Responsible Use of Web Mining Techniques: Insights from a BoxNovel to EPUB Conversion Project

Advait Deochakke - 20BCE1143

Review 3 – Web Mining

Introduction

This project aims to explore the field of web mining through the development of a Python script that converts web novel chapters from the BoxNovel website to the EPUB ebook format. The project seeks to examine standardization on the web, including adherence to HTML and EPUB standards, and ethical considerations related to data extraction. The author will investigate best practices for web scraping, data extraction, and ebook creation techniques. Through this project, the author hopes to contribute to the ongoing discussion about responsible use of web mining techniques and generate insights and recommendations for future research in this field.

What is an Epub?

An EPUB file is a popular file format used for storing eBooks and other types of content. EPUB stands for electronic publication and was named the official standard of the International Digital Publishing Forum (IDPF) in September 2007. EPUB files can store words, images, stylesheets, fonts, metadata details, and tables of content. They are considered layout agnostic, meaning that screen size doesn't affect the formatting—EPUB files can display content on screens as small as 3.5″, making it well suited for devices such as the Amazon Kindle. This and the fact it's a freely available standard is why a majority of eReaders support EPUB files.

What is a webnovel?

A web novel is a novel that is published online. It can be free-to-read or pay-to-read. Web novels are usually serialized, which means the author publishes the story in installments — often one chapter at a time and usually released on a schedule. This allows writers to interact directly with their audience and readers can comment on each new addition to the story.

Web novels have become extremely popular in recent years, with many websites and platforms dedicated to publishing and sharing them. They offer a new and exciting way for writers to share their work with a global audience and for readers to discover new stories and authors.

Prominent websites include WuxiaWorld, RoyalRoad, and Webnovel for English language books; WoopRead and Kakao for Korean languages ones; and Webnovel, ZhongHeng, and 17k for Chinese language ones. The vast majority of English webnovels are translated from Chinese or Korean, and this gives the books a unique feel compared to Western literature.

How does the script work?

The script works by first prompting the user to enter the link to the novel on the Boxnovel website. The link can be to the novel's main page or any chapter page. The script then sanitizes the link and exits if it is invalid. Next, the user is prompted to enter the desired chapter range. If the user enters -1 for either the start or end chapter, the script will retrieve the full novel. Otherwise, it will retrieve only the specified range of chapters.

The script then retrieves the novel's title and cover image from the Boxnovel website and creates an Epub book object with this information. It also sets the book's language to English and adds a table of contents, and a book spine.

Next, the script enters a loop that iterates over each chapter in the specified range. For each chapter, it retrieves the chapter text from the Boxnovel website and formats it as an HTML string. The script also retrieves and formats the chapter title and subtitle if present. The formatted chapter text is then added to the Epub book object as a new chapter.

Once all chapters have been added, the script writes the Epub book object to a file with a name based on the novel's title and stores it in the current working directory.

For keeping the code relatively clear, we use a separate ‘my_funcs.py’ file to write custom functions. The functions in ‘my_funcs.py’ generally perform tasks related to retrieving and processing data from the Boxnovel website. These tasks include checking if a file exists, getting the current time, formatting chapter titles, retrieving and parsing web pages, downloading images, retrieving chapter file names, validating titles for use in Windows file names, extracting data from URLs, and sanitizing user input.

General logic behind processing Chapters into text

In the loop that iterates over each chapter in the specified range, the script retrieves the chapter text from the Boxnovel website and formats it as an HTML string.

First, the script navigates to the next chapter on the Boxnovel website using the `page` variable. It then retrieves the page's HTML content and parses it using BeautifulSoup.

Next, the script retrieves the next chapter link from the parsed HTML content and updates the `page` variable accordingly. This ensures that the script will navigate to the correct chapter on the next iteration of the loop.

The script then retrieves and formats the chapter title and subtitle if present. It does this by searching for `h1`, `h2`, or `h3` elements within a `div` element with class `"text-left"`. If one of these elements is found, its text content is extracted and formatted as a title or subtitle. The formatted title or subtitle is then appended to the `chapter_text` list as an HTML string.

Next, the script retrieves all `p` elements within the same `div` element with class `"text-left"`. These elements contain the main text of the chapter. The script iterates over each `p` element and appends its HTML content to the `chapter_text` list.

Once all relevant text has been retrieved and formatted, the script concatenates all strings in the `chapter_text` list to form a single HTML string representing the entire chapter. This HTML string is then added to the Epub book object as a new chapter.

Chapter Text processing – Standardization on the internet

One of the key challenges we encountered when writing the script is the issue related to parsing the correct HTML tags from the correct places to form a readable and coherent whole. Since the website's, and the original publisher's only goal is presenting the text on a webpage, formatting and standardization happens very leniently. This gives rise to issues such as the chapter's first few lines being forced into the title tag accompanied by a line break, no rules for what tag the title should have, the completely arbitrary inclusion or omission of translator names, and even completely different tag classes between chapters of the same novel.

The issue of unstructured data and non-standardized formats is a fundamental problem in web mining. This can lead to problems with data accuracy, integrity, and consistency, as well as issues with data retrieval and management. In certain versions of our code, this lead to chapter titles being so long that the program raised an error, and the novel titles having characters or naming which mismatched with the Windows file naming conventions.

In general, the lack of standardization across websites can pose a significant challenge for data mining and analysis. Without common formatting standards, data from different sources may be challenging to compare or integrate. This can be especially problematic when data is being collected from multiple websites, each with its own unique formatting conventions and data structures. This problem is long recognized in the industry and has given rise to Big Data mining tools such as Google's MapReduce and subsequently the open-source Hadoop. AWS and Microsoft's Azure suite further have specific tools available to deal with this issue.

Moreover, web mining activities can raise ethical concerns around privacy, data ownership, and consent. As web mining involves collecting data without explicit user consent, it can infringe on individual privacy rights. Furthermore, the use of data for purposes other than those for which it was originally collected can raise issues around data ownership and consent. As such, it is important to consider ethical considerations when engaging in web mining activities and to ensure that appropriate measures are taken to protect user privacy and obtain informed consent where necessary.

Concerns and delving into ethics and morality of web scraping

Before delving into the specifics of web scraping, it is important to acknowledge the ethical and legal concerns that arise when harvesting data from websites. While these issues may be beyond the scope of the current project, it is vital to recognize their significance in the broader context of data collection and usage. In an era where personal data is more valuable than ever, web scraping presents a unique set of challenges that must be approached with care and consideration.

Web scraping, which involves the automated extraction of data from websites, can raise concerns about the legality and ethicality of the practice. One of the most significant issues with web scraping is that it may violate the terms of service or privacy policy of the websites being scraped. These policies are put in place by website owners to protect their intellectual property rights, user data, and other sensitive information.

When a website's terms of service or privacy policy explicitly prohibit web scraping, doing so can result in legal consequences such as civil lawsuits, injunctions, or even criminal charges. Violating these policies can also damage the reputation of the scraper and their organization, as it can be viewed as a breach of trust and privacy.

Furthermore, web scraping can also result in unintended consequences, such as disrupting the normal operation of websites, reducing their performance, or causing technical issues. This can result in inconvenience and financial loss for the website owners and users alike.

Therefore, it is important for web scrapers to understand and respect the policies of websites they are scraping, and to obtain proper consent before doing so. In some cases, scraping may be legal or ethical, such as when the data being scraped is publicly available and used for non-commercial or academic purposes. However, even in these cases, it is essential to carefully consider the potential implications and risks of scraping, and to ensure that the practice is conducted in a responsible and transparent manner.

Web scraping can also raise concerns about the infringement of intellectual property rights or copyrights held by the website owners or content creators. Intellectual property rights include patents, trademarks, and copyrights, which are legal protections for original works of authorship, inventions, and other forms of creativity.

When web scraping involves the extraction of copyrighted content, such as text, images, or videos, without proper authorization or attribution, it can lead to infringement of the content creator's copyright. This can result in legal consequences such as civil lawsuits, injunctions, or even criminal charges.

Moreover, web scraping may also infringe on the intellectual property rights of website owners, such as their trademarks or trade secrets. For example, scraping product descriptions, pricing information, or customer reviews from e-commerce websites can lead to the violation of trade secrets and unfair competition laws.

Therefore, web scrapers must be aware of the legal implications of infringing on intellectual property rights and copyrights, and ensure that their scraping activities do not lead to such violations. It is essential to obtain proper authorization and attribution for the content being scraped, and to respect the website owners' trademarks and trade secrets.

One major concern is that web scraping can overload the website's infrastructure and cause operational disruptions. This can result in inconvenience and frustration for users, as well as reputational damage for the organization involved.

Another issue is the potential for web scraping to collect personally identifiable information (PII) that is protected by data protection regulations such as GDPR or CCPA. This can result in serious legal and ethical implications for the scraper and the organization, as well as harm to the individuals affected.

Additionally, web scraping can be used for malicious activities such as spamming, phishing, and fraud. This can cause discrimination or bias, resulting in serious legal and ethical consequences.

To avoid these issues, web scrapers must take the necessary precautions to ensure that their activities are responsible and transparent. This can include measures such as rate limiting, respecting robots.txt files, obtaining consent from users, and anonymizing or de-identifying data where necessary.

Prominent cases regarding web mining

In 2000, eBay filed a lawsuit against Bidder’s Edge, an online price comparison website for consumers, for scraping content from eBay’s website. The court sided with eBay and issued an order preventing Bidder’s Edge from scraping eBay’s content again. The main argument eBay won over was that Bidder’s Edge was exhausting their system and others following Bidder’s Edge could cause more harm to eBay’s system.

In another case, in 2009, Facebook sued Power Ventures for scraping content from its websites that its users had uploaded. This set an example for a case where web scraping was evaluated from an intellectual property standpoint. The court sided with Facebook and ordered a fiscal penalty for Power Ventures.

On the other hand, in a case between LinkedIn and hiQ Labs, the U.S. Ninth Circuit Court of Appeals ruled that web scraping is legal. HiQ Labs is a data company that uses data scraped from public sections of LinkedIn to create reports for corporate customers. LinkedIn attempted to stop hiQ Labs from accessing its user profiles and claimed that web scraping endangers user privacy. However, the court affirmed its 2019 preliminary injunction stopping LinkedIn from blocking hiQ Labs from accessing publicly visible LinkedIn member profiles. But, the decision was later put under further review following a decision in another case, following which it was reaffirmed in April 2022. But following incessant pressure from LinkedIn, a November 2022 decision (the fourth in this case) lead to a settlement between the parties involved as the US Court took the side of LinkedIn.

This case illustrates that while web scraping can raise legal concerns, it is not always considered illegal. It is important for corporations and individuals alike to ensure that their web scraping activities are conducted in an ethical and legal manner to avoid hapless and needless legal trouble.

Best Practices

Web scraping can be a powerful tool for extracting valuable data from websites. However, it is important to ensure that your web scraping activities are conducted in an ethical and legal manner. Here are some best practices to follow when engaging in web scraping:

1. Respect the website's terms of service and privacy policy: While it may be tempting to ignore a website's terms of service and privacy policy, doing so can put you at risk of legal action or other consequences. Make sure to read these documents carefully and comply with any rules or restrictions they impose. Additionally, be aware that terms of service and privacy policies can change over time, so it is important to periodically check for updates
2. Obtain consent: Obtaining consent is particularly important when scraping personal data or data that is not publicly available. This includes data such as email addresses, phone numbers, and other personally identifiable information. Before proceeding with any scraping activities, make sure to obtain explicit consent from the data subjects.
3. Be transparent: Being transparent about your web scraping activities can help you build trust with website owners and reduce the likelihood of legal action or other negative consequences. One way to be transparent is to provide a user agent string that includes your contact information, such as an email address or phone number. This allows website owners to contact you if they have any questions or concerns.
4. Do not overload the website's infrastructure: Overloading a website's infrastructure can cause operational disruptions and even lead to legal action. To avoid this, make sure to scrape data at a reasonable rate. This may involve setting limits on the number of requests you make per second or per hour.
5. Respect intellectual property rights: Intellectual property rights can be complex, but generally speaking, it is important to obtain proper authorization and attribution for any content you scrape. This may involve obtaining permission from content creators or website owners, or providing proper attribution when using scraped content in your own work.
6. Handle data responsibly: Handling data responsibly is crucial for ethical and legal web scraping. This includes ensuring the accuracy, integrity, and security of any data you scrape. Additionally, it may involve taking measures to protect personal data and sensitive information, such as encrypting data or using secure storage systems. It is also important to be aware of any laws or regulations that govern the handling of specific types of data, such as health information or financial data.

Grey area of web mining

One more example of such a controversial web mining practice is Sci-Hub, a website that provides free access to millions of academic papers and books that are otherwise paywalled or restricted by publishers. Sci-Hub claims to support open access and to fight inequality in knowledge access by enabling researchers, students, and the general public to access scientific literature without paying fees or subscriptions. Sci-Hub obtains its content by using leaked or stolen user credentials, phishing attacks, or other methods to bypass the paywalls of publishers' websites.

Sci-Hub has been sued by several publishers in different countries for violating their intellectual property rights and has been blocked or ordered to be blocked by some courts. However, Sci-Hub has also received support from some researchers, academics, activists, and lawyers who argue that Sci-Hub is providing a valuable service for the scientific community and society at large. They contend that Sci-Hub is not infringing any rights but rather exercising fair use or fair dealing exceptions under copyright law. They also assert that Sci-Hub is promoting scientific progress and innovation by facilitating knowledge dissemination and collaboration among researchers across the world.

The case of Sci-Hub illustrates the grey area in web mining where legal and ethical norms are not clear-cut or universally agreed upon. The case also reflects the tension between two competing values: intellectual property rights versus open access. On one hand, intellectual property rights are meant to protect the interests and incentives of authors and publishers who invest time, effort, and money in producing and distributing scientific content. On the other hand, open access is meant to promote the interests and benefits of users and society who seek to access and use scientific content for research, education, or personal purposes.

The question of how to balance these two values is not easy to answer. It depends on various factors such as the nature and purpose of web mining; the type and source of content being mined; the impact and outcome of web mining; the rights and responsibilities of web miners; the rights and expectations of content owners; the rights and needs of content users; and the legal and ethical frameworks governing web mining. These factors may vary across different contexts, cultures, disciplines, domains, stakeholders, etc.

References

https://www.howtogeek.com/362592/what-is-an-epub-file-and-how-do-i-open-one/

https://medium.com/fiction-friends/whats-a-web-novel-and-why-should-you-be-excited-about-them-1181ae02be3b

https://law.justia.com/cases/federal/district-courts/FSupp2/100/1058/2478126/

https://www.natlawreview.com/article/hiq-and-linkedin-reach-proposed-settlement-landmark-scraping-case

Krotov, V., & Silva, L. (2018). Legality and ethics of web scraping.

Luscombe, A., Dick, K., & Walby, K. (2022). Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences. Quality & Quantity, 56(3), 1023-1044.

Van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. Ethics and Information Technology, 6(2), 129-140.

Sweet, C. (2018). An Introduction to Sci-Hub and Its Ethical Dilemmas. LOEX Quarterly, 45(3), 4.

https://docs.sourcefabric.org/projects/ebooklib/en/latest/

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

The Ethics of Sci-Hub (exhibita.in)

An Introduction to Sci-Hub and its Ethical Dilemmas (emich.edu)

Appendix

Code Repository : https://github.com/AdvaitDeochakke/BoxnovelToEpub