CS50's Introduction to Programming with Python
Final Project: Extract metadata from Instagram

Pengyu Jin(Gyro)
Shanghai, China

Github: Pengyu-Jin
edX: Annie_Leonhart

December 6, 2024

Background

The information means everything

- In this era dominated by short videos and traffic, information means everything. Therefore, acquiring useful information and processing important data are fundamental skills for a programmer.
- I've long heard of terms like web scraping and script.
- Inspired by the Youtuber Hany's CS50P final project[1].

Figure: generated by DALL · E

Introduction

This script automates the process of fetching Instagram posts and their associated metadata, saving them locally, and storing the extracted information in a SQLite database[1]. It leverages the following mainly Python libraries:

- **Selenium**: For interacting with the Instagram website to save user cookies.
- **Instaloader**: For accessing and downloading Instagram profile data and posts.
- **SQLite3**: For persisting the metadata of downloaded posts.
- **Requests**: For downloading videos directly from URLs.
- **OS**: For filesystem operations like creating directories and saving files.

The script is user-friendly, requiring minimal manual intervention, and provides a clean structure for downloading and storing Instagram data.

[1] Take .db as the extension

Usage

- When you use the script for the first time, you need to log in manually to save the cookies.🙂
- For subsequent runs, the script will automatically load the locally saved cookies in JSON format, allowing you to extract data directly without manual intervention.😍

Convenient for multiple inputs

When running the script with different Instagramer, separate folders will be generated for each account, making management more convenient.

Each folder contains post folders and database file with the same name.

Let Me Show You!

Features

1. Cookie Management
   - Saves user cookies using a browser session
   - Loads cookies into ‘Instaloader‘ for authenticated access
2. Database Integration
   - Stores post metadata such as title, hashtags, URL, likes, comments, etc., in a SQLite database.
3. Post Downloading
   - Downloads videos, single image, and multi-image posts.
   - Organizes each Instagramer’s metadata into separate directories.
4. Customizations
   - Allows the user to specify the Instagram account and the number of posts to download.

Lessons Learned

- As the project progressed, the number of desired features kept increasing, leading to more and more code, which easily became messy.😵‍💫
- Implementing each feature required reading the relevant official library documentation. However, due to the library's extensive functionality and sheer volume of information, it often felt overwhelming.

Initially, I planned to create a downloader for X(formerly Twitter) videos and images. However, the strict anti-scraping measures, network restrictions, and the instability of VPNs made automation extremely challenging.

Eventually, I decided to switch to **Instagram**.

Conclusion

Thank you Professor Malan and the CS50 team for making this course available for free.🥳

It was a great Journey of knowledge for my lifetime.

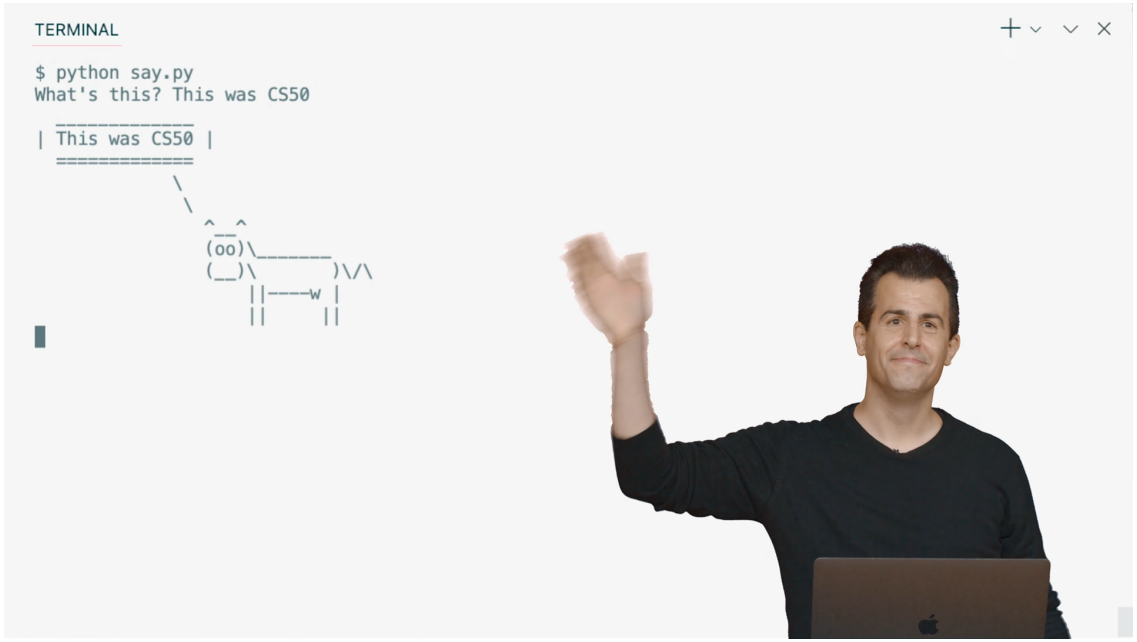

Figure: CS50P-Lecture 9 screenshot

Reference

[1] Hany. *CS50P Final Project: Instagram Scraper*. Website. https://www.youtube.com/watch?v=rD8VCxQsC5w&ab_channel=Hany. 2023.

[2] Instaloader. *Instaloader is a tool to download pictures (or videos) along with their captions and other metadata from Instagram.* https://instaloader.github.io/index.html.

[3] Adam Noel. *UBC blue beamer theme*. Website. https://ramblingacademic.com/2015/12/08/how-to-quickly-overhaul-beamer-colors/. 2015.

[4] Selenium. *Selenium is a suite of tools for automating web browsers.* https://www.selenium.dev/documentation/overview/.

Thank You!