The Data Analysis Report of Douban Movie Short Review Dataset

1. **Significance of Crawling This Data**
   1. **Data Source**

This dataset is sourced from the **Douban Movie** platform, one of China's largest movie rating and review platforms, which boasts a large user base and rich user-generated content. On Douban, users rate movies and share personal opinions, covering various aspects of the movies, including plot, actors, directors, special effects, and storyline depth. The platform not only provides users with an interactive movie community but also offers valuable feedback data for filmmakers, marketing teams, and academic researchers.

By crawling this data, we obtained audience feedback information about movies, including movie titles, ratings, review content, likes, and other relevant fields. These data represent the real emotions and attitudes of the audience, providing rich raw material for subsequent data analysis.

* 1. **Data Scale and Value**

We used web scraping techniques to gather **280,000 movie short review data** from the Douban platform. This dataset involves a wide range of movie types and audience ratings, making it highly representative. By crawling this data, we are able to obtain authentic user feedback on various movies, including but not limited to ratings (1-5 stars), review content, like counts, and comment timestamps. This dataset provides invaluable data support for future movie analysis, recommendation system construction, and sentiment monitoring.

The significance of crawling these data lies in several aspects:

**Sentiment Analysis**:

By performing sentiment analysis on the review content, we can understand the emotional tendencies of the audience toward a movie (e.g., positive, negative, sarcastic). Sentiment analysis helps assess the public reception of a movie, providing feedback to filmmakers, helping them gauge a movie's popularity and the emotional responses it elicits from viewers. Based on the emotional tendencies of the reviews, further analysis of the movie's market performance can be conducted, offering data support for marketing strategies.

**Recommendation System Construction**:

By analyzing the content of reviews and user ratings, combined with information about the movie's theme, actors, and directors, we can build a personalized recommendation system. This system can recommend similar movies based on users' past ratings and reviews, enhancing the viewing experience. The combination of review content and ratings makes the recommendation system more precise, effectively improving user satisfaction and platform engagement.

**Market Feedback Analysis**:

By analyzing review content and rating data, movie companies can gain insights into audience feedback, enabling timely adjustments to movie promotional strategies. By tracking positive or negative sentiment trends in the comments, filmmakers can identify strengths and weaknesses in a movie, making optimizations or changes to maximize the film's market impact. Furthermore, the number of likes and interaction frequency of reviews are important indicators of the success of movie marketing.

**Academic Research**:

For researchers in sentiment analysis, text mining, and natural language processing (NLP), this dataset provides real-world textual data that can serve as a foundation for verifying and improving relevant algorithms. Using this dataset, scholars can explore sentiment classification, topic modeling, sentiment trend analysis, and other techniques, advancing research in these fields. Additionally, the text content of the dataset can be used to train and test sentiment analysis models, helping improve the application of machine learning in movie review analysis.

1. **Crawling Process**

In this section, we will detail the specific process of crawling the Douban movie short review data, including the tools and techniques used, the steps involved in data crawling, and the challenges faced during the crawling process and their solutions.

* 1. **Crawling Tools and Techniques**

To crawl movie short review data from the Douban platform, we used **Python** as the programming language, combined with the following crawling tools and techniques:

**requests**:

* The requests library was used to send HTTP requests and retrieve the webpage source code. By simulating browser behavior, it sends GET requests to fetch the HTML content of the Douban movie review pages.

**BeautifulSoup**:

* The BeautifulSoup library was used to parse the HTML content and extract movie reviews, ratings, and other relevant information. It conveniently handles HTML tags, allowing us to extract the necessary data.

**Scrapy**:

* The **Scrapy** framework was used for bulk data crawling, especially for tasks that require extracting data from multiple pages or nested web structures. Scrapy can automatically handle pagination, concurrent requests, and more, significantly improving the efficiency and stability of data crawling.
  1. **Crawling Steps**

The data crawling process can be broken down into the following major steps:

1. **Building the Crawler Script**:
   * First, we wrote the crawler script and identified the target pages for Douban movie reviews. The URL structure of these target pages is fixed, and we can access different movie review pages by modifying the page number.
   * The script uses the **requests** library to send HTTP requests and simulates a browser request by setting the **User-Agent** to avoid being detected by Douban's anti-crawling system as a bot.
2. **Sending Requests and Simulating Browser Behavior**:
   * In addition to simulating the **User-Agent**, we set headers in the request to mimic a real browser's request. This helps bypass Douban's anti-crawling mechanisms.
   * To avoid IP blocking due to frequent requests, we used an **IP proxy pool**, sending requests from multiple proxy IP addresses to distribute the load and avoid blocking a single IP.
3. **Data Extraction**:
   * We used **BeautifulSoup** to parse the HTML structure of the web pages and extract key information such as review content, movie name, user rating, and review timestamp.
   * Specifically, we focused on extracting the following key fields:
     + **Movie\_Name\_EN**: Movie's English name
     + **Movie\_Name\_CN**: Movie's Chinese name
     + **Star**: User rating (1-5 stars)
     + **Comment**: The content of the review
     + **Like**: Number of likes for the comment
     + **Date**: Timestamp of the review
     + **Username**: Reviewer's username
4. **Data Storage**:
   * To ensure the data could be easily processed and analyzed later, we stored the extracted data in **CSV** format. The CSV format allows for easy data handling, analysis, and importing into databases or other analysis tools.
5. **Pagination Handling**:
   * When crawling multiple pages, we set up a pagination mechanism. By parsing the pagination information on the page, we recursively crawled the reviews from different page numbers until all target data was collected.
   1. **Challenges and Solutions**

During the crawling process, we encountered several common challenges and implemented corresponding solutions:

1. **Anti-crawling Mechanisms**:
   * Douban has strict anti-crawling mechanisms, and frequent requests can lead to IP blocking. To address this, we used an **IP proxy pool**, randomly switching proxy IPs to avoid detection by the anti-crawling system. Additionally, we set reasonable intervals between requests (i.e., controlling the request frequency) to avoid making excessive requests and reduce the risk of blocking.
   * Besides the proxy pool, we also randomized the **User-Agent** in the request headers to simulate requests from different browsers, further improving the likelihood of successfully crawling the data.
2. **Data Extraction Accuracy**:
   * The structure of Douban's web pages can change at any time, requiring periodic checks of the crawler's accuracy and the precision of data extraction. Whenever the web structure changes, the crawler code needs to be adjusted to accommodate the new HTML tags and structure.
   * To improve the accuracy of data extraction, we used **regular expressions** and **XPath** techniques to ensure precise extraction of key fields like movie names, review content, ratings, and review timestamps.
3. **Data Integrity and Consistency**:
   * During the crawling process, some reviews may have incomplete data (such as missing ratings or empty comment fields). This was handled during the data cleaning phase. We strictly filtered the raw data, retaining only those records that contained complete and essential information.
4. **Request Frequency Control**:
   * To prevent IP blocking by Douban's anti-crawling system, we set an interval between requests. A reasonable request interval (usually a few seconds) not only helps avoid IP blocking but also prevents overloading the server, ensuring the stability of the data crawling process.
5. **Data Cleaning**

Data cleaning is a critical step to ensure the quality and consistency of the data, especially when crawling a large amount of data from multiple pages. The cleaning process removes unnecessary information, standardizes data formats, and fills in missing values, preparing the data for subsequent analysis, modeling, and visualization. This dataset originally contained 280,000 review records, but after cleaning, approximately 210,000 valid records were retained. Below are the specific steps and technical details of the data cleaning process.

* 1. **Purpose of Data Cleaning**

The primary purpose of data cleaning is to improve data quality and consistency, ensuring that noise and irrelevant data are removed. Through cleaning, we can standardize data formats, handle missing values, and preprocess the data, providing high-quality input for subsequent tasks such as sentiment analysis, recommendation system development, and visualization.

* 1. **Steps in Data Cleaning**
     1. **Removing Duplicate Data**:
* During the data crawling process, duplicate movie reviews might be retrieved. We removed these duplicate records by comparing fields such as **movie name**, **review content**, and **reviewer ID**, ensuring that each review is unique in the dataset.
  + 1. **Handling Missing Values**:
* Some fields (e.g., review content or ratings) may have missing data. For records missing **review content** (Comment) or **ratings** (Star), we used a strategy to **drop records with missing values**, ensuring that each record contains complete information. This approach guarantees data integrity and consistency.
  + 1. **Data Formatting**:
* **Date Field Handling**: The **review date** (Date) may contain various date formats. We converted the field to a standardized date format (e.g., yyyy-mm-dd) and then converted it into a Python datetime object for easier time-series analysis.
* **Rating Field Handling**: We ensured that all **ratings** (Star) are numeric and handled any invalid ratings (e.g., 0 or NaN). Invalid ratings were either deleted or corrected, ensuring the accuracy and consistency of the rating data.
  + 1. **Review Content Cleaning**:
* **Text Data**: Review content (Comment) is textual and often contains extra symbols, emojis, or HTML tags. We cleaned the review content by removing unnecessary characters to retain clean textual data, making it easier for sentiment analysis or keyword extraction.
  + We used regular expressions (re library) to remove unnecessary symbols and emojis, ensuring only text information is retained.
  + Due to the complexity of the review content, which may contain complex syntax or special characters, we applied custom cleaning rules that successfully removed most invalid information and retained meaningful reviews.
    1. **Removing Meaningless Reviews**:
* To avoid interference with subsequent analysis, we filtered out overly short or empty reviews. These reviews usually contain little meaningful information and would negatively impact the performance of sentiment analysis and other tasks. By removing these, we ensure that only useful reviews remain in the dataset.
  + 1. **Data Storage and Output**:
* The cleaned data was standardized and saved to a new **CSV file** to ensure its usability and facilitate subsequent processing.
  1. **Challenges and Strategies in Data Cleaning**
     1. **Diversity in Review Content**:

Review content is highly diverse in expression, sometimes containing emojis, slang, or complex sentence structures. This posed a challenge for text cleaning. To address this, we combined regular expressions with the **nltk**library to successfully remove most invalid characters and noise data.

* + 1. **Diversity in Sentiment Expression**:

The sentiment expressed in reviews can be very complex, with some using sarcasm, irony, or other nuanced expressions. This creates challenges for sentiment analysis. During the data cleaning process, we manually checked sample data and optimized it according to the subsequent sentiment analysis model to ensure that the cleaned data would effectively support sentiment analysis.

* 1. **Cleaned Data**

After data cleaning, about **70,000 records** were removed due to redundancy or invalid information, leaving **approximately 210,000 valid records**. These records have undergone deduplication, missing value handling, data formatting, and text cleaning, ensuring the quality and consistency of the data. This cleaned dataset is now suitable for further tasks, such as sentiment analysis, recommendation system construction, and other types of data mining.