Project Report: MovieMatcher

Introduction: In this data science project, we built a simple content-based movie recommendation system using the TMDB movie dataset. The system recommends movies based on the similarity of their plot summaries. We used the TfidfVectorizer and sigmoid kernel from scikit-learn to calculate the similarity scores between the movies.

Overview: We started by loading the TMDB movie dataset which consists of two CSV files: two datasets. We merged the two datasets to create a single dataframe and dropped unnecessary columns. We then used the TfidfVectorizer to convert the text data from the overview column into a matrix of numerical features. We calculated the similarity between each pair of movies using the sigmoid kernel. Finally, we created a function that takes a movie title as input and returns the top 10 recommended movies based on their similarity scores.

Implementation: We used the pandas library to load and manipulate the dataset. We used the TfidfVectorizer and sigmoid kernel from scikit-learn to calculate the similarity scores. We also used the pandas series to create a reverse mapping of indices and movie titles. Finally, we created a function to recommend movies based on their similarity scores.

Summary: In this project, we built a simple content-based movie recommendation system using the TMDB movie dataset. We used the TfidfVectorizer and sigmoid kernel to calculate the similarity between each pair of movies. The system recommends movies based on their similarity scores. Although this model is not complex, it provides a good starting point for building more complex models in the future. In the future, we could improve this model by incorporating more features such as genres, actors, and directors. We could also use collaborative filtering to incorporate user preferences and ratings.