ScrapBook
The Web Application Based On Web Scraping

A Thesis / Project Submitted in Partial Fulfillment of the Requirements for the Degree of
Bachelor in Computer Science & Engineering

by

Shaharia Islam Rabby
CSE 053 06715

Supervised by: Adnan Ferdous Ashrafi
Lecturer

Department of Computer Science and Engineering
STAMFORD UNIVERSITY BANGLADESH

November 2017

Abstract

Internet is a source of live data that is constantly updating with data of almost any field we can imagine. Having tools that can automatically detect these updates and can select that information that we are interested in are becoming of utmost importance nowadays. That is the reason why I focus on some economic websites, studying their structures and identifying a common type of website in this field: Dynamic Websites. Even when there are many tools that allow to extract information from the Internet, not many tackle these kind of websites. For this reason I study and implement some tools that allow the developers to address these pages from a different perspective.Web scraping refers to a software program that mimics human web surfing behavior by pointing to a website and collecting large amounts of data that would otherwise be difficult for a human to extract. A typical program will extract both unstructured and semi-structured data, as well as images, and convert the data into a structured format.

Approval

The Project Report 'ScrapBook' the web application based on web scraping " submitted by Shaharia Islam Rabby ID: CSE 053 06715, to the Department of Computer Science & Engineering, Stamford University Bangladesh, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science & Engineering and approved as to its style and contents.

Board of Examiner's Name, Signature, and Date:

....	....	....
(Tamjid Rahman)	**(Zonayed Ahmed)**	**(Aiasha Siddika)**
Date:	Date:	Date:

Supervisor's Signature and Date:

....
Adnan Ferdous Ashrafi

Date:

Declaration

I, hereby, declare that the work presented in this Project is the outcome of the investigation performed by me under the supervision of Adnan Ferdous Ashra, Lecturer, Department of Computer Science & Engineering, Stamford University Bangladesh. I also declare that no part of this Project and thereof has been or is being submitted elsewhere for the award of any degree or Diploma.

Signature and Date:

....
Shaharia Islam Rabby
ID: CSE 053 06715
Date:

Dedicated to ...
Beloved Parents and Teachers

Acknowledgments

First of all, I would like to express my highest gratitude to the Almighty Allah for HIS kindness on me that made it possible for me to complete the study and preparation of this project. I also giving thank you to my parents and teachers for their dedication. Also thank you to the faculty of Computer Science and Engineering for their contribution to complete the project. I am indebted to the Supervisor of my project, Adnan Ferdous Ashrafi, Lecturer, Department of Computer Science and Engineering for giving me patient hearing, sufficient time for discussions and continuous suggestions and guidance for preparation of this project. He helped me a lot in many ways ultimately not only in preparation of this project but also to defense the project for earning the degree of B.Sc. in Computer Science and Engineering.

Table of Contents

List of Figures

1 Introduction

There is a lot of data flowing everywhere. Not structured, not useful pieces of data moving here and there. Getting this data and structuring, processing can make it really expensive. There are companies making billions of dollars just for scraping web content and showing in a nice form. There are many reasons why people and organizations want to scrape websites, and there are numerous web scraping programs available today. Organizations often seek web-based information that increases business value, including harnessing sales leads,market intelligence, news, creative content, company and sector performance data,enhanced e-commerce operations, and information for use in marketing and promotional campaigns. In the modern era of big data and the need for data and information, people and companies alike are going to great lengths to gather relevant data and information. For example, Shopzilla operates a portfolio of shopping websites that aggregate product availability and provides price comparisons for retail consumers; a half-billion-dollar company built on web scraping. Other companies like Nextag and PriceGrabber provide similar services. A quick Internet search will yield numerous web scraping tools, from free and paid desktop applications to web-browser add-ins. While many websites provide an application program interface (API) or web-services to provide data to the client, many simply do not. Even when the API or web-service is provided, many individuals and smaller organizations do not have the technology and/or programming skill resources that are available in larger organizations. For example, both an API and a web-service require the client to write their own program according to the servers protocols and specifications. In the event the client doesnt have the capacity to consume the services then web scraping may be the only alternative. And while one may program their own web scraper, there is really no need to since so many programs are already available. In this case the user simply adopts a web scraper, shows the web scraper how to navigate the website and the data to extract, and then the web scraper does the rest [1].

1.1 Data scraping

Data scraping is most often done either to interface to a legacy system which has no other mechanism which is compatible with current hardware, or to interface to a third-party system which does not provide a more convenient API. In the second case, the operator of the third-party system will often see screen scraping as unwanted, due to reasons such as increased system load, the loss of advertisement revenue, or the loss of control of the information content. Data scraping is generally considered an ad hoc, inelegant technique, often used only as a ”last resort” when no other mechanism for data interchange is available. Aside from the higher programming and processing overhead, output displays intended for human consumption often change structure frequently. Humans can cope with this easily, but a computer program may report nonsense, have been told to read data in a particular format or from a particular place, and with no knowledge of how to check its results for validity[2].

1.2 Types of data scraping

There are three ways to data scraping.

1. Screen scraping

2. Report mining

3. Web scraping

1.2.1 Screen scraping

Screen scraping is normally associated with the programmatic collection of visual data from a source, instead of parsing data as in Web scraping. Originally, screen scraping referred to the practice of reading text data from a computer display terminal's screen. This was generally done by reading the terminal's memory through its auxiliary port, or by connecting the terminal output port of one computer system to an input port on another. The term screen scraping is also commonly used to refer to the bidirectional exchange of data. This could be the simple cases where the controlling program navigates through the user interface, or more complex scenarios where the controlling program is entering data into an interface meant to be used by a human.As a concrete example of a classic screen scraper, consider a hypothetical legacy system dating from the 1960s the dawn of computerized data processing[3].

1.2.2 Web scraping

Web pages are built using text-based mark-up languages (HTML and XHTML), and frequently contain a wealth of useful data in text form. However, most web pages are designed for human end-users and not for ease of automated use. Because of this, tool kits that scrape web content were created. A web scraper is an API or tool to extract data from a web site. Companies like Amazon AWS and Google provide web scraping tools, services and public data available free of cost to end users. Newer forms of web scraping involve listening to data feeds from web servers. For example, JSON is commonly used as a transport storage mechanism between the client and the web server. Recently, companies have developed web scraping systems that rely on using techniques in DOM parsing, computer vision and natural language processing to simulate the human processing that occurs when viewing a web page to automatically extract useful information[4].

1.2.3 Report mining

Report mining is the extraction of data from human readable computer reports. Conventional data extraction requires a connection to a working source system, suitable connectivity standards or an API, and usually complex querying. By using the source system's standard reporting options, and directing the output to a spool file instead of to a printer, static reports can be generated suitable for of-line analysis via report mining. This approach can avoid intensive CPU usage during business hours, can minimize end-user license costs for ERP customers, and can offer very rapid prototyping and development of custom reports. Whereas data scraping and web scraping involve interacting with dynamic output, report mining involves extracting data from files in a human readable format, such as HTML, PDF, or text. These can be easily generated from almost any system by intercepting the data feed to a printer. This approach can provide a quick and simple route to obtaining data without needing to program an API to the source syste[5].

1.3 Purpose of the work

In this paper, I propose a news portal based web application through the web scraping system that expresses a how to work web scraping and how to use the web scraping method to generate a web application .My system is web based news portal application in a unique way to see user various news. My system build a unique system to make user happy.

Build a system that is individual to every person and make personal life a lit easer. To Gather information in a efficient way in the web method and reduce web surfing time and reduce time in gathering important news.

1.4 Objective

1. Make personal life a lit easier.

2. Build a system that is individual to every person.

3. Gather information in a better way.

4. Reduce web surfing time.

5. Reduce time in gathering important news.

2 Literature Review

2.1 Background Study

Web scraping is the process of extracting and creating a structured representation of data from a web site. A company may for instance want to autonomously monitor its competitors product prices, or an enterprising student may want to unify information on parties from all campus bar and dormitory web sites and present them in a calendar on her own web site. If the owner of the information does not provide an open API, the remedy is to write a program that targets the markup of the web page. A common approach is to parse the web page to a tree representation and evaluate an XPath expression on it. An XPath denotes a path, possibly with wildcards, and when evaluated on a tree, the result is the set of nodes at the end of any occurence of the path in the tree. HTML, the markup language used to structure data on web pages, is intended for creating a visually appealing interface for humans. The drawback of the existing techniques used for web scraping is that the markup is subject to change either because the web site is highly dynamic or simply because the look-and-feel is updated. Even XPaths with wildcards are vulnerable to these changes because a given change may be to a tag which can not be covered by a wildcard[6].

2.2 Techniques

Web scraping is the process of automatically mining data or collecting information from the World Wide Web. It is a field with active developments sharing a common goal with the semantic web vision, an ambitious initiative that still requires breakthroughs in text processing, semantic understanding, artificial intelligence and human-computer interactions. Current web scraping solutions range from the ad-hoc, requiring human effort, to fully automated systems that are able to convert entire web sites into structured information, with limitations.

2.2.1 Human copy-and-paste

Sometimes even the best web-scraping technology cannot replace a humans manual examination and copy-and-paste, and sometimes this may be the only workable solution when the websites for scraping explicitly set up barriers to prevent machine automation[7].

2.2.2 Text pattern matching

A simple yet powerful approach to extract information from web pages can be based on the UNIX grep command or regular expression-matching facilities of programming languages The phrase regular expressions (and consequently, regexes) is often used to mean the specific, standard textual syntax (distinct from the mathematical notation described below) for representing patterns that matching text need to conform to. Each character in a regular expression (that is, each character in the string describing its pattern) is understood to be a metacharacter (with its special meaning), or a regular character (with its literal meaning). For example, in the regex a. a is a literal character which matches just 'a' and . is a meta character which matches every character except a newline. Therefore, this regex would match for example 'a ' or 'ax' or 'a0'. Together, metacharacters and literal characters can be used to identify textual material of a given pattern, or process a number of instances of it. Pattern-matches can vary from a precise equality to a very general similarity (controlled by the metacharacters). For example, . is a very general pattern, [a-z] (match all letters from 'a' to 'z') is less general and a is a precise pattern (match just 'a'). The metacharacter syntax is designed specifically to represent prescribed targets in a concise and flexible way to direct the automation of text processing of a variety of input data, in a form easy to type using a standard ASCII keyboard.

A very simple case of a regular expression in this syntax would be to locate the same word spelled two different ways in a text editor, the regular expression seriali[sz]e matches both "serialise" and "serialize". Wildcards could also achieve this, but are more limited in what they can pattern (having fewer metacharacters and a simple language-base).

The usual context of wildcard characters is in globbing similar names in a list of files, whereas regexes are usually employed in applications that pattern-match text strings in general[**?**].

2.2.3 HTTP programming

Static and dynamic web pages can be retrieved by posting HTTP requests to the remote web server using socket programming.

Static web pages A static web page (sometimes called a flat page/stationary page) is a web page that is delivered to the user exactly as stored, in contrast to dynamic web pages which are generated by a web application. Consequently, a static web page displays the same information for all users, from all contexts, subject to modern capabilities of a web server to negotiate content-type or language of the document where such versions are available and the server is configured to do so. Static web pages are often HTML documents stored as files in the file system and made available by the web server over HTTP (nevertheless URLs ending with ".html" are not always static). However, loose interpretations of the term could include web pages stored in a database, and could even include pages formatted using a template and served through an application server, as long as the page served is unchanging and presented essentially as stored[8].

Dynamic web pages A server-side dynamic web page is a web page whose construction is controlled by an application server processing server-side scripts. In server-side scripting, parameters determine how the assembly of every new web page proceeds, including the setting up of more client-side processing.

A client-side dynamic web page processes the web page using HTML scripting running in the browser as it loads. JavaScript and other scripting languages determine the way the HTML in the received page is parsed into the Document Object Model, or DOM, that represents the loaded web page. The same client-side techniques can then dynamically update or change the DOM in the same way.

A dynamic web page is then reloaded by the user or by a computer program to change some variable content. The updating information could come from the server, or from changes made to that page's DOM. This may or may not truncate the browsing history or create a saved version to go back to, but a dynamic web page update using Ajax technologies will neither create a page to go back to, nor truncate the web browsing history forward of the displayed page. Using Ajax technologies the end user gets one dynamic page managed as a single page in the web browser while the actual web content rendered on that page can vary. The Ajax engine sits only on the browser requesting parts of its DOM, the DOM, for its client, from an application server[9].

2.2.4 HTML parsing

Many websites have large collections of pages generated dynamically from an underlying structured source like a database. Data of the same category are typically encoded into similar pages by a common script or template. In data mining, a program that detects such templates in a particular information source, extracts its content and translates it into a relational form, is called a wrapper. Wrapper generation algorithms assume that input pages of a wrapper induction system conform to a common template and that they can be easily identified in terms of a URL common scheme.[3] Moreover, some semi-structured data query languages, such as XQuery and the HTQL, can be used to parse HTML pages and to retrieve and transform page content[10].

Requirement and Features A HTML DOM parser written in PHP5+ let you manipulate HTML in a very easy way

- Require PHP 5+.
- Supports invalid HTML.
- Find tags on an HTML page with selectors just like jQuery.
- Extract contents from HTML in a single line.[10]

2.2.5 DOM parsing

The Document Object Model (DOM) is a cross-platform and language-independent application programming interface that treats an HTML, XHTML, or XML document as a tree structure wherein each node is an object representing a part of the document. The objects can be manipulated pro-grammatically and any visible changes occurring as a result may then be reflected in the display of the document.

The principal standardization of DOM was handled by the World Wide Web Consortium, which last developed a recommendation in 2004. WHATWG took over development of the standard, publishing it as a living document. The W3C now publishes stable snapshots of the WHATWG standard.

Document Object Model, or DOM, defines the style, structure and the contents contained within the XML files. DOM parsers are generally used by scrapers that want to get an in-depth view of the structure of the web page. One can use the DOM parser to get the nodes containing information, and then use a tool like XPath to scrape web pages[11].

3 Problem And System Overview

In this section we formally define the extraction problem and briefly overview our solution

3.1 Problem Definition

We define the problem as follows:

Given a set of labeled DOM trees D parsed from pages of a particular website, a group of wrappers (w1, w1, ..., wn) should be learned from D. And the target is to maximize the overall extraction accuracy P when generated wrappers are tested on another DOM-tree set D that comes from the same website.

In this paper, we use manually labeled training data to explain and verify our ideas. Although the idea of joint optimization of wrapper induction and template detection is not constrained to labeled data, we do so for several reasons. First, our main focus is not on the algorithm of wrapper induction but on how to detect similarity-based templates and how the detection influences extraction performance. Labeled data can simplify the evaluation of extraction results. Second, using labeled data to generate wrappers is commonly used in some scenarios such as comparison shopping. As the accuracy of price is required to be close to 100 percent, automatic attributes labeling methods cannot meet the requirement. Furthermore, inducing wrappers based on labeling data is selectively used for only a few of head sites. For each site, as few as tens of pages are enough to train a robust wrapper set. Thus, the cost of labeling is acceptable.

3.2 System Overview

A flowchart of our system is shown in Figure 3.1 To begin with, training pages are parsed into DOM trees before they are processed by our system. We will not discuss the HTML parsing technique since it is beyond the scope of this paper. Second, the DOM trees will be fed to the wrapper-oriented page clustering module that combines template detection and wrapper generation into one step and outputs a set of wrappers. A by-product in the step is that the training DOM trees are also clustered into similarity-based page classes.

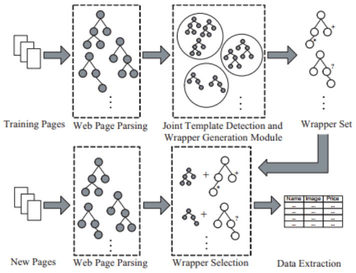


Figure 3.1: System overview

When a new Web page is introduced, it will be parsed into a DOM tree first. Then, our system can automatically select a wrapper from the generated wrapper set, which makes a best match with the DOM tree. At last, data is extracted and saved in a structured format like a relational database[12].

4 Implementation

4.1 Materials

4.1.1 Sublime Text3

Sublime Text is a proprietary cross-platform source code editor with a Python application programming interface (API). It natively supports many programming languages and markup languages, and functions can be added by users with plug-ins, typically community-built and maintained under free-software licenses. A full-featured package manager that helps discovering, installing, updating and removing packages for Sublime Text 2. It features an automatic upgrade and supports Git-hub, Bit-bucket and a full channel/repository system.

Sublime Text 3 (ST3) is lightweight, cross-platform code editor known for its speed, ease of use, and strong community support. Its an incredible editor right out of the box, but the real power comes from the ability to enhance its functionality using Package Control and creating custom settings.

Lime Text is a powerful and elegant text editor primarily developed in Go. It aims to be a free and open-source software successor to Sublime Text. Lime has a few front-ends (QML, command-line interface) that can be selectively used with the pluggable backend.

In common use, sublime is an adjective meaning ”awe-inspiring grand, excellent, or impressive,” like the best chocolate fudge sundae you’ve ever had. You might describe a spine-tingling piece of music as ”a work of sublime beauty”[13].

4.1.2 WampServer

WAMP refers to a set of free applications combined with Microsoft Windows, which are commonly used in Web server environments. WAMP is acronym for the combination of

Windows, Apache, MySQL and any one of PHP, Perl or Python. The WAMP stick up the four key elements of a Web server environments ie an operating system, Web server, database and a web scripting language. In WAMP Windows is the operating system, Apache is the web server, MySQL is the database and PHP is a scripting language (or the alternative scripting languages like Python or Perl can be used instead of PHP). The equivalent version of WAMP package for macintosh operating system is called MAMP and for the Linux operating system its called LAMP.

The acronym WAMP refers to a set of free (open source) applications, combined with Microsoft Windows, which are commonly used in Web server environments. The WAMP stack provides developers with the four key elements of a Web server: an operating system, database, Web server and Web scripting software. The combined usage of these programs is called a server stack. In this stack, Microsoft Windows is the operating system (OS), Apache is the Web server, MySQL handles the database components, while PHP, Python, or PERL represents the dynamic scripting languages[14].

4.1.3 HTML

Hypertext Markup Language (HTML) is the standard markup language for creating web pages and web applications. With Cascading Style Sheets (CSS) and JavaScript it forms a triad of cornerstone technologies for the World Wide Web. Web browsers receive HTML documents from a web server or from local storage and render them into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document.

HTML elements are the building blocks of HTML pages. With HTML constructs, images and other objects, such as interactive forms, may be embedded into the rendered page. It provides a means to create structured documents by denoting structural semantics for text such as headings, paragraphs, lists, links, quotes and other items. HTML elements are delineated by tags, written using angle brackets. Tags such as ¡img /¿ and ¡input /¿ introduce content into the page directly. Others such as ¡p¿...¡/p¿ surround and provide information about document text and may include other tags as sub-elements. Browsers do not display the HTML tags, but use them to interpret the content of the page.

HTML documents imply a structure of nested HTML elements. These are indicated in the document by HTML tags, enclosed in angle brackets thus: ¡p¿

In the simple, general case, the extent of an element is indicated by a pair of tags: a ”start tag” ¡p¿ and ”end tag” ¡/p¿. The text content of the element, if any, is placed between these tags.

Tags may also enclose further tag markup between the start and end, including a mixture of tags and text. This indicates further (nested) elements, as children of the parent element.

The start tag may also include attributes within the tag. These indicate other information, such as identifiers for sections within the document, identifiers used to bind style information to the presentation of the document, and for some tags such as the ¡img¿ used to embed images, the reference to the image resource.

Some elements, such as the line break ¡br¿, do not permit any embedded content, either text or further tags. These require only a single empty tag (akin to a start tag) and do not use an end tag.

Many tags, particularly the closing end tag for the very commonly used paragraph element ¡p¿, are optional. An HTML browser or other agent can infer the closure for the end of an element from the context and the structural rules defined by the HTML standard. These rules are complex and not widely understood by most HTML coders.

The general form of an HTML element is therefore: ¡tag attribute1="value1" attribute2="value2"¿"content"¡/tag¿. Some HTML elements are defined as empty elements and take the form ¡tag attribute1="value1" attribute2="value2"¿. Empty elements may enclose no content, for instance, the ¡br¿ tag or the inline ¡img¿ tag. The name of an HTML element is the name used in the tags. Note that the end tag's name is preceded by a slash character, "/", and that in empty elements the end tag is neither required nor allowed. If attributes are not mentioned, default values are used in each case[15].

4.1.4 CSS

Cascading Style Sheets (CSS) is a stylesheet language used to describe the presentation of a document written in HTML or XML (including XML dialects such as SVG or XHTML). CSS describes how elements should be rendered on screen, on paper, in speech, or on other media.

This module gets you started with the basics of how CSS works, including selectors and properties, writing CSS rules, applying CSS to HTML, how to specify length, color, and other units in CSS, cascade and inheritance, box model basics, and debugging CSS.[16].

Styling text Here we look at text styling fundamentals, including setting font, boldness, and italics, line and letter spacing, and drop shadows and other text features. We round off the module by looking at applying custom fonts to your page, and styling lists and links.

Styling boxes Next up, we look at styling boxes, one of the fundamental steps towards laying out a web page. In this module we recap the box model then look at controlling box layouts by setting padding, borders and margins, setting custom background colors, images and other features, and fancy features such drop shadows and filters on boxes.

CSS layout At this point we've already looked at CSS fundamentals, how to style text, and how to style and manipulate the boxes that your content sits inside. Now it's time to look at how to place your boxes in the right place in relation to the viewport, and one another. We have covered the necessary prerequisites so you can now dive deep into CSS layout, looking at different display settings, traditional layout methods involving float and positioning, and new fangled layout tools like flexbox[16].

4.1.5 Bootstrap

Bootstrap is a free and open-source front-end web framework for designing websites and web applications. It contains HTML- and CSS-based design templates for typography, forms, buttons, navigation and other interface components, as well as optional JavaScript extensions. Unlike many web frameworks, it concerns itself with front-end development only.

Bootstrap is modular and consists of a series of Less stylesheets that implement the various components of the toolkit. These stylesheets are generally compiled into a bundle and included in web pages, but individual components can be included or removed. Bootstrap provides a number of configuration variables that control things such as color and padding of various components.

Since Bootstrap 2, the Bootstrap documentation has included a customization wizard which generates a customized version of Bootstrap based on the requested components and various settings.

As of Bootstrap 4, Sass is used instead of Less for the stylesheets.

Each Bootstrap component consists of an HTML structure, CSS declarations, and in some cases accompanying JavaScript code.

Grid system and responsive design comes standard with an 1170 pixel wide grid layout. Alternatively, the developer can use a variable-width layout. For both cases, the toolkit has four variations to make use of different resolutions and types of devices: mobile phones, portrait and landscape, tablets and PCs with low and high resolution. Each variation adjusts the width of the columns[17].

4.1.6 PHP

Hypertext Preprocessor (earlier called, Personal Home Page) PHP is an HTML-embedded, server-side scripting language designed for web development. It is also used as a general purpose programming language. It was created by Rasmus Lerdorf in 1994 and appeared in the market in 1995. PHP (recursive acronym for PHP: Hypertext Preprocessor) is a widely-used open source general-purpose scripting language that is especially suited for web development and can be embedded into HTML.

PHP is mainly focused on server-side scripting, so you can do anything any other CGI program can do, such as collect form data, generate dynamic page content, or send and receive cookies. But PHP can do much more. There are three main areas where PHP scripts are used. Server-side scripting.

PHP is a general-purpose scripting language that is especially suited to server-side web development, in which case PHP generally runs on a web server. Any PHP code in a requested file is executed by the PHP runtime, usually to create dynamic web page content or dynamic images used on websites or elsewhere.

PHP developers develop programs, applications, and web sites using the dynamic scripting language PHP. PHP is known for web development and business applications. Depending on job function, PHP developers may be classified as software developers or web developers.

Secondly, apart from server side scripting, PHP web development tools can also be used from stand alone client side or command line scripting GUI applications. As PHP website development software is freely available, thus it can easily be embedded into HTML.

Basically, there is nothing that makes a language a scripting language except that it is called such, especially by its creators. The major set of modern scripting languages is PHP, Perl, JavaScript, Python, Ruby and Lua. ... So Factor is a scripting language (or at least was when that was written), but, say, Java is not[18].

4.1.7 MySql

MySQL (officially pronounced as /mał skjul/ ”My S-Q-L”,) is an open-source relational database management system (RDBMS). Its name is a combination of ”My”, the name of co-founder Michael Widenius’s daughter, and ”SQL”, the abbreviation for Structured Query Language. The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL was owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Oracle Corporation. For proprietary use, several paid editions are available, and offer additional functionality.

MySQL is a central component of the LAMP open-source web application software stack (and other ”AMP” stacks). LAMP is an acronym for ”Linux, Apache, MySQL, Perl/PHP/Python”. Applications that use the MySQL database include: TYPO3, MODx, Joomla, WordPress, phpBB, MyBB, and Drupal. MySQL is also used in many high-profile, large-scale websites, including Google (though not for searches), Facebook, Twitter, Flickr, and YouTube.

MySQL is written in C and C++. The code written in MySQL is not case sensitive.Its SQL parser is written in yacc, but it uses a home-brewed lexical analyzer. MySQL works on many system platforms, including AIX, BSDi, FreeBSD, HP-UX, eComStation, i5/OS, IRIX, Linux, macOS, Microsoft Windows, NetBSD, Novell NetWare, OpenBSD, OpenSolaris, OS/2 Warp, QNX, Oracle Solaris, Symbian, SunOS, SCO OpenServer, SCO UnixWare, Sanos and Tru64. A port of MySQL to OpenVMS also exists.

The MySQL server software itself and the client libraries use dual-licensing distribution. They are offered under GPL version 2, beginning from 28 June 2000 (which in 2009 has been extended with a FLOSS License Exception) or to use a proprietary license.

Support can be obtained from the official manual. Free support additionally is available in different IRC channels and forums. Oracle offers paid support via its MySQL Enterprise products. They differ in the scope of services and in price. Additionally, a number of third party organisations exist to provide support and services, including MariaDB and Percona.

MySQL has received positive reviews, and reviewers noticed it ”performs extremely well in the average case” and that the ”developer interfaces are there, and the documentation (not to mention feedback in the real world via Web sites and the like) is very, very good”. It has also been tested to be a ”fast, stable and true multi-user, multi-threaded sql database server”[19].

4.1.8 Text pattern matching

You may be familiar with searching for text by pressing CTRL-F and typing in the words youre looking for. Regular expressions go one step further: They allow you to specify a pattern of text to search for. You may not know a businesss exact phone number, but if you live in the United States or Canada, you know it will be three digits, followed by a hyphen, and then four more digits (and optionally, a three-digit area code at the start). This is how you, as a human, know a phone number when you see it: 415-555-1234 is a phone number, but 4,155,551,234 is not.

Regular expressions are helpful, but not many non-programmers know about them even though most modern text editors and word processors, such as Microsoft Word or OpenOffice, have find and find-and-replace features that can search based on regular expressions. Regular expressions are huge time-savers, not just for software users but also for programmers. In fact, tech writer Cory Doctorow argues that even before teaching programming, we should be teaching regular expressions:

Knowing [regular expressions] can mean the difference between solving a problem in 3 steps and solving it in 3,000 steps. When youre a nerd, you forget that the problems you solve with a couple keystrokes can take other people days of tedious, error-prone work to slog through.

In this chapter, we will start by writing a program to find text patterns without using regular expressions and then see how to use regular expressions to make the code much less bloated. we will show you basic matching with regular expressions and then move on to some more powerful features, such as string substitution and creating your own character classes. Finally, at the end of the chapter, we will write a program that can automatically extract phone numbers and email addresses from a block of text[20].

4.2 Method

4.2.1 HTML Parsing in PHP using Simple HTML DOM parser

HTML parsing is the process of extracting relevant information like title of the web page, paragraphs, headings, links etc.

HTML parsing is very easy task with the help of SimpleHtmlDom library. For users who are unfamiliar with SimpleHtmlDom, It is a PHP library that allows to parse HTML files.

In this article we will learn how to get started with HTML parsing and certain frequently used SimpleHtmlDom library code . The article also includes the code to demonstrate how HTML parsing can be implemented easily using this library in PHP.

Installing simplehtmldom.:

Simplehtmldom is a PHP library that facilitates the process of creating web scrapers. It is a HTML DOM parser written in PHP5 that let you manipulate HTML in a quick and easy way.

First download the library from sourceforge. Unzip the library in you PHP includes directory or a directory where you will be testing the code.

Writing our first scraper.

Now we are ready with the tools, lets write our first web scraper. For our initial idea let us see how to grab the sponsored links section from a google search page.

Before we can retrieve the required data we need to know the HTML structure of the page so that we can know precisely where the required information is located[11].

1. Load the HTML and Create DOM object

Listing 4.1: DOM object from an HTML string

```
// Create a DOM object from an HTML string
$html = str_get_html('<html><body><p>Hello Web Scrapers!</p><p>We are here to learn SimpleHtmlDom library</p></body></html>');

// Create DOM object from an HTML file or URL
$html = file_get_html('http://www.webscrapingblog.com/');
```

2. How to find HTML elements?

Using find() method, we can find elements with particular name, ID, class, attributes etc. Below are the examples:

Listing 4.2: find HTML elements

```
// Find all images, returns an array of element objects
$images = $html->find('img');

```

```
// Find (N)th image element, returns element object or null if not found (zero based)
$image = $html->find('img', 0);

// Find latest image, returns element object or null if not found (zero based)
$image = $html->find('img', -1);

// Find all <div> elements with attribute id=section
$image = $html->find('div[id=section]');
```

Practical Example:

Listing 4.3: Example of web scrap

```
 //include the simple_html_dom library
require_once 'simple_html_dom.php';

// create DOM element from URL or file
$html = file_get_html('http://www.imdb.com/chart/top');

// find all the rows of the table
$rows = $html->find('table[class="chart full-width"] tbody[class="lister-list"] tr');

// create array to store extracted movie information
$movies = array();

if(count($rows)>1){
 // loop through each rows
 for($i=1; $i<count($rows); $i++){
  //extract poster image url
  $poster = $rows[$i]->find('td[class="posterColumn"] a img',0)->src;

  // extract movie title
```

```
    $title = $rows[$i]->find('td[class="titleColumn"] a',0)->plaintext;

    // extract movie detail url
    $url = "http://www.imdb.com".$rows[$i]->find('td[class="titleColumn"] a',0)->href;

    // extract movie release year
    $year = $rows[$i]->find('td[class="titleColumn"] span',0)->plaintext;

    // extract movie rating
    $rating = $rows[$i]->find('td[class="ratingColumn imdbRating"]',0)->plaintext;

    // extract movie rating title attribute
    $rating_title = $rows[$i]->find('td[class="ratingColumn imdbRating"] strong',0)->title;

    // create movie array item
    $movie = array("rank"=>$i, "poster"=>trim($poster), "title"=>trim($title), "url"=>trim($url), "year"=>trim($year), "rating"=>trim($rating), "rating_title"=>trim($rating_title));

    // store the movie item into movies array
    array_push($movies, $movie);
  }
}

// clear the dom object to avoid memory leak
$html->clear();
unset($html);
```

The code and comments are self-explanatory. After parsing movie information, we can easily display it on web page or store it into database or export it to CSV/Excel Spreadsheet. The code may not work well if website structure changes.

4.2.2 Web scraping

Listing 4.4: first scrap

```
 $data =  $html->find(’td[id=rhsline]’);
echo $data[0]−>children(1);
```

The complete source is given below.

Listing 4.5: ource Code

```
 <?php

/* update your path accordingly */
include_once ’libs/simplehtmldom/simple_html_dom.php’;

$search_term = ”mobiles”;

$url = ”http://www.google.co.in/search?hl=en&q={$search_term}”;

$html = file_get_html($url);

/*
Get all table rows having the id attribute named ’rhsline’.
As the list of sponsored links is in the ’ol’ tag; as can be
seen from the DOM tree above; we use the ’children’ function
on the $data object to get the sponsored links.
*/
$data =  $html->find(’td[id=rhsline]’);

/*
  Make sure that sponsors ads are available,
  Some keywords do not have sponsor ads.
*/
```

```
if(isset($data[0]))
    echo $data[0]->children(1);

?>
```

In the next example we will grab the list of contents from the latest Wired magazine issue.

Listing 4.6: grab the list of contents

```
 $ret = $html->find(’div[id=this_month] div[class=story]’);

foreach($ret as $story)
    echo $story->find(’a’, 0) . ”<br>”;
```

This will return all the content section links from the page. To just return the links as text we can use the plaintext modifier.

Listing 4.7: links the page

```
 echo $story->find(’a’, 0)->plaintext . ”<br>”;
```

The source code is given below..

Listing 4.8: complete source

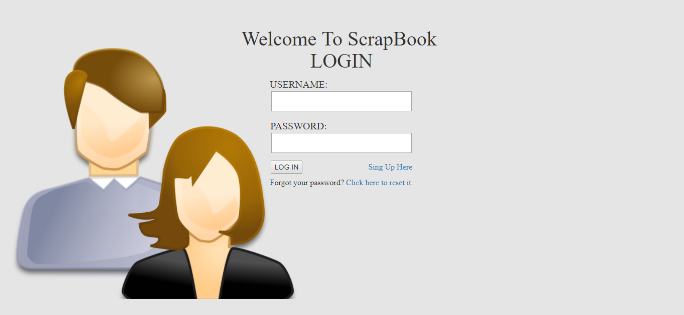
```
 <?php

/* update your path accordingly */
include_once ’libs/simplehtmldom/simple_html_dom.php’;

$url = ”http://www.wired.com/wired/”;

$html = file_get_html($url);

```

```
$ret = $html->find(’div[id=this_month] div[class=story]’);

foreach($ret as $story)
    echo $story->find(’a’, 0)->plaintext . ”<br>”;

?>
```

4.3 Project Demonstration

4.3.1 View Login Page

View the login page if the user dose not have account then you must have register.if already have register then login.



Welcome To ScrapBook
LOGIN

USERNAME:

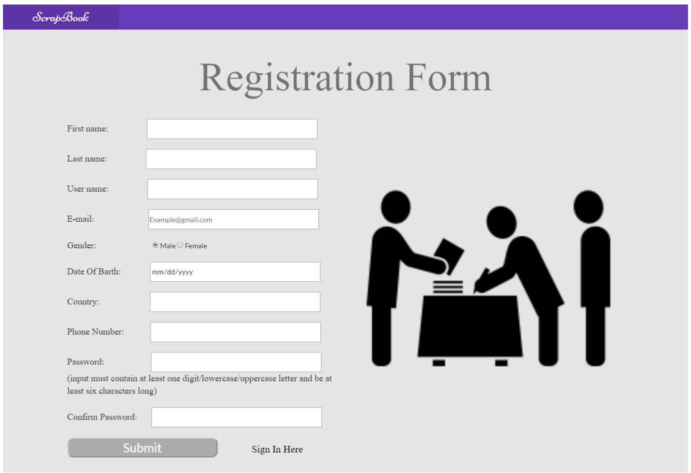
PASSWORD:

Forgot your password? Click here to reset it.

Figure 4.1: System view Login Page

4.3.2 View Register Page

In this page you have to register your account.



Registration Form

First name:

Last name:

User name:

E-mail: Example@gmail.com

Gender: ● Male ○ Female

Date Of Barth: mm/dd/yyyy

Country:

Phone Number:

Password:

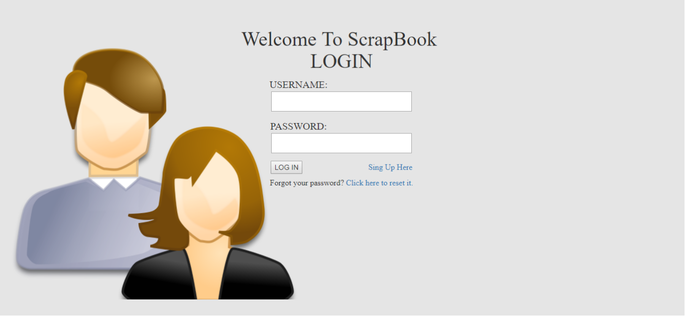
(input must contain at least one digit/lowercase/uppercase letter and be at least six characters long)

Confirm Password:

Figure 4.2: System view Register Page

4.3.3 View Login Page

In this section you have to Login in your account and see your profile.



Welcome To ScrapBook
LOGIN

USERNAME:

PASSWORD:

Forgot your password? Click here to reset it.

Figure 4.3: System view Login Page

4.3.4 View User Page

In this section user see the user page where he/she see the other features of this application.

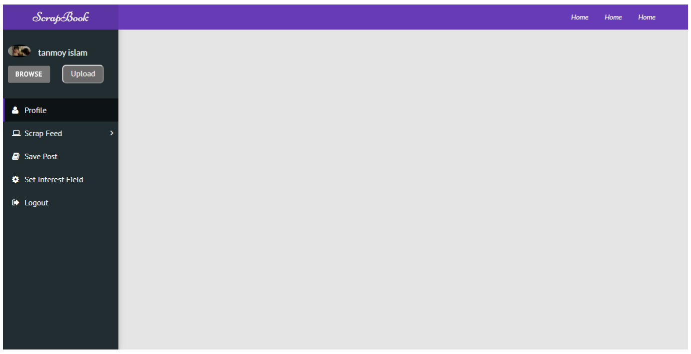


Figure 4.4: System view User Home Page

4.3.5 Set Profile Picture

In this section user can set his profile picture on his profile.

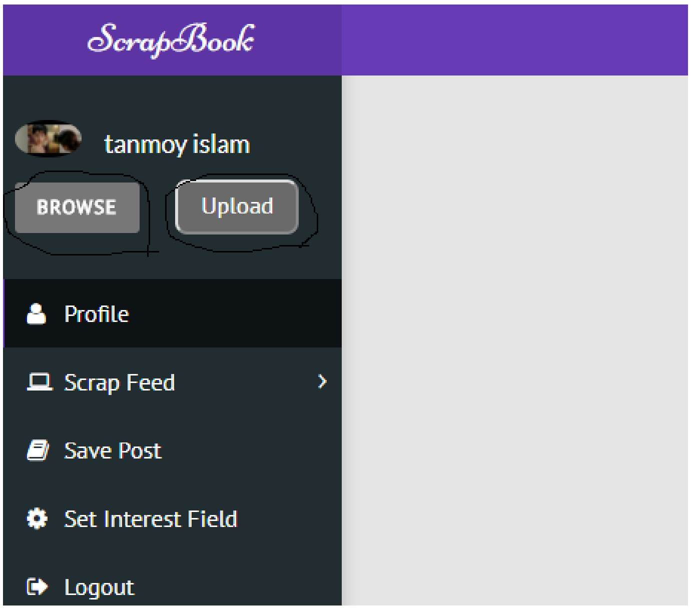
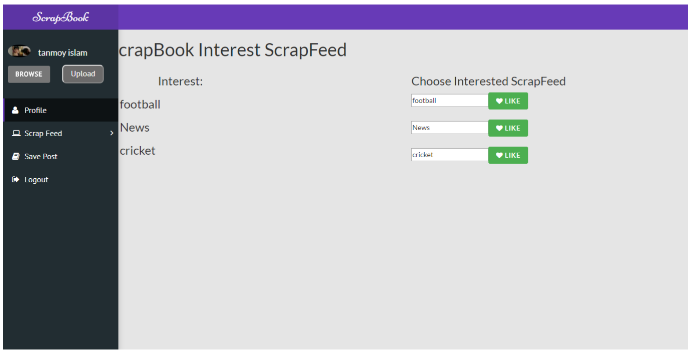


Figure 4.5: System view Set Profile Picture Page

4.3.6 Set Interest Field

In this section user can choose his interest and set to his menu.



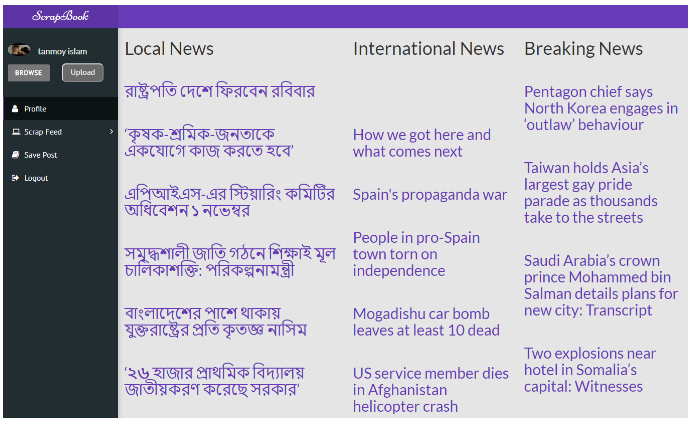
crapBook Interest ScrapFeed

Interest:	Choose Interested ScrapFeed
football	football ♥ LIKE
News	News ♥ LIKE
cricket	cricket ♥ LIKE

Figure 4.6: System view Interest Page

4.3.7 *View Interest News Page*

In this page user can show his interest news



Local News

রাষ্ট্রপতি দেশে ফিরবেন রবিবার

‘কৃষক-শ্রমিক-জনতাকে একযোগে কাজ করতে হবে’

এপিআইএস-এর স্টিয়ারিং কমিটির অধিবেশন ১ নভেম্বর

সমৃদ্ধশালী জাতি গঠনে শিক্ষাই মূল চালিকাশক্তি: পরিকল্পনামন্ত্রী

বাংলাদেশের পাশে থাকায় যুক্তরাষ্ট্রের প্রতি কৃতজ্ঞ নাসিম

‘২৬ হাজার প্রাথমিক বিদ্যালয় জাতীয়করণ করেছে সরকার’

International News

How we got here and what comes next

Spain's propaganda war

People in pro-Spain town torn on independence

Mogadishu car bomb leaves at least 10 dead

US service member dies in Afghanistan helicopter crash

Breaking News

Pentagon chief says North Korea engages in ‘outlaw’ behaviour

Taiwan holds Asia’s largest gay pride parade as thousands take to the streets

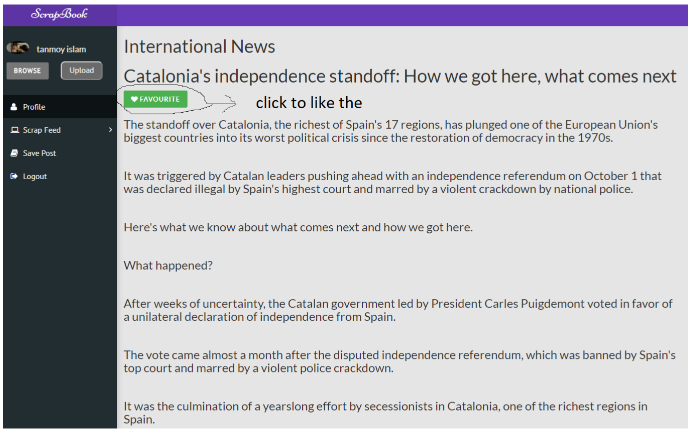
Saudi Arabia’s crown prince Mohammed bin Salman details plans for new city: Transcript

Two explosions near hotel in Somalia’s capital: Witnesses

Figure 4.7: System view News Page

4.3.8 *Save Your Favorite News*

In this section user can see the specific news which he/she can see and she/he can save it on his database and read the news letter.



International News

Catalonia's independence standoff: How we got here, what comes next

click to like the

The standoff over Catalonia, the richest of Spain's 17 regions, has plunged one of the European Union's biggest countries into its worst political crisis since the restoration of democracy in the 1970s.

It was triggered by Catalan leaders pushing ahead with an independence referendum on October 1 that was declared illegal by Spain's highest court and marred by a violent crackdown by national police.

Here's what we know about what comes next and how we got here.

What happened?

After weeks of uncertainty, the Catalan government led by President Carles Puigdemont voted in favor of a unilateral declaration of independence from Spain.

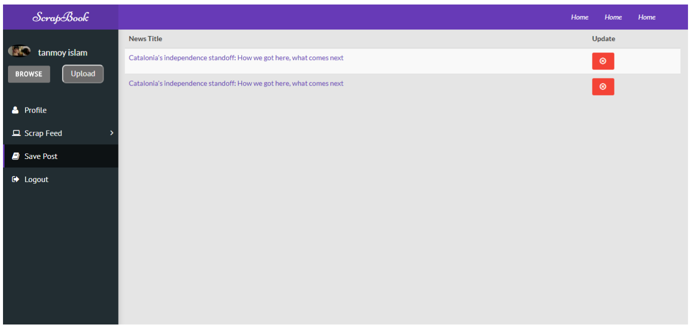
The vote came almost a month after the disputed independence referendum, which was banned by Spain's top court and marred by a violent police crackdown.

It was the culmination of a yearslong effort by secessionists in Catalonia, one of the richest regions in Spain.

Figure 4.8: System view News Page

4.3.9 Show Save Favorite News

In this section user can see the save post which is he/she can save in past.



News Title	Update
Catalonia's independence standoff: How we got here, what comes next	⊗
Catalonia's independence standoff: How we got here, what comes next	⊗

Figure 4.9: System view Save Post Page

4.4 Chapter Summary

Web scraping is the process of automatically mining data or collecting information from the World Wide Web. It is a field with active developments sharing a common goal with the semantic web vision, an ambitious initiative that still requires breakthroughs in text processing, semantic understanding, artificial intelligence and human-computer interactions. Current web scraping solutions range from the ad-hoc, requiring human effort, to fully automated systems that are able to convert entire web sites into structured information, with limitations.

All those code and comments are self-explanatory. After parsing movie information, you can easily display it on web page or store it into database or export it to CSV/Excel Spreadsheet. The code may not work well if website structure changes.

5 Conclusion

This project focused on how to create a web scraping application using the simple html dom library function ,and the tools that the framework provided, as well as the use of PHP tools that were available. The main intention was to create a web application using the web scraping and retrieved useful information from a website. This goal was met, and a robust application was created.

The program was able to gather most of the data that I had defined. That being said, I was not able to extract author names for every book. This was due to the fact that I had only selected the hyperlink element that displayed the name of a books author. After taking another look at the list of search results I realized that some of the hyperlinks that contained the book author were not active.

5.1 Limitations

Web-scraping can be also challenging if you don't have the proper tools. Largely, you're completely at the mercy of the target website, and that website can change at anytime - without notice. Or, it may contain faulty JavaScript that causes it to crash and exhibit surprising behavior. The server that hosts the website may crash, or the website may undergo maintenance. Many potential problems can occur during a lengthy web-scraping session, and you have very little influence on any of them.

5.2 Future Works

In the realm of future work, I hope to continue research of web scraping with more advanced software. This would allow me to use a more precise method of parsing that would ensure that every piece of data I wanted was collected. The ideal tool would allow me to create a unique program to extract the information that I needed. Overall, it is hoped that this body of work was able to expose academicians, students and practitioners to the concept and necessity of web scraping and demonstrate a tool that the average computer user could utilize on their own.

References

[1] “web scraping 2017,” 2017. [Online]. Available: https://www.codediesel.com/php/web-scraping-in-php-tutorial/

[2] Hemenway, Kevin and Calishain, “data scraping 2017,” 2017. [Online]. Available: http://searchdatacenter.techtarget.com/definition/data-scraping/

[3] Yeh, Tom (2009, “Screen scraping,” 2017. [Online]. Available: http://searchdatacenter.techtarget.com/definition/screen-scraping/

[4] @Kenneth, Hirschey, Jeffrey, “data scraping 2017,” 2017. [Online]. Available: http://webscraper.io/

[5] Margaret Rouse, “Report mining,” 2017. [Online]. Available: http://searchdatacenter.techtarget.com/definition/data-scraping/

[6] X. Legaspi, 2017. [Online]. Available: http://lnu.diva-portal.org/smash/get/diva2:1032894/FULLTEXT01.pdf

[7] ——, 2017. [Online]. Available: http://lnu.diva-portal.org/smash/get/diva2:1032894/FULLTEXT01.pdf

[8] M. Spaanem, “Static vs. dynamic websites: What are they and which is better? — rocket media,” 2017. [Online]. Available: https://rocketmedia.com/blog/static-vs-dynamic-websites

[9] ——, “Static vs. dynamic websites: What are they and which is better? — rocket media,” 2017. [Online]. Available: https://rocketmedia.com/blog/static-vs-dynamic-websites

[10] "Php simple html dom parser," 2017. [Online]. Available: http://simplehtmldom.sourceforge.net/

[11] View, "Html parsing in php using simple html dom parser - web scraping blog," 2017. [Online]. Available: http://www.webscrapingblog.com/html-parsing-php/

[12] S. Zheng, D. Wu, R. Song, and J.-R. Wen, "Wrapping oriented classification of web pages," 2017. [Online]. Available: http://blog.endpoint.com/2016/07/scrape-web-content-with-php-no-api-no.html

[13] J. Skinner, "sublimetext," 2017. [Online]. Available: https://www.sublimetext.com/blog/articles/sublime-text-3-point-0

[14] S. Yegulalp, "wampserver," 2017. [Online]. Available: https://www.infoworld.com/article/2616867/web-development/review–wamp-stacks-for-web-developers.html

[15] W. Dave Raggett, "HTML." [Online]. Available: https://www.w3.org/MarkUp/html3/CoverPage

[16] J. h. mrhands, rolfedh, "CSS," 2017. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/CSS

[17] ——, "BOOTSTRAP," 2017. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/bootstrap

[18] PHP.net, "PHP," 2017. [Online]. Available: http://php.net/manual/en/intro-whatis.php

[19] M. R. Rob McCormack, "MYSQL SERVER," 2017. [Online]. Available: http://searchoracle.techtarget.com/definition/MySQL

[20] X. Legaspi, 2017. [Online]. Available: http://lnu.diva-portal.org/smash/get/diva2:1032894/FULLTEXT01.pdf