious chessboard.
130	Rebel Without a Cause	After moving to a new town, troublemaking teen Jim Stark is supposed to have a clean slate, although being the new kid in town brings its own problems. While searching for some stability, Stark forms a bond with a disturbed classmate, Plato, and falls for local girl Judy. However, Judy is the girlfriend of neighborhood tough, Buzz. When Buzz violently confronts Jim and challenges him to a drag race, the new kid's real troubles begin.

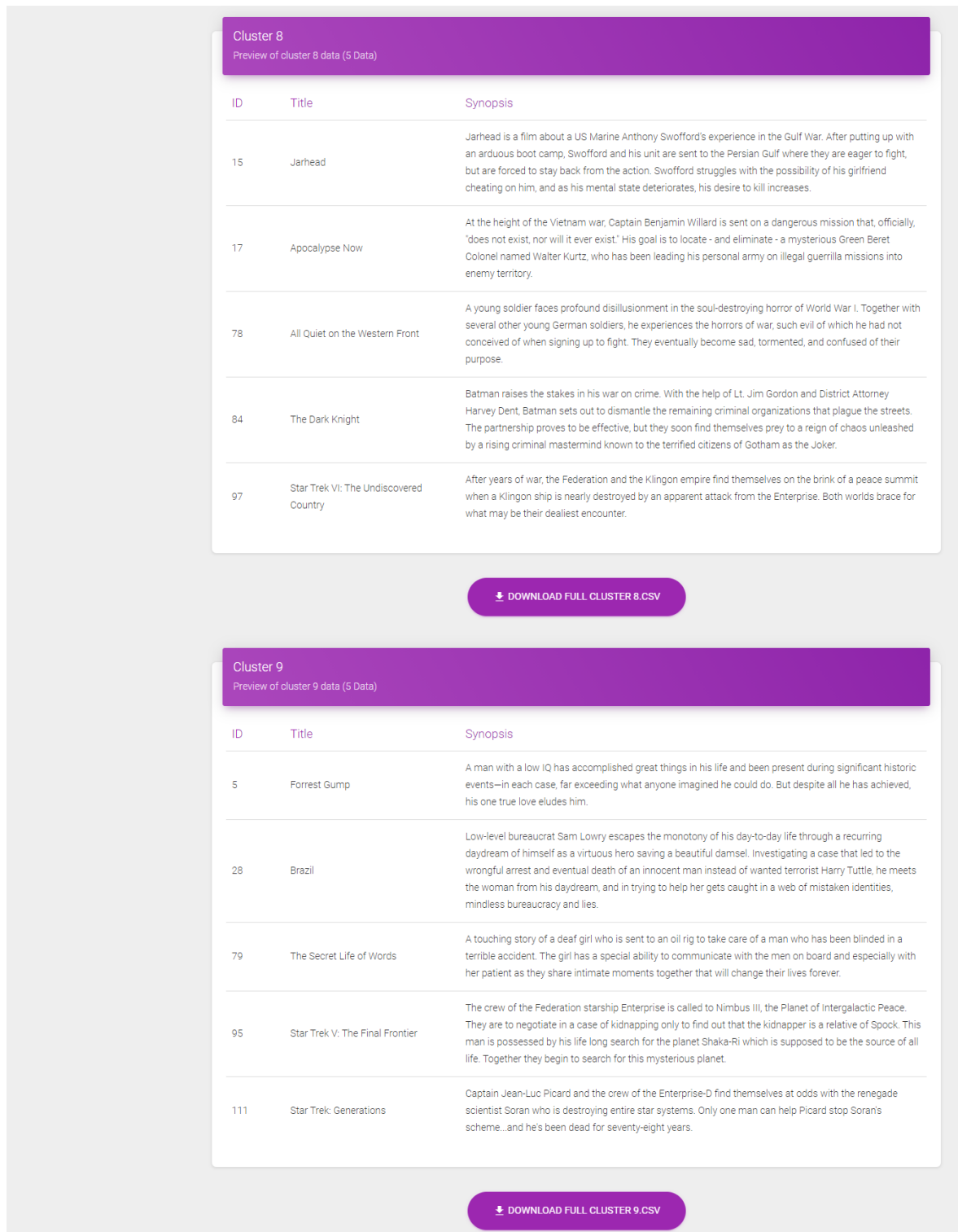
Download Full Cluster 6.CSV

Cluster 7
Preview of cluster 7 data (5 Data)

ID	Title	Synopsis
4	Finding Nemo	Nemo, an adventurous young clownfish, is unexpectedly taken from his Great Barrier Reef home to a dentist's office aquarium. It's up to his worrisome father Marlin and a friendly but forgetful fish Dory to bring Nemo home -- meeting vegetarian sharks, surfer dude turtles, hypnotic jellyfish, hungry seagulls, and more along the way.
7	Citizen Kane	Newspaper magnate, Charles Foster Kane is taken from his mother as a boy and made the ward of a rich industrialist. As a result, every well-meaning, tyrannical or self-destructive move he makes for the rest of his life appears in some way to be a reaction to that deeply wounding event.
9	The Dark	Adèle and her daughter Sarah are traveling on the Welsh coastline to see her husband James when Sarah disappears. A different but similar looking girl appears who says she died in a past time. Adèle tries to discover what happened to her daughter as she is tormented by Celtic mythology from the past.
10	The Fifth Element	In 2257, a taxi driver is unintentionally given the task of saving a young girl who is part of the key that will ensure the survival of humanity.
11	My Life Without Me	A fatally ill mother with only two months to live creates a list of things she wants to do before she dies without telling her family of her illness.

Download Full Cluster 7.CSV

Ganbar 4. 12 Antar Muka Halaman Data Cluster: 6 dan 7



Ganbar 4. 13 Antar Muka Halaman Data Cluster: 8 dan 9

Cluster 10
Preview of cluster 10 data (5 Data)

ID	Title	Synopsis
18	Unforgiven	William Munny is a retired, once-ruthless killer turned gentle widower and hog farmer. To help support his two motherless children, he accepts one last bounty-hunter mission to find the men who brutalized a prostitute. Joined by his former partner and a cocky greenhorn, he takes on a corrupt sheriff.
36	Memento	Leonard Shelby is tracking down the man who raped and murdered his wife. The difficulty of locating his wife's killer, however, is compounded by the fact that he suffers from a rare, untreatable form of short-term memory loss. Although he can recall details of life before his accident, Leonard cannot remember what happened fifteen minutes ago, where he's going, or why.
80	48 Hrs.	A hard-nosed cop reluctantly teams up with a wise-cracking criminal temporarily paroled to him, in order to track down a killer.
94	28 Days Later	Twenty-eight days after a killer virus was accidentally unleashed from a British research facility, a small group of London survivors are caught in a desperate struggle to protect themselves from the infected. Carried by animals and humans, the virus turns those it infects into homicidal maniacs – and it's absolutely impossible to contain.
98	Saw	Obsessed with teaching his victims the value of life, a deranged, sadistic serial killer abducts the morally wayward. Once captured, they must face impossible choices in a horrific game of survival. The victims must fight to win their lives back, or die trying...

Download Full Cluster 10.CSV

Cluster 11
Preview of cluster 11 data (5 Data)

ID	Title	Synopsis
62	Match Point	Match Point is Woody Allen's satire of the British High Society and the ambition of a young tennis instructor to enter into it. Yet when he must decide between two women - one assuring him his place in high society, and the other that would take him far from it - palms start to sweat and a dark psychological match in his head begins.
122	Dead Poets Society	At an elite, old-fashioned boarding school in New England, a passionate English teacher inspires his students to rebel against convention and seize the potential of every day, courting the disdain of the stern headmaster.
257	Klute	A high-priced call girl is forced to depend on a reluctant private eye when she is stalked by a psychopath.
258	The Hole	Four teenagers at a British private school secretly uncover and explore the depths of a sealed underground hole created decades ago as a possible bomb shelter.
315	Spider-Man	After being bitten by a genetically altered spider, nerdy high school student Peter Parker is endowed with amazing powers to become the Amazing superhero known as Spider-Man.

Download Full Cluster 11.CSV

Gambar 4. 14 Antar Muka Halaman Data Cluster: 10 dan 11

4.1.5. Tampilan Antar Muka Halaman Cluster Prediction

Pada halaman ini pengguna dapat melakukan prediksi cluster dengan memasukkan data sinopsis suatu film. Sebagai contoh sinopsis yang dimasukkan adalah sinopsis film Finding Dory, yaitu "After his son is captured in the Great Barrier Reef and taken to Sydney, a timid clownfish sets out on a journey to bring him home". Setelah meinputkan sinopsis tersebut didapatkan prediksi cluster-nya adalah cluster 7.

Pada halaman ini juga menampilkan 10 film related movies, di mana 10 film tersebut merupakan film yang terkait dengan sinopsis yang dimasukkan (ditampilkan juga nilai cosine similarity berurutan dari yang terbesar ke terkecil, semakin besar nilai cosine similarity maka akan semakin mirip dengan sinopsis yang dimasukkan). Berikut merupakan tampilan antar muka halaman cluster prediction:

MOVIE SYNOPSIS CLUSTERING

- K-Means Model Training
- Clustering Result
- Feature Names
- Data Per Clusters
- Cluster Prediction**
- Cluster Prediction (By Title)

Predict Cluster from Synopsis
With movie recommendation from related synopsis

Synopsis :

Friendly but forgetful blue tang Dory begins a search for her long-lost parents, and everyone learns a fe

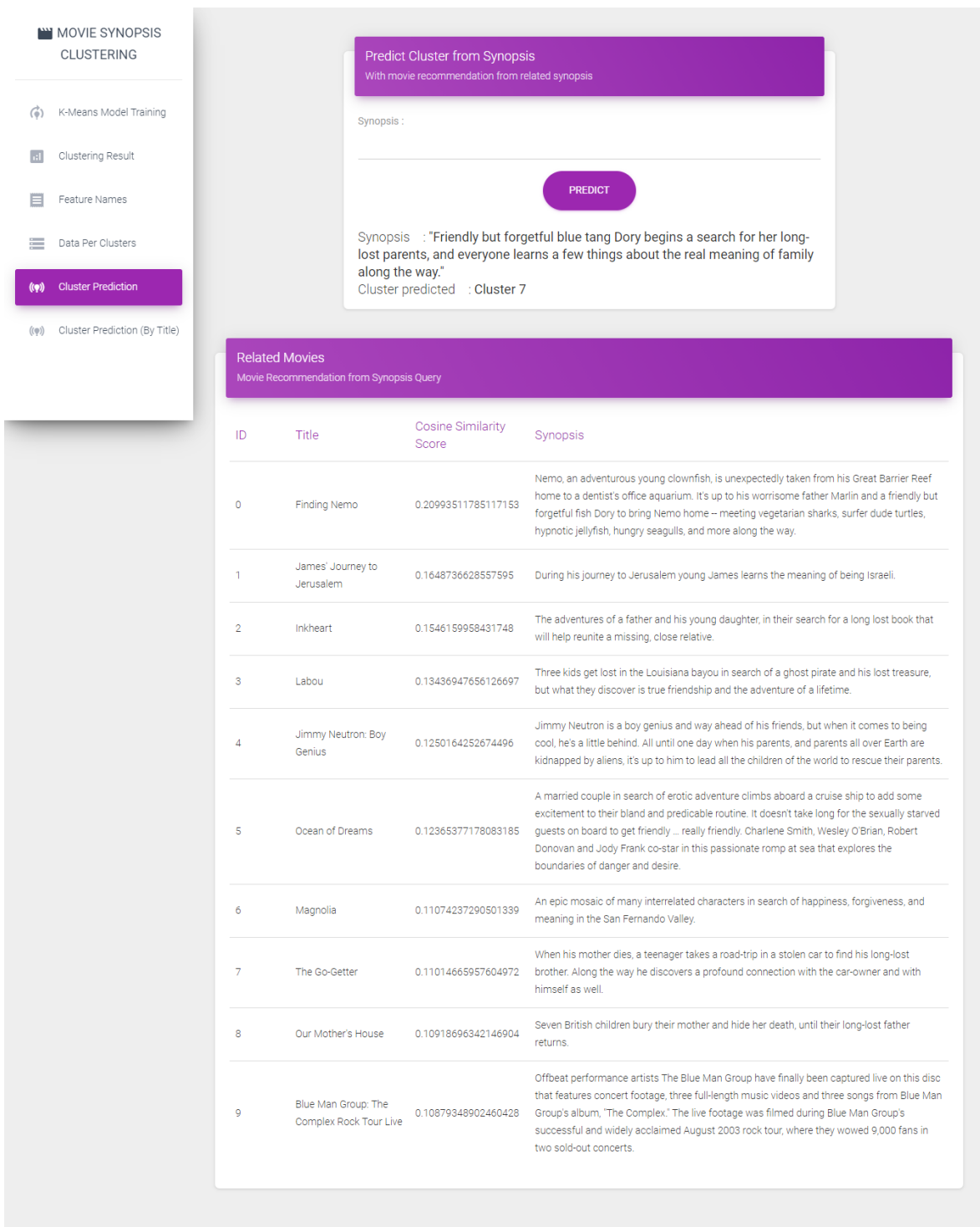
PREDICT

Synopsis : ""
Cluster predicted : Cluster

Related Movies
Movie Recommendation from Synopsis Query

ID	Title	Cosine Similarity Score	Synopsis

Ganbar 4. 15 Antar Muka Halaman Cluster Prediction



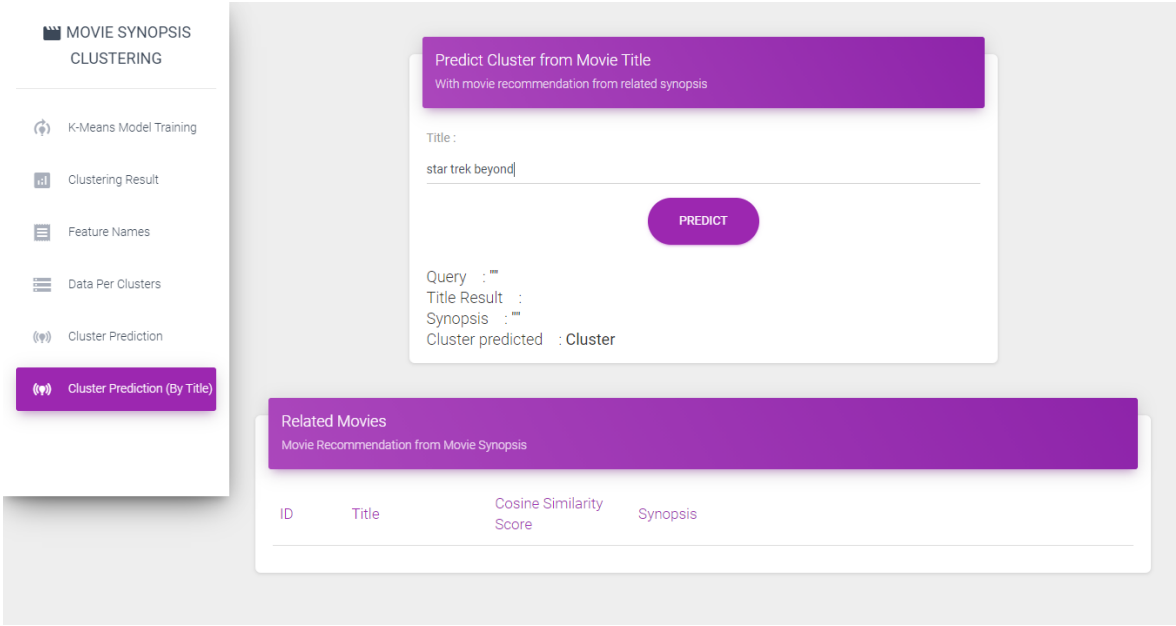
Gambar 4. 16 Antar Muka Halaman Cluster Prediction: Related Movies

4.1.6. Tampilan Antar Muka Halaman Cluster Prediction (By Title)

Halaman ini merupakan halaman untuk menentukan prediksi cluster. Namun berbeda dengan halaman sebelumnya yang menentukan prediksi cluster dengan memasukkan sinopsis film sebagai masukkannya, halaman prediksi cluster ini menggunakan judul film sebagai data masukkannya. Sebagai contoh, dimasukkan judul film "star trek beyond" kemudian

didapatkan hasil pencarian judul dan sinopsis dari database. Setelah itu data judul dan sinopsis yang telah ditemukan akan ditampilkan pada halaman aplikasi disertai dengan tampilan prediksi cluster yang didapatkan yaitu cluster 6.

Pada halaman ini juga menampilkan 10 film related movies, di mana 10 film tersebut merupakan film yang terkait dengan judul film yang dimasukkan (ditampilkan juga nilai cosine similarity berurutan dari yang terbesar ke terkecil, semakin besar nilai cosin similarity maka akan semakin mirip dengan sinopsis yang dimasukkan). Berikut merupakan tampilan antar muka halaman cluster prediction (by title):



Ganbar 4. 17 Antar Muka Halaman Cluster Prediction (By Title)

4.2. Hasil Penelitian

Hasil penelitian berupa aplikasi text clustering untuk menentukan genre film berdasarkan sinopsisnya. Aplikasi text clustering ini dibuat dalam bentuk website sehingga dapat mempermudah user untuk menggunakannya. Aplikasi dibuat menggunakan bahasa pemrograman Python dan untuk halaman antar muka dibuat menggunakan bahasa pemrograman HTML.

Sesuai dengan fungsinya aplikasi ini memiliki enam halaman utama sebagai halaman fungsionalnya. Keenam halaman tersebut yaitu:

1. Halaman K-Means Model Training
2. Halaman Clustering Result
3. Halaman Feature Names
4. Halaman Data Per Cluster
5. Halaman Cluster Prediction
6. Halaman Cluster Prediction (By Title)

Aplikasi ini memberikan informasi mengenai clustering film berdasarkan sinopsisnya dan menampilkan hasil clustering tersebut. Aplikasi ini juga akan menampilkan evaluasi hasil clustering yang telah dilakukan serta dapat menampilkan visualisasi hasil clustering dalam bentuk visualisasi 2 dimensi dan 3 dimensi.

4.3. Analisa Hasil

Berdasarkan dari hasil penelitian yang telah dijelaskan pada bagian sebelumnya, Aplikasi Clustering Film Berdasarkan Sinopsisnya sudah dapat berjalan dengan baik sesuai dengan fungsionalitasnya. Clustering yang dihasilkan sudah sesuai dengan jumlah cluster yang dimasukkan oleh user pada halaman K-Means Model Training. Pada halaman clustering result Aplikasi Clustering Film Berdasarkan Sinopsisnya juga sudah menampilkan hasil cluster yang sesuai. Hasil evaluasi cluster yang ditampilkan pada halaman clustering result juga sudah sesuai, juga dengan tampilan visualisasi data hasil clustering yang ditampilkan sudah cukup baik.

Pada halaman feature names Aplikasi Clustering Film Berdasarkan Sinopsisnya, nama-nama fitur yang ditampilkan juga sudah sesuai dengan genre film yang terkait. Halaman data per cluster yang ditampilkan pada Aplikasi Clustering Film Berdasarkan Sinopsisnya juga sudah dapat menampilkan data judul film dan sinopsisnya sesuai dengan cluster masing-masing.

Fungsi yang ada di halaman cluster prediction pada Aplikasi Clustering Film Berdasrkan Sinopsisnya juga sudah berjalan dengan baik. Hal ini dibuktikan dengan hasil prediksi cluster yang dihasilkan pada Aplikasi Clustering Film Berdasrkan Sinopsisnya, baik dengan masukan berupa sinopsis film maupun dengan masukan judul film. Selain itu, 10 Film yang ditampilkan sebagai related movies juga sudah sesuai dengan film yang terkait.

BAB V

PENUTUP

Bab ini merupakan penutup yang berisi kesimpulan dari pembuatan Aplikasi Text Clustering untuk Menentukan Genre Film Berdasarkan Sinopsisnya.

5.1. Kesimpulan

Aplikasi Clustering Film Berdasarkan Sinopsisnya merupakan text clustering untuk menentukan genre film berdasarkan sinopsisnya. Aplikasi ini dibuat dalam bentuk website dengan menggunakan bahasa pemrograman Python dan untuk halaman antar muka dibuat menggunakan bahasa pemrograman HTML. Data yang digunakan dalam aplikasi ini menggunakan data judul film dan sinopsis dari website <https://www.themoviedb.org/>.

Berdasarkan fungsionalitasnya aplikasi ini memiliki enam halaman utama sebagai halaman fungsionalnya, yaitu halaman K-Means Model Training, halaman Clustering Result, halaman Feature Names, halaman Data Per Cluster, halaman Cluster Prediction, halaman Cluster Prediction (By Title). Aplikasi Clustering Film Berdasarkan Sinopsisnya menampilkan hasil clustering memberikan informasi mengenai clustering film berdasarkan sinopsisnya sesuai dengan jumlah cluster yang diinputkan oleh user. Aplikasi ini juga akan menampilkan evaluasi hasil clustering yang telah dilakukan dengan menggunakan Evaluasi Silhouette dan Elbow Method, serta dapat menampilkan visualisasi hasil clustering dalam bentuk visualisasi 2 dimensi dan 3 dimensi.

Sejauh ini Aplikasi Clustering Film Berdasarkan Sinopsisnya sudah berjalan dengan baik. Semua fitur yang ada pada Aplikasi Clustering Film Berdasarkan Sinopsisnya juga sudah bekerja sesuai dengan fungsinya. Penulis berharap Aplikasi Clustering Film Berdasarkan Sinopsisnya ini dapat tetap berjalan dengan baik sehingga dapat menghasilkan cluster yang sesuai dengan inputan yang dimasukan user dan dapat memberikan informasi sesuai dengan apa yang diharapkan oleh user.

DAFTAR PUSTAKA

- Adriani, Mirna. 2016. *Pemrosesan Teks*. Jakarta: Fasilkom, Universitas Indonesia.
- Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer.
- Jain, Anil K. 1988. *Algorithms for Clustering Data*. Prentice Hall.
- Nedjah, Nadia, et al. 2009. *Intelligent Text Categorization and Clustering*. Springer.

LAMPIRAN

1. Source Code Program

Import Dependencies

```
import json
import requests
import os, shutil
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import pickle
from google.colab import drive
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from wordcloud import WordCloud
from sklearn.metrics.pairwise import cosine_similarity,
euclidean_distances
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
```

```
drive.mount('/content/gdrive/')
path = "/content/gdrive/MyDrive/Upload/Akademik/Tugas Besar PBA"
path_api = "/content/gdrive/MyDrive/Upload/API/"
```

Data Collecting

Retrieve Genres

```
# Get All Genres
```

```
genres = []
for i in range(20000) :
    url = ("https://api.themoviedb.org/3/movie/%d?api_key=63fea4c7
09da1f1496b7a1ca7a3c6083" % i)
    r = requests.get(url)
    json_data = json.loads(r.text)
    try:
        if (json_data['genres'] != "") :
```

```

        # print(json_data['genres'])
        for j in json_data['genres'] :
            genre = j.get('name')
            if (genre not in genres) :
                genres.append(genre)
    except Exception:
        pass
print(genres)
len(genres)

```

Retrieve Titles & Synopsis

```

titles = []
synopsis = []

for i in range(20000) :
    url = ("https://api.themoviedb.org/3/movie/%d?api_key=63fea4c709da1f1496b7a1ca7a3c6083" % i)
    r = requests.get(url)
    json_data = json.loads(r.text)
    try:
        if (json_data['overview'] != "" and json_data['overview'] !=
            "No overview found." and json_data['original_language'] == 'en'
        ) :
            titles.append(json_data['title'])
            synopsis.append(json_data['overview'])
    except Exception:
        pass
import pandas as pd

df_movies = pd.DataFrame({'title': titles, 'synopsis': synopsis}
)
df_movies.head()
df_movies.info()
with open('/content/gdrive/MyDrive/Upload/movie_synopsis.csv', '
w') as f:
    df_movies.to_csv(f)

```

Data Preprocessing

Import Data

```

from google.colab import drive

drive.mount('/content/gdrive/')

import pandas as pd

```

```
df_movies = pd.read_csv('/content/gdrive/MyDrive/Upload/Akademik
/Tugas Besar PBA/movie_synopsis_8000_no overview.csv', linetermi
nator='\n')
df_movies.head()
df_movies.info()
```

Delete Missing Value

```
# Get names of indexes for which column synopsis has no overview
index_drop = df_movies[df_movies['synopsis'] == "No overview fou
nd."].index
```

```
# Delete these row indexes from dataFrame
df_movies.drop(index_drop , inplace=True)
```

```
# Reset the index
df_movies.reset_index(drop =True, inplace=True)
```

```
df_movies.info()
```

Case Folding

```
df_movies["synopsis"] = df_movies["synopsis"].str.lower()
df_movies.head()
```

TF-IDF Training

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
vectorizer = TfidfVectorizer(stop_words='english')
features = vectorizer.fit_transform(df_movies['synopsis'].values
) # training vector TF-IDF pada tiap data
```

```
# vector TF-
IDF tiap kalimat (features), pasangan index kata dan nilainya
for i in features :
    print(i)
    print('-----')
```

```
import pickle
```

```
# Save Model
pickle.dump(vectorizer, open('/content/gdrive/MyDrive/Upload/Aka
demik/Tugas Besar PBA/vectorizer.sav', 'wb'))
```

```
import pickle
```

```
# Load Model
vectorizer = pickle.load(open('/content/gdrive/MyDrive/Upload/Akademik/Tugas Besar PBA/vectorizer.sav', 'rb'))
```

KMeans Model

KMeans Model Training

```
from sklearn.cluster import KMeans

k = 19 # jumlah cluster
kmeans_model = KMeans(n_clusters = k, n_init = 3, max_iter = 500
)
synopsis_clusters = kmeans_model.fit(features)

import pickle

# Save Model
pickle.dump(kmeans_model, open('/content/gdrive/MyDrive/Upload/Akademik/Tugas Besar PBA/kmeans_model.sav', 'wb'))

import pickle

# Load Model
kmeans_model = pickle.load(open('/content/gdrive/MyDrive/Upload/Akademik/Tugas Besar PBA/kmeans_model.sav', 'rb'))
```

Data Labelling

```
df_movies['label'] = kmeans_model.labels_
df_movies.head()

with open('/content/gdrive/MyDrive/Upload/movie_synopsis_labeled.csv', 'w') as f:
    df_movies.to_csv(f)
```

Data Distribution

```
movie_clusters = df_movies.groupby('label')

clusters_count = []
for cluster in movie_clusters :
    clusters_count.append(len(cluster[1]))
```

```
plt.figure(figsize=(14, 7))
sns.set_theme(style="whitegrid")
graph = sns.barplot(x = np.arange(k), y = clusters_count,
palette='hls')
```

Data from each Cluster

Data Preview for each Cluster

```
clusters = df_movies.groupby('label')

for cluster in clusters.groups :
    print("Cluster %d : " % cluster)
    data_cluster = clusters.get_group(cluster)[['title','synopsis']]
    for i in range(5) :
        data = data_cluster.iloc[i]
        print('    ', data['title'], ': ', data['synopsis'])
    print('\n')
```

Save All Data from each Cluster

```
clusters = df_movies.groupby('label')

for cluster in clusters.groups :
    f = open('cluster'+str(cluster)+ '.csv', 'w') # buat file csv
    v untuk tiap cluster
    data = clusters.get_group(cluster)[['title','synopsis']] # j
    udul dan sinopsis tiap data pada tiap cluster
    f.write(data.to_csv(index_label='id')) # simpan ke csv
    f.close()
```

Feature Names of each Cluster

```
order_centroids = kmeans_model.cluster_centers_.argsort()[:, :-
1] # diurutkan berdasarkan indeks -> lalu di-reversed
terms = vectorizer.get_feature_names()
n_terms = 10

for i in range(k) :
    print("Cluster %d : " % i)
    for j in order_centroids [i, :n_terms] :
        print('    %s' % terms[j])
    print('-----')

from wordcloud import WordCloud
import matplotlib.pyplot as plt
```

```

c = []
keywords = []
for i in range(k) :
    c.append(i)
    key = ""
    for j in order_centroids [i, :n_terms] :
        key = key+(" ")+(terms[j])
    keywords.append(key)

for i in range(k) :
    print('Cluster: %d' % c[i])
    text = keywords[i]
    wordcloud = WordCloud(max_font_size=50, background_color="white").generate(text)
    plt.figure()
    plt.imshow(wordcloud, interpolation="bilinear")
    plt.axis("off")
    plt.show()

```

Cluster Prediction and Recommendation

```

def predict_cluster(sentence) :
    Y = vectorizer.transform([sentence])
    prediction = kmeans_model.predict(Y)
    cluster_prediction = prediction[0]
    print("Sentence          : ", sentence)
    print("Cluster predicted : ", cluster_prediction)

def movie_recommendation(n, query, features, vectorizer, df_movies) :
    query = [query.lower()]
    query = vectorizer.transform(query)

    cosine_score = []
    index = 0
    for i in features :
        cosine_score.append(cosine_similarity(query, i))
        index +=1

    cosine_score_update = []
    for i in cosine_score :
        cosine_score_update.append(i[0][0])
    cosine_score_update = np.array(cosine_score_update)
    cosine_score_update
    indices = np.argsort(cosine_score_update)[:, :-1]

    for i in range(n) :

```



```

        index_now = indices[i]
        if (cosine_score_update[index_now] != 0) :
            print(i, " . ", df_movies['title'][index_now], " : ", cosine_score_update[index_now], " : ", df_movies['synopsis'][index_now])

def movie_recommendation_by_title(n, title, features, vectorizer, df_movies) :
    index = 0
    movie_index = 0
    for i in df_movies['title'] :
        if (i == title) :
            movie_index = index
            break
        index +=1

    synopsis_ori = ""
    title_new = ""
    url = ("https://api.themoviedb.org/3/search/movie?api_key=63fea4c709da1f1496b7a1ca7a3c6083&language=en-US&query=%s&page=1&include_adult=false" % title)
    r = requests.get(url)
    json_data = json.loads(r.text)
    try:
        if (json_data['results'][0]['overview'] != "") :
            title_new = json_data['results'][0]['title']
            synopsis_ori = json_data['results'][0]['overview']
    except Exception:
        pass

    synopsis = [synopsis_ori.lower()]
    synopsis = vectorizer.transform(synopsis)

    cosine_score = []
    index = 0
    for i in features :
        cosine_score.append(cosine_similarity(synopsis, i))
        index +=1

    cosine_score_update = []
    for i in cosine_score :
        cosine_score_update.append(i[0][0])
    cosine_score_update = np.array(cosine_score_update)
    cosine_score_update
    indices = np.argsort(cosine_score_update)[::-1]

    print("Title Query : ", title)
    print(title_new, " : ", synopsis_ori, "\n")

```

```

for i in range(n) :
    index_now = indices[i]
    if (cosine_score_update[index_now] != 0) :
        print(i, " . ", df_movies['title'][index_now], " : ", cosine_score_update[index_now], " : ", df_movies['synopsis'][index_now])

```

Predict Sentences

```

sentence = "Britain declared war on Germany in september 1939 and begin world war 2"

```

```

predict_cluster(sentence)

```

```

sentence = "There are a lot of aliens at space"

```

```

predict_cluster(sentence)

```

```

sentence = "Johnstone's Family is very nice family"

```

```

predict_cluster(sentence)

```

Movie Recommendation (by Query/ Synopsis)

```

n = 10
query = "Johnstone's Family is very nice family"

print("Synopsis : ", query, "\n")
movie_recommendation(n, query, features, vectorizer, df_movies)

```

Movie Recommendation (by Title)

```

n = 10
title = "Star Trek Beyond"

movie_recommendation_by_title(n, title, features, vectorizer, df_movies)

```

Evaluation

Elbow Method (SSE)

```

SSE = []
K = range(3,30)
for n_k in K:
    kmeans_model_iteration = KMeans(n_clusters = n_k).fit(features)
    SSE.append(kmeans_model_iteration.inertia)

```

```
import matplotlib.pyplot as plt

plt.plot(K, SSE, 'bx-')
plt.xlabel('k')
plt.ylabel('SSE')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```

Silhouette Score

```
from sklearn.metrics import silhouette_score

# nilai silhouette score antara -
1 dan 1, semakin tinggi semakin bagus
print(f'Silhouette Score : {silhouette_score(features, labels =
kmeans_model.labels_)})')
```

Data Visualization

2 Dimensi

PCA Dimensionality Reduction

```
from sklearn.decomposition import PCA

pca = PCA(n_components = 2)
reduced_features = pca.fit_transform(features.toarray())
reduced_cluster_centers = pca.transform(synopsis_clusters.cluste
r_centers_)

import matplotlib.pyplot as plt

plt.figure(figsize=(30, 15), dpi=80)
plt.scatter(reduced_features[:, 0], reduced_features[:,1], c = k
means_model.predict(features))
plt.scatter(reduced_cluster_centers[:, 0], reduced_cluster_cente
rs[:,1], marker='x', s=150, c='b')
```

t-SNE Dimensionality Reduction

```
from sklearn.manifold import TSNE

tsne = TSNE(n_components=2)
reduced_features_tsne = tsne.fit_transform(features.toarray())

import pickle
```

```
# Save Model
pickle.dump(tsne, open('/content/gdrive/MyDrive/Upload/Akademik/
Tugas Besar PBA/tsne.sav', 'wb'))

import matplotlib.pyplot as plt

plt.figure(figsize=(30, 15), dpi=80)
plt.scatter(reduced_features_tsne[:, 0], reduced_features_tsne[:, 1], c = kmeans_model.predict(features))
```

3 Dimensi

PCA Dimensionality Reduction

```
from sklearn.decomposition import PCA

pca_3d = PCA(n_components = 3)
reduced_features_3d = pca_3d.fit_transform(features.toarray())
reduced_cluster_centers_3d = pca_3d.transform(synopsis_clusters.
cluster_centers_)

%matplotlib notebook
%matplotlib inline
from mpl_toolkits.mplot3d import Axes3D

sns.set(style = "darkgrid")

fig = plt.figure(figsize=(30, 15), dpi=80)
ax = fig.add_subplot(111, projection = '3d')

x = reduced_features_3d[:, 0]
y = reduced_features_3d[:, 1]
z = reduced_features_3d[:, 2]

ax.scatter(x, y, z, c = synopsis_clusters.labels_)
ax.scatter(reduced_cluster_centers_3d[:, 0], reduced_cluster_centers_3d[:, 1], reduced_cluster_centers_3d[:, 2], marker='x', s=500, c='b')

plt.show()
```

t-SNE Dimensionality Reduction

```
from sklearn.manifold import TSNE

tsne_3d = TSNE(n_components=3)
```

```

reduced_features_3d_tsne = tsne_3d.fit_transform(features.toarray())

import pickle

# Save Model
pickle.dump(tsne_3d, open('/content/gdrive/MyDrive/Upload/Akademik/Tugas Besar PBA/tsne_3d.sav', 'wb'))

import matplotlib.pyplot as plt

from mpl_toolkits.mplot3d import Axes3D

sns.set(style = "darkgrid")

fig = plt.figure(figsize=(30, 20), dpi=80)
ax = fig.add_subplot(111, projection = '3d')

x = reduced_features_3d_tsne[:, 0]
y = reduced_features_3d_tsne[:, 1]
z = reduced_features_3d_tsne[:, 2]

ax.scatter(x, y, z, c = synopsis_clusters.labels_)

plt.show()

```