IMPLEMENTASI METODE *IMPROVED K-MEANS* DENGAN ALGORITMA *DBSCAN* UNTUK PENGELOMPOKAN FILM

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat
Untuk Memperoleh Gelar Sarjana Jurusan Statistika

Disusun oleh:

Annisa Ayunda Permata Sari

16611007

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2020

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul	**:**	**Implementasi Metode *Improved K-means* dengan Algoritma *Dbscan* untuk Pengelompokan Film**
Nama Mahasiswa	**:**	**Annisa Ayunda Permata Sari**
Nomor Mahasiswa	**:**	**16 611 007**

TUGAS AKHIR INI TELAH DIPERIKSAN DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta, 17 Maret 2020

Pembimbing

(Muhammad Muhajir, S.Si., M.Sc)

HALAMAN PENGESAHAN

TUGAS AKHIR

IMPLEMENTASI METODE *IMPROVED K-MEANS* DENGAN ALGORITMA *DBSCAN* UNTUK PENGELOMPOKAN FILM

Nama Mahasiswa : Annisa Ayunda Permata Sari

NIM : 16 611 007

TUGAS AKHIR INI TELAH DIUJIKAN
PADA TANGGAL: 3 April 2020

Nama Penguji:	**Tanda Tangan**
1. Ir. Ali Parkhan, M.T	.....
2. Tuti Purwaningsih, S. Stat, M.Si	.....
3. Muhammad Muhajir, S.Si., M.Sc	.....

Mengetahui,
Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

(Dr. Ir. Kryanto, S.Pd., M.Si., Ph.D.)

KATA PENGANTAR

بسم الله الرحمن الرحيم

Assalamu'alaikum Wr, Wb

Puji Syukur Kehadirat Allah SWT yang telah melimpahkan rahmat dan hidayah-Nya sehingga tugas akhir yang berjudul "**Implementasi Metode *Improved K-means* dengan Algoritma *Dbscan* untuk Pengelompokan Film**" dapat diselesaikan. Shalawat serta salam semoga selalu tercurah kepada junjungan Nabi Besar Muhammad SAW serta para sahabat dan pengikutnya sampai akhir jaman.

Tugas akhir ini disusun sebagai salah satu persyaratan yang harus dipenuhi dalam menyelesaikan jenjang Strata Satu atau S1 di Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia. Penyelesaian tugas akhir ini tidak terlepas dari bantuan, arahan, dan bimbingan dari berbagai pihak. Untuk itu Pada kesempatan ini penulis ingin mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Bapak Prof. Riyanto, S.Pd., M.Si., Ph.D. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia.
2. Bapak Dr. Edy Widodo S.Si., M.Si selaku Ketua Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Univeritas Islam Indonesia.
3. Bapak Muhammad Muhajir, S.Si., M.Sc, selaku Dosen pembimbing tugas akhir yang dengan kesabaran memberikan dukungan yang begitu besar dari awal bimbingan persiapan penulisan tugas akhir sampai pada selesainya penulisan tugas akhir ini.
4. Keluarga tercinta, Ibu Sri Marhaeningsih, S.H., Mas Very Perdana Saputra, Mbak Anggun Dwi W, S. Psg., Mas Pungky Marhendra P.P, S. Ars., Oktavia Fadriza Divatama, serta keluarga besar yang senantiasa memberikan dukungan, doa, dan semangat sehingga penulis dapat menyelesaikan kuliah dari awal masuk hingga laporan tugas akhir yang tersusun lancar ini.

5. Para calon istri idaman (Havidzah, Yolla, Nadya, Rahma, Widya, Azizah), Alfi Indah, dan Makamdowo geng unit 138 (Iqbal, Saif, Ali, Nico, Nisa, Risma, Bella) terima kasih karena selalu ada untuk mendukung, menemani, dan mendengarkan segala keluh kesah penulis selama ini.
6. Teman-teman statistika UII angkatan 2016, terima kasih untuk hari-hari indah dan pelajaran berharga bersama kalian.
7. Teman-teman SMA ku, Bella, Ina, Yohana, Tiara, Eden terima kasih untuk semangat dan dukungan kalian.
8. Sabahat-Sahabat UII, terimakasih atas kebersamaannya selama ini.
9. Semua pihak yang telah mendukung dan ikut membantu penulis tidak bisa disebutkan satu-satu, terima kasih.

Semoga segala bantuan, bimbingan dan pengajaran yang telah diberikan kepada penulis mendapatkan imbalan dari Allah SWT. Penulis memohon maaf apabila selama dalam proses penyusunan tugas akhir ini terdapat kekhilafan dan kesalahan. Penulis menyadari sepenuhnya akan keterbatasan kemampuan dalam penulisan tugas akhir ini, oleh karena itu penulis mengharapkan adanya kritik dan saran yang membangun demi kesempurnaan penyusunan dan penulisan tugas akhir ini. Semoga tugas akhir ini dapat bermanfaat bagi semua yang membaca dan membutuhkan.

Wassalamualaikum Wr.Wb

Yogyakarta, 17 Maret 2020

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan orang lain, kecuali yang diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 17 Maret 2020

METERAI TEMPEL
5A5C9AHF401468809
6000
ENAM RIBU RUPIAH

Penulis

ABSTRAK

IMPLEMENTASI METODE *IMPROVED K-MEANS* DENGAN ALGORITMA *DBSCAN* UNTUK PENGELOMPOKAN FILM

Annisa Ayunda Permata Sari

Program Studi Statistika, Fakultas MIPA

Universitas Islam Indonesia

Industri perfilman Indonesia terus mengalami peningkatan dilihat dari banyaknya film-film yang muncul di bioskop saat ini dengan peningkatan *box office* sebesar 28 persen setiap tahun nya dalam kurun waktu empat tahun terakhir. Salah satu media yang digunakan untuk mendistribusikan film adalah internet. Informasi-informasi mengenai film seperti tema, genre, aktor, rating, sutradara, dll dapat ditemukan dengan mudah melalui internet. Beberapa sumber informasi untuk film adalah *IMDb, Netflix, TMDb,* dan *Rotten tomatoes. Internet Movie Database* (IMDb) adalah situs web yang menyediakan informasi mengenai film dari seluruh dunia, termasuk orang-orang yang terlibat di dalamnya mulai dari aktor/aktris, sutradara, penulis sampai penata rias dan soundtrack. IMDb merupakan sumber informasi paling populer dan terpercaya baik untuk film, TV, dan konten selebritas lain. Dalam hal ini peneliti ingin melakukan penelitian mengenai karakteristik film dan faktor yang membuat sebuah film dapat masuk dalam IMDb Top 250. Data yang digunakan pada penelitian ini menggunakan data hasil *scraping* dari website. Metode yang digunakan adalah metode pengelompokan *cluster non-hierarki*, yaitu *kmeans* dan *Dbscan*. Dimana algoritma *Dbscan* digunakan untuk menentukan jumlah *cluster* optimum kemudian dilanjutkan dengan mengelompokkan data berdasarkan *centroid* dengan algoritma *k-means*. Dari hasil analisis diperoleh bahwa faktor yang dapat memengaruhi suatu film masuk dalam IMDB Top 250 adalah durasi, jumlah votes, dan film yang disutradarai oleh Rajkumar Hirani dan untuk jumlah cluster optimal menggunakan algoritma *Dbscan* diperoleh sebanyak enam *cluster*, diperoleh 3 film untuk *cluster 1, 5* film untuk *cluster 2*, 9 film untuk *cluster 3*, dan berturut-turut pada *cluster* 4, 5, dan 6 adalah sebanyak 7, 223, dan 3 film. Dengan algoritma *improved k-means* didapatkan nilai akurasi untuk hasil *cluster* sebesar 87.2%.

Kata Kunci: Film, IMDb, K-Means, Dbscan

ABSTRACT

IMPLEMENTATION OF IMPROVED K-MEANS METHOD WITH DBSCAN ALGORITHM FOR FILM GROUPING

Annisa Ayunda Permata Sari

Program Studi Statistika, Fakultas MIPA

Universitas Islam Indonesia

The Indonesian film industry continues to experience an increase seen from the number of films that appear in theaters today with a box office increase of 28 percent each year in the past four years. One of the media used to distribute films is the internet. Information about films such as themes, genres, actors, ratings, directors, etc. can be found easily through the internet. Some sources of information for films are IMDb, Netflix, TMDb, and Rotten tomatoes. Internet Movie Database (IMDb) is a website that provides information about films from around the world, including the people involved in it from actors / actresses, directors, writers to makeup artists and soundtracks. IMDb is the most popular and trusted source of information for movies, TV and other celebrity content. In this case the researcher wants to conduct research on the characteristics of the film and the factors that make a film to be included in the IMDb Top 250. The data used in this study uses scraped data from the website. The method used is a non-hierarchical clustering method, namely kmeans and Dbscan. Where the Dbscan algorithm is used to determine the optimum number of clusters then proceed by grouping data based on centroids with the k-means algorithm. From the analysis it was found that the factors that could influence a film included in the IMDB Top 250 were duration, number of votes, and films directed by Rajkumar Hirani and for the optimal number of clusters using the Dbscan algorithm obtained as many as six clusters, obtained 3 films for clusters 1, 5 films for cluster 2, 9 films for cluster 3, and respectively in clusters 4, 5, and 6 are 7, 223, and 3 films. With the improved k-means algorithm, the accuracy value for the cluster results is 87.2%.

Keywords: *Movie, IMDb, K-Means, Dbscan*

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Dunia perfilman memiliki kisah perjalanan yang cukup panjang, mulai dari film bisu dan 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 bagus, berbeda dengan film Indonesia yang bermain aman dengan hanya memproduksi film dengan genre tertentu yang akan banyak ditonton masyarakat. Dan juga modal serta dukungan pemerintah terhadap industri ini masih kurang (Wandira, 2018).

Meskipun begitu, Industri perfilman Indonesia terus mengalami peningkatan dilihat dari banyaknya film-film yang muncul di bioskop saat ini dengan peningkatan *box office* sebesar 28 persen setiap tahun nya dalam kurun waktu empat tahun terakhir. Pada tahun 2017, Indonesia menempati posisi ke-16 pasar film terbesar di dunia. Konvensi film tahunan terbesar di Asia, *CineAsia*, menilai bahwa Indonesia merupakan pasar film paling potensial di kawasan Asia Pasifik. Pada 2017, data yang ada mencatat bahwa Asia Pasifik memberikan sumbangan *box office* sebanyak 16 miliar dolar AS atau meningkat 44 persen dalam kurun waktu lima tahun. Dimana Indonesia menjadi negara Asia Pasifik yang memiliki perkembangan yang paling signifikan sehingga membuat *CineAsia 2018* menyebut Indonesia sebagai *The Rise of the Sleeping Giant*. Berbagai genre film kini hadir baik dilayar kaca ataupun di bioskop yang akhirnya memberikan warna lain dalam industri ini. Beberapa film Indonesia sukses memenangkan penghargaan di ajang

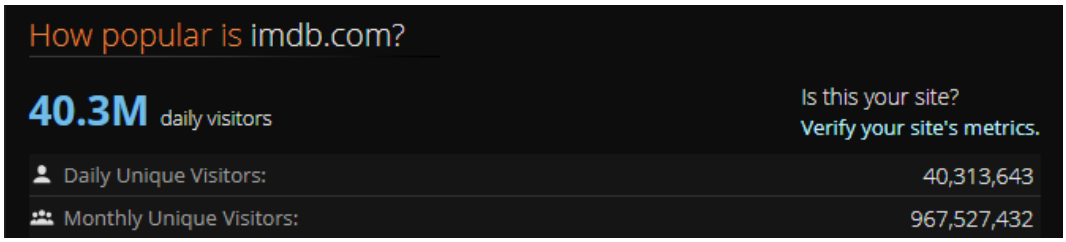
film Internasional, seperti Wiro Sableng dan Marlina si Pembunuh dalam Empat Babak (Portal Informasi, 2019).

Hal tersebut membuktikan bahwa pasar film Indonesia mempunyai potensi yang layak untuk dapat bersaing dengan film luar lainnya. Sehingga perlu diamati bagaimana perkembangan film-film dari seluruh penjuru dunia yang populer atau banyak disukai saat ini. Dari awal lahirnya film hingga sekarang, film Hollywood telah merajai industri perfilman secara global. Meskipun saat ini film-film Asia mulai dapat bersaing dengan film Hollywood.

Salah satu media yang digunakan untuk mendistribusikan film adalah internet. Dimana internet sendiri berperan sebagai media komunikasi antara pembuat film dengan penikmat film. Informasi-informasi mengenai film seperti tema, genre, aktor, rating, sutradara, dll dapat ditemukan dengan mudah melalui internet. Beberapa sumber informasi untuk film adalah *IMDb, Netflix, TMDb,* dan *Rotten tomatoes. Internet Movie Database* (IMDb) adalah situs web yang menyediakan informasi mengenai film dari seluruh dunia, termasuk orang-orang yang terlibat di dalamnya mulai dari aktor/aktris, sutradara, penulis sampai penata rias dan soundtrack. IMDb merupakan sumber informasi paling populer dan terpercaya baik untuk film, TV, dan konten selebritas lain (HypeStat, 2020). Berbagai informasi mengenai film, seperti: judul, *review* atau penilaian, jadwal tayang, daftar pemain, sutradara, penulis, pendapatan film, durasi, dll tersedia dalam situs ini. Berkaitan dengan hal tersebut, akan diambil beberapa informasi pokok yang akan dijadikan variabel untuk diteliti seperti halnya judul film yang menjadi objek utama, tahun *release*, durasi film, deskripsi atau sinopsis film, *votes*, *rating* atau hasil penilaian dari penonton yang berupa skor, *genre* atau aliran film, *actor* yang merupakan pemain utama dalam film, dan *director* atau sutradara untuk film yang berada di Top 250 versi IMDb.

Hal tersebut merupakan preferensi penonton terhadap film berdasarkan atribut filmnya. Pemasaran seperti sinopsis atau *trailer* film, sebelum menonton penonton cenderung mencari informasi terhadap film yang akan disaksikan. Setelah memilih genre dan mendapatkan informasi dari pemasaran, penonton akan

mengutamakan film yang memiliki sekuel dibandingkan dengan film baru. Lalu penonton akan menentukan film berdasarkan sutradaranya, karena setiap sutradara memiliki kredibilitasnya dalam membuat film. Adapun penonton yang menentukan film berdasarkan pemain kesukaannya. Saat ini latar film sudah mulai diperhatikan oleh penonton, namun atribut ini cenderung bisa diketahui melalui judul film atau pemasaran.



Gambar 1.1 Jumlah Pengunjung IMDb

Sumber: https://hypestat.com/info/IMDb.com#info

Berdasarkan Gambar 1.1 menunjukkan bahwa jumlah pengunjung harian IMDb mencapai 40.3 juta pengunjung dan jumlah pengunjung bulanan mencapai 967.5 juta. Pengunjung IMDb sebagian besar bukan merupakan pengguna *unique*, yang artinya pengguna hanya sekedar mengunjungi situs untuk mencari beberapa informasi tentang film dan tidak menjadi member. Dan berdasarkan pada Tabel 1.1 menunjukkan bahwa pengguna IMDb terbanyak berasal dari Amerika diikuti oleh India, Jepang, dan Korea. Sedangkan banyak pengguna IMDb yang berasal dari Indonesia sebesar 1% dari seluruh dunia, yang mana membuat Indonesia berada di peringkat ke-24.

Tabel 1.1 Pengguna IMDB berdasarkan Negara

No	Negara	Pengguna (%)
1.	Amerika Serikat	28
2.	India	8
3.	Jepang	6.1
4.	Korea	3.4
⋮	⋮	⋮
24.	Indonesia	1

Sebagian besar film yang berperingkat tinggi dalam IMDB telah diakui secara kritis dan sudah terjamin menjadi sesuatu hal yang aman dalam hal keberhasilan komersial. Dari pemikiran tersebut, peneliti ingin menyelidiki lebih lanjut apakah film-film yang masuk dalam list "Top" ini mempunyai beberapa fitur menarik atau fitur unik yang menjadikan film tersebut memiliki peringkat tinggi sehingga masuk dalam list IMDb Top 250. Penelitian ini bertujuan untuk mengetahui jenis *cluster* alami yang terbentuk dalam 250 film teratas dari IMDb. Penggunaan metode *cluster* diharapkan dapat menentukan kelompok yang melekat dalam data dan memberi informasi untuk mengamati pola yang berulang. Metode *cluster* non-hierarki digunakan karena jumlah objek cukup besar dan agar objek dalam satu kelompok lebih mirip satu sama lain dibandingkan dengan objek pada kelompok lain. Sehingga untuk dapat mengelompokkan film berdasarkan kesamaan karakteristiknya digunakan metode *Dbscan* dan *K-means.* Algoritma *Dbscan* digunakan untuk menentukan jumlah *cluster* optimum kemudian dilanjutkan dengan algoritma *K-means* untuk mengelompokkan data berdasarkan titik pusat (*centroid*).

Dalam hal ini peneliti ingin melakukan penelitian mengenai karakteristik film dan faktor yang membuat sebuah film dapat masuk dalam IMDb Top 250, sehingga oleh penikmat film dapat digunakan sebagai referensi untuk memilih film sesuai dengan karakteristik yang diinginkan dan oleh para pembuat film dapat menjadi bahan masukan mengenai bagaimana karakter film yang banyak disukai oleh masyarakat dunia. IMDb merupakan salah satu website yang menjadi pilihan masyarakat global dalam mencari informasi lebih lanjut mengenai dunia film atau layar kaca. Berdasarkan beberapa sumber penelitian terdahulu, ada beberapa metode *cluster* yang digunakan, seperti: *agglomerative, birch, k-means, mean shift*, dan *Dbscan*. Sedangkan pada penelitian ini metode yang digunakan adalah metode pengelompokan *cluster non-hierarki,* yaitu *Dbscan* untuk menentukan jumlah *cluster* optimum dan *K-means* untuk mengelompokkan. Data yang digunakan pada penelitian ini menggunakan data hasil *scraping* dari website.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disebut diatas, maka diperoleh rumusan masalah untuk penelitian ini adalah sebagai berikut:

1. Faktor apa yang paling berpengaruh yang menjadikan suatu film dapat masuk dalam IMDb Top 250?
2. Bagaimanakah hasil pengelompokan film dalam IMDb Top 250?

1.3 Batasan Masalah

Dari latar belakang tersebut ditetapkan batasan bahwa penelitian ini dibatasi dengan sumber data yang diperoleh dari *website* IMDb dengan teknik *web scraping* menggunakan program *RStudio 1.1.447*. Analisis penelitian menggunakan metode *Principal Component Analysis* (PCA) dilanjutkan dengan metode peng*cluster*-an *Dbscan* dan *K-means* dengan bantuan program *Phyton.* Dimana variabel yang dianalisis, yaitu: judul film, tahun *release*, durasi, deskripsi, *votes, rating, genre, actor*, dan *director* film dari list IMDb Top 250.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk menjawab rumusan masalah, yaitu:

1. Mengetahui faktor yang paling berpengaruh yang mempengaruhi suatu film masuk dalam IMDb Top 250.
2. Mengetahui hasil pengelompokan film dalam IMDb Top 250.

1.5 Manfaat Penelitian

Dari penelitian ini, diperoleh manfaat sebagai berikut:

1. Sebagai evaluasi untuk industri perfilman Indonesia ke depan nya agar dapat bersaing dalam dunia perfilman global.
2. Mampu memberi pemahaman secara rinci atribut atau faktor apa saja yang dipertimbangkan oleh konsumen dalam menonton film.

BAB II
TINJAUAN PUSTAKA

Tinjauan pustaka sangat bermanfaat untuk dijadikan referensi dan pembanding dalam penelitian ini sehingga dalam bab ini akan memberikan beberapa penelitian terdahulu yang berkaitan dengan penelitian film yang dilakukan oleh sejumlah peneliti di berbagai negara, antara lain sebagai berikut:

Tabel 2.1 Tabel Penelitian Sebelumnya

No	Nama Peneliti (Tahun)	Judul	Variabel	Metode	Hasil Penelitian
1.	Bulut dan Korukoglu (2011)	*Analysis and Clustering of Movie Genres*	Judul film, kata kunci, genre.	Analisis *Clustering* Hierarki	Dengan mengklasifikasikan genre menjadi lima *cluster* dan menetapkan pasangan dua atau tiga genre yang paling mendekati dan membandingkan hasil yang didapat dengan metode *cluster* hierarki dan PCA. Hasil dari kedua analisis tersebut dekat dengan masing-masing lainnya: klasifikasi 24 dari 27 genre (88.9%) cocok satu sama lain.

2.	Aditya, Rajaraman, Subashini (2018)	*Comparative Analysis of Clustering Techniques for Movie Recommendation*	Judul film, id film, tahun *release*, cast total facebook *likes*, *content rating*.	*Clustering: agglomerative, birch, k-means*, dan *mean shift*.	Beberapa film (titik data menyimpang) yang sama sekali berbeda dari kelompok utama ditangani secara berbeda oleh algoritma yang berbeda dan dianalisis secara singkat. Kombinasi fitur memberikan output yang berbeda dan dapat memperoleh film yang sama tergantung pada atribut yang dapat dipilih pengguna.
3.	Budiman, Safitri, dan Ispriyanti (2016)	Perbandingan Metode *K-means* dan Metode *Dbscan* pada Pengelompokan Rumah Kos Mahasiswa di Kelurah	Rumah kost, harga, fasilitas.	Analisis kelompok dengan *K-means* dan *Dbscan*	Berdasarkan penelitian yang telah dilakukan diketahui bahwa metode *K-Means* bekerja lebih baik daripada *Dbscsan* untuk mengklasifikasikan rumah kost sebagaimana dibuktikan oleh nilai indeks *silhouette* pada *k-means* sebesar 0.463 lebih tinggi dari nilai indeks *silhouette* di

		an Tembalang Semarang			*Dbscan* yang sama dengan 0.281.
4.	Ajit (2018)	*Clustering of the Top 250 movies from IMDB*	*Title, Year, Rated, Runtime, Genre, Director, Writer, Actors, Plot, Language, Country*	*Dbscan clustering* dan *k-means clustering*	*Dbscan* memberikan jumlah *cluster* optimal dalam dataset dan kemudian *K-Means* diterapkan untuk mendapatkan *cluster* yang benar-benar dapat ditafsirkan. Peneliti hanya menunjukkan 5 dari total 13 cluster. Hasil analisis dapat digunakan dan direferensikan silang dengan dataset film yang besar untuk melihat *cluster* mana yang lebih baik yang dapat diperoleh.
5.	Guan, Yuen, Chen (2017)	*Towards A Hybrid Approach of K-means and*	*Image segmentation*	*Kmeans-Dbscan*	Karena kompleksitas komputasi yang tinggi dari *Dbscan* dan ukuran gambar *dataset* yang besar, *k-means* diterapkan untuk mengurangi ukuran

		Density-Based Spatial Clustering of Applications with Noise for Image Segmentation			gambar *dataset* dalam pendekatan tertentu. Empat gambar terpilih dari *brenchmarking datasets* digunakan untuk mengevaluasi kegunaan dari metode yang ditetapkan. Hasil dari metode yang ditetapkan lebih masuk akal dibandingkan hasil segmentasi *Dbscan* maupun *kmeans.* Penelitian kedepannya akan memperluas struktur metode yang diusulkan dengan lebih banyak eksperimen.

Terdapat penelitian sebelumnya oleh (Bulut & Korukoglu, 2011) yang berkaitan dengan clustering film IMDB yang mengelompokkan film dengan bahasa inggris antara tahun 2006 dan 2010 dengan 27 genre, 48483 judul, dan 19561 kata kunci. Metode yang digunakan adalah cluster hierarki *complete linkage* dan *Principal Component Factor Analysis* (PCFA). Dalam penelitian ini, peneliti menggunakan lebih banyak variabel dari penelitian sebelumnya, dimana variabel yang digunakan merupakan informasi-informasi yang muncul dalam *website* IMDb.

Terkait dengan penelitian terdahulu, penulis melakukan penelitian yang berkaitan dengan film dimana IMDb merupakan salah satu *website* populer yang menyediakan informasi rinci mengenai film atau acara televisi. Dalam *website*

tersebut, Top 250 merupakan halaman yang cukup menarik bagi peneliti untuk ditelusuri lebih dalam.

Berdasarkan pemikiran tersebut, peneliti menggunakan variabel yang muncul dalam halaman website IMDb, seperti: judul film, tahun *release*, durasi, *votes*, genre, *rating*, *description*, *director*, dan *actor*. Peneliti menggunakan beberapa metode dalam penelitian diantaranya PCA untuk mereduksi dimensi, *cluster Dbscan* untuk menentukan jumlah *cluster* optimal dan *K-means* untuk mengelompokkan.

BAB III
LANDASAN TEORI

3.1 Film

3.1.1 Pengertian

Berdasarkan Kamus Besar Bahasa Indonesia, film mempunyai dua artian. Pertama, film disebut sebagai gambar (cerita) yang hidup. Kedua, film diartikan sebagai selaput tipis yang dibuat dari seluloid untuk tempat gambar negatif dan positif, dimana gambar negatif untuk gambar yang akan dibuat potret dan positif untuk gambar yang akan dimainkan dibioskop.

Dalam industri, film merupakan bagian dari produksi ekonomi suatu masyarakat dan dalam hubungan dengan produk-produk lainnya harus dipandang. Dalam komunikasi, film adalah bagian penting dari sistem yang digunakan baik oleh para individu maupun kelompok untuk saling bertukar (mengirim dan menerima) pesan (Ibrahim, 2011). Film telah menjadi media komunikasi audio visual yang akrab dinikmati oleh masyarakat dari berbagai usia dan latar belakang sosial. Kekuatan dan kemampuan film dalam menjangkau banyak segmen sosial tersebut lantas membuat para ahli berpikir bahwa film memiliki potensi untuk mempengaruhi khalayaknya (Sobur, 2004). Film mampu memengaruhi bahkan mengubah dan membentuk karakter penonton melalui pesan yang terkandung di dalamnya.

Pada hakikatnya, semua film adalah dokumen sosial dan budaya yang membantu mengkomunikasikan zaman ketika film itu dibuat bahkan sekalipun ia tak pernah dimaksudkan untuk itu (Ibrahim, 2011).

3.1.2 Sejarah dan Perkembangan Film

Film pertama kali diperlihatkan pada khalayak umum dengan membayar berlangsung di *Grand Cafe Boulevard de Capucines*, Paris, Perancis pada 28 Desember 1895. Peristiwa ini sekaligus menandai lahirnya film dan bioskop di dunia. Meskipun usaha untuk membuat "citra bergerak"

atau film ini sendiri sudah dimulai jauh sebelum tahun 1895, bahkan sejak 130 Masehi (Sumarno, 1996).

Perjalanan film terus mengalami perkembangan besar bersamaan dengan perkembangan kemajuan teknologi pendukungnya. Diawali dengan film hitam putih dan tanpa suara atau biasa disebut “film bisu” yang berakhir masanya pada tahun 1920-an karena mulai ditemukan film bersuara. Pada tahun 1927 diproduksi film bersuara pertama dengan judul “Jazz Singer” dan diikuti dengan ditemukannya film berwarna pada tahun 1930-an. Perubahan dalam industri perfilman terlihat pada teknologi yang digunakan, yang awalnya hanya berupa gambar hitam putih hingga berkembang sampai saat ini sesuai dengan sistem penglihatan manusia, berwarna dengan segala efek film yang membuatnya terlihat lebih nyata. Seiring berjalannya waktu, film tidak hanya dapat dinikmati di bioskop atau televisi saja, melainkan dapat dinikmati di rumah dengan menggunakan VCD atau DVD (*Blue-Ray*).

Selain itu di beberapa negara Eropa, film digunakan sebagai media penyampai produk kebudayaan. Sehingga berdampak pada film yang dipandang sebagai artefak budaya yang perlu dikembangkan, kajian film yang skala nya membesar serta eksperimen-eksperimen yang didukung oleh negara. Dan di sisi lain, film juga digunakan sebagai aset politik untuk media propaganda negara. Hal tersebut menjadi sebab mengapa di Indonesia film berada dibawah pengawasan departemen penerangan dengan konsep lembaga sensor film. Di Amerika sendiri, film-film menjadi ladang ekspor yang memberikan keuntungan cukup besar, meskipun diproduksi dengan latar belakang budaya mereka (Dolfi, 2017).

3.1.3 Klasifikasi Film

Dalam konteks film, genre didefinisikan sebagai jenis atau klasifikasi dari sekelompok film yang memiliki pola ataupun karakter yang sama, seperti *setting*, isi dan subjek cerita, tema, struktur cerita, aksi atau peristiwa, periode, gaya, situasi, ikon, *mood*, serta karakter. Dari klasifikasi

tersebut menghasilkan beberapa genre populer seperti *action, adventure, drama, comedy, horror,* dan lain lain. Fungsi genre sendiri adalah untuk memudahkan klasifikasi sebuah film sesuai dengan spesifikasinya (Himawan, 2008, p. 10). Suatu film memiliki genre yang dominan, meskipun kebanyakan film merupakan gabungan dari beberapa genre sekaligus. Gabungan genre dalam sebuah film dinamakan genre hibrida (campuran). Berikut merupakan beberapa genre umum (Himawan, 2008, pp. 11-12):

1. *Drama*

   Pada umumnya berhubungan dengan cinta, setting cerita, karakter dan suasana yang menggambarkan kehidupan nyata. Karena jangkauan ceritanya yang luas, genre ini menjadi genre yang paling banyak diproduksi.

2. *Action*

   Berkaitan dengan adegan-adegan yang mengandung aksi fisik yang seru, menegangkan, berbahaya, nonstop dengan tempo yang cepat. Merupakan genre yang paling adaptif dengan genre lainnya.

3. *Crime*

   Berhubungan dengan aksi-aksi kriminal atau kejahatan seperti perampokan, pencurian, perjudian, pembunuhan, persaingan antar kelompok, serta aksi kelompok bawah tanah yang bekerja diluar sistem hukum.

4. *Comedy*

   Film yang mengundang tawa bagi penonton, yang biasanya berupa drama ringan yang melebih-lebihkan aksi, situaso, bahasa, hingga karakternya.

5. *Adventure*

   Mengangkat kisah tentang perjalanan, eksplorasi, atau ekspedisi ke suatu wilayah asing yang belum pernah tersentuh. Ciri khas dari film ini adalah menampilkan panorama alam yang eksotis seperti hutan rimba, savana, pengunungan, atau pulau terpencil.

6. *Biography*

   Genre biografi merupakan pengembangan dari genre drama dan sejarah yang menceritakan pengalaman kisah nyata ataupun kisah hidup seorang tokoh yang berpengaruh dimasa lalu maupun masa kini.

7. *Western*

   Merupakan genre asli milik Amerika, yang umumnya mengangkat konflik antara pihak baik dan jahat.

8. *Horror*

   Biasa menggambarkan karakter antagonis non-manusia dengan fisik yang menyeramkan. Film dengan genre ini mempunyai tujuan membuat penonton merasakan efek takut, kejutan, serta teror.

9. *Film-Noir*

   Noir sendiri mempunyai arti ”gelap” atau ”suram” yang merupakan turunan dari genre kriminal dan gangster yang populer pada tahun 1940an-1950-an. Tema dalam film genre ini selalu berhubungan dengan tindak kriminal seperti pembunuhan, pecurian serta pemerasan (Himawan, 2008).

10. *Animation*

    Dalam pembuatan atau pengolahannya, film ini menggunakan bantuan grafik komputer yang dapat menghasilkan efek 2D dan 3D (Redaksi, 2013).

11. *Mystery*

3.2 *Web Scraping*

Web Scraping adalah pengumpulan data secara otomatis dari internet yang hampir setua internet itu sendiri. Meskipun *web scraping* bukan suatu hal yang baru, dalam beberapa tahun ini praktiknya lebih diketahui sebagai *screen scraping, data mining, web harvesting,* dan sejenisnya. Secara teori, *web scraping* merupakan praktik mengumpulkan data melalui cara apapun selain dengan program yang berinteraksi dengan API (atau dengan kata lain melalui manusia dengan menggunakan *web browser*). Hal ini paling umum dilakukan dengan menuliskan program otomatis yang menanyakan server untuk meminta data (biasanya dalam

format HTML dan file lain meliputi halaman web), kemudian mengurai data untuk diekstrak dalam format yang dibutuhkan (Mitchell, 2015). *Web scraping* digunakan dalam berbagai bidang untuk mengumpulkan data yang tidak dengan mudah tersedia dalam sebuah format (Jarmul & Lawson, 2017).

3.3 *Internet Movie Database (IMDb)*

IMDb merupakan sumber terpercaya dan paling terkenal didunia untuk film, TV, dan informasi mengenai selebriti, yang diciptakan untuk membantu pengunjung menjelajahi dunia film dan pertunjukan untuk menentukan apa yang harus di tonton. Database *IMDb* dapat menelusuri jutaan film, program TV, dan program hiburan serta para pemain dan kru. *IMDb* membantu pengunjung menyalurkan ingatan tentang film, pertunjukan, membantu memberi rekomendasi film untuk ditonton selanjutnya dan saling berbagi pengetahuan dan pendapat dengan komunitas penggemar terbesar di dunia. *IMDb* menciptakan sebuah konten video original baru setiap minggunya, termasuk: *The IMDb Show*, IMDbrief, dan *Casting Calls* diantara wawancara lain dan video aktul harian. *IMDb* pertama diluncurkan secara online pada 1990 dan sudah menjadi anak perusahaan dari *Amazon.com* sejak tahun 1998 (What is IMDb?, 2019).

3.4 *Text Mining*

Text mining adalah suatu proses analisis data berupa teks dimana sumber datanya didapatkan dari dokumen. Konsep *text mining* biasa digunakan untuk mengklasifikasikan dokumen-dokumen sesuai dengan topiknya. Sehingga dapat diketahui jenis kategorinya melalui kata-kata yang terdapat dalam dokumen tersebut. Tahapan dalam melakukan analisis pada *text mining* adalah mengumpulkan data kemudian melakukan ekstraksi terhadap fitur yang digunakan. Ekstraksi fitur dilakukan dengan melakukan pembersihan data mulai dari *tokenizing, stop words removal,* dan *stemming*. Kemudian data ditransformasi dengan pembobotan *term* yang telah dibersihkan. Dilanjutkan dengan mereduksi data dan terakhir melakukan analisis data terhadap proses klasifikasi untuk merepresentasikan hasil informasi yang ditemukan.

Preprocessing merupakan tahap awal untuk mempersiapkan teks menjadi data *numeric* yang akan diolah lebih lanjut (Feldman, Ronen, Sanger, & dkk, 2007). Terdapat beberapa tahap dalam proses ini diantaranya (Rizki, Dhidik, & Supraptono, 2017):

3.4.1 *Tokenizing*

Tokenizing merupakan proses memecah dokumen menjadi kumpulan kata. Dilakukan dengan menghilangkan karakter tertentu atau tanda baca, memisahkan setiap spasinya, dan mengubah setiap token menjadi bentuk huruf kecil (*lower case*).

3.4.2 *Stopwords Removal*

Stopwords removal adalah proses penghilangkan kata yang tidak penting pada deskripsi melalui pengecekan kata-kata apakah termasuk di dalam daftar kata tidak penting (*stoplist*) atau tidak.

3.4.3 *Stemming*

Stemming adalah sebuah teknik untuk menghilangkan tanda hubung dari sebuah kata yang diakhiri dengan asal kata nya (Hardeniya, Perkins, & Deepti, 2016).

3.5 Algoritma *Term Frequency Inverse Document Frequency* (TF-IDF)

Metode TF-IDF merupakan metode yang digunakan untuk menghitung bobot setiap kata yang paling umum, sehingga metode ini dikenal efisien, mudah, dan akurat (Ma'arif, 2015). Metode TF-IDF adalah sebuah cara pemberian bobot hubungan suatu kata *(term)* terhadap dokumen. Dimana TF-IDF ini merupakan ukuran statistik yang digunakan untuk mengevaluasi seberapa penting peran kata tersebut dalam sebuah dokumen atau kelompok kata, dilihat dari frekuensi kemunculan kata dalam dokumen tersebut. Jika frekuensi muncul kata dalam suatu dokumen, maka bobot kata nya semakin besar dan semakin kecil jika muncul dalam banyak dokumen (Putra, 2016). Menurut Fitri (2013) yang dikutip dari Herwijayanti, dkk. (2018) metode ini menggabungkan dua buah konsep yakni frekuensi kemunculan suatu kata dalam dokumen dan inverse frekuensi dokumen yang mengandung kata tersebut.

Digunakan rumus untuk menghitung bobot (W) masing-masing dokumen terhadap kata kunci pada algoritma TF-IDF:

$$W_{dt} = tfdt * log\left(\frac{N}{df}\right) \tag{3.1}$$

Dimana:

W_{dt} = bobot dokumen ke-d terhadap kata ke-t

$tfdt$ = banyaknya kata yang dicari pada sebuah dokumen

N = total dokumen

df = banyak dokumen yang mengandung kata yang dicari.

3.6 Analisis Faktor

Analisis faktor merupakan suatu metode yang digunakan untuk mereduksi atau meringkas data, dari variabel yang banyak menjadi variabel yang lebih sedikit tanpa menghilangkan informasi yang terkandung dalam variabel asli (Supranto, 2004). Terdapat tiga fungsi umum dari analisis faktor (Dillon & Goldsten, 1984).:

1. Dapat mereduksi jumlah variabel untuk penelitian lebih lanjut dengan mempertahankan sebanyak mungkin informasi asli yang mungkin. Kumpulan variabel asli dapat diubah menjadi lebih sedikit jumlahnya yang menyumbang sebagian besar variansi dalam data.
2. Dalam situasi dimana jumlah data yang tersedia sangat besar diluar batasan pemahaman, analisis faktor dapat membedakan antara data kualitatif dan kuantitatif.
3. Dapat digunakan untuk menguji hipotesis mengenai perbedaan kualitatif dan kuantitatif pada data. Jika peneliti mempunyai hipotesis awal mengenai jumlah dimensi atau karakter dimensi, hipotesis ini dalam kondisi tertentu dapat menjadi pengujian statistik.

3.7 *Principal Component Analysis* (PCA)

PCA merupakan teknik dalam analisis multivariat yang tertua dan paling banyak diketahui yang pertama kali diperkenalkan oleh *Pearson* (1901) dan dikembangkan secara umum oleh *Hotelling* (1933). Tujuan utama dari PCA adalah untuk mereduksi dimensi dari kumpulan data *(data set)* yang mengandung banyak

variabel yang saling berhubungan, dengan mempertahankan sebanyak mungkin variasi dalam *data set*. Hasil reduksi diperoleh dengan mentransformasikan kumpulan variabel baru, komponen utama, yang tidak saling berhubungan, dan diurutkan sehingga beberapa yang pertama mempertahankan variasi paling besar yang ada di semua variabel asli. Perhitungan komponen utama yang tereduksi menjadi solusi dari permasalahan *eigenvalue-eigenvector* untuk matrik positif-*semidefinite*. Syarat variabel yang akan dianalisis dengan PCA adalah harus distandarisasi terlebih dahulu untuk mengatasi jika ada nilai skala yang berbeda jauh (Supranto, 2004).

$$X = \frac{(X_i - \bar{X})}{S} \qquad ;i = 1,2,3, \dots, p \tag{3.2}$$

Dengan: X = variabel yang sudah distandarisasi

X_i = variabel ke-i

$\bar{X}$ = rata-rata variabel ke-i

S = standar deviasi variabel ke-i

Dengan menggunakan PCA, variabel yang sebelumnya sebanyak n variabel akan diseleksi menjadi k variabel baru yang disebut *principal component*, dengan jumlah k lebih kecil dari n. Dengan hanya menggunakan k *principal component* akan menghasilkan nilai yang sama dengan n variabel. Dimana variabel hasil dari seleksi disebut sebagai *principal component* (Kotu & Deshpande, 2015).

Menurut (Jolluffe, 2002), prosedur PCA bertujuan untuk menyederhanakan dan menghilangkan faktor atau indikator skrining yang kurang dominan dan kurang relevan tanpa mengurangi arti dan tujuan dari data asli dari variabel random x (matriks berukuran n x n), dimana baris-baris yang berisi observasi sebanyak n dari variabel acak x) adalah sebagai berikut:

1. Menghitung matrik varians dan kovarian dari data observasi.

   Varians (Var (x)) untuk menemukan penyebaran data dalam dataset untuk menentukan penyimpangan data dalam data sampel. Matriks kovarian (Cov (x, y)) adalah matriks yang nilai-nilai kovariansi pada tiap *cell*-nya didapatkan dari sampel. Misalkan x dan y adalah suatu variabel random.

$$Var\ (x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Z_{ij} - \mu_{ij}\right)^2 \quad (3.3)$$

$$Cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_{xj})(y_{ij} - \mu_{yj}) \quad (3.4)$$

dimana: μ_x, μ_y = rata-rata (*mean*) sampel dari variabel x dan y,

x_i, y_i = nilai observasi ke-i dari variabel x dan y.

Dari data nilai yang digunakan, maka diperoleh matrik kovarian berukuran n x n.

2. Mencari *eigenvalues* dan *eigenvector* dari varian-kovarian yang telah didapat. Nilai eigen yang dikomputasi kemudian ditransformasikan (*rotation orthogonal varimax*) menggunakan persamaan berikut (Johnson & Winchern, 2007):

$$Det\ (A - \lambda I) = 0 \quad (3.5)$$

dimana:

A = matriks n x n

λ = nilai *eigenvalue*

I = matriks identitas (matriks persegi dengan elemen diagonal utama bernilai 1 sedangkan elemen lain bernilai 0)

3. Menentukan nilai proporsi *principal component* (%) dengan persamaan:

$$PC\ (\%) = \frac{Nilai\ eigen}{VarianceCovarian}\ x\ 100\% \quad (3.6)$$

4. Menghitung bobot faktor (*factor loading*) berdasarkan *eigenvector* dengan persamaan (Johnson & Winchern, 2007):

$$Ax = \lambda x \quad (3.7)$$

Sehingga diperoleh kombinasi linear yaitu:

a. $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n$ adalah *eigenvalue* matrik A

b. $x_1, x_2, x_3, \ldots, x_n$ adalah *eigenvector* sesuai eigenvalue-nya (λ_n).

Persamaan *eigenvalue* & *eigenvector* merupakan *Eigen Value Decomposition* (EVD), dengan persamaan sebagai berikut:

$$AX = XD$$

$$A = X\ D\ X^{-1} \quad (3.8)$$

dimana:

A = matrik n x n yang memiliki n *eigenvalue* (λ_n)

D = *eigenvalue* dari *eigenvector*-nya

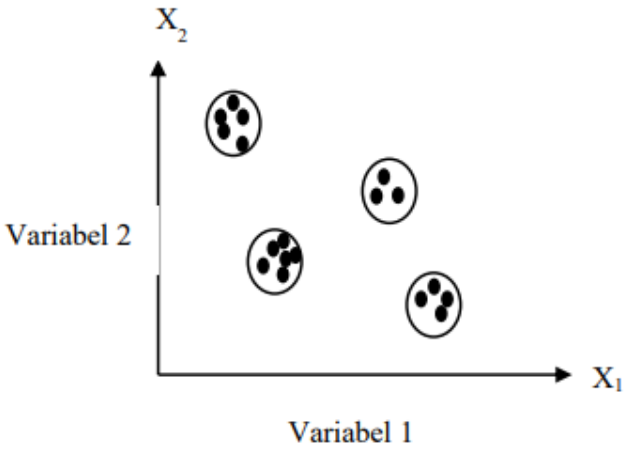
X = *eigenvector* dari matrik A

X^{-1} = invers dari *eigenvector* X

5. Setiap nilai akan mewakili *weight* (bobot) dan akan disimpan sebagai vektor.

3.8 Analisis Clustering

Menurut Supranto (2004) analisis *clustering* merupakan suatu kelas teknik yang dipergunakan untuk mengklasifikasikan objek kedalam kelompok yang relatif homogen. Dalam setiap kelompok, objek atau kasus nya cenderung mirip satu sama lain dan berbeda jauh dengan objek dari kelompok lainnya. Dimana setiap objek hanya dapat masuk kedalam satu kelompok saja, sehingga tidak terjadi tumpang tindih (*overlapping).*



Gambar 3.1 Pengelompokan Ideal

Gambar 3.1 menunjukkan hasil peng*cluster*an yang ideal, dimana setiap objek hanya masuk dalam satu kelompok saja (tidak mungkin menjadi anggota dua kelompok atau lebih) dan menunjukkan situasi dimana *cluster* dipisahkan secara berbeda pada dua variabel (Supranto, 2004).

Analisis *cluster* dibagi menjadi dua, yaitu metode hierarki dan non-hierarki. Metode hierarki digunakan untuk objek yang kecil dan jumlah kelompok yang akan dibentuk belum diketahui. Pengelompokan ini disajikan dalam bentuk dendogram atau diagram pohon (Usman & Sobari, 2013). Metode hierarki terbagi menjadi dua, yaitu *Agglomerative* (Penggabungan) dan *Divisive* (Pembagian). Beberapa metode

pengelompokan hierarki adalah *single linkage, complete linkage, average linkage,* dan metode *ward* (Prasetyo, 2012). Metode non-hierarki menghasilkan partisi dari data sehingga objek dalam satu *cluster* lebih mirip satu sama lain dibandingkan dengan objek dalam *cluster* lain (Triyanto, 2015). Prosedur peng*cluster*an dalam metode non-hierarki tidak dilakukan secara bertahap seperti pada metode hierarki dan jumlah *cluster* yang akan dibuat sudah ditentukan terlebih dahulu (Machfudhoh & Wahyuningsih, 2013). Contoh metode ini adalah *k-means, k-medoids, Dbscan,* dll (Supranto, 2004).

3.9 *Dbscan Clustering*

Dbscan adalah metode pengelompokan berdasarkan tingkat kepadatan data (*density-based*). Tidak seperti *k-means clustering*, *Dbscan* tidak perlu menentukan jumlah kelompok secara manual. Namun, diperlukan jumlah minimum tetangga untuk dipertimbangkan dalam kelompok dan jarak maksimum yang diperbolehkan antara titik manapun untuk menjadi bagian dari kelompok yang sama. Dalam pengguna tertentu, ditentukan jarak disekitar sampel, *Dbscan* akan menghitung jumlah tetangga. Ketika jumlah tetangga antar jarak melebihi ambang batas, *Dbscan* akan mengelompokan titik data sebagai satu kelompok (Misra, Li, & He, 2019). *Dbscan* merupakan algoritma *cluster* populer yang digunakan sebagai alternatif dari *k-means.* Dimana input jumlah *cluster* tidak perlu dilakukan untuk menjalankannya. Namun sebagai gantinya, perlu memasukkan dua parameter lain. Penerapan *scikit-learn* menyediakan *default* untuk parameter *eps* dan *min_samples,* namun umumnya perlu disetel oleh peneliti. Parameter *eps* merupakan jarak maksimum antara dua titik data untuk dapat dipertimbangkan dalam lingkungan yang sama. Parameter *min_samples* merupakan jumlah minimum dari titik data dalam lingkungan untuk dipertimbangkan menjadi *cluster*. *Dbscan* sangat sensitif terhadap skala karena *epsilon* merupakan nilai tetap untuk jarak maksimum antara titik (Maklin, 2019).

Salah satu keuntungan *Dbscan* dibanding *k-means* adalah *Dbscan* tidak terbatas pada jumlah *cluster* yang ditetapkan saat inisialisasi. Algoritmanya akan menentukan jumlah *cluster* berdasarkan kepadatan suatu daerah. Algoritma *Dbscan*

dibangun dalam konsep kebisingan (*noise*). Pada umumnya digunakan untuk mendeteksi *outliers* dalam data, seperti aktivitas kecurangan dalam kartu kredit, *e-commerce*, atau klaim asuransi (Bari, Chaouchi, & Jung, 2014).

Setiap titik x dalam dataset, dengan jumlah tetangga lebih dari atau sama dengan *min_samples*, ditandai sebagai titik pusat. Dikatakan x adalah titik batas, jika jumlah tetangganya kurang dari *min_samples*, tetapi milik lingkungan *eps* dari beberapa titik pusat z. Dengan demikian, jika suatu titik bukanlah titik pusat atau titik batas, maka titik tersebut merupakan *outlier* (Ester, Kriegel, Sander, & Xu, 1996).

Konsep kepadatan dalam *Dbscan* ini adalah banyaknya data (*minPts*) yang berada dalam radius *eps* (ε) dari setiap data. Konsep kepadatan seperti ini menghasilkan tiga macam status dari setiap data, yaitu inti (*core*), batas (*border*), dan *noise* (Prasetyo, 2012). Data inti merupakan data yang jumlah data di dalam radius *eps* lebih dari *minPts*, data noise merupakan data yang jumlah data di dalam radius *eps* kurang dari *minPts*, dan data batas merupakan data yang jumlah data di dalam radius *eps* kurang dari *minPts* tetapi menjadikan data tetangganya menjadi data inti. Proses pengelompokan *Dbscan* adalah menghitung jarak titik pusat (p) ke titik yang lain menggunakan jarak *Euclidean* dan dinyatakan dalam persamaan (3.9) (Eriyanto, 2007).

Konsep utama dari algoritma *Dbscan* adalah untuk menemukan daerah dengan kepadatan tinggi yang terpisah satu sama lain dengan daerah dengan kepadatan rendah. Jadi , cara mengukur kepadatan suatu wilayah adalah dengan dua langkah:

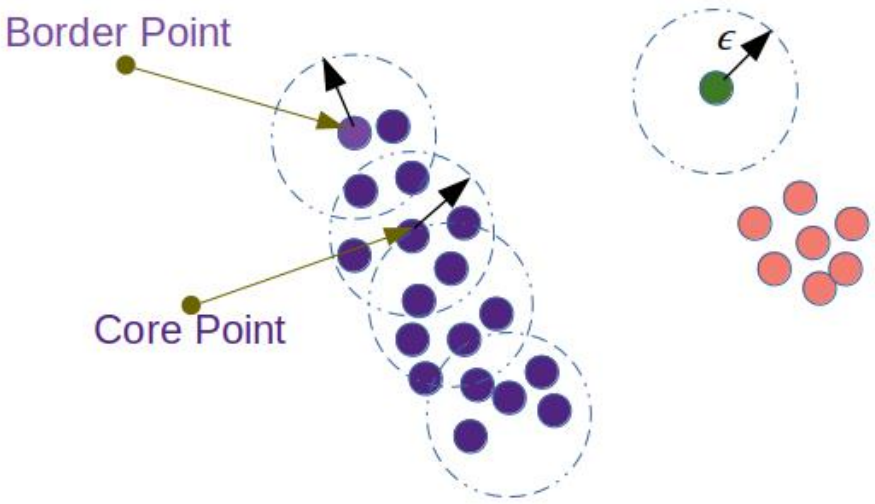
- Densitas pada titik p: jumlah titik dalam lingkaran radius *eps* (ε) dari titik p.
- Wilayah padat: untuk setiap titik di *cluster*, lingkaran dengan jari-jari (ϵ) berisi setidaknya jumlah titik minimum (*minPts*).

Lingkungan *epsilon* dari titik p kedalam *database* D didefinisikan sebagai:

$$N(p) = \{q \in D \mid dist\,(p,q) \leq \epsilon\} \tag{3.13}$$

Mengikuti definisi wilayah padat, suatu titik dapat diklasifikasikan sebagai Titik Inti jika $|\,N\,(p)\,| \geq MinPts$. *Core Points*, seperti namanya, biasanya terletak di dalam interior cluster. *Border Point* memiliki kurang dari *MinPts* dalam

lingkungan-lingkungannya (N), tetapi terletak di lingkungan titik inti lainnya. Noise adalah titik data apa pun yang bukan inti atau titik perbatasan. Lihat gambar di bawah untuk pemahaman yang lebih baik.



Gambar 3.2 Konsep Kepadatan

Sumber: https://towardsdatascience.com/dbscan-algorithm-complete-guide-and-application-with-python-scikit-learn-d690cbae4c5d

Mengacu pada Gambar 3.2, epsilon merupakan jari-jari yang diberikan untuk menguji jarak antara titik data. Jika suatu titik berada dalam jarak epsilon dari titik lain, kedua titik tersebut akan berada dalam kelompok yang sama. Selanjutnya, jumlah minimum titik yang dibutuhkan dalam skenario ini misal 4. Ketika melewati setiap titik data, selama *Dbscan* menemukan 4 titik dalam jarak epsilon satu sama lain, sebuah cluster terbentuk. Algoritma yang digunakan dalam *Dbscan* (Lutins, 2017):

1. Menentukan data sebagai nilai X.
2. Memasukan berbagai nilai parameter *eps* dan *min_samples* untuk memilih nilai *eps* dan *min_samples* terbaik.
3. Menginput nilai *eps* dan *min_samples* terbaik pada model *Dbscan.*
4. Menyimpan *label* yang dibentuk oleh *Dbscan.*
5. Mengidentifikasi titik mana yang membentuk *core points.*
6. Menghitung jumlah *cluster* dan nilai *silhoutte score*.

Silhouette Score dapat dihitung dengan rumus berikut:

$$S[i] = \frac{b[i]-a[i]}{\max\ (a[i],b[i])}$$

Dengan:

$b\ [i]$ = rata-rata ketidakmiripan terendah cluster ke-i dengan cluster lain atau *neighbouring cluster*.

$a\ [i]$ = rata-rata ketidakmiripan dari semua data yang ada pada cluster ke-i.

Silhoutte score dihitung dengan menggunakan rata-rata jarak antar *cluster* dengan titik dan jarak rata-rata *cluster* terdekat. Misal, sebuah *cluster* dengan banyak titik data yang sangat dekat satu sama lain (kepadatan tinggi) dan jauh dari *cluster* terdekat berikutnya (menunjukkan bahwa *cluster* sangat unik dibandingkan dengan yang terdekat berikutnya), akan memiliki *silhoutte score* yang kuat. Nilai *silhoutte score* berkisar dari -1 hingga 1, dengan -1 merupakan skor terburuk dan 1 menjadi skor terbaik. Sedangkan nilai *silhoutte score* 0 menunjukkan *cluster* yang tumpang tindih.

3.10 *K-Means Clustering*

K-means adalah suatu metode penganalisaan data atau metode *data mining* yang melakukan proses pemodelan tanpa supervisi (*unsupervised*) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi (Agusta, 2007). *K-means* merupakan jenis algoritma yang digunakan untuk mengelompokkan data berdasarkan titik pusat *cluster* (*centroid)* data. Pengelompokkan data dilakukan dengan cara memaksimalkan kesamaan data pada satu *cluster* dan meminimalkan kesamaan data antar *cluster*. Fungsi jarak dalam *cluster* digunakan sebagai ukuran kemiripan. Sehingga proses pemaksimalan kemiripan data didapatkan dari jarak terpendek antara data terhadap titik *centroid* (Asroni & Adrian, 2015).

Proses *clustering* dimulai dengan mengidentifikasi data yang akan di*cluster*, X_{ij} $(i = 1, \dots, n\ ; j = 1, \dots, m)$ dengan $n =$ jumlah data yang akan di *cluster* dan $m =$ jumlah variabel. Pada awal iterasi, pusat (*centroid*) setiap *cluster* ditetapkan secara bebas (sembarang), $C_{kj}(k = 1, \dots, p\ ; j = 1, \dots, m)$. Selanjutnya dihitung jarak antara setiap data dengan setiap *centroid*. Untuk melakukan perhitungan jarak

data ke-i (x_i) pada pusat cluster ke-k (c_k), atau bisa disebut (d_{ik}), dapat digunakan rumus *Euclidean* (Han, Jiawei, & Kamber, 2001) seperti pada persamaan berikut:

$$d_{ik} = \sqrt{\sum_{j=1}^{m} \left(x_{ij} - c_{kj}\right)^2} \tag{3.9}$$

Suatu data akan menjadi anggota dari suatu *cluster* ke-k apabila jarak data ke pusat ke-k bernilai minimum jika dibandingkan dengan jarak ke pusat lainnya. Hal ini dapat dihitung dengan menggunakan persamaan (3.10). Kemudian mengelompokkan data-data yang menjadi anggota pada setiap *cluster*.

$$Min \sum_{k=1}^{k} d_{ik} = \sqrt{\sum_{j=1}^{m} \left(x_{ij} - c_{kj}\right)^2} \tag{3.10}$$

Nilai pusat *cluster* yang baru dapat dihitung dengan cara mencari nilai rata-rata dari data-data yang menjadi anggota pada *cluster* tersebut, dengan menggunakan rumus pada persamaan (3.11):

$$c_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p} \tag{3.11}$$

Dimana: $x_{ij} \in$ *cluster* ke-k

$p =$ banyaknya anggota *cluster* ke-k.

3.11 *Improved K-Means*

Konsep dasar yang digunakan pada *improved k-means* adalah metode *k-means clustering*. Pada metode *k-means*, pencarian titik pusat awal dilakukan dengan cara acak. Namun pada *improved k-means*, dilakukan modifikasi pada tahapan algoritma dengan menambahkan beberapa tahapan dalam pencarian titik pusat awal sehingga titik pusat awal ditemukan tanpa pengacakan. Pencarian titik pusat awal dilakukan dengan mencari jarak terjauh antar *dataset* kemudian dua data tersebut dijadikan sebagai titik pusat awal *cluster* (Guang-ping & Wen-peng, 2012).

Berikut merupakan langkah-langkah dalam algoritma *k-means*:

1. Memasukkan jumlah *cluster* (k) yang telah diperoleh dengan algoritma *Dbscan*, dan menetapkan pusat *cluster* sembarang.
2. Menghitung jarak setiap data ke pusat *cluster* menggunakan persamaan (3.9).
3. Mengelompokkan data ke dalam *cluster* dengan jarak paling dekat menggunakan persamaan (3.10).

4. Menghitung pusat *cluster* yang baru dengan persamaan (3.11).
5. Mengulangi langkah 2 sampai 4 sampai sudah tidak ada lagi data yang berpindah ke *cluster* yang lain (Narwati, 2010).

BAB IV
METODOLOGI PENELITIAN

4.1 Obyek Penelitian

Populasi penelitian ini adalah seluruh informasi film di website IMDb. Sampel data yang digunakan adalah data judul film, tahun *release*, durasi, deskripsi, *votes, rating, genre, actor*, dan *director* film dalam IMDb Top 250 per 5 Desember 2019.

4.2 Jenis dan Sumber Data Penelitian

Data yang digunakan dalam penelitian ini adalah berupa data sekunder yang diperoleh dari website *Internet Movie Database* (IMDb) dengan teknik *web scraping*.

4.3 Metode Analisis Data

Peneliti melalukan analisis data menggunakan metode *PCA* dan *clustering.* Metode PCA digunakan untuk mereduksi dimensi dari kumpulan data *(data set)* yang mengandung banyak variabel yang saling berhubungan, dengan mempertahankan sebanyak mungkin variasi dalam *data set*.

Analisis *clustering* dipergunakan untuk mengklasifikasikan objek kedalam kelompok yang relatif homogen. Metode *clustering* yang digunakan oleh peneliti adalah metode non-hierarki *Dbscan* dan *k-means. Dbscan* digunakan untuk menentukan jumlah optimal kelompok dan algoritma *k-means* untuk mengelompokkan obyek film.

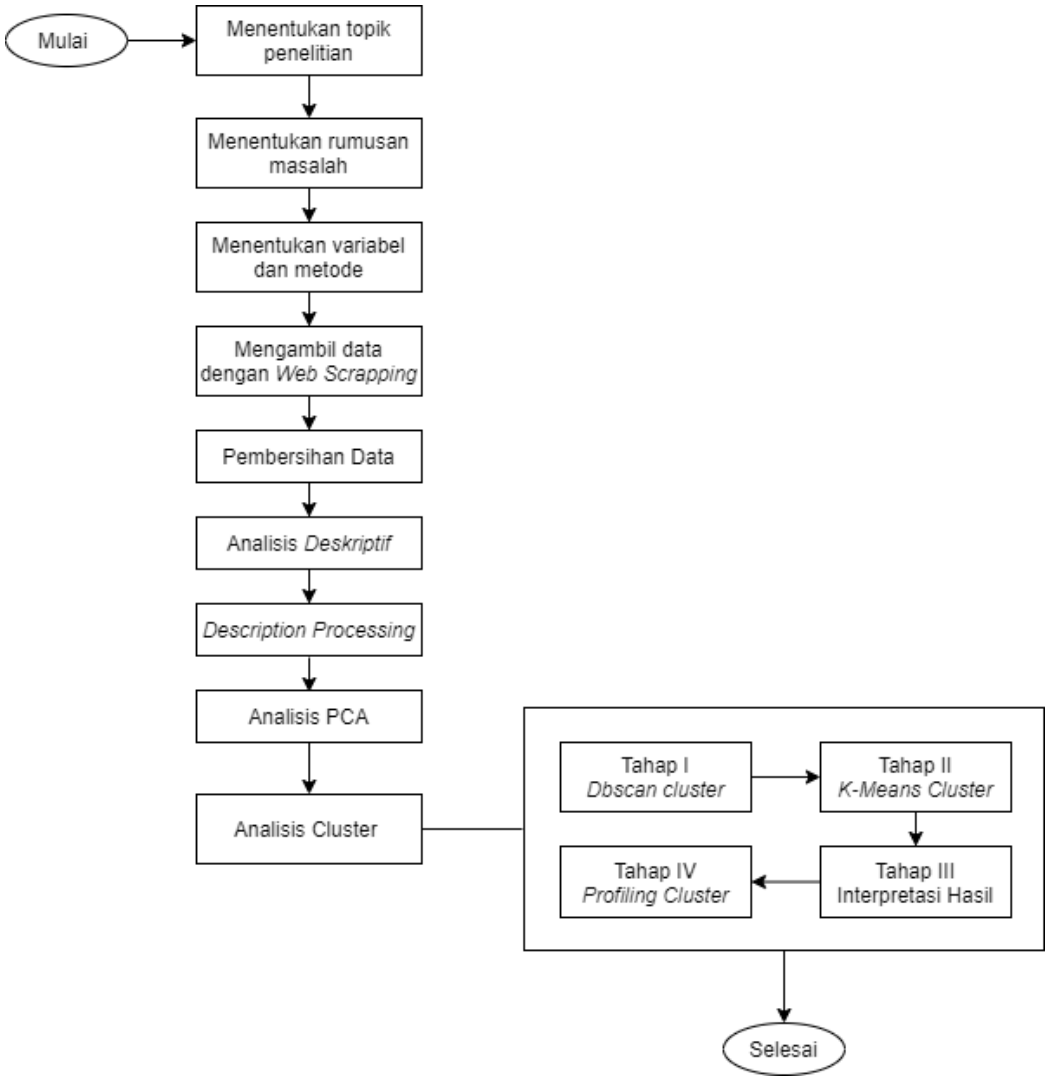
4.4 Definisi Operasional Peubah

Variabel yang akan dianalisis adalah variabel *title, year, runtime, votes, rating, genre, description, actor*, dan *director* film sebagai variabel independen dan variabel *title* sebagai variabel dependen atau obyek pengelompokan.

Tabel 4.1 Definisi Operasional Peubah

No	Nama Variabel	Definisi
1.	*Title*	Nama atau judul film
2.	*Year*	Tahun *release* film. 2 kategori: *Recent* dan *non-recent*
3.	*Runtime*	Durasi film dari awal sampai akhir
4.	*Votes*	Banyaknya jumlah *votes* atau dukungan terhadap film
5.	*Genre*	Jenis/ Aliran film
6.	*Rating*	*Score* atau hasil penilaian penonton terhadap film
7.	*Description*	Deskripsi mengenai film
8.	*Actor*	Pemeran utama film
9.	*Director*	Sutradara atau pembuat film

4.5 Langkah-langkah Penelitian

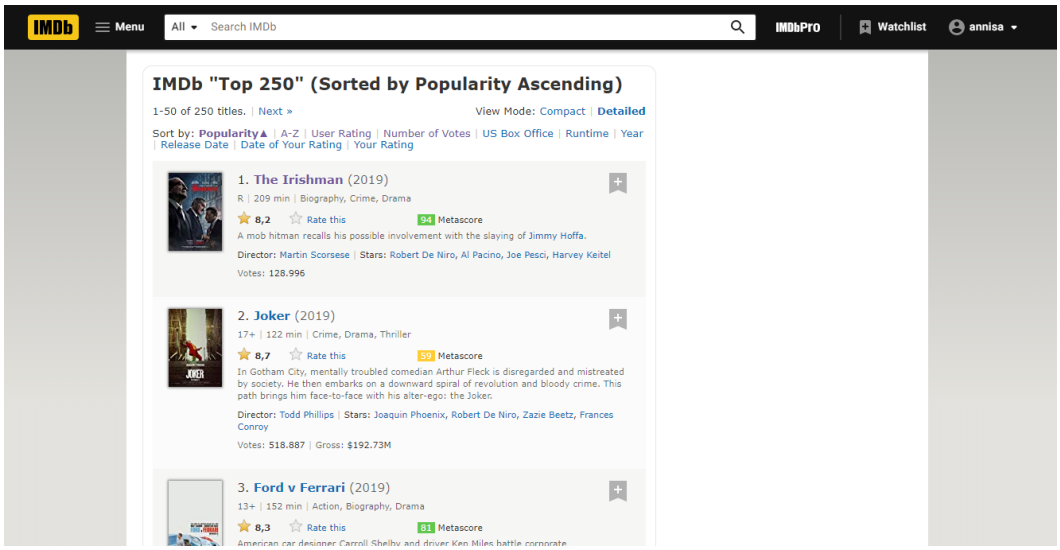


Gambar 4.1 Alur Penelitian

BAB V
HASIL DAN PEMBAHASAN

5.1 Gambaran Data IMDB

Data dalam penelitian diambil dari *website Internet Movie Database (IMDb)*. Cara pengambilan data dilakukan dengan melakukan *web scrapping* dengan *rvest* dalam program R.



Gambar 5.1 *IMDB Charts Top 250*

Tampilan tersebut memuat banyak informasi mengenai film yang memiliki popularitas terbaik berdasarkan rating pengguna. Masing-masing film mempunyai detail informasi yang berbeda, seperti: judul film, tahun *release*, *runtime*, *votes*, *genre*, *rating*, *gross*, *plot*, *director*, dan *actor*. Sehingga data awal dalam penelitian ini terdapat 250 film dan 10 variabel. Berikut merupakan data film dengan popularitas tertinggi dalam *web IMDb* (Lihat Tabel 5.1):

Tabel 5.1 Data Film *IMDB Top 250*

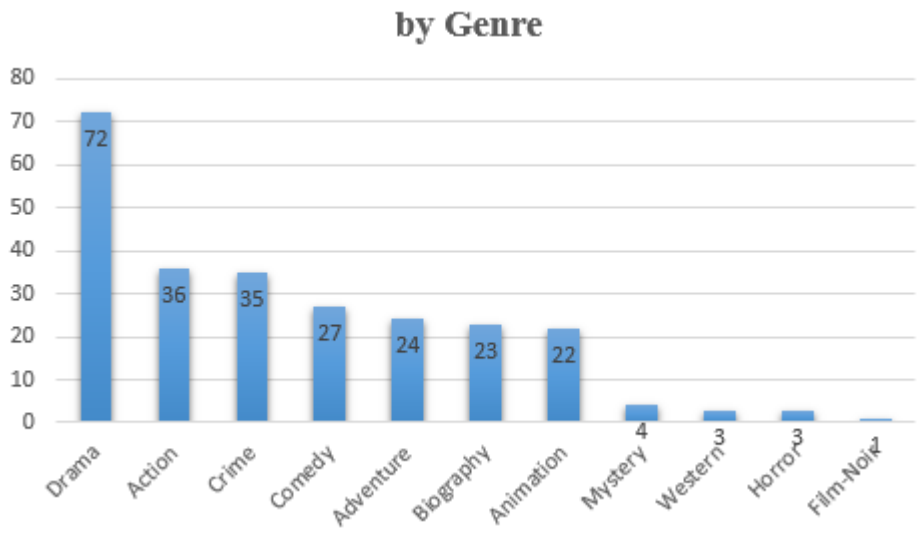
No	Title	Year	Runtime	…	Director	Actor
1.	The Irishman	2019	209	…	Martin Scorsese	Robert De Niro
2.	Joker	2019	122	…	Todd Phillips	Joaquin Phoenix
3.	Ford v Ferrari	2019	152	…	James Mangold	Matt Damon
4.	Marriage Story	2019	136	…	Noah Baumbach	Adam Driver
5.	Gisaengchung	2019	132	…	Bong Joon Ho	Kang-ho Song
6.	Avengers: Endgame	2019	96	…	Anthony Russo	Robert Downey Jr.
7.	Goodfellas	1990	181	…	Martin Scorsese	Robert De Niro
8.	The Godfather	1972	146	…	Francis Ford Coppola	Marlon Brando
9.	It's a Wonderful Life	1946	175	…	Frank Capra	James Stewart
10.	Die Hard	1988	130	…	John McTiernan	Bruce Willis
11.	Taxi Driver	1976	132	…	Martin Scorsese	Robert De Niro
12.	Star Wars	1977	114	…	George Lucas	Mark Hamill
13.	The Wolf of Wall Street	2013	121	…	Martin Scorsese	Leonardo DiCaprio
⋮	⋮	⋮	⋮		⋮	⋮

No	Title	Year	Runtime	…	Director	Actor
245.	Munna Bhai M.B.B.S.	2003	156	…	Rajkumar Hirani	Sanjay Dutt
246	Babam ve Oglum	2005	108	…	Ã‡agan Irmak	Ã‡etin Tekindor
247.	Ace in the Hole	1951	111	…	Billy Wilder	Kirk Douglas
248.	The Gold Rush	1925	95	…	Charles Chaplin	Charles Chaplin
249.	The General	1926	67	…	Clyde Bruckman	Buster Keaton
250.	Sherlock Jr.	1924	45	…	Buster Keaton	Buster Keaton

Setelah semua data terkumpul, perlu dilakukan pembersihan data yaitu dengan mengubah data variabel *Year* (Tahun) menjadi kategorik. Tahun 1990 dipilih sebagai batas yang tepat. Film yang rilis sebelum tahun 1990 dikategorikan menjadi 0 dan setelah 1990 menjadi 1. Kemudian variabel *Year* akan menjadi variabel *dummy* (0,1) pada *dataframe*. Dalam fase pembersihan data ini, variabel *Gross* dari hasil *scraping* tidak digunakan oleh peneliti karena terdapat beberapa data *missing* untuk beberapa judul film. Sehingga terdapat sembilan variabel akhir dalam penelitian ini.

Variabel lain yang perlu diubah menjadi kategorik adalah *Runtime (minutes)*. Durasi film 125 menit menjadi batasan yang tepat karena merupakan nilai median dari semua titik data. Film yang memiliki durasi kurang dari 125 menit menjadi 0 dan lebih dari 125 menit menjadi 1.

Dalam penelitian ini terdapat sebelas macam genre (*Drama, Action, Crime, Comedy, Adventure, Biography, Animation, Mystery, Western, Horror, Film-Noir)* dengan frekuensi dapat dilihat pada plot histogram dibawah.



Gambar 5.2 Plot *Genre*

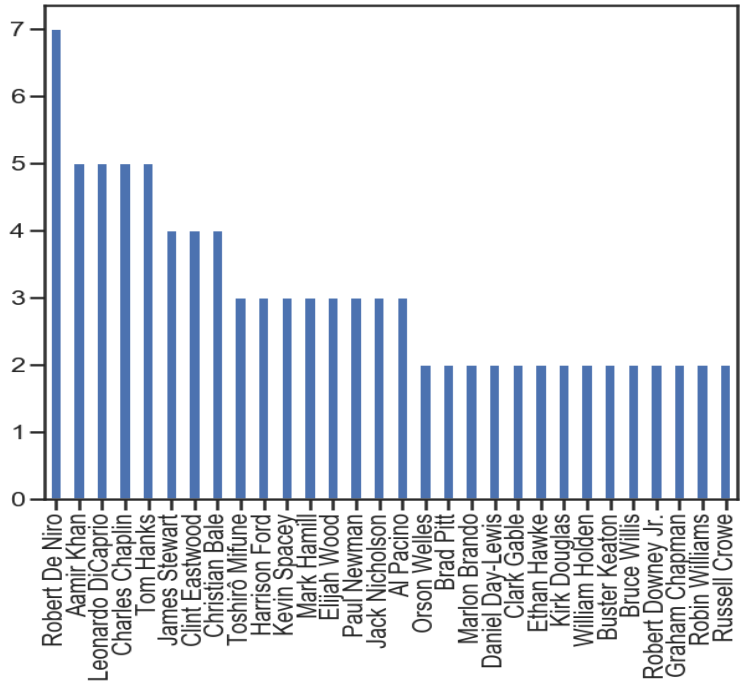
Semua *genre* dalam penelitian akan digunakan sebagai *predictor* dalam dataset. Diawali dengan memilih *genre* unik yang akan masuk dalam *dataframe* kemudian menambahkan satu kolom untuk setiap *genre* dalam *dataframe.*

Berdasarkan **Gambar 5.2** dapat dilihat frekuensi *genre* film yang masuk dalam *Top 250.* Film dengan *genre Drama* merupakan yang paling banyak muncul dalam *Top 250* dengan lebih dari 60 film, diikuti dengan *genre Action* dan *Crime* yang masing-masing muncul sebanyak 36 dan 35 film. Sedangkan frekuensi *genre* yang muncul paling sedikit dalam *Top 250* adalah film dengan *genre Film-Noir* sebanyak 1 film.

	Title	Year	Runtime	Votes	Genre	Rating	Gross	Plot	Director	Actor
0	The Irishman	1	1	119597	Biography	8.2	192.7	A mob hitman recalls his possible involvem...	Martin Scorsese	Robert De Niro
1	Joker	1	0	514253	Crime	8.7	NaN	In Gotham City, mentally troubled comedian...	Todd Phillips	Joaquin Phoenix
2	Ford v Ferrari	1	1	55920	Action	8.3	NaN	American car designer Carroll Shelby and d...	James Mangold	Matt Damon
3	Marriage Story	1	1	34510	Comedy	8.4	NaN	Noah Baumbach's incisive and compassionate...	Noah Baumbach	Adam Driver
4	Gisaengchung	1	1	95385	Comedy	8.6	192.7	All unemployed, Ki-taek and his family tak...	Bong Joon Ho	Kang-ho Song

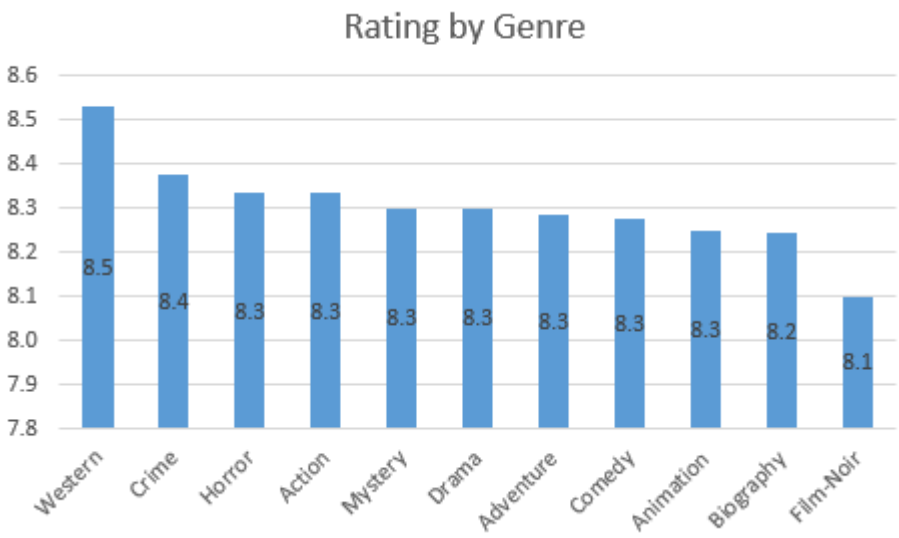
Gambar 5.3 *Dataframe* Baru

Hal diatas juga dilakukan pada variabel *Actor* dan *Director*. Namun untuk variabel *Actor* ini hanya akan diambil 30 *actor* dan untuk 20 D*irector* yang mempunyai frekuensi muncul terbanyak dalam judul film berbeda.



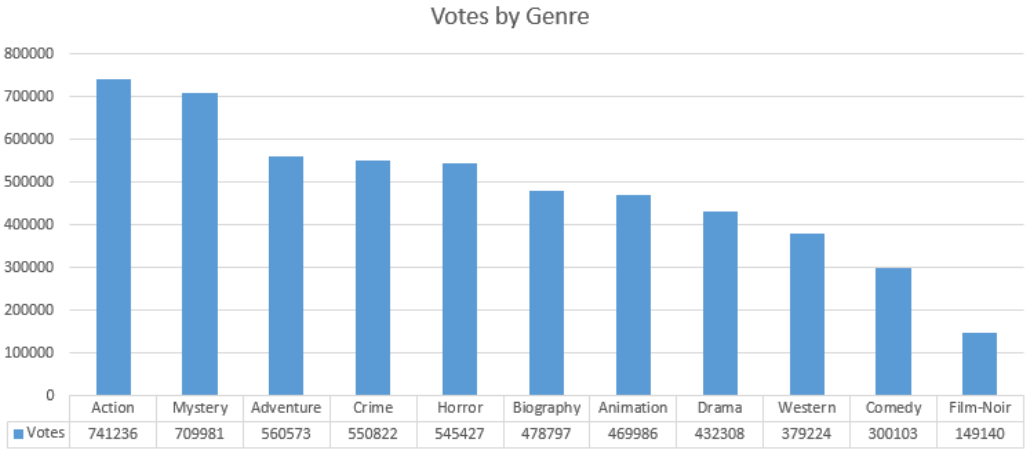
Gambar 5.4 Plot *Actors*

Gambar 5.4 merupakan plot untuk aktor yang membintangi film terbanyak yang masuk dalam list Top IMDb. Aktor yang film nya paling banyak masuk dalam list adalah *Robert De Niro* dengan total 7 film diikuti aktor *Aamir Khan, Leonardo DiCaprio, Charles Chaplin, Tom Hanks* dengan masing-masing 5 judul film. Variabel *Genre, Actors,* dan *Directors* dipilih sebagai *predictor* karena ketiganya mampu memberikan perbedaan karakteristik untuk *cluster* nantinya.



Gambar 5.5 Rata-rata *Rating* Berdasarkan Genre

Berdasarkan **Gambar 5.5**, film dengan genre *Western* memiliki rata-rata rating yang paling tinggi diantara genre lainnya. Hal ini dikarenakan hanya terdapat tiga film dengan genre barat dan ketiga nya memiliki rating yang cukup baik. Diikuti oleh film genre *Crime* yang memiliki rata-rata rating 8.4. Untuk film genre *Horror, Action, Mystery, Drama, Adventure, Comedy,* dan *Animation* masing-masing memiliki rata-rata rating 8.3 dengan selisih tipis. Film genre *Biography* memiliki rata-rata rating 8.2 dengan banyaknya film dengan genre ini sebanyak 23 judul film. Dan diposisi terakhir terdapat film dengan genre *Film-Noir* dengan rating 8.1 dengan hanya satu judul film.



Gambar 5.6 Rata-rata *Votes* Berdasarkan Genre

Berdasarkan jumlah *Votes* setiap genre, film dengan genre *Action* memiliki suara terbanyak dengan 741,236 *Votes* dan untuk genre *Film-Noir* memiliki suara terendah dibandingkan genre lainnya. Posisi kedua dengan jumlah suara terbanyak terdapat genre *Mystery* dengan 709,981 *votes*.

5.2 *Description Processing*

Tantangan utama pada penelitian ini adalah mengekstrak kata-kata relevan dari variabel *description* yang dapat digunakan sebagai kelompok atau variabel *(features)* dalam *cluster*. Langkah awal dalam *description processing* adalah memperbaiki teks dan menghilangkan tanda baca dan gangguan lainnya. Kemudian, menghapus semua kata yang tidak memberikan informasi. Setelah pembersihan awal teks, digunakan *TF-IDF* untuk mendapatkan *features* penting dan relevan dari teks untuk dapat digunakan dalam pengelompokan. *TF-IDF* sendiri

merupakan statistik yang mencerminkan seberapa penting suatu kata bagi suatu dokumen dalam kumpulan dokumen. Peneliti memilih metode ini karena *TF-IDF* dapat berfungsi sebagai alat deskriptif untuk variabel *description. TF* adalah banyaknya istilah muncul dalam *description* film dan *IDF* mengukur sedikitnya istilah yang muncul dalam *description*. Dengan demikian, *TF-IDF* memberikan ukuran keunikan dan menampilkan kata-kata yang unik dalam *description*. *TF-IDF* memberikan bobot tinggi pada kata-kata yang mewakili *description* film tetapi tidak terlalu sering muncul di semua *description*. Hal ini memberikan keunikan karena *description* yang memiliki kata-kata dengan indikator yang serupa harus dikelompokkan bersama.

Tabel 5.2 Tabel *TF-IDF* Teratas

Words	*Scores*
life	14.36
find	11.78
man	11.29
young	10.50
help	10.26
year	8.48
family	8.33
war	8.21
woman	7.19
murder	7.16

Words	*Scores*
old	6.56
friend	6.55
struggle	6.54
child	6.48
go	5.98
world	5.79
son	5.74
new	5.74
boy	5.68
story	5.61

Berdasarkan *TF-IDF*, semua kata tersebut penting. Terdapat beberapa tema yang umum seperti *war, life, murder.* Setelah mengidentifikasi kata-kata penting, satu pengkodean dilakukan untuk setiap kata dan dibuat kolom *dummy* masing-masing, jika sebuah film memiliki kata itu diberi nilai 1 dan 0 jika tidak. Untuk kolom *Actor* dan *Director* diambil masing-masing 30 dan 20 terbaik berdasarkan jumlahnya dan membuat kolom *dummy.* Sehingga didapatkan *dataframe* dengan

250 film dan 86 kolom. Memodelkan *dataset* tersebut sangat sulit karena ukuran dimensi yang menjadikan dasar komputasi dan teknik statistik tidak dapat digunakan.

```
In [24]: df_dbscan=df_final2

         (250, 86)
```

Gambar 5.7 Ukuran *Dataframe*

5.3 *Principal Component Analysis (PCA)*

Sebelum melakukan pencocokan dengan algoritma *clustering* perlu melakukan reduksi jumlah *features* dengan *PCA*. Metode ini digunakan untuk mendekomposisi suatu matriks yang besar menjadi komponen vektor orthogonal menggunakan nilai *eigenvectors* dan *eigenvalues* dari matriks tersebut. Langkah awal *PCA* adalah menstandarisasi *features* atau kolom.

```
45
['0.04', '0.03', '0.03', '0.03', '0.03', '0.03', '0.03', '0.03', '0.02', '0.02', '0.02', '0.02', '0.02', '0.02', '0.02', '0.0
2', '0.02', '0.02', '0.02', '0.02', '0.02', '0.02', '0.02', '0.02', '0.02', '0.02', '0.02', '0.01', '0.01', '0.01', '0.01', '0.01', '0.
01', '0.01', '0.01', '0.01', '0.01', '0.01', '0.01', '0.01', '0.01', '0.01', '0.01', '0.01', '0.01', '0.01']
0.8154522712037137
```

Gambar 5.8 Komponen Terbaik dan Nilai Persen Varians

Dalam penelitian ini, dimasukkan juga komponen yang menjelaskan lebih dari 1% variansi data. Sehingga diperoleh 45 komponen yang menjelaskan 81.54% variansi dalam *dataset*. Berikut merupakan *features* yang paling penting dalam *principal component* pertama.

Tabel 5.3 Tabel *Principal Component*

Features	*Abs_weights*
Runtime	0.44846
Votes	0.326861
Director:Rajkumar Hirani	0.242888
Actor:Mark Hamill	0.242888
Director:Stanley Kubrick	0.237781
...	...
woman	0.008577
Actor:Graham Chapman	0.008337

genre:Mystery	0.008262
Actor:Paul Newman	0.005478
Actor:Buster Keaton	0.001358

Pada tabel diatas, variabel *Runtime* dan *Votes* merupakan *features* yang paling berpengaruh dibanding *features* lain pada komponen utama ini dikarenakan kedua variabel tersebut memiliki nilai bobot lebih dari 0.3. *Runtime* menempati posisi pertama dengan variansi terbesar, yang artinya durasi sebuah film memiliki peran penting untuk dapat masuk dalam kategori *IMDB'*s *Top 250*. *Votes* muncul dengan nilai variansi terbesar kedua, dimana variabel tersebut merupakan satu-satunya variabel kontinu sehingga cukup mengganggu. Wajar jika *votes* memiliki variansi besar karena kontinu, tetapi fakta bahwa bobot nya secara signifikan jauh lebih tinggi dibanding *features* lainnya juga berarti bahwa *PCA* tidak bekerja dengan baik dengan *dataset* yang didominasi oleh kolom *dummy*.

Hal ini perlu diingat untuk penggunaan *PCA* kedepannya. Mungkin lebih baik jika semua kolom *dummy* daripada satu atau dua kolom saja. Dari **Tabel 5.2** didapatkan juga bahwa *Rajkumar Hirani* dan *Stanley Kubrick* mempunyai bobot yang tinggi diantara *director* film lainnya, karena karya film keduanya paling banyak masuk dalam *IMDB's Top 250.*

5.4 *Dbscan Clustering*

Langkah selanjutnya setelah melakukan *PCA* adalah memasukkan algoritma *Dbscan cluster* pada data. Karena belum diketahui jumlah *cluster* yang cocok untuk data penelitian ini, maka perlu digunakan algoritma *Dbscan.* Untuk mendapatkan konsistensi dari *cluster* yang dibentuk oleh *Dbscan*, digunakan *shilhouette score* sebagai metriks untuk mengukur kinerja *Dbscan*. *Shilhouette score* mengukur jumlah kohesi dalam sebuah *cluster* dibandingkan dengan *cluster* lain (pemisahan). Nilai *shilhouette score* antara -1 hingga +1 dan nilai yang lebih tinggi menunjukkan bahwa *datapoint* sangat mirip dengan *cluster* sendiri dan berbeda dengan *cluster* tetangga lainnya. *Shilhouette score* ini dihitung dengan metrik jarak Eucledian. Untuk menemukan jumlah *cluster* yang optimal, perlu mengacak nilai epsilon dan *min_samples* untuk mendapatkan *shilhouette score* yang baik.

Tabel 5.4 Tabel Paramater *Dbscan*

eps	*min_samp*	*shilhouette score*
0.5	1	0.0423
0.5	2	-0.3081
1	1	0.0477
1	2	-0.2970
2	1	0.0618
2	2	-0.2581
3	1	0.0660
3	2	-0.1382
3	3	-0.1905
3	4	-0.2079
3	5	-0.2160
4	1	0.0746
4	2	0.0351
4	3	-0.0152
5	1	0.1333
5	2	0.1303

eps	*min_samp*	*shilhouette score*
5	3	0.0561
6	1	0.1707
6	2	0.2176
6	3	0.1885
7	1	0.2211
7	2	0.2431
7	3	0.1944
8	1	0.2446
8	2	0.2580
8	3	0.1957
9	1	0.2608
9	2	0.2698
9	3	0.2252
10	1	0.2437
10	2	0.2491
11	1	0.2708

Pasangan nilai yang memiliki *shilhouette score* terbaik dan dapat dijelaskan lah yang dipilih. Output untuk model terbaik untuk *eps*=11 dan *min_samples*=1. Karena data nya cukup menyebar, diketahui bahwa pengelompokan *Dbscan* tidak bekerja dengan baik dalam pengelompokan film yang serupa. Kurang padatnya data film ini membuat *output Dbscan* belum dapat diolah dalam bentuk laporan, namun hasil *cluster* dapat dicetak dari *notebook IPynb*. Dengan demikian, *Dbscan* hanya digunakan untuk menemukan jumlah cluster yang optimal dalam dataset ini.

Berdasarkan performa model terbaik algoritma *Dbscan* mengelompokkan data menjadi 6 kelompok (*cluster).* Jumlah *cluster* yang diperoleh dari *Dbscan* ini akan digunakan atau diterapkan pada *K-Means Clustering.*

```
Estimated number of clusters: 6
Silhouette Coefficient: 0.271
```

Gambar 5.9 *Number* of *Dbscan Cluster*

5.5 *K-Means Clustering*

Setelah mendapatkan jumlah *cluster* terbaik, digunakan metode *kmeans* untuk mengelompokkan dataset. Meskipun kolom kategorik, tetap digunakan *kmeans* untuk mengelompokkan dataset dibanding dengan *k-modes,* karena *k-modes* mengharuskan *centroid* untuk mengambil nilai *features* terbesar tanpa memperhatikan apakah titik data dalam *cluster* berada dalam hubungan yang kuat. Untuk menghindarinya maka hanya dianalisis dengan *kmeans.* Bagaimanapun, diperoleh enam *cluster* terbaik yang memiliki hubungan atau karakteristik yang jelas.

Dalam penelitian ini perlu ditentukan karakteristik umum setiap *cluster* film yang dibentuk dengan mencocokan dan menafsirkan atribut dari film yang dapat dikelompokkan bersama.

Cluster 1:

- Titles: The Godfather: Part II, Heat, Scarface
- Kelompok ini merupakan kelompok "***Al Pacino*-Crime***". Dimana film dalam kelompok ini bergenre *Crime* dan dibintangi oleh aktor *Al Pacino*. Film dalam kelompok ini juga mempunyai durasi >125 menit.

Cluster 2:

- Titles: Modern Times, The Great Dictator, City Lights, The Kid, The Gold Rush.
- Kelompok ini merupakan kelompok "***Charlie Chaplin***" karena semua film dalam kelompok ini dibintangi dan disutradarai oleh *Charlie Chaplin.* Dengan genre *Comedy* dan merupakan film-film lama yang di rilis sebelum tahun 1990.

- Di masa sekarang semakin jarang film bergenre *Comedy* sehingga disarankan bagi pembuat film untuk dapat memproduksi film-film *Comedy* atau me-*remake* nya.

Cluster 3:

- Titles: The Irishman, Goodfellas, Taxi Driver, The Wolf of Wall Street, Casino, The Departed, Raging Bull, Once Upon a Time in America, Shutter Island.
- Merupakan kelompok "***Martin Scorsese***". Hampir seluruh film dalam kelompok ini disutradarai oleh *Martin Scorsese* dan semua film masuk dalam ranking 60 teratas.

Cluster 4:

- Titles: The Dark Knight, Interstellar, The Dark Knight Rises, Inception, Batman Begins, The Prestige, Memento.
- Merupakan kelompok "**Most Voted-Christopher Nolan**". Karena semua film dalam kelompok ini memiliki jumlah *votes* lebih dari 1 juta suara dan disutradarai oleh *Christopher Nolan*. Kelompok ini didominasi oleh film dengan genre *Action.*

Cluster 5:

- Titles: Joker, Ford v Ferrari, Marriage Story, Gisaengchung, Avengers: Endgame, The Godfather, It's a Wonderful Life, Die Hard, Star Wars, The Shawshank Redemption, The Shining, Star Wars: Episode VI - Return of the Jedi, Pulp Fiction, Harry Potter and the Deathly Hallows: Part 2, The Matrix, Star Wars: Episode V - The Empire Strikes Back, Avengers: Infinity War, Apocalypse Now, The Lion King, Guardians of the Galaxy, Spider-Man: Into the Spider-Verse, Gladiator, Inglourious Basterds, Green Book, Se7en, Mad Max: Fury Road, Fight Club, Forrest Gump, Schindler's List, Léon, The Silence of the Lambs, Gone Girl, Good Will Hunting, Logan, Django Unchained, Back to the Future, Stand by Me, Once Upon a Time in the West, The Green Mile, American History X, Requiem for a Dream, Jurassic Park, Coco, Whiplash, The Deer Hunter, Full Metal Jacket, American Beauty, Blade Runner, Alien, Reservoir Dogs, Catch Me If You

Can, The Big Lebowski, Snatch, Terminator 2: Judgment Day, Saving Private Ryan, Three Billboards Outside Ebbing, Missouri, The Terminator, A Clockwork Orange, Hacksaw Ridge, One Flew Over the Cuckoo's Nest, The Princess Bride, 12 Angry Men, The Grand Budapest Hotel, Prisoners, Room, No Country for Old Men, Inside Out, Raiders of the Lost Ark, Aladdin, Kill Bill: Vol. 1, Toy Story, Braveheart, Monsters, Inc., Il buono, il brutto, il cattivo, The Help, 2001: A Space Odyssey, Sen to Chihiro no kamikakushi, There Will Be Blood, The Usual Suspects, Toy Story 3, Aliens, A Beautiful Mind, Spotlight, Shichinin no samurai, The Pianist, Psycho, Unforgiven, Oldeuboi, Eternal Sunshine of the Spotless Mind, Ah-ga-ssi, Ben-Hur, Trainspotting, Kimi no na wa., Indiana Jones and the Last Crusade, Into the Wild, Up, Gone with the Wind, The Sixth Sense, 12 Years a Slave, La vita è bella, V for Vendetta, Rocky, Pan's Labyrinth, Capharnaüm, Fargo, Cidade de Deus, The Thing, Finding Nemo, Rush, WALL·E, The Intouchables, Chinatown, Casablanca, Amadeus, Citizen Kane, Amélie, Platoon, The Lives of Others, Salinui chueok, Lock, Stock and Two Smoking Barrels, Warrior, The Truman Show, Stalker, Rear Window, How to Train Your Dragon, Vertigo, Jagten, Dead Poets Society, 3 Idiots, Lawrence of Arabia, Million Dollar Baby, Gran Torino, L.A. Confidential, Monty Python and the Holy Grail, Hauru no ugoku shiro, The Great Escape, Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb, Some Like It Hot, Dangal, Per qualche dollaro in più, Hotaru no haka, Cool Hand Luke, Incendies, Metropolis, Der Untergang, To Kill a Mockingbird, Mononoke-hime, Before Sunrise, Akira, El secreto de sus ojos, Das Boot, Andhadhun, Barry Lyndon, Butch Cassidy and the Sundance Kid, Life of Brian, Idi i smotri, North by Northwest, Seppuku, Singin' in the Rain, The Elephant Man, Nuovo Cinema Paradiso, Paris, Texas, White Heat, Faa yeung nin wa, Tonari no Totoro, Jodaeiye Nader az Simin, The Apartment, Relatos salvajes, Hachi: A Dog's Tale, Det sjunde inseglet, The Sting, Sunset Blvd., Before Sunset, La battaglia di Algeri, Paths of Glory, In the Name of the Father, M - Eine Stadt sucht einen

Mörder, PK, Rebecca, Hotel Rwanda, On the Waterfront, All About Eve, Mou gaan dou, La haine, Network, Ran, Amores perros, Taare Zameen Par, The Bridge on the River Kwai, The Third Man, Gangs of Wasseypur, Persona, Witness for the Prosecution, Rashômon, Kaze no tani no Naushika, Tenkû no shiro Rapyuta, The Red Shoes, Andrei Rublev, The Treasure of the Sierra Madre, Double Indemnity, Mr. Smith Goes to Washington, Les quatre cents coups, Mary and Max, Dial M for Murder, Ladri di biciclette, Judgment at Nuremberg, Lagaan: Once Upon a Time in India, Ikiru, Yôjinbô, It Happened One Night, Bacheha-Ye aseman, Smultronstället, Drishyam, Tôkyô monogatari, La passion de Jeanne d'Arc, Eskiya, Le salaire de la peur, Rang De Basanti, Munna Bhai M.B.B.S., Babam ve Oglum, Ace in the Hole, The General, Sherlock Jr.

- Berdasarkan list film diatas, merupakan kelompok Drama.

Cluster 6:

- Titles: The Lord of the Rings: The Fellowship of the Ring, The Lord of the Rings: The Return of the King, The Lord of the Rings: The Two Towers.
- Kelompok "**Lord of the Rings**" dengan genre *Adventure*. Kelompok ini berisi series film *Lord of the Rings* yang merupakan film terkini, sehingga kelompok ini memiliki sutradara dan aktor film yang sama yaitu *Peter Jackson* dan aktor *Elijah Wood.*

Berdasarkan hasil *cluster* dengan algoritma *improved k-means* didapatkan nilai akurasi sebesar 0.872 atau 87.2%, yang artinya hasil *cluster* cukup baik.

BAB VI
PENUTUP

6.1 Kesimpulan

Dari analisis, diperoleh kesimpulan sebagai berikut:

1. Faktor yang dapat memengaruhi suatu film dapat masuk dalam IMDB Top 250 adalah Durasi dan Jumlah *Votes*. Dimana kedua variabel tersebut merupakan *features* yang paling berpengaruh dibanding *features* lain pada komponen utama ini.
2. Hasil pengelompokan film dalam list IMDB Top 250 adalah sebagai berikut:

   Cluster 1 merupakan kelompok "***Al Pacino-Crime***". Dimana film dalam kelompok ini bergenre kejahatan dan dibintangi oleh aktor *Al Pacino*. Film dalam kelompok ini juga mempunyai durasi >125 menit.

   Cluster 2: kelompok "***Charlie Chaplin***" karena semua film dalam kelompok ini dibintangi dan disutradarai oleh *Charlie Chaplin.* Dengan genre *Comedy-Adventure* dan merupakan film-film lama yang di rilis sebelum tahun 1990.

   Cluster 3 dinamakan kelompok "***Martin Scorsese***". Hampir seluruh film dalam kelompok ini disutradarai oleh *Martin Scorsese* dan semua film masuk dalam ranking 60 teratas.

   Cluster 4 adalah kelompok "**Most Voted-Christopher Nolan**". Karena semua film dalam kelompok ini memiliki jumlah *votes* lebih dari 1 juta suara dan disutradarai oleh *Christopher Nolan*. Kelompok ini didominasi oleh film dengan genre *Action.*

   Cluster 5 merupakan kelompok **Drama**.

   Cluster 6 disebut kelompok "**Lord of the Rings**" dengan genre *Adventure.* Kelompok ini berisi series film *Lord of the Rings* yang merupakan film terkini, sehingga kelompok ini memiliki sutradara dan aktor film yang sama yaitu *Peter Jackson* dan aktor *Elijah Wood.*

6.2 Saran

Adapun saran yang dapat diberikan penulis berdasarkan hasil penelitian guna menambah kesempurnaan penelitian selanjutnya yaitu:

1. Menggunakan database film yang lebih besar dan menggunakan algoritma cluster lain untuk menentukan jumlah cluster optimum selain *dbscan* dan membandingkan hasilnya.
2. Dapat memproduksi film-film *Comedy* atau me-remake nya karena di masa sekarang semakin jarang film bergenre *Comedy* yang populer.
3. Dapat menjadikan tokoh-tokoh yang karya nya telah berhasil masuk kategori Top IMDB sebagai acuan dalam membuat karya film yang lebih baik.
4. Dapat lebih mengenali preferensi film kesukaan penonton berdasarkan atribut filmnya.
5. Lebih memperhatikan durasi film dilihat dari faktor yang paling berpengaruh yang menjadikan suatu film dapat masuk dalam IMDB Top 250 ini dan berusaha untuk mendapatkan *votes* dari penonton karena dengan banyaknya *votes* dari penonton dapat meyakinkan pengunjung IMDB lain untuk menonton film tersebut.

DAFTAR PUSTAKA

Asroni, & Adrian, R. (2015). *Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang*. Jurnal Ilmiah Semesta Teknika, 78.

Agusta, Y. 2007. *K-Means-Penerapan, Permasalahan dan Metode Terkait*. Jurnal Sistem dan Informatika Vol.3, 47- 60.

Bari, A., Chaouchi, M., & Jung, T. (2014). *Predictive Analytics For Dummies.* New Jersey: John Wiley & Sons, Inc.

Bulut, H., & Korukoglu, S. (2011). Analysis and Clustering of Movie Genres. *JOURNAL OF COMPUTING, VOLUME 3*.

Dillon, W. R., & Goldsten, M. (1984). *Multivariate Analysis Methods and Applications.* Canada: John Wiley & Sons, Inc.

Eriyanto. (2007). *Teknik Sampling Analisis Opini Publik.* Yogyakarta: LKis Yogyakarta.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).* Institute for Computer Science, University of Munich.

Feldman, Ronen, Sanger, & dkk. (2007). *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data.* New York: Cambridge University Press.

Fitri, Meisya. (2013). *Perancangan Sistem Temu Balik Informasi Dengan Metode Pembobotan Kombinasi Tf-Idf Untuk Pencarian Dokumen Berbahasa Indonesia*. Universitas Tanjungpura: Semarang.

Guang-ping, C., & Wen-peng, W. (2012). Improved K-means Algorithm with Meliorated Initial Center. *The 7th International Conference on Computer Science & Education*, Volume 12, 150-153.

Han, Jiawei, & Kamber, M. (2001). *Data Mining Concepts and Techniques Second Edition.* San Francisco: Morgan Kauffman.

Hardeniya, N., Perkins, J., & Deepti. (2016). *Natural Language Processing: Python and NLTK.* Birmingham, UK: Packt Publishing Ltd.

Herwijayanti, B., Ratnawati, D. E., & Muflikhah, L. (2018). Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 306-312.

Himawan, P. (2008). *Memahami Film.* Yogyakarta: Homerian Pustaka.

Hype Stat. (2020). *Imdb.Com – Info*. https://hypestat.com/info/imdb.com#info. Diakses pada 13 Februari 2020.

Ibrahim, I. S. (2011). *Budaya Populer sebagai Komunikasi; Dinamika Popscape dan.* Yogyakarta: Jalasutra.

Jarmul, K., & Lawson, R. (2017). *Phyton Web Scraping Second Edition.* Birmingham: Packt Publishing Ltd.

Johnson, I. T., & Winchern, D. W. (2007). *Applied Multivariate Statistical Analysis 6th Edition.* New Jersey: Pearson Prentice Hall.

Jolluffe, I. T. (2002). *Principal Component Analysis 2nd Edition.* New York: Springer-Verlag.

Joseph, Dolfi. *Pusat Apresiasi Film di Yogyakarta*. Artikel diakses pada 09 Desember 2019 dari http://e-journal.uajy.ac.id/821/3/2TA11217.pdf

Kotu, V., & Deshpande, B. (2015). *Predictive Analytics and Data Mining.* San Francisco: Morgan Kauffman Publisher.

Lutins, Evan. (2017). *DBSCAN: What is it? When to Use it? How to use it.* https://medium.com/@elutins/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818. Diakses pada 2 Maret 2020.

Ma'arif, A. A. (2015). *Penerapan Algoritma TF-IDF Untuk Pencarian Karya Ilmiah*.

Machfudhoh, S., & Wahyuningsih, N. (2013). Analisis Cluster Kabupaten/Kota. *Jurnal Sains dan Seni*, Vol. 2 No.1 1-8.

Maklin, C. (2019, Juni 30). *Medium*. Retrieved from DBSCAN Python Example: The Optimal Value For Epsilon (EPS):

https://towardsdatascience.com/machine-learning-clustering-dbscan-determine-the-optimal-value-for-epsilon-eps-python-example-3100091cfbc

Misra, S., Li, H., & He, J. (2019). *Machine Learning for Subsurface Characterization.* Cambridge: Elsevier Inc.

Mitchell, R. (2015). *Web Scraping With Phyton.* Sebastopol: O'Reilly Media Inc.

Narwati. (2010). Pengelompokan Mahasiswa Menggunakan Algoritma K-Means. *Jurnal Dinamika Informatika*, Vol 2 No. 2.

Portal Informasi Indonesia. (2019). *Tren Positif Film Indonesia.* https://indonesia.go.id/ragam/seni/sosial/tren-positif-film-indonesia. Diakses pada 13 Februari 2020.

Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi menggunakan MATLAB.* Yogyakarta: Andi.

Prasetyo, Lambang Agung. 2012. Aplikasi Sistem Informasi Geografis pada Kerusakan Bangunan Akibat Erupsi Merapi Tahun 2010 di Kabupaten Sleman Daerah Istimewa Yogyakarta Berbasis WEB. *Skripsi*. UII.

Putra, A. A. (2016). *Implementasi Text Summarization Menggunakan Metode Vector Space Model Pada Artikel Berita Berbahasa Indonesia*.

Rizki, Dhidik, & Supraptono, E. (2017). *Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen.* Semarang: Universitas Negeri Semarang.

Rohmawati,W.N., Defiyanti,S., Jajuli,M. 2015. Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa. *Jurnal Ilmiah Teknologi Informasi Terapan*, 1(2), 62-68.

Sobur, A. (2004). *Analisis Teks Media; Suatu Pengantar untuk Analisis Wacana, Analisis Semiotik,*. Bandung: PT. Remaja Rosdakarya.

Sumarno, M. (1996). *Dasar-Dasar Apresiasi Film.* Jakarta: PT.Grasindo.

Supranto, J. (2004). *Analisis Multivariat: Arti dan Interpretasi.* Jakarta: PT. Rineka Cipta.

Tim Penyusun Kamus Pusat Pembinaan dan Pengembangan Bahasa. 1990. *Kamus Besar Bahasa Indonesia*. Jakarta: Balai Pustaka. hlm. 242.

Tim Redaksi. *Cinemags*. Edisi 171. 2013

Triyanto, W. A. (2015). Algoritma K-Medoids untuk Penentuan Strategi. *Jurnal SIMETRIS*, Vol. 6 No.1 April 2015 183-188.

Usman, H., & Sobari, N. (2013). *Aplikasi Teknik Multivariat Untuk Riset Pemasaran.* Jakarta: PT. Raja Grafindo Persada.

Wandira, A. (2018). *Perkembangan Film Luar dan Indonesia*. Kompasiana: https://www.kompasiana.com/wandira/5c0e0cf1ab12ae46fc0b59d2/perkembangan-film-luar-dan-indonesia. Diakses pada 15 Februari 2020.

What is IMDb? (2019). Retrieved from IMDB Help Center: https://help.imdb.com/article/imdb/general-information/what-is-imdb/G836CY29Z4SGNMK5?ref_=helpart_nav_1#

LAMPIRAN

Lampiran 1 Data List Film IMDB Top 250

Title	Year	Runtime	Votes	Genre	Rating	Plot	Director	Actor
The Irishman	1	1	119597	Biography	8.2	A mob hitman	Martin Scorsese	Robert De Niro
Joker	1	0	514253	Crime	8.7	In Gotham City	Todd Phillips	Joaquin Phoenix
Ford v Ferrari	1	1	55920	Action	8.3	American car d	James Mangold	Matt Damon
Marriage Story	1	1	34510	Comedy	8.4	Noah Baumbac	Noah Baumbach	Adam Driver
Gisaengchung	1	1	95385	Comedy	8.6	All unemploye	Bong Joon Ho	Kang-ho Song
Avengers: Endgame	1	1	617281	Action	8.5	After the deva	Anthony Russo	Robert Downey Jr.
Goodfellas	1	1	939407	Biography	8.7	The story of He	Martin Scorsese	Robert De Niro
The Godfather	0	1	1487901	Crime	9.2	The aging patri	Francis Ford Coppola	Marlon Brando
It's a Wonderful Life	0	1	368445	Drama	8.6	An angel is sen	Frank Capra	James Stewart
Die Hard	0	1	737557	Action	8.2	An NYPD office	John McTiernan	Bruce Willis
Taxi Driver	0	0	664925	Crime	8.3	A mentally uns	Martin Scorsese	Robert De Niro
Star Wars	0	0	1151636	Action	8.6	Luke Skywalke	George Lucas	Mark Hamill
The Wolf of Wall Street	1	1	1074244	Biography	8.2	Based on the t	Martin Scorsese	Leonardo DiCaprio
Casino	1	1	428398	Crime	8.2	A tale of greed	Martin Scorsese	Robert De Niro
The Shawshank Redemption	1	1	2166909	Drama	9.3	Two imprisone	Frank Darabont	Tim Robbins
The Dark Knight	1	1	2141838	Action	9	When the men	Christopher Nolan	Christian Bale
The Shining	0	1	821563	Drama	8.4	A family heads	Stanley Kubrick	Jack Nicholson
The Godfather: Part II	0	1	1036306	Crime	9	The early life a	Francis Ford Coppola	Al Pacino
Interstellar	1	1	1354266	Adventure	8.6	A team of expl	Christopher Nolan	Matthew McConaughey
Star Wars: Episode VI - Return of the Jedi	0	1	886230	Action	8.3	After a daring r	Richard Marquand	Mark Hamill
Pulp Fiction	1	1	1701356	Crime	8.9	The lives of tw	Quentin Tarantino	John Travolta
The Departed	1	1	1107960	Crime	8.5	An undercover	Martin Scorsese	Leonardo DiCaprio
Harry Potter and the Deathly Hallows: Part 2	1	1	710564	Adventure	8.1	Harry (Daniel R	David Yates	Daniel Radcliffe
The Lord of the Rings: The Fellowship of the Ring	1	1	1553663	Adventure	8.8	A meek Hobbit	Peter Jackson	Elijah Wood
The Matrix	1	1	1559951	Action	8.7	A computer ha	Lana Wachowski	Keanu Reeves
Star Wars: Episode V - The Empire Strikes Back	0	0	1082923	Action	8.7	After the Rebe	Irvin Kershner	Mark Hamill
Avengers: Infinity War	1	1	730523	Action	8.5	The Avengers a	Anthony Russo	Robert Downey Jr.
Raging Bull	0	1	298329	Biography	8.2	The life of box	Martin Scorsese	Robert De Niro
The Dark Knight Rises	1	1	1425690	Action	8.4	Eight years afte	Christopher Nolan	Christian Bale
Inception	1	1	1899427	Action	8.8	A thief who ste	Christopher Nolan	Leonardo DiCaprio
Apocalypse Now	0	1	565266	Drama	8.4	A U.S. Army of	Francis Ford Coppola	Martin Sheen
The Lion King	1	0	875219	Animatior	8.5	A Lion cub crov	Roger Allers	Matthew Broderick
Guardians of the Galaxy	1	0	984643	Action	8	A group of inte	James Gunn	Chris Pratt
Spider-Man: Into the Spider-	1	0	280582	Animatior	8.4	Teen Miles Mo	Bob Persichetti	Shameik Moore
Gladiator	1	1	1250776	Action	8.5	A former Roma	Ridley Scott	Russell Crowe
Inglourious Basterds	1	1	1166566	Adventure	8.3	In Nazi-occupie	Quentin Tarantino	Brad Pitt
Green Book	1	1	266052	Biography	8.2	A working-clas	Peter Farrelly	Viggo Mortensen
Se7en	1	1	1332412	Crime	8.6	Two detectives	David Fincher	Morgan Freeman
Mad Max: Fury Road	1	0	808545	Action	8.1	In a post-apoca	George Miller	Tom Hardy
Heat	1	1	537724	Crime	8.2	A group of prof	Michael Mann	Al Pacino
Fight Club	1	1	1731235	Drama	8.8	An insomniac o	David Fincher	Brad Pitt
Forrest Gump	1	1	1668886	Drama	8.8	The presidenci	Robert Zemeckis	Tom Hanks
Schindler's List	1	1	1124449	Biography	8.9	In German-occ	Steven Spielberg	Liam Neeson
Once Upon a Time in	0	1	287141	Crime	8.4	A former Prohi	Sergio Leone	Robert De Niro
LÃ©on	1	0	959095	Action	8.5	Mathilda, a 12-	Luc Besson	Jean Reno
The Silence of the Lambs	1	0	1174214	Crime	8.6	A young F.B.I. c	Jonathan Demme	Jodie Foster
Gone Girl	1	1	787431	Drama	8.1	With his wife's	David Fincher	Ben Affleck
Batman Begins	1	1	1228692	Action	8.2	After training v	Christopher Nolan	Christian Bale
Good Will Hunting	1	1	793186	Drama	8.3	Will Hunting, a	Gus Van Sant	Robin Williams
Scarface	0	1	685022	Crime	8.3	In 1980 Miami,	Brian De Palma	Al Pacino
Logan	1	1	589190	Action	8.1	In a future whe	James Mangold	Hugh Jackman
Django Unchained	1	1	1253806	Drama	8.4	With the help o	Quentin Tarantino	Jamie Foxx
Back to the Future	0	0	969000	Adventure	8.5	Marty McFly, a	Robert Zemeckis	Michael J. Fox
The Lord of the Rings: The Return of the King	1	1	1538478	Adventure	8.9	Gandalf and Ar	Peter Jackson	Elijah Wood
Stand by Me	0	0	342046	Adventure	8.1	After the death	Rob Reiner	Wil Wheaton
Once Upon a Time in the	0	1	280436	Western	8.5	A mysterious s	Sergio Leone	Henry Fonda
Shutter Island	1	1	1040201	Mystery	8.1	In 1954, a U.S. I	Martin Scorsese	Leonardo DiCaprio
The Green Mile	1	1	1054451	Crime	8.6	The lives of gu	Frank Darabont	Tom Hanks
American History X	1	0	976887	Drama	8.5	A former neo-r	Tony Kaye	Edward Norton

Requiem for a Dream	1	0	718729	Drama	8.3	The drug-indu	Darren Aronofsky	Ellen Burstyn
Jurassic Park	1	1	805082	Action	8.1	A pragmatic Pa	Steven Spielberg	Sam Neill
Coco	1	0	309363	Animatior	8.4	Aspiring music	Lee Unkrich	Anthony Gonzalez
Whiplash	1	0	645365	Drama	8.5	A promising yo	Damien Chazelle	Miles Teller
The Deer Hunter	0	1	291575	Drama	8.1	An in-depth ex	Michael Cimino	Robert De Niro
Full Metal Jacket	0	0	628970	Drama	8.3	A pragmatic U.	Stanley Kubrick	Matthew Modine
American Beauty	1	0	1011954	Drama	8.3	A sexually frus	Sam Mendes	Kevin Spacey
Blade Runner	0	0	647350	Action	8.1	A blade runner	Ridley Scott	Harrison Ford
Alien	0	0	735881	Horror	8.4	After a space n	Ridley Scott	Sigourney Weaver
Reservoir Dogs	1	0	856771	Crime	8.3	When a simple	Quentin Tarantino	Harvey Keitel
Catch Me If You Can	1	1	748546	Biography	8.1	A seasoned FB	Steven Spielberg	Leonardo DiCaprio
The Big Lebowski	1	0	681428	Comedy	8.1	Jeff "The Dude	Joel Coen	Jeff Bridges
Snatch	1	0	736639	Comedy	8.3	Unscrupulous b	Guy Ritchie	Jason Statham
The Prestige	1	1	1098905	Drama	8.5	After a tragic a	Christopher Nolan	Christian Bale
Terminator 2: Judgment Day	1	1	938091	Action	8.5	A cyborg, ident	James Cameron	Arnold Schwarzenegger
Saving Private Ryan	1	1	1147919	Drama	8.6	Following the l	Steven Spielberg	Tom Hanks
Three Billboards Outside Ebbing, Missouri	1	0	370818	Comedy	8.2	A mother pers	Martin McDonagh	Frances McDormand
The Terminator	0	0	754335	Action	8	In 1984, a hum	James Cameron	Arnold Schwarzenegger
A Clockwork Orange	0	1	709708	Crime	8.3	In the future, a	Stanley Kubrick	Malcolm McDowell
Hacksaw Ridge	1	1	386448	Biography	8.1	World War II A	Mel Gibson	Andrew Garfield
One Flew Over the Cuckoo's	0	1	855554	Drama	8.7	A criminal plea	Milos Forman	Jack Nicholson
The Princess Bride	0	0	369506	Adventure	8.1	While home si	Rob Reiner	Cary Elwes
12 Angry Men	0	0	620492	Drama	8.9	A jury holdout	Sidney Lumet	Henry Fonda
The Grand Budapest Hotel	1	0	652899	Adventure	8.1	The adventure	Wes Anderson	Ralph Fiennes
Prisoners	1	1	545720	Crime	8.1	When Keller D	Denis Villeneuve	Hugh Jackman
Room	1	0	331489	Drama	8.1	Held captive fo	Lenny Abrahamson	Brie Larson
No Country for Old Men	1	0	789072	Crime	8.1	Violence and n	Ethan Coen	Tommy Lee Jones
Inside Out	1	0	558707	Animatior	8.2	After young Ril	Pete Docter	Amy Poehler
Raiders of the Lost Ark	0	0	832891	Action	8.4	In 1936, archae	Steven Spielberg	Harrison Ford
Aladdin	1	0	340081	Animatior	8	A kindhearted	Ron Clements	Scott Weinger
Kill Bill: Vol. 1	1	0	935332	Action	8.1	After awakenir	Quentin Tarantino	Uma Thurman
Toy Story	1	0	826043	Animatior	8.3	A cowboy doll	John Lasseter	Tom Hanks
Braveheart	1	1	911181	Biography	8.3	When his secre	Mel Gibson	Mel Gibson
Monsters, Inc.	1	0	762856	Animatior	8	In order to pov	Pete Docter	Billy Crystal
Il buono, il brutto, il cattivo	0	1	643199	Western	8.8	A bounty hunti	Sergio Leone	Clint Eastwood
The Lord of the Rings: The Two Towers	1	1	1391565	Adventure	8.7	While Frodo ar	Peter Jackson	Elijah Wood
The Help	1	1	401605	Drama	8.1	An aspiring aut	Tate Taylor	Emma Stone
2001: A Space Odyssey	0	1	560220	Adventure	8.3	After discoveri	Stanley Kubrick	Keir Dullea
Sen to Chihiro no	1	1	579428	Animatior	8.6	During her fam	Hayao Miyazaki	Daveigh Chase
There Will Be Blood	1	1	477034	Drama	8.2	A story of fami	Paul Thomas Anderso	Daniel Day-Lewis
The Usual Suspects	1	0	931243	Crime	8.5	A sole survivor	Bryan Singer	Kevin Spacey
Toy Story 3	1	0	709747	Animatior	8.3	The toys are m	Lee Unkrich	Tom Hanks
Aliens	0	1	616020	Action	8.4	Ellen Ripley is	James Cameron	Sigourney Weaver
A Beautiful Mind	1	1	792398	Biography	8.2	After John Nas	Ron Howard	Russell Crowe
Spotlight	1	1	380080	Biography	8.1	The true story	Tom McCarthy	Mark Ruffalo
Shichinin no samurai	0	1	293287	Action	8.6	A poor village	Akira Kurosawa	ToshirÃ´ Mifune
The Pianist	1	1	668064	Biography	8.5	A Polish Jewish	Roman Polanski	Adrien Brody
Psycho	0	0	557266	Horror	8.5	A Phoenix secr	Alfred Hitchcock	Anthony Perkins
Unforgiven	1	1	350619	Drama	8.2	Retired Old We	Clint Eastwood	Clint Eastwood
Oldeuboi	1	0	474372	Action	8.4	After being kid	Chan-wook Park	Min-sik Choi
Eternal Sunshine of the Spotless Mind	1	0	846976	Drama	8.3	When their rel	Michel Gondry	Jim Carrey
Ah-ga-ssi	1	1	87699	Drama	8.1	A woman is hir	Chan-wook Park	Min-hee Kim
Ben-Hur	0	1	206200	Adventure	8.1	When a Jewish	William Wyler	Charlton Heston
Trainspotting	1	0	602942	Drama	8.1	Renton, deepl	Danny Boyle	Ewan McGregor
Memento	1	0	1055660	Mystery	8.4	A man with sho	Christopher Nolan	Guy Pearce
Kimi no na wa.	1	0	150458	Animatior	8.4	Two strangers	Makoto Shinkai	RyÃ»nosuke Kamiki
Indiana Jones and the Last	0	1	650672	Action	8.2	In 1938, after h	Steven Spielberg	Harrison Ford
Into the Wild	1	1	536740	Adventure	8.1	After graduatir	Sean Penn	Emile Hirsch
Up	1	0	872545	Animatior	8.2	78-year-old Ca	Pete Docter	Edward Asner
Gone with the Wind	0	1	270475	Drama	8.1	A manipulative	Victor Fleming	Clark Gable
The Sixth Sense	1	0	863825	Drama	8.1	A boy who com	M. Night Shyamalan	Bruce Willis
12 Years a Slave	1	1	593978	Biography	8.1	In the antebell	Steve McQueen	Chiwetel Ejiofor
La vita Ã¨ bella	1	0	573021	Comedy	8.6	When an open	Roberto Benigni	Roberto Benigni
V for Vendetta	1	1	973498	Action	8.2	In a future Brit	James McTeigue	Hugo Weaving
Rocky	0	0	485989	Drama	8.1	A small-time b	John G. Avildsen	Sylvester Stallone

Pan's Labyrinth	1	0	585347	Drama	8.2	In the Falangis	Guillermo del Toro	Ivana Baquero
CapharnaÃ¼m	1	1	33074	Drama	8.4	While serving a	Nadine Labaki	Zain Al Rafeea
Fargo	1	0	576153	Crime	8.1	Jerry Lundegaa	Joel Coen	William H. Macy
Cidade de Deus	1	1	661055	Crime	8.6	In the slums of	Fernando Meirelles	Alexandre Rodrigues
The Thing	0	0	343135	Horror	8.1	A research tea	John Carpenter	Kurt Russell
Finding Nemo	1	0	892558	Animatior	8.1	After his son is	Andrew Stanton	Albert Brooks
Rush	1	0	404029	Biography	8.1	The merciless	Ron Howard	Daniel BrÃ¼hl
WALLÂ·E	1	0	934506	Animatior	8.4	In the distant f	Andrew Stanton	Ben Burtt
The Intouchables	1	0	700960	Biography	8.5	After he becon	Olivier Nakache	FranÃ§ois Cluzet
Chinatown	0	1	273589	Drama	8.2	A private dete	Roman Polanski	Jack Nicholson
Casablanca	0	0	489169	Drama	8.5	A cynical Amer	Michael Curtiz	Humphrey Bogart
Amadeus	0	1	346380	Biography	8.3	The life, succe	Milos Forman	F. Murray Abraham
Citizen Kane	0	0	373057	Drama	8.3	Following the	Orson Welles	Orson Welles
AmÃ©lie	1	0	667261	Comedy	8.3	AmÃ©lie is an	Jean-Pierre Jeunet	Audrey Tautou
Platoon	0	0	359800	Drama	8.1	A young soldie	Oliver Stone	Charlie Sheen
The Lives of Others	1	1	333034	Drama	8.4	In 1984 East Be	Florian Henckel von D	Ulrich MÃ¼he
Salinui chueok	1	1	108915	Action	8.1	In a small Kore	Bong Joon Ho	Kang-ho Song
Lock, Stock and Two Smoking	1	0	503955	Comedy	8.2	A botched card	Guy Ritchie	Jason Flemyng
Warrior	1	1	412148	Drama	8.2	The youngest s	Gavin O'Connor	Tom Hardy
The Truman Show	1	0	864276	Comedy	8.1	An insurance s	Peter Weir	Jim Carrey
Stalker	0	1	102677	Drama	8.2	A guide leads t	Andrei Tarkovsky	Alisa Freyndlikh
Rear Window	0	0	411872	Mystery	8.4	A wheelchair-b	Alfred Hitchcock	James Stewart
How to Train Your Dragon	1	0	627264	Animatior	8.1	A hapless your	Dean DeBlois	Jay Baruchel
Vertigo	0	1	332190	Mystery	8.3	A former polic	Alfred Hitchcock	James Stewart
Jagten	1	0	252158	Drama	8.3	A teacher lives	Thomas Vinterberg	Mads Mikkelsen
Dead Poets Society	0	1	385027	Comedy	8.1	English teache	Peter Weir	Robin Williams
3 Idiots	1	1	312238	Comedy	8.4	Two friends ar	Rajkumar Hirani	Aamir Khan
Lawrence of Arabia	0	1	249786	Adventure	8.3	The story of T.E	David Lean	Peter O'Toole
Million Dollar Baby	1	1	599467	Drama	8.1	A determined	Clint Eastwood	Hilary Swank
Gran Torino	1	0	684251	Drama	8.1	Disgruntled Ko	Clint Eastwood	Clint Eastwood
L.A. Confidential	1	1	504174	Crime	8.2	As corruption g	Curtis Hanson	Kevin Spacey
Monty Python and the Holy	0	0	473468	Adventure	8.2	King Arthur an	Terry Gilliam	Graham Chapman
Hauru no ugoku shiro	1	0	292738	Animatior	8.2	When an unco	Hayao Miyazaki	Chieko BaishÃ´
The Great Escape	0	1	210049	Adventure	8.2	Allied prisoner	John Sturges	Steve McQueen
Dr. Strangelove or: How I Learned to Stop Worrying	0	0	425127	Comedy	8.4	An insane gen	Stanley Kubrick	Peter Sellers
Some Like It Hot	0	0	227322	Comedy	8.2	When two mal	Billy Wilder	Marilyn Monroe
Dangal	1	1	132705	Action	8.4	Former wrestle	Nitesh Tiwari	Aamir Khan
Per qualche dollaro in piÃ¹	0	1	214037	Western	8.3	Two bounty hu	Sergio Leone	Clint Eastwood
Hotaru no haka	0	0	207192	Animatior	8.5	A young boy ar	Isao Takahata	Tsutomu Tatsumi
Cool Hand Luke	0	1	152339	Crime	8.1	A laid back Sou	Stuart Rosenberg	Paul Newman
Incendies	1	1	130054	Drama	8.3	Twins journey	Denis Villeneuve	Lubna Azabal
Metropolis	0	1	148968	Drama	8.3	In a futuristic c	Fritz Lang	Brigitte Helm
Der Untergang	1	1	312935	Biography	8.2	Traudl Junge, t	Oliver Hirschbiegel	Bruno Ganz
To Kill a Mockingbird	0	1	278956	Crime	8.3	Atticus Finch, a	Robert Mulligan	Gregory Peck
Mononoke-hime	1	1	305361	Animatior	8.4	On a journey t	Hayao Miyazaki	YÃ´ji Matsuda
Before Sunrise	1	0	246281	Drama	8.1	A young man a	Richard Linklater	Ethan Hawke
Akira	0	0	147545	Animatior	8.1	A secret milita	Katsuhiro Ã"tomo	Mitsuo Iwata
El secreto de sus ojos	1	1	178039	Drama	8.2	A retired legal	Juan JosÃ© Campanel	Ricardo DarÃ-n
Das Boot	0	1	216668	Adventure	8.3	The claustroph	Wolfgang Petersen	JÃ¼rgen Prochnow
Andhadhun	1	1	52882	Crime	8.4	A series of mys	Sriram Raghavan	Ayushmann Khurrana
Barry Lyndon	0	1	136983	Adventure	8.1	An Irish rogue	Stanley Kubrick	Ryan O'Neal
Butch Cassidy and the	0	0	191839	Biography	8	Wyoming, earl	George Roy Hill	Paul Newman
Life of Brian	0	0	343743	Comedy	8.1	Born on the ori	Terry Jones	Graham Chapman
Idi i smotri	0	1	46130	Drama	8.3	After finding a	Elem Klimov	Aleksey Kravchenko
North by Northwest	0	1	279559	Adventure	8.3	A New York Cit	Alfred Hitchcock	Cary Grant
Seppuku	0	1	28507	Drama	8.7	When a ronin r	Masaki Kobayashi	Tatsuya Nakadai
Singin' in the Rain	0	0	203341	Comedy	8.3	A silent film pr	Stanley Donen	Gene Kelly
The Elephant Man	0	0	205238	Biography	8.1	A Victorian sur	David Lynch	Anthony Hopkins
Nuovo Cinema Paradiso	0	1	209426	Drama	8.5	A filmmaker re	Giuseppe Tornatore	Philippe Noiret
Paris, Texas	0	1	78461	Drama	8.1	Travis Henders	Wim Wenders	Harry Dean Stanton
White Heat	0	0	26236	Action	8.2	A psychopathic	Raoul Walsh	James Cagney
Faa yeung nin wa	1	0	110353	Drama	8.1	Two neighbors	Kar-Wai Wong	Tony Chiu-Wai Leung
Tonari no Totoro	0	0	248206	Animatior	8.2	When two girls	Hayao Miyazaki	Hitoshi Takagi
Jodaeiye Nader az Simin	1	0	203855	Drama	8.3	A married coup	Asghar Farhadi	Payman Maadi
The Apartment	0	1	149583	Comedy	8.3	A man tries to	Billy Wilder	Jack Lemmon

Relatos salvajes	1	0	157577	Comedy	8.1	Six short storie	DamiÃ¡n Szifron	DarÃ-o Grandinetti
Hachi: A Dog's Tale	1	0	231531	Drama	8.1	A college profe	Lasse HallstrÃ¶m	Richard Gere
Det sjunde inseglet	0	0	148411	Drama	8.2	A man seeks ar	Ingmar Bergman	Max von Sydow
The Sting	0	1	226395	Comedy	8.3	Two grifters te	George Roy Hill	Paul Newman
Sunset Blvd.	0	0	185474	Drama	8.4	A screenwriter	Billy Wilder	William Holden
Before Sunset	1	0	215332	Drama	8	Nine years afte	Richard Linklater	Ethan Hawke
La battaglia di Algeri	0	0	48641	Drama	8.1	In the 1950s, fe	Gillo Pontecorvo	Brahim Hadjadj
Paths of Glory	0	0	162489	Drama	8.4	After refusing	Stanley Kubrick	Kirk Douglas
In the Name of the Father	1	1	145405	Biography	8.1	A man's coerce	Jim Sheridan	Daniel Day-Lewis
M - Eine Stadt sucht einen	0	0	132511	Crime	8.3	When the poli	Fritz Lang	Peter Lorre
PK	1	1	144700	Comedy	8.1	An alien on Ear	Rajkumar Hirani	Aamir Khan
Rebecca	0	1	113793	Drama	8.1	A self-conscio	Alfred Hitchcock	Laurence Olivier
Hotel Rwanda	1	0	319332	Biography	8.1	Paul Rusesaba	Terry George	Don Cheadle
On the Waterfront	0	0	132162	Crime	8.1	An ex-prize fig	Elia Kazan	Marlon Brando
All About Eve	0	1	112532	Drama	8.2	An ingÃ©nue i	Joseph L. Mankiewicz	Bette Davis
Mou gaan dou	1	0	113493	Crime	8.1	A story betwee	Andrew Lau	Andy Lau
La haine	1	0	135770	Crime	8.1	24 hours in the	Mathieu Kassovitz	Vincent Cassel
Network	0	0	135411	Drama	8.1	A television ne	Sidney Lumet	Faye Dunaway
Ran	0	1	102601	Action	8.2	In Medieval Ja	Akira Kurosawa	Tatsuya Nakadai
Amores perros	1	1	211174	Drama	8.1	A horrific car a	Alejandro G. IÃ±Ã¡rritu	Emilio EchevarrÃ-a
Taare Zameen Par	1	1	149539	Drama	8.4	An eight-year-	Aamir Khan	Darsheel Safary
The Bridge on the River Kwai	0	1	190597	Adventure	8.1	British POWs b	David Lean	William Holden
The Third Man	0	0	149140	Film-Noir	8.1	Pulp novelist H	Carol Reed	Orson Welles
Gangs of Wasseypur	1	1	72088	Action	8.2	A clash betwee	Anurag Kashyap	Manoj Bajpayee
Persona	0	0	90869	Drama	8.1	A nurse is put i	Ingmar Bergman	Bibi Andersson
Modern Times	0	0	198951	Comedy	8.5	The Tramp stru	Charles Chaplin	Charles Chaplin
Witness for the Prosecution	0	0	96867	Crime	8.4	A veteran Briti	Billy Wilder	Tyrone Power
The Great Dictator	0	1	187346	Comedy	8.5	Dictator Adenc	Charles Chaplin	Charles Chaplin
RashÃ´mon	0	0	140077	Crime	8.2	The rape of a b	Akira Kurosawa	ToshirÃ´ Mifune
Kaze no tani no Naushika	0	0	132383	Animatior	8.1	Warrior and pa	Hayao Miyazaki	Sumi Shimamoto
TenkÃ» no shiro Rapyuta	0	1	132196	Animatior	8	A young boy ar	Hayao Miyazaki	Anna Paquin
The Red Shoes	0	1	26532	Drama	8.2	A young ballet	Michael Powell	Anton Walbrook
Andrei Rublev	0	1	40985	Biography	8.2	The life, times	Andrei Tarkovsky	Anatoliy Solonitsyn
The Treasure of the Sierra	0	1	106438	Adventure	8.2	Two American	John Huston	Humphrey Bogart
Double Indemnity	0	0	132468	Crime	8.3	An insurance r	Billy Wilder	Fred MacMurray
Mr. Smith Goes to	0	1	100487	Comedy	8.1	A naive man is	Frank Capra	James Stewart
Les quatre cents coups	0	0	95377	Crime	8.1	A young boy, le	FranÃ§ois Truffaut	Jean-Pierre LÃ©aud
Mary and Max	1	0	154723	Animatior	8.1	A tale of frienc	Adam Elliot	Toni Collette
Dial M for Murder	0	0	144790	Crime	8.2	A tennis player	Alfred Hitchcock	Ray Milland
City Lights	0	0	153325	Comedy	8.5	With the aid of	Charles Chaplin	Charles Chaplin
Ladri di biciclette	0	0	132825	Drama	8.3	In post-war Ita	Vittorio De Sica	Lamberto Maggiorani
Judgment at Nuremberg	0	1	63794	Drama	8.2	In 1948, an Am	Stanley Kramer	Spencer Tracy
Lagaan: Once Upon a Time in	1	1	96232	Adventure	8.1	The people of	Ashutosh Gowariker	Aamir Khan
Ikiru	0	1	60908	Drama	8.3	A bureaucrat tr	Akira Kurosawa	Takashi Shimura
YÃ´jinbÃ´	0	0	101617	Action	8.2	A crafty ronin c	Akira Kurosawa	ToshirÃ´ Mifune
It Happened One Night	0	0	87015	Comedy	8.1	A spoiled heire	Frank Capra	Clark Gable
Bacheha-Ye aseman	1	0	56864	Drama	8.3	After a boy los	Majid Majidi	Mohammad Amir Naji
SmultronstÃ¤llet	0	0	87034	Drama	8.2	After living a li	Ingmar Bergman	Victor SjÃ¶strÃ¶m
The Kid	0	0	102165	Comedy	8.3	The Tramp care	Charles Chaplin	Charles Chaplin
Drishyam	1	1	27366	Crime	8.5	A man goes to	Jeethu Joseph	Mohanlal
TÃ´kyÃ´ monogatari	0	1	45947	Drama	8.2	An old couple	YasujirÃ´ Ozu	ChishÃ» RyÃ»
La passion de Jeanne d'Arc	0	0	42453	Biography	8.2	In 1431, Jeanne	Carl Theodor Dreyer	Maria Falconetti
Eskiya	1	1	57746	Crime	8.3	Baran the Banc	Yavuz Turgul	Sener Sen
Le salaire de la peur	0	1	49935	Adventure	8.1	In a decrepit S	Henri-Georges Clouzo	Yves Montand
Rang De Basanti	1	1	103493	Comedy	8.2	The story of six	Rakeysh Omprakash M	Aamir Khan
Munna Bhai M.B.B.S.	1	1	67644	Comedy	8.1	A gangster sets	Rajkumar Hirani	Sanjay Dutt
Babam ve Oglum	1	0	71125	Drama	8.3	The family of a	Ã‡agan Irmak	Ã‡etin Tekindor
Ace in the Hole	0	0	26370	Drama	8.2	A frustrated fo	Billy Wilder	Kirk Douglas
The Gold Rush	0	0	92319	Adventure	8.2	A prospector g	Charles Chaplin	Charles Chaplin
The General	0	0	74057	Action	8.1	When Union sp	Clyde Bruckman	Buster Keaton
Sherlock Jr.	0	0	35427	Action	8.2	A film projecti	Buster Keaton	Buster Keaton

Lampiran 2 Hasil Cluster dengan *Phyton*

```
Top terms per cluster:

Cluster 0 words: Year, Actor:Clint Eastwood, genre:Action, Runtime, genre:Adventure, Actor:Aamir Khan,

Cluster 0 titles: The Godfather: Part II, Heat, Scarface,

Cluster 1 words: Director:Rajkumar Hirani, Actor:Clint Eastwood, Actor:Mark Hamill, war, genre:Biography, man,

Cluster 1 titles: Modern Times, The Great Dictator, City Lights, The Kid, The Gold Rush,

Cluster 2 words: war, genre:Biography, Runtime, Director:Christopher Nolan, Actor:Daniel Day-Lewis, Director:Akira Kurosawa,

Cluster 2 titles: The Irishman, Goodfellas, Taxi Driver, The Wolf of Wall Street, Casino, The Departed, Raging Bull, Once Upon
a Time in America, Shutter Island,

Cluster 3 words: Runtime, Actor:Leonardo DiCaprio, Director:Stanley Kubrick, genre:Crime, Title, Actor:Daniel Day-Lewis,

Cluster 3 titles: The Dark Knight, Interstellar, The Dark Knight Rises, Inception, Batman Begins, The Prestige, Memento,

Cluster 4 words: Actor:Aamir Khan, Votes, Director:Stanley Kubrick, Title, man, Director:David Fincher,

Cluster 4 titles: Joker, Ford v Ferrari, Marriage Story, Gisaengchung, Avengers: Endgame, The Godfather, It's a Wonderful Life,
Die Hard, Star Wars, The Shawshank Redemption, The Shining, Star Wars: Episode VI - Return of the Jedi, Pulp Fiction, Harry Pot
ter and the Deathly Hallows: Part 2, The Matrix, Star Wars: Episode V - The Empire Strikes Back, Avengers: Infinity War, Apocal
ypse Now, The Lion King, Guardians of the Galaxy, Spider-Man: Into the Spider-Verse, Gladiator, Inglourious Basterds, Green Boo
k, Se7en, Mad Max: Fury Road, Fight Club, Forrest Gump, Schindler's List, Léon, The Silence of the Lambs, Gone Girl, Good Will
Hunting, Logan, Django Unchained, Back to the Future, Stand by Me, Once Upon a Time in the West, The Green Mile, American Histo
ry X, Requiem for a Dream, Jurassic Park, Coco, Whiplash, The Deer Hunter, Full Metal Jacket, American Beauty, Blade Runner, Al
ien, Reservoir Dogs, Catch Me If You Can, The Big Lebowski, Snatch, Terminator 2: Judgment Day, Saving Private Ryan, Three Bill
boards Outside Ebbing, Missouri, The Terminator, A Clockwork Orange, Hacksaw Ridge, One Flew Over the Cuckoo's Nest, The Prince
ss Bride, 12 Angry Men, The Grand Budapest Hotel, Prisoners, Room, No Country for Old Men, Inside Out, Raiders of the Lost Ark,
Aladdin, Kill Bill: Vol. 1, Toy Story, Braveheart, Monsters, Inc., Il buono, il brutto, il cattivo, The Help, 2001: A Space Ody
ssey, Sen to Chihiro no kamikakushi, There Will Be Blood, The Usual Suspects, Toy Story 3, Aliens, A Beautiful Mind, Spotlight,
Shichinin no samurai, The Pianist, Psycho, Unforgiven, Oldeuboi, Eternal Sunshine of the Spotless Mind, Ah-ga-ssi, Ben-Hur, Tra
inspotting, Kimi no na wa., Indiana Jones and the Last Crusade, Into the Wild, Up, Gone with the Wind, The Sixth Sense, 12 Year
s a Slave, La vita è bella, V for Vendetta, Rocky, Pan's Labyrinth, Capharnaüm, Fargo, Cidade de Deus, The Thing, Finding Nemo,
Rush, WALL·E, The Intouchables, Chinatown, Casablanca, Amadeus, Citizen Kane, Amélie, Platoon, The Lives of Others, Salinui chu
eok, Lock, Stock and Two Smoking Barrels, Warrior, The Truman Show, Stalker, Rear Window, How to Train Your Dragon, Vertigo, Ja
gten, Dead Poets Society, 3 Idiots, Lawrence of Arabia, Million Dollar Baby, Gran Torino, L.A. Confidential, Monty Python and t
he Holy Grail, Hauru no ugoku shiro, The Great Escape, Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb, So
me Like It Hot, Dangal, Per qualche dollaro in più, Hotaru no haka, Cool Hand Luke, Incendies, Metropolis, Der Untergang, To Ki
ll a Mockingbird, Mononoke-hime, Before Sunrise, Akira, El secreto de sus ojos, Das Boot, Andhadhun, Barry Lyndon, Butch Cassid
y and the Sundance Kid, Life of Brian, Idi i smotri, North by Northwest, Seppuku, Singin' in the Rain, The Elephant Man, Nuovo
Cinema Paradiso, Paris, Texas, White Heat, Faa yeung nin wa, Tonari no Totoro, Jodaeiye Nader az Simin, The Apartment, Relatos
salvajes, Hachi: A Dog's Tale, Det sjunde inseglet, The Sting, Sunset Blvd., Before Sunset, La battaglia di Algeri, Paths of Gl
ory, In the Name of the Father, M - Eine Stadt sucht einen Mörder, PK, Rebecca, Hotel Rwanda, On the Waterfront, All About Eve,
Mou gaan dou, La haine, Network, Ran, Amores perros, Taare Zameen Par, The Bridge on the River Kwai, The Third Man, Gangs of Wa
sseypur, Persona, Witness for the Prosecution, Rashômon, Kaze no tani no Naushika, Tenkû no shiro Rapyuta, The Red Shoes, Andre
i Rublev, The Treasure of the Sierra Madre, Double Indemnity, Mr. Smith Goes to Washington, Les quatre cents coups, Mary and Ma
x, Dial M for Murder, Ladri di biciclette, Judgment at Nuremberg, Lagaan: Once Upon a Time in India, Ikiru, Yôjinbô, It Happene
d One Night, Bacheha-Ye aseman, Smultronstället, Drishyam, Tôkyô monogatari, La passion de Jeanne d'Arc, Eskiya, Le salaire de
la peur, Rang De Basanti, Munna Bhai M.B.B.S., Babam ve Oglum, Ace in the Hole, The General, Sherlock Jr.,

Cluster 5 words: Runtime, Director:Rajkumar Hirani, Year, Votes, Director:Quentin Tarantino, Actor:Mark Hamill,

Cluster 5 titles: The Lord of the Rings: The Fellowship of the Ring, The Lord of the Rings: The Return of the King, The Lord
of the Rings: The Two Towers,
```

Lampiran 3 Contoh Perhitungan TF-IDF

Misal:

Sebuah data mengandung salah satu variabel yang berisi kumpulan kata (dokumen) dengan banyak 250 data, jika kata "*life*" muncul dalam dokumen sebanyak 94 kali, dimana kata tersebut muncul dalam 33 dokumen. Berapakah bobot kata "*life*" tersebut?

Perhitungan:

Sesuai dengan persamaan (3.1) digunakan rumus bobot sebagai berikut:

$$W_{dt} = tfdt * log\left(\frac{N}{df}\right)$$

$$W = 33 * log\frac{250}{93}$$

$$= 33 * \ 0.43$$

$$= \ 14.17208$$

Yang artinya kata "*life*" mempunyai bobot sebesar 14.17 dalam keseluruhan dokumen tersebut.