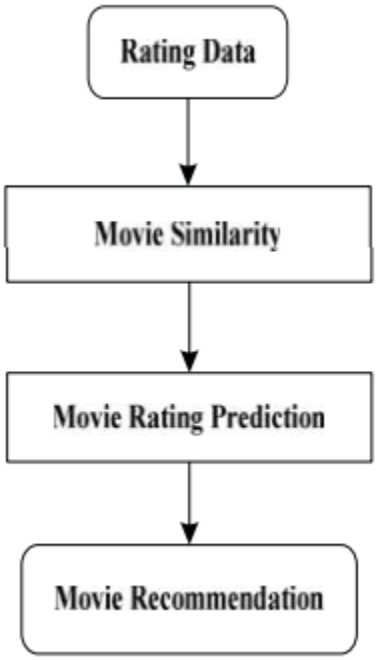
Metode Penelitian

A. Collaborative Filtering

Collaborative Filtering merupakan pendekatan yang berdasarkan pada pola rating yang diberikan oleh user terhadap film yang sudah ditonton. CF memiliki 2 fase utama untuk menghasilkan rekomendasi, yaitu similaritas film, dan prediksi rating pada film. berikut merupakan alur dalam pendekatan CF.



1. Menghitung Similaritas Film

similaritas film merupakan pendekatan untuk menghitung similaritas antara film berdasarkan rating pada film. algoritma yang diimplementasikan adalah adjusted-cosine similarity, fungsi umum yang digunakan di pendekatan collaborative filtering yaitu :

$$Sim(i,j) = \frac{\sum_{u \in U_i \cap U_j} (r_{ui} - \mu_u).(r_{uj} - \mu_u)}{\sqrt{\sum_{u \in U_i \cap U_j} (r_{ui} - \mu_u)^2} . \sqrt{\sum_{u \in U_i \cap U_j} (r_{uj} - \mu_u)^2}}$$

Input : Data Rating Film

Proses :

- menghitung rata-rata rating dari masing-masing user

- mendapatkan list user yang sudah memberi rating pada film
- implementasi algoritma adjusted-cosine similarity untuk mendapatkan similaritas antar film

Output : matrix similaritas film

2. Prediksi Rating dan Rekomendasi Film

   prediksi rating adalah tahap untuk memprediksi rating yang telah diberikan oleh user target untuk film yang belum diberi rating. hasil prediksi ini kemudian diurutkan secara descending untuk mengahasilkan list film yang direkomendasikan. prediksi rating film untuk user target dapat dihitung dengan rumus berikut :

$$\hat{r}_{ui} = \frac{\sum_{j \in Y_u(i)} Sim(i,j).r_{uj}}{\sum_{j \in Y_u(i)} |Sim(i,j)|}$$

Input : matrix similaritas film

Proses :

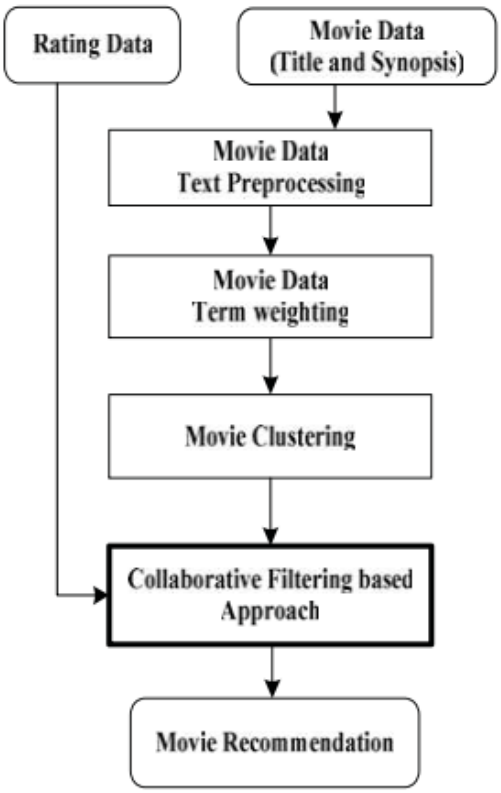
- Menentukan jumlah $\bar{k}$ untuk menentukan ukuran film yang serupa.
- Menciptakan daftar $\bar{k}$ film yang memiliki kesamaan tertinggi dengan setiap film.
- Mengimplementasikan fungsi prediksi (persamaan nomor 2) untuk menghitung peringkat film yang belum dinilai

Output : Prediksi Rating Film

B. Hybrid Filtering

Hybrid filtering merupakan pendekatan yang menambahkan manfaat Content based filtering pada pendekatan collaborative filtering. oleh karena itu, pendekatan ini juga membutuhkan data film dan proses tambahan. Pendekatan Hybrid filtering memiliki 4 fase utama untuk menghasilkan rekomendasi film, yaitu text processing, term weighting,

movie clustering, dan collaborative filtering. proses tersebut digambarkan dalam diagram berikut.



1. Teks Preprocessing

   Pra-pemrosesan teks adalah teknik untuk mengubah teks tak terstruktur menjadi bentuk yang lebih mudah dibaca untuk merepresentasikan film. disini akan diterapkan pra-pemrosesan teks untuk mengubah data film menjadi daftar istilah dalam setiap film. Data film mencakup judul dan sinopsis film. disini akan digunakan Porter Stemmer karena merupakan algoritma stemming yang paling umum digunakan untuk bahasa Inggris.

   Input : data judul film dan sinopsis

   Proses :

   - case folding, mengubah semua huruf menjadi huruf kecil(lowercase)
   - tokenisasi, memisahkan per kata, dari data yang sudah diolah diproses case folding.
   - filtering, menghapus stop words (cth: kata sambung, dan kata yang

kurang memiliki makna) dari hasil tokenisasi, sehingga yang tersisa hanya kata yang bermakna

- stemming, mendapat akar setiap kata dalam himpunan teks yang telah difilter.

Output : Daftar frasa atau istilah yang terdapat dalam setiap film.

2. Pembobotan TF-IDF

   Term weighting adalah teknik yang umum digunakan dalam studi informasi retrieval yang menghitung bobot setiap istilah dalam setiap dokumen. disini akan diterapkan teknik pembobotan TF-IDF untuk menghitung bobot setiap istilah dalam setiap film berdasarkan daftar istilah dalam setiap film.

input : Daftar frasa atau istilah yang terdapat dalam setiap film

proses :

- Term Frequency (TF), mengukur seberapa sering suatu kata kunci muncul dalam suatu item atau dokumen. contoh, Jika kata "movie" muncul 5 kali dalam suatu ulasan film yang terdiri dari 100 kata, maka TF untuk kata "movie" dalam ulasan tersebut adalah 0.05.
- Inverse Document Frequency (IDF), mengukur seberapa unik atau jarang suatu kata kunci muncul di seluruh dataset. contoh, Jika ada 1000 dokumen dalam dataset dan kata "movie" muncul di 100 dokumen, maka IDF untuk kata "movie" adalah log(1000/100)
- TF-IDF, produk dari TF dan IDF dan digunakan untuk memberikan bobot pada suatu kata kunci dalam suatu dokumen terkait dengan keseluruhan dataset. contoh, Jika TF untuk kata "movie" dalam suatu dokumen adalah 0.05 dan IDF untuk kata "movie" adalah 2, maka TF-IDF untuk kata "movie" dalam dokumen tersebut adalah 0.1.

Output : matrix pembobotan istilah

3. K-Means Clustering

Clustering adalah teknik yang mengelompokkan data yang dianggap mirip ke dalam klaster yang sama. Proses ini diperlukan karena ukuran data film yang besar dapat menyebabkan masalah skalabilitas. disini akan diterapkan teknik pengelompokan K-Means untuk mengelompokkan film berdasarkan matriks bobot istilah-film. Teknik K-Means mengelompokkan data berdasarkan jaraknya ke centroid.

Input : Jumlah kluster, matrix pembobotan istilah

proses :

- Inisialisasi nilai C centroid secara acak
- Hitung jarak setiap film ke setiap titik centroid. Tentukan setiap film ke centroid terdekatnya.
- Perbarui setiap centroid dengan mengambil rata-rata dari poin-poin film yang diassign ke klaster terkait.
- ulangi step 2 dan 3

output : C cluster

4. Collaborative Filtering

   Daftar rekomendasi film dalam pendekatan berbasis Hybrid dihasilkan dengan menerapkan pendekatan berbasis Collaborative Filtering (sebagaimana yang sudah dijelaskan sebelumnya). Perlu diperhatikan bahwa kesamaan film dihitung hanya antara film-film dalam klaster yang sama.