Home Assignment Report: Video Downloading and Metadata Extraction – part 1

Overview

In this assignment, I explored several news and media websites to download videos and extract their metadata. I used various tools and approaches, including manual investigation with browser developer tools, FFmpeg for video handling, and AI assistance (ChatGPT) for understanding streaming protocols and metadata extraction methods.

CNN Video

Approach

I started by inspecting the CNN webpage's HTML source to locate a direct <video> element or MP4 file but found none. Using browser developer tools, I monitored network requests during video playback and noticed the actual video content loaded only after starting playback. The video segments appeared as multiple small MP4 chunks, which were too small individually to be useful.

Based on this observation, I consulted ChatGPT, which explained that CNN delivers video via HLS streaming using .m3u8 playlist files — plain text files that reference segmented media files. Using this insight, I learned I could download the full video by feeding the .m3u8 URL to FFmpeg.

Tools and Libraries

I chose **FFmpeg**, a widely-used, cross-platform multimedia framework capable of processing video streams from .m3u8 playlists and converting or merging them into single video files. I downloaded FFmpeg and added it to my system PATH for command-line access.

Metadata Extraction

Most metadata such as publication date was inferred from the URL structure, while categories/tags were not directly visible on the page. ChatGPT helped me discover that CNN embeds metadata in <script type="application/ld+json"> blocks using JSON-LD schema. Inspecting this JSON-LD revealed additional metadata fields, including video categories and higher resolution video link, which I eventually used instead of the video I extracted with the ffmpeg.

Mako Video

Approach

The Mako video was embedded differently, and the direct .m3u8 playlist trick used for CNN did not apply straightforwardly. Inspecting the JSON-LD metadata gave basic video information, but downloading required accessing the streaming playlist.

Using the same FFmpeg method, I located the master .m3u8 playlist URL, which references variant streams at multiple resolutions. Downloading via the master playlist gave the best quality video.

Challenges

Unlike CNN, the video segments were served with dynamic URLs, sometimes represented as blob: URLs in the network tab — these cannot be directly downloaded. The .m3u8 playlists were essential to access the stream.

CBS Sports Video

Initial Findings

The CBS Sports video page also presented only a master .m3u8 playlist, similar to previous sites.

Overcoming Geo-Restrictions

Both Fox and CBS Sports sites possibly restricted video content by region (United States only). To access these, I used **ProtonVPN**, a free VPN service, to simulate a US-based IP address.

Fox Video Retrieval

With the VPN enabled, I was able to locate a direct MP4 video URL within the JSON-LD metadata script for the Fox video, simplifying download.

CBS Sports Video Retrieval

The CBS Sports .m3u8 link was more complex and included CMCD (Common Media Client Data) headers, which complicate direct downloads. I copied the network request details and used ChatGPT to help me construct the proper FFmpeg command, including the full .m3u8 URL with query parameters to successfully download the stream.

Limitations and Edge Cases

- **Geo-restrictions:** Videos may only be accessible from specific regions, necessitating VPN usage.

AI Contribution

- **Protocol and format understanding:** ChatGPT helped clarify the nature of .m3u8 playlists and how segmented video streams work.
- **Tool usage advice:** It guided the correct use of FFmpeg commands for downloading and merging video segments.
- **Metadata extraction tips:** Helped identify JSON-LD scripts as a source of rich metadata beyond visible page content.

Summary

Using a combination of browser inspection, FFmpeg streaming tools, VPN access, and AI assistance, I successfully downloaded and extracted metadata from several news video sources. I documented the rationale behind each tool and approach, addressed potential edge cases, and explained AI's role in accelerating and clarifying the solution.