OSTBAYERISCHE
TECHNISCHE HOCHSCHULE
REGENSBURG

Vladyslav Keidaliuk

IMDB SOFTWARE OF HOLLYWOOD ACTORS AND ACTRESSES. TASK 1

Faculty:	Informatik und Mathematik
Degree:	Computer Science
Submission deadline:	26.11.2024
Supervisor/examiner:	Istvan Lengyel

Table of Contents

Problem Overview

The goal of this project is to develop user-friendly software to manage and display information about the top 50 popular Hollywood actors and actresses. The software will provide various functionalities, including listing actors and actresses, detailed information about them, their movies, awards, movie genres, average movie ratings, and top-performing movies. This report outlines the approach, tools and data structures chosen for the implementation.

The following functionality needs to be implemented:

1. List of all available actors and actresses
2. About the actor/actresses
3. All time movie names and years
4. Awards to actor/actresses in different years
5. Movie genre of actor/actresses
6. Average rating of their movies (overall and each year)
7. Top 5 movies, their respective years and genre

The problem is that it is necessary not only to implement the application, but also to solve the problem of obtaining data from the site.

Proposed Solution

Tools and Technologies

1. **Selenium WebDriver**: Selenium WebDriver will be used for web scraping to extract data from IMDb as there is no free official API for IMDb data, but there are third-party ones that do not provide the functionality I need. Selenium is robust, allows for dynamic content extraction, and handles JavaScript-rendered pages effectively, unlike libraries such as BeautifulSoup, which are limited to static content. Additionally, getting all the information you need would require a lot of repeated actions (like clicks) because some of the information is not visible in the DOM without these actions.

2. **Tkinter / CustomTkinter**: These libraries will be used to create a graphical user interface (GUI). Tkinter is lightweight and built into Python, while CustomTkinter provides a modernized appearance and more customization options, ensuring a user-friendly experience.
3. **JSON for Data Storage**: This lightweight data-interchange format is suitable for small-scale storage and can easily handle hierarchical data like actor profiles. For this project,

JSON will be used to store data related to actors, their profiles, movies, genres, awards, and ratings. Each actor's profile will be structured as a JSON object with nested fields for movies, awards, and other relevant details. This approach ensures that data is clear and easy to retrieve, even with complex relationships between entities. But there is also the possibility of using MySQL | MongoDB + SQLAlchemy.

Data Structures

1. **Dictionaries**: For organizing scraped data into key-value pairs (e.g., actor names as keys, their details as values).
2. **Lists**: To store collections of movies, awards, and genres.
3. **JSON Objects**: For data storage and exchange.

Steps to Implementation

1. **Data Collection**: Use Selenium to scrape IMDb data.
2. **Data Storage**: Save the cleaned data in JSON files or insert it into MongoDB collections.
3. **GUI Development**: Design a Tkinter/CustomTkinter interface with menus, search functionality, and display widgets.
4. **Feature Implementation**: Implement functions for listing actors, displaying profiles, calculating ratings, and identifying top movies.
5. **Finalization**: Package the software into an executable format for easy distribution.

Using the described tools and data structures, the project will create a user-friendly and functional application that meets all the stated requirements.

Conclusion

In conclusion, this project aims to develop a comprehensive, user-friendly software application for managing and displaying detailed information about the top 50 Hollywood actors and actresses. By utilizing powerful tools like Selenium WebDriver for web scraping, Tkinter/ CustomTkinter for creating a visually appealing and interactive GUI, and JSON for efficient data storage, the project addresses key challenges in data collection, organization, and presentation.

The chosen approach ensures that dynamic content from IMDb can be extracted reliably and efficiently, enabling users to easily access a variety of information such as actor profiles, their movies, awards, genres, and ratings. With data being stored in a structured format (JSON),

we have ensured both flexibility and scalability, allowing for future enhancements and the addition of new features.

The user-friendly interface created using Tkinter/ CustomTkinter will provide intuitive navigation and a smooth experience, empowering users to explore detailed actor data, view movie lists, and analyze performance trends.