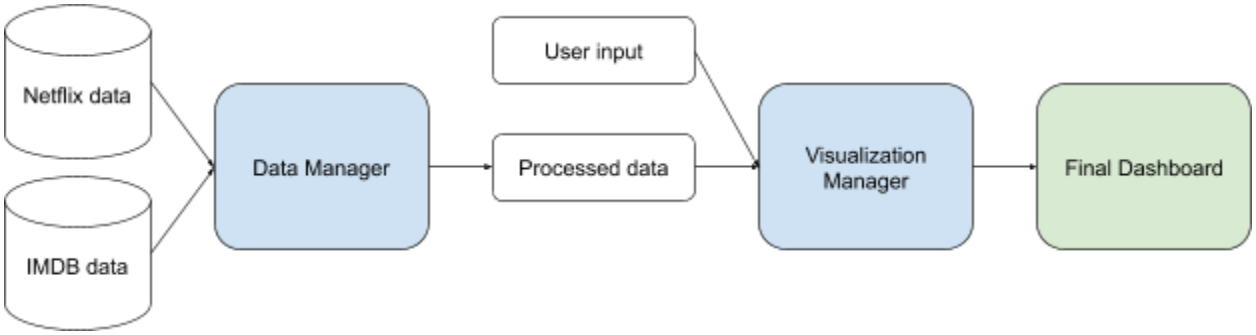
Component Specification

Software Components



Data Manager

The Data Manager loads the raw data files with the user movie ratings, the table with information about each Netflix movie, as well as the IMDB dataset. It merges and formats the data into usable dataframes, and then transforms them in order to reach a format fit for use in the final dashboards. The Netflix data will be used for the recommendation system based on an input movie, and the IMDB data for the recommendation of the most popular movies/series by year and genre.

Inputs:

- Netflix data:
    - combined_data_1.txt, combined_data_2.txt, combined_data_3.txt, combined_data_4.txt: user movie ratings
    - movie_titles.csv: movie titles and release years
- IMDB data:
    - title.ratings.tsv: title: average movie ratings and number of votes
    - title.basics.tsv: movie titles, type, release year, genre

Outputs:

- Movie recommendation dictionary: for each movie, we get the list of recommended movies.
- Association between movie_id and title.
- Average rating and number of votes by movie, release year, and genre.

Visualization Manager

The Visualization Manager is in charge of presenting the final results, based on the use case selected by the user.

- Case 1: If the user wants to have a recommendation based on an existing movie, the Visualization Manager will get the recommendations for that specific movie from the movie recommendation dictionary, and will output them as a list.
- Case 2: If the user selects a specific year, the Visualization Manager will use that to filter the aggregated data obtained from the Data Manager, and will output a list of the top 10 most popular movies or series released that year.
- Case 3: If the user selects a specific genre, the Visualization Manager will use that to filter the aggregated data obtained from the Data Manager, and will output a list of the top 10 most popular movies or series from that genre.

Inputs:

- Processed data defined as output from the Data Manager.
- User selection: base movie, year, or genre.

Outputs:
A dashboard that changes based on use case selected by the user:

- Movie recommendation based on an existing movie: List with recommended movies.
- Top movies by year: List with the most popular movies from that year.
- Top movies by genre: List with the most popular movies from that genre.

Interactions to accomplish use cases

In the first use case, a user wants to see movies similar to another movie they liked. The user will first interact with the movie recommendation dashboard to input their base movie. The visualization manager will display to the user a list of top 10 recommended movies, ordered from the highest-matching movie to the lowest, based on a query to the processed data. The processed data is obtained through an interaction with the data manager component, which produces a set of the most highly-rated movies ranked by users who also liked the base movie. In turn, the data manager component formats the raw data from Netflix to create a matrix based on the cosine similarity between movies, outputting processed data of movies with the highest similarity with the base movie.

In the second and third use cases, when a user wishes to receive a recommendation based on a particular genre or year, they will input the desired features of the movies they want to watch, and the visualization manager will display a plot of the top 10 movies based on processed data of movies matching these criteria. The data manager creates the processed data based on input data from the IMDB dataset, subset to movies matching the features from the user criteria, and generates a weighted rating calculated as the product of movie ratings and number of ratings to determine the top movies to recommend.

Preliminary plan

Research possible visualization methods.
Define the repository structure.
Create functions to load and structure the data.
Create aggregated output for the second and third use cases.
Create the movie recommendation matrix for the first use case.
Define and implement tests.
Implement the visualization method.