**Crawling Implications**

**User preference insight and recommendation system optimisation:**

Rotten Tomatoes is a world-renowned movie review aggregation site that contains ratings and reviews of movies by fans and critics, and this data is ideal for analysing user preferences. By crawling the details of the movies (e.g., director, actors, genre, etc.) as well as the ratings and comments of the audience and critics, a more accurate personalised recommendation system can be built. For example, analysing highly rated user reviews in specific genres of films can help understand user preferences in specific categories, which can then be used for film recommendations.

**Market-driven analysis of film production and distribution:**

Crawling Rotten Tomatoes data can help film production companies and distributors understand market trends. By analysing data on film genres, ratings and audience's emotional response, it is possible to determine which genres are more popular and thus guide future film production and market placement. In particular, user ratings and critics' comments can directly reflect the strengths and weaknesses of a film, which is an important reference for film production, improvement, and marketing strategy development.

**User sentiment analysis and text mining:**

The content of reviews on Rotten Tomatoes is perfect for sentiment analysis and text mining to get positive and negative feedback from fans about a particular film. The review text can be analysed using Natural Language Processing (NLP) techniques to extract the popular elements of the film or the viewers' dissatisfaction. This kind of information is important for film improvement and publicity direction adjustment.

**Difference analysis between critics‘ and viewers’ reactions:**

On Rotten Tomatoes there are not only the ratings of ordinary viewers, but also the reviews of professional film critics. These data can be used to compare the differences between critics' and general audience's ratings of a film, helping to understand the relationship between professional reviews and public acceptance, so as to choose appropriate media strategies for film promotion.

**Film critic trust modelling:**

By crawling film critics‘ reviews, it is possible to build a model of film critics’ trust. Certain critics‘ reviews may be more consistent with the audience's evaluations, while other critics’ ratings may deviate from the public's preferences. By analysing this information, it is possible to provide the audience with recommendations of trusted critics and enhance the reference value of film reviews.

**Crawling Process**

The first step of the crawling task is to select the target page of Rotten Tomatoes.

Links to each film and each review on the page are extracted using the CSS selector in the Scrapy framework. Once these links are extracted, the corresponding detail pages can be further requested to get more detailed information from them.

In the movie detail page, the content to be extracted includes the movie title, description, genre, duration, director, actors, user ratings and critic ratings. This information can provide basic data for the film recommendation system and market analysis.

In the film review page, information such as the name of the reviewer, the content of the review, the date, and the rating need to be extracted. These contents can be used to build a film critic's trust model and help users choose reviews that are worth referring to.

The Rotten Tomatoes website limits the number of requests from the same IP address per unit of time. To avoid triggering this rate limitation, I set a reasonable download delay as well as randomised request intervals to simulate irregular user access patterns.

To prevent a single IP address from being blocked, I used a pool of proxy IPs to ensure that each request originated from a different IP address. This effectively bypasses the site's IP blocking mechanism and also improves the stability of data collection.

Parts of Rotten Tomatoes are loaded dynamically in JavaScript, making it impossible for traditional HTML parsing tools to capture data directly. To solve this problem, I used the Selenium automation tool to simulate a browser loading a web page and extracting dynamic content.

Finally, the crawled data needs to be saved in CSV format.

**Data Cleaning Process**

The crawled data contains redundant information or empty data, data cleansing is an essential step to remove invalid data and standardise the field format.

The process is:

1. Remove missing values (remove rows with missing comment content) respectively.

2. Remove duplicate comments.

3. fill in missing values for ratings, e.g. missing ratings can be replaced with a default value such as 0.