Web Scraping Report

## Introduction

The IMDB website is a popular database for movies, tv series, drama series and similar media. It houses information regarding the movie name, cast, director, year of release as well as a rating for the movie. The ratings are gathered from registered IMDB users and are pooled to give potential viewers an idea about the quality of the movie.

For this project, the movie name, year of release, rating, and URL links for the top 250 movies were automatically collected through the use of web scraping tools such as Beautiful Soup.

## Data Characteristics

The collected data are web resources written in html format. Python’s default html parser were used to analyze or organize the text into logical syntactic format.

## Data Processing

No data processing involved.

## Data Output

The scraped data was formatted into a DataFrame using the Panda’s Library and was then saved externally as a csv file.

## Project Findings

The web scraping project was able to gather data regarding the top 250 movies listed in the IMDB website. The top movie is Shawshank Redemption with an average rating of 9.2. The movie was released in 1992.

The lowest ranked movies in the top 250 have a rating of 8. In total there are 40 movies with this rating, with the latest movie to join being Jhai Bhim which was released in 2021.

The newest movie to be part of the top 250 is *Spiderman: Across the Spiderverse* which was released in 2023; the movie currently ranks 10th. The oldest movie on the other hand is *The Kid* released in 1921 which sits at rank 129.