Pattern Matching-based scraping of news websites

Conference Paper · October 2020

CITATIONS
2

READS
1,060

2 authors:

Hamza moh. Salem
Innopolis University
24 PUBLICATIONS **63** CITATIONS
SEE PROFILE

Manuel Mazzara
Innopolis University
436 PUBLICATIONS **5,101** CITATIONS
SEE PROFILE

Pattern Matching-based scraping of news websites

Hamza Salem, Manuel Mazzara

Software and Service Engineering lab, Innopolis University, Republic of Tatarstan, Russia

E-mail: h.salem@innopolis.university,m.mazzara@innopolis.ru

Abstract. Web Scraping is the process of extracting content from human-readable websites in order to import it into local storage such as databases or CSV Files. The process of data extraction and its design is time-consuming requiring an analysis of the website, data representation of the objects comprising its structure (DOM), HTML tags, and the Cascading Style Sheets (CSS) classes. To support this process we aim at providing automation. In this paper, we propose a pattern mining technique to scrap news and blog websites by recognizing title and body based on a content structure pattern. This approach consists of three steps, i.e.: extracting news website structure, constructing a pattern of HTML content, and implementing the pattern as a set of rules in web scraping. Our approach is a simple, general, and straightforward way to extract articles that consist of the title, the body of any blogs, or news websites.

Keywords— pattern mining, HTML structure, web scrapping , Web data extraction, Beautiful Soup.

1. Introduction

While Web scraping is attracting increasing attention from specialists of data science, data mining, and data analysis, it is also raising controversy due to the legal issues arising around the opportunity of extracting data with ownership and copyright using automated scripts or tools [1]. Web scraping, however, can also be used to simply extract public data. Applications are many: from extracting the content of a website to use it for Data Mining, Data Indexing, to extracting offers of competitors in order to make a comparison for online analysis of E-commerce websites [2].

The process of web scraping includes three steps; First, fetching or downloading of a page and it is similar to what the browser does when you view the page using one of the scraping libraries. Second, web crawling process or fetching pages for later processing [3]. Finally, the content of the selected page will be parsed, searched, or reformatted, all data will be copied into a spreadsheet or JSON files, and so on [4]. There are two methods for web scraping; using software or writing code. Web scraping software can be categorized into two main categories on-premises(local) or cloud software. The other method is writing code done by developers, for example, you can hire a developer to build custom data extraction software for your specific requirement [5]. In this paper, we will introduce a new technique to scrap news and blog websites using pattern mining. With this technique, we have to remove human interaction by writing customized code for these websites and we managed to scarp the title and body of a news article.

This paper is organized as follows, the first section after the Introduction is the Terminology and it is a brief overview of terms related to web development technology such as HTML and

CSS. The third section examines and analyzes related work on case studies that have been done before, related to combining some Machine learning and pattern mining with scraping methods. In the fourth section, we presented methods, architecture, and the technology stack used in our solution. Our implementation is described in the fifth section by showing the real implementation and the design for all components in the system from inside and how they achieve the main goal. The sixth section shows our results based on the data that has been collected in our use case. Finally, our conclusions are drawn in the final section, and questions regarding implementing pattern mining in different categories are discussed in that section.

2. Terminology & Literature Review

Here we have some important terms we will use in the paper:
Tag: The term 'Tag' is generally understood to mean HTML tag like (p) paragraph, (div) division or (H1) header one.
div:The 'div' tag defines a division or a section in an HTML document.
Website : News, Blogs and articles website.
Pattern : Repeated or regular way in which something happens or is done [6].
Web scraping Rules : Set of rules done by developer in any scraper to scrap specific element in web page.
A growing body of literature has investigated the process of Web content extraction. Various approaches have been proposed to combine patterns or Machine learning and Web Scraping technology. In [6] the authors revisit the different existing Web Scraping approaches, categories, tools, and areas of application and show how it works with different use-cases especially HTML-aware web scraping techniques. In [7] the authors proposed an approach classifies text blocks using a mixture of visual and language-independent features and a pipeline is devised to automatically label data points through clustering where each cluster is scored based on its relevance to the web page description extracted from the meta tags, and data points in the best cluster are selected as positive training examples. They draw our attention to focus on understanding how humans can identify the main content, without recognizing the language of the website. A well-known criticism of [6] work is websites that do not visually distinguish content from other parts of the web page, the visual features such as Tag Path or CSS selectors alone are not sufficient.
Another direction for web scraping with news website was taken in [8], the authors proposed a model to perform categorization which extracts useful information for classifying a document into category by referring to URL. The main weakness in their study is that they make no attempt to create a general scraper that can work with any news website.

3. Methods

Several Web Scraping methods have been applied and documented in the literature review, from simple manual human examination and copy-pasting to client-side scripts parsing the contents of the web page into DOM tree to parsing metadata. We are here instead of implementing a new scraping method enhanced by pattern mining.
According to our observation, any news web page has some basic visual structure. The biggest text size on the page is a title and the biggest (div) character length is the body. This pattern has been taken from forty different website in three languages (Arabic, English, Russian). From this pattern, rules have been written to our scraper. The quality of the result will be evaluated using other test data from 10 different websites.
This method was chosen because it is one of the most practical ways that can be generalized in multiple fields. Patterns ins scarping are unique methods that can be extended based on the target category for a website. For example, we have analyzed website for blogs and news and we have come up with method can be applied for any website in the same category, in the same

way, it can be done for e-commerce websites or social media.
This paper focuses on how to recognize both title and body of any articles based using pattern mining in HTML content. We have used forty news and blog websites to extract this pattern and we have created a simple scraper that accepts any website as input and retrieves title and body as output. In the next sections, we will demonstrate the process of creating our scraper and shows how the rules have been written for it to be general.

4. Design & Implementation

We have used Beautiful Soup as the main component that parses content, it is Python library for pulling data out of HTML and XML files. This library provides idiomatic ways of navigating, searching, and modifying the parse tree of HTML files [10]. However, before start searching inside the HTML page you need to do requests through the code to retrieve the content, so we used Python library called Requests that allows you to send HTTP/1.1 requests extremely easily [11]. The traditional approach to data scrap data Beautiful Soup and Requests are enough to do the job. However, in our case, we are developing a rules-based scraping process and we need to make sure the scraper is smart enough to know the title and body of the article without human interaction. As seen in Figure 1, the title is the biggest text on the page and body is the biggest div length too. From this observation, we will conclude our rules in the next section for our data sample.

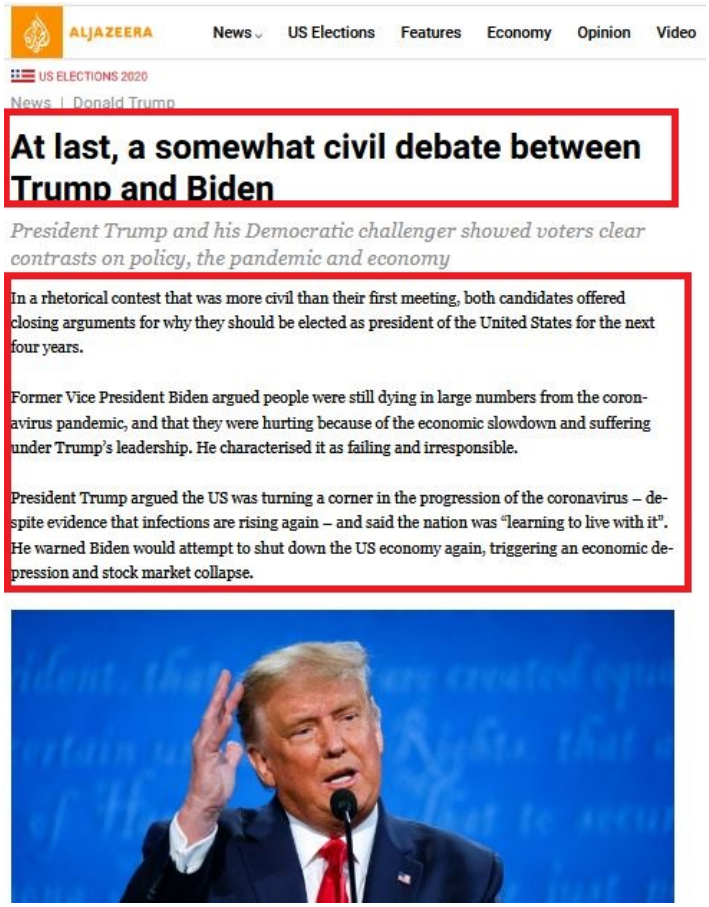

Figure 1. Example for news page shows Title and Body of an Article.

5. Results & Discussion

We had analyzed our sample websites to conclude the results shown in Table 1, "Headers" column represents how many headers in the news page, and "Header Title Order" represents how many the order of the title in that page. Finally, the last column answers the question if the biggest div in the HTML tree includes the article's body.

Table 1. Our Sample Data: Forty News and Blogs Websites

Website	Headers	Header Title Order	Is Biggest Div Length Include Body Text?
22 Websites	1	1	Yes
10 Websites	2	1	Yes
6 Websites	3	1	Yes
1 Website	more than 3	1	Yes
1 Website	0	0	Yes

As seen in Table 1, 97.5% from our sample websites follows two rules or pattern in HTML structure:

(i) First Header is the title of an article.

(ii) Biggest Div include body always.

Despite that, we had one website not following the pattern, and we have analyzed that website especially and we discovered that this website is not containing any headers and it used custom CSS to make the font size bigger for a regular tag like 'span'.

The following code in Code Snippet 1 is used to apply the first rule to get the first and biggest header on any news page.

```
for a in ["h1","h2"]:
    headers=soup.find_all(a)
    title=headers[0]
    Break
print(title)
```

Code Snippet 1: Biggest and first header

And you can add all headers from 1-6 but from our sample, we just saw that h1 and h2 are used as title only. For the body part, the following code in Code Snippet 2 applies the second rule, retrieve all (Div)s find the biggest inner div.

```
divs=soup.find_all("div")
i=0
max=1
for i in range(len(divs)):
    if divs[i]>max:
        body=divs[i]
        max=len(divs[i])
print(max)
```

Code Snippet 2: Biggest length for inner divs

This result has further strengthened our confidence in using pattern mining techniques in web scraping. Our technique shows that a smart scraper can be built to recognize the main title and body of an article by defining some rules based on the category of the website. This result can be replicated for multiple categories such as E-commerce and social media profiles. Our work share similarities with solutions and techniques presented in [6]. This paper specifically introduces pattern mining for news and blog websites and covers 97.5% of our sample. However, our findings are currently based on a limited number of websites and the analysis should be extended to a larger number as future work. This number of the data sample is slightly fewer than the value we anticipated and there is certainly room for improvement. Because we had this category and we declare the website languages from the beginning and we had limited time and limited domain to do the experiment.

6. Conclusions

In this paper, we have presented a new method to enhance web scraping for news and blog websites using pattern mining. In this method, we have used pattern mining scraper to recognize title and body in news and blog websites without any information about the page itself like HTML structure or CSS classes. By using the pattern inside our sample, this scarper can recognize the title and body of news or blog based on parameters collected when we scrap any general information such as Number of Headers, title header order, and Length for biggest div in the page. The results of this study indicate that from this pattern the first header is the title and the biggest div length is the body.

Our work clearly has some limitations. The most important limitation is a result of the fact that we had a small sample size, so caution must be taken. However, Our paper provides a blueprint to build an enhanced scraping method based on the page information and structure using pattern mining. Our investigations into this area are still in progress and seem likely to confirm our hypothesis.

References

[1] Krotov, V., Silva, L. (2018). Legality and ethics of web scraping.

[2] Parikh, K., Singh, D., Yadav, D., & Rathod, M. (2018). DETECTION OF WEB SCRAPING USING MACHINE LEARNING.

[3] Thelwall, M. (2001). A web crawler design for data mining. Journal of Information Science, 27(5), 319-325.

[4] Wed, J. (n.d.). Can Scraping Non-Infringing Content Become Copyright Infringement... Because Of How Scrapers Work? Retrieved October 20, 2020, from https://www.techdirt.com/articles/20090605/2228205147.shtml

[5] SysNucleus. (n.d.). WebHarvy Web Scraper. Retrieved October 20, 2020, from https://www.webharvy.com/articles/what-is-web-scraping.html

[6] Cooley, R., Mobasher, B., Srivastava, J. (1997, November). Web mining: Information and pattern discovery on the world wide web. In Proceedings ninth IEEE international conference on tools with artificial intelligence (pp. 558-567). IEEE

[7] Kadam, V. B., Pakle, G. K. (2014). A survey on HTML structure aware and tree based web data scraping technique. International Journal of Computer Science and Information Technologies, 5(2), 1655-1658

[8] Sundaramoorthy, K., Durga, R., Nagadarshini, S. (2017, April). Newsone—an aggregation system for news using web scraping method. In 2017 International Conference on Technical Advancements in Computers and Communications (ICTACC) (pp. 136-140). IEEE.

[9] Rozenfeld, B., Feldman, R. (2008). Self-supervised relation extraction from the web. Knowledge and Information Systems, 17(1), 17-33.

[10] Richardson, L. (2007). Beautiful soup documentation. Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018].

[11] Van Rossum, G., Drake, F. L. (2009). Introduction To Python 3: Python Documentation Manual Part 1. CreateSpace.