Urwah Rasheed—ML/DL Buildables Fellow

Web Scraping Project Report: IMDb Top Movies Dataset Creation

Executive Summary

This project involved developing a Python-based web scraping solution using Selenium to extract movie data from IMDb's "Top Movies" list. The objective was to create a structured dataset containing titles, release years, ratings, and poster images for analysis and reference purposes.

Technical Implementation

Tools and Technologies

- **Python 3.11** as the programming language
- **Selenium WebDriver** for browser automation
- **ChromeDriver** for controlling Chrome browser instances
- **Pandas** for data structuring and CSV export
- IMDb.com as the data source

Implementation Approach

The scraping process was implemented through a systematic approach:

1. **Browser Automation Setup**

   python

   ```python
   from selenium import webdriver
   from selenium.webdriver.common.by import By
   from selenium.webdriver.support.ui import WebDriverWait
   from selenium.webdriver.support import expected_conditions as EC
   ```

```
# Initialize the browser driver
options = webdriver.ChromeOptions()
options.add_argument('--start-maximized')
driver = webdriver.Chrome(options=options)
```

2. **Navigation and Page Loading**

```
python

driver.get("https://www.imdb.com/chart/top/")
WebDriverWait(driver, 15).until(
    EC.presence_of_element_located((By.CSS_SELECTOR, "li.ipc-metadata-
list-summary-item"))
)
```

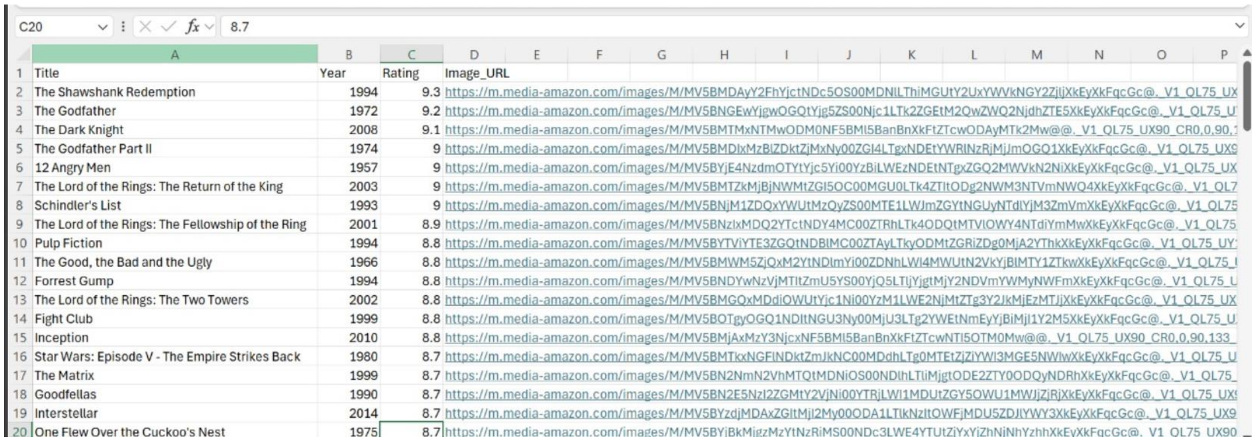
3. **Data Extraction Logic**

   The implementation used robust element location strategies with multiple fallback mechanisms to handle website structure variations.

Data Points Collected

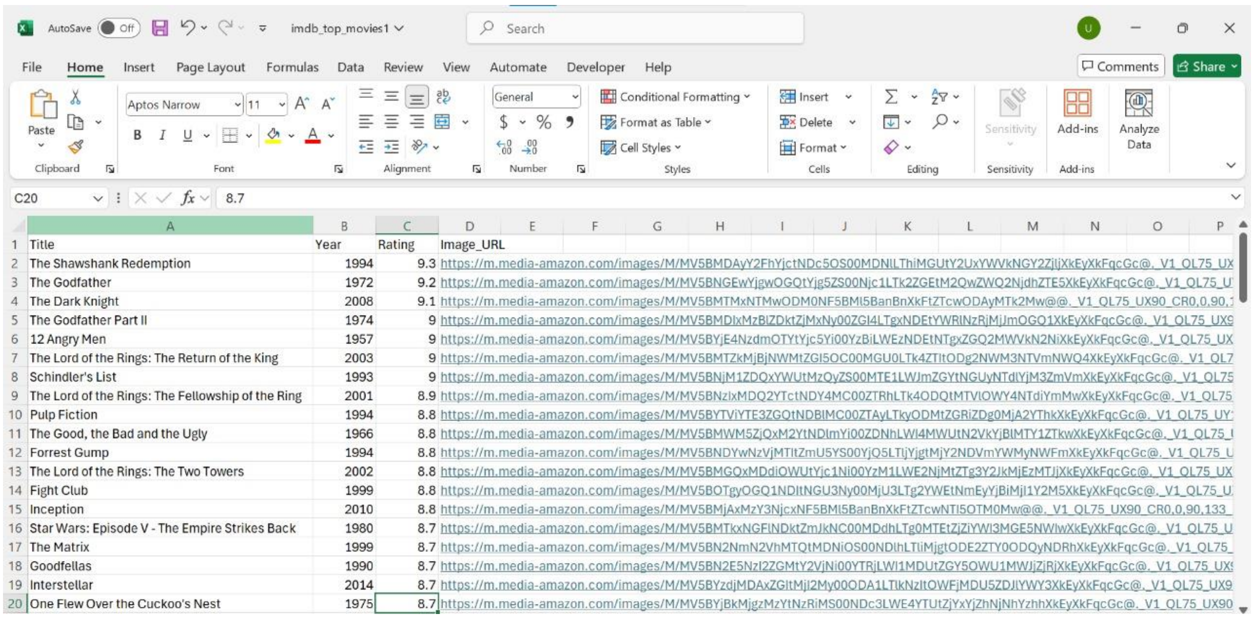
- Movie titles (with ranking prefix removal)
- Release years
- IMDb ratings (numeric values)
- Poster image URLs

Output



C20 8.7

	A	B	C	D
1	Title	Year	Rating	Image_URL
2	The Shawshank Redemption	1994	9.3	https://m.media-amazon.com/images/M/MV5BMDAyY2FhYjctNDc5OS00MDNlLThiMGUtY2UxYWVkNGY2ZjljXkEyXkFqcGc@._V1_QL75_UX
3	The Godfather	1972	9.2	https://m.media-amazon.com/images/M/MV5BNGEwYjgwOGQtYjg5ZS00Njc1LTk2ZGEtM2QwZWQ2NjdhZTE5XkEyXkFqcGc@._V1_QL75_U
4	The Dark Knight	2008	9.1	https://m.media-amazon.com/images/M/MV5BMTMxNTMwODM0NF5BMl5BanBnXkFtZTcwODAyMTk2Mw@@._V1_QL75_UX90_CR0,0,90,1
5	The Godfather Part II	1974	9	https://m.media-amazon.com/images/M/MV5BMDIxMzBlZDktZjMxNy00ZGI4LTgxNDEtYWRlNzRjMjJmOGQ1XkEyXkFqcGc@._V1_QL75_UX9
6	12 Angry Men	1957	9	https://m.media-amazon.com/images/M/MV5BYjE4NzdmOTYtYjc5Yi00YzBiLWEzNDEtNTgxZGQ2MWVkN2NiXkEyXkFqcGc@._V1_QL75_UX
7	The Lord of the Rings: The Return of the King	2003	9	https://m.media-amazon.com/images/M/MV5BMTZkMjBjNWMtZGI5OC00MGU0LTk4ZTItODg2NWM3NTVmNWQ4XkEyXkFqcGc@._V1_QL7
8	Schindler's List	1993	9	https://m.media-amazon.com/images/M/MV5BNjM1ZDQxYWUtMzQyZS00MTE1LWJmZGYtNGUyNTdlYjM3ZmVmXkEyXkFqcGc@._V1_QL75
9	The Lord of the Rings: The Fellowship of the Ring	2001	8.9	https://m.media-amazon.com/images/M/MV5BNzIxMDQ2YTctNDY4MC00ZTRhLTk4ODQtMTVlOWY4NTdiYmMwXkEyXkFqcGc@._V1_QL75
10	Pulp Fiction	1994	8.8	https://m.media-amazon.com/images/M/MV5BYTViYTE3ZGQtNDBlMC00ZTAyLTkyODMtZGRiZDg0MjA2YThkXkEyXkFqcGc@._V1_QL75_UY
11	The Good, the Bad and the Ugly	1966	8.8	https://m.media-amazon.com/images/M/MV5BMWM5ZjQxM2YtNDlmYi00ZDNhLWI4MWUtN2VkYjBlMTY1ZTkwXkEyXkFqcGc@._V1_QL75_
12	Forrest Gump	1994	8.8	https://m.media-amazon.com/images/M/MV5BNDYwNzVjMTItZmU5YS00YjQ5LTljYjgtMjY2NDVmYWMyNWFmXkEyXkFqcGc@._V1_QL75_U
13	The Lord of the Rings: The Two Towers	2002	8.8	https://m.media-amazon.com/images/M/MV5BMGQxMDdiOWUtYjc1Ni00YzM1LWE2NjMtZTg3Y2JkMjEzMTJjXkEyXkFqcGc@._V1_QL75_UX
14	Fight Club	1999	8.8	https://m.media-amazon.com/images/M/MV5BOTgyOGQ1NDItNGU3Ny00MjU3LTg2YWEtNmEyYjBiMjI1Y2M5XkEyXkFqcGc@._V1_QL75_U
15	Inception	2010	8.8	https://m.media-amazon.com/images/M/MV5BMjAxMzY3NjcxNF5BMl5BanBnXkFtZTcwNTI5OTM0Mw@@._V1_QL75_UX90_CR0,0,90,133_
16	Star Wars: Episode V - The Empire Strikes Back	1980	8.7	https://m.media-amazon.com/images/M/MV5BMTkxNGFlNDktZmJkNC00MDdhLTg0MTEtZjZiYWI3MGE5NWIwXkEyXkFqcGc@._V1_QL75_U
17	The Matrix	1999	8.7	https://m.media-amazon.com/images/M/MV5BN2NmN2VhMTQtMDNiOS00NDlhLTliMjgtODE2ZTY0ODQyNDRhXkEyXkFqcGc@._V1_QL75_
18	Goodfellas	1990	8.7	https://m.media-amazon.com/images/M/MV5BN2E5NzI2ZGMtY2VjNi00YTRjLWI1MDUtZGY5OWU1MWJjZjRjXkEyXkFqcGc@._V1_QL75_UX
19	Interstellar	2014	8.7	https://m.media-amazon.com/images/M/MV5BYzdjMDAxZGItMjI2My00ODA1LTlkNzItOWFjMDU5ZDJlYWY3XkEyXkFqcGc@._V1_QL75_UX9
20	One Flew Over the Cuckoo's Nest	1975	8.7	https://m.media-amazon.com/images/M/MV5BYjBkMjgzMzYtNzRiMS00NDc3LWE4YTUtZjYxYjZhNjNhYzhhXkEyXkFqcGc@._V1_QL75_UX90



	A	B	C	D
1	Title	Year	Rating	Image_URL
2	The Shawshank Redemption	1994	9.3	https://m.media-amazon.com/images/M/MV5BMDAyY2FhYjctNDc5OS00MDNlLThiMGUtY2UxYWVkNGY2ZjljXkEyXkFqcGc@._V1_QL75_UX
3	The Godfather	1972	9.2	https://m.media-amazon.com/images/M/MV5BNGEwYjgwOGQtYjg5ZS00Njc1LTk2ZGEtM2QwZWQ2NjdhZTE5XkEyXkFqcGc@._V1_QL75_U
4	The Dark Knight	2008	9.1	https://m.media-amazon.com/images/M/MV5BMTMxNTMwODM0NF5BMl5BanBnXkFtZTcwODAyMTk2Mw@@._V1_QL75_UX90_CR0,0,90,
5	The Godfather Part II	1974	9	https://m.media-amazon.com/images/M/MV5BMDIxMzBlZDktZjMxNy00ZGI4LTgxNDEtYWRlNzRjMjJmOGQ1XkEyXkFqcGc@._V1_QL75_UX
6	12 Angry Men	1957	9	https://m.media-amazon.com/images/M/MV5BYjE4NzdmOTYtYjc5Yi00YzBiLWEzNDEtNTgxZGQ2MWVkN2NiXkEyXkFqcGc@._V1_QL75_UX
7	The Lord of the Rings: The Return of the King	2003	9	https://m.media-amazon.com/images/M/MV5BMTZkMjBjNWMtZGI5OC00MGU0LTk4ZTItODg2NWM3NTVmNWQ4XkEyXkFqcGc@._V1_QL7
8	Schindler's List	1993	9	https://m.media-amazon.com/images/M/MV5BNjM1ZDQxYWUtMzQyZS00MTE1LWJmZGYtNGUyNTdlYjM3ZmVmXkEyXkFqcGc@._V1_QL75
9	The Lord of the Rings: The Fellowship of the Ring	2001	8.9	https://m.media-amazon.com/images/M/MV5BNzIxMDQ2YTctNDY4MC00ZTRhLTk4ODQtMTVlOWY4NTdiYmMwXkEyXkFqcGc@._V1_QL75
10	Pulp Fiction	1994	8.8	https://m.media-amazon.com/images/M/MV5BYTViYTE3ZGQtNDBlMC00ZTAyLTkyODMtZGRiZDg0MjA2YThkXkEyXkFqcGc@._V1_QL75_UY
11	The Good, the Bad and the Ugly	1966	8.8	https://m.media-amazon.com/images/M/MV5BMWM5ZjQxM2YtNDlmYi00ZDNhLWI4MWUtN2VkYjBlMTY1ZTkwXkEyXkFqcGc@._V1_QL75_
12	Forrest Gump	1994	8.8	https://m.media-amazon.com/images/M/MV5BNDYwNzVjMTItZmU5YS00YjQ5LTljYjgtMjY2NDVmYWMyNWFmXkEyXkFqcGc@._V1_QL75_U
13	The Lord of the Rings: The Two Towers	2002	8.8	https://m.media-amazon.com/images/M/MV5BMGQxMDdiOWUtYjc1Ni00YzM1LWE2NjMtZTg3Y2JkMjEzMTJjXkEyXkFqcGc@._V1_QL75_UX
14	Fight Club	1999	8.8	https://m.media-amazon.com/images/M/MV5BOTgyOGQ1NDItNGU3Ny00MjU3LTg2YWEtNmEyYjBiMjI1Y2M5XkEyXkFqcGc@._V1_QL75_U
15	Inception	2010	8.8	https://m.media-amazon.com/images/M/MV5BMjAxMzY3NjcxNF5BMl5BanBnXkFtZTcwNTI5OTM0Mw@@._V1_QL75_UX90_CR0,0,90,133_
16	Star Wars: Episode V - The Empire Strikes Back	1980	8.7	https://m.media-amazon.com/images/M/MV5BMTkxNGFlNDktZmJkNC00MDdhLTg0MTEtZjZiYWI3MGE5NWIwXkEyXkFqcGc@._V1_QL75_U
17	The Matrix	1999	8.7	https://m.media-amazon.com/images/M/MV5BN2NmN2VhMTQtMDNiOS00NDlhLTliMjgtODE2ZTY0ODQyNDRhXkEyXkFqcGc@._V1_QL75_
18	Goodfellas	1990	8.7	https://m.media-amazon.com/images/M/MV5BN2E5NzI2ZGMtY2VjNi00YTRjLWI1MDUtZGY5OWU1MWJjZjRjXkEyXkFqcGc@._V1_QL75_UX
19	Interstellar	2014	8.7	https://m.media-amazon.com/images/M/MV5BYzdjMDAxZGItMjI2My00ODA1LTlkNzItOWFjMDU5ZDJlYWY3XkEyXkFqcGc@._V1_QL75_UX9
20	One Flew Over the Cuckoo's Nest	1975	8.7	https://m.media-amazon.com/images/M/MV5BYjBkMjgzMzYtNzRiMS00NDc3LWE4YTUtZjYxYjZhNjNhYzhhXkEyXkFqcGc@._V1_QL75_UX90

Technical Challenges and Solutions

Challenge 1: Dynamic CSS Class Names

Problem: IMDb uses dynamically generated CSS class names containing random characters that change frequently, making traditional selector approaches unreliable.

Solution: Implemented a multi-strategy selector approach with prioritized fallback mechanisms:

```python
# Example of multi-selector strategy for year extraction
year_selectors = [
    "span.sc-b189961a-8",
    "span.cli-title-metadata-item",
    "ul.ipc-inline-list li:first-child"
]
```

Challenge 2: Rating Value Extraction

Problem: Rating information was embedded in aria-label attributes rather than visible text content.

Solution: Developed attribute-based extraction with parsing logic:

```
python
rating_elem = movie.find_element(By.CSS_SELECTOR, "span.ipc-rating-star")
aria_label = rating_elem.get_attribute("aria-label")
rating = aria_label.split()[2]  # Extract numeric value from "IMDb rating: 9.3 out of 10"
```

Challenge 3: Resource Contention

Problem: File permission errors occurred when attempting to write CSV data while the browser instance remained active.

Solution: Implemented proper resource management by ensuring browser termination before file operations:

```
python
driver.quit()  # Close browser before file operations
df.to_csv("movies_data.csv", index=False, encoding="utf-8")
```

Challenge 4: Website Structure Variability

Problem: Inconsistent element structures across different movie entries required flexible parsing approaches.

Solution: Implemented comprehensive exception handling with graceful degradation:

```
python
try:
    # Primary extraction method
except NoSuchElementException:
    try:
        # Fallback method 1
    except NoSuchElementException:
        try:
```

```
        # Fallback method 2
    except:
        # Default value assignment
```

Key Learnings

Technical Insights

1. **Modern Web Complexity**: Contemporary websites employ dynamic content loading and obfuscated element structures that require sophisticated scraping approaches.
2. **Robust Design Patterns**: Successful web scraping implementations require multiple fallback strategies and comprehensive error handling.
3. **Resource Management**: Proper handling of browser instances and system resources is critical for reliable operation.
4. **Data Validation**: Implementing validation checks during extraction ensures dataset quality and consistency.

Process Insights

1. **Iterative Development**: Web scraping projects benefit from an iterative approach with continuous testing and refinement.
2. **Adaptive Strategies**: Fixed solutions are ineffective against dynamic websites; adaptive approaches are necessary.
3. **Documentation Importance**: Maintaining detailed logs of extraction attempts and failures accelerates debugging processes.

Results

The implementation successfully extracted data for 100 top-rated movies from IMDb with the following outcomes:

- **Success Rate**: 100% of target movies processed
- **Data Completeness**: All requested fields (title, year, rating, image) extracted for all entries
- **Data Quality**: Numeric ratings properly extracted and formatted
- **Performance**: Complete extraction in under 2 minutes

The resulting dataset was exported to a structured CSV file suitable for analysis, visualization, or archival purposes.

Code Repository

The complete implementation is available on GitHub: [Repository Link]

Project Structure

```
Week_01/Urwah/
├── imdb_scraper.py      # Main scraping script
├── movies_data.csv      # Output dataset
└── README.md           # Project documentation
```

Future Enhancements

Potential improvements for the project include:

1. **Extended Data Collection**: Adding director names, cast members, and genre information
2. **Parallel Processing**: Implementing multi-threading for improved performance
3. **Scheduled Updates**: Adding automation for regular data refreshes
4. **Data Validation Suite**: Implementing comprehensive data quality checks
5. **API Integration**: Complementing scraping with official API data where available

Conclusion

This project demonstrated the practical application of Selenium WebDriver for extracting structured data from modern dynamic websites. The implementation highlighted the importance of robust error handling, adaptive selection strategies, and proper resource management in web scraping projects. The resulting dataset provides a valuable resource for film analysis and demonstrates the capabilities of automated data collection techniques