відмінно і 58,8% - добре; при використанні презинтації 29,4% - відмінно і 41,7% - добре; на традиційній лекції матеріал 56,9% матеріал засвоюють частково, 23,5% - добре і тільки 13,7% - відмінно. Отже, для подання студентам нового матеріалу викладачеві необхідно застосовувати на парах інтерактивні дошки та презентації. Варто відзначити, що інформаційні технології значною мірою підвищують мотивацію студентів до навчання, проведе нию різних науково-дослідних робіт, експериментів, створення інноваційних проектів і статей. У наш XXI століття - століття комп'ютерів використання інформаційних технологій у вищій освіті є необхідністю, здатної підготувати студентів до життя і роботи в сучасному інформаційному суспільстві.

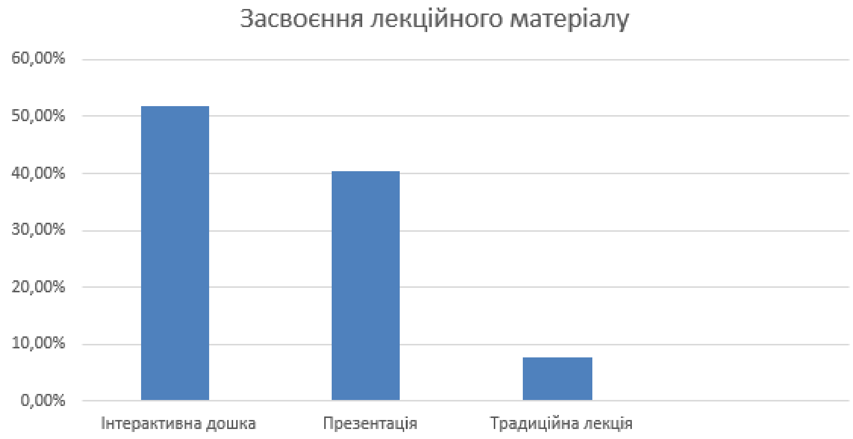


Рис. 1. Діаграма «Засвоєння лекційного матеріалу студентами»

ДЖЕРЕЛА

1. Широкова Е.А. Хмарні технології. Сучасні тенденції технічних наук. Уфа, 2011 року.

2. Зайцева С.А., Іванов В.В. Інформаційні технології.

3. Грибан О.Н. Інформаційні технології в освіті. – Доступ.: http://griban.ru/ blog/14-informacionnye-tehnologii-v-processe-obuchenija.html

AUTOMATION OF INFORMATION RETRIEVAL FOR PROJECT MANAGEMENT PURPOSES

Rudenko Vitaly, PhD Teslenko Pavlo
Odessa national polytechnic university
Ukraine, Odesa
vitalyrudenko.pers@gmail.com

We propose an automatic web-based tool for analyzing similar projects, such as their features, targeted service information, services they use, and so on. This information can be used to make decisions at every stage of the project life cycle.

Keywords: web-scraping, html-extractors, content analysis

Web-scraping (also known as web harvesting and web data extraction) is data scraping used for extracting certain information from website for later analysis. Web-scraping software is basically a program (sometimes called HTML-extractor – Hyper Text Markup Language extractor), which takes one or multiple URLs (Uniform Resource Locators) on the input, downloads website content, performs certain operations – removing useless parts, finding necessary information. The output of the program is often a tabular data (for example, a csv file), which can already be used for analysis and aggregation [1, 2].

HTML-extractors (web-scrapers) usually provide user interface or API (application programming interface) to manage data selectors and other settings, so user can write his own rules of page scraping. For example, user can write rules for mobile numbers and select html elements, in which to look up these numbers. For example, ScreamingFrog Web Scraping application allows you to specify multiple regular and X-Path expressions (syntax for defining parts of XML documents – Extensible Markup Language) to scrape specific information out of web page.

The main problem of most HTML-extractors, that they're not designed to scrape multiple websites with the same rules. This leads to uselessness of extractors if they're used for project management purposes – mainly analysis of the available market by taking as much info from many pages as possible [3,4].

The other problem with web-scraping is that user needs to know how to write rules and have basic knowledge of HTML and JavaScript. Therefore, precious time and sometimes money should be spent just to write extracting rules, which may be useless in the end.

Let's take some look at existing solutions. ScreamingFrog Web Scraping has already been mentioned before. Its limits are bounded by specific regular and X-Path expressions, which may be confusing for a user without programming experience. The next solution is Webscraper.io, which allows user to visually select blocks of website's interface and extract data from all similar elements. It also has deep-scraping feature, which allows scraper to navigate necessary links to extract as much information as possible. The problem of this solution is that only one website can be scrapped, so this software couldn't be used in market analysis purposes.

The other popular solution is ParseHub – has a comprehensive UI, which allows user to specify associations between html nodes and use them to extract only required data. This solution also has a single-website restriction.

It has been decided to develop new HTML-extractor to simplify process of website analysis. The main feature of the scraper is modularity, which allows user to select certain analyzer providers (modules) to extract only essential information.

Modules of the developed HTML-extractor analyze elements' contents rather than plain HTML nodes and attributes, so internal rules of each module are mostly universal and can be applied to different websites.

Main modules have been developed:

- mobile phone numbers extractor module,
- addresses extractor module,
- pricing extractor module.

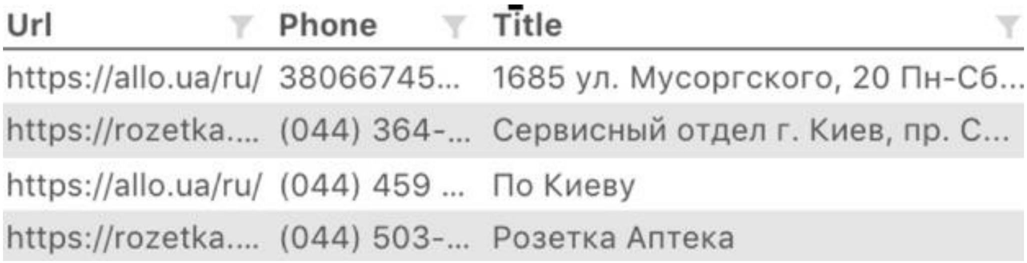
Each of them performs its own operation to retrieve required information. For example, mobile phone numbers extractor searches for elements, which contain phone numbers (in different formats), then collects surrounding it text and returns the phone number and it's presumptive (sometimes can be misleading and incorrect) title. It also performs navigation to 'About us' and 'Contacts' pages if they're available and scrapes information from them too.

Addresses extractor is similar to mobile phone number extractor, except that it's looking for addresses (for example, of shop's stocks). Firstly, it looks for 'About us', 'Contacts' or 'Shops' links and navigates to them. Then the addresses are being collected (also in different formats: full, short, with and without city etc.). This is a basic module, which can hardly be used to collect useful data. But it can be used as an example for creating other modules.

The most powerful module is the pricing extractor. It searches for 'Pricing' or 'Plans' link and navigates to the pricing page via the links to exclude redundant data from the results. Finally, the module looks up for all prices on page and forms a table with their list, where each price is mapped to the website URLs.

Basically, scrapping algorithms, used in the developed HTML-extractor modules, could be used to analyze multiple websites with similar structure and collect necessary data. It's the main feature of the extractor.

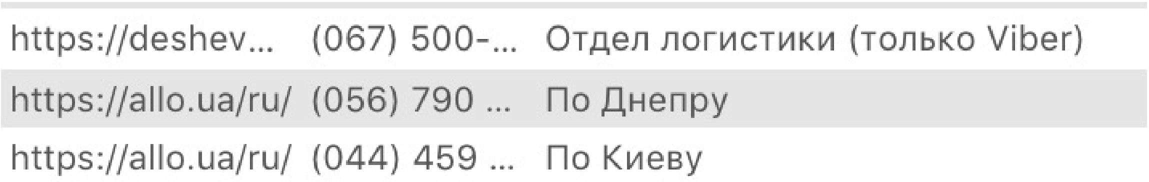
To prove the above statement, URLs of popular web-services where given to the program's input to extract mobile phone numbers and addresses. Then the program has been executed to collect the data (pic. 1).



Url	Phone	Title
https://allo.ua/ru/	38066745...	1685 ул. Мусоргского, 20 Пн-Сб...
https://rozetka....	(044) 364-...	Сервисный отдел г. Киев, пр. С...
https://allo.ua/ru/	(044) 459 ...	По Киеву
https://rozetka....	(044) 503-...	Розетка Аптека

Picture 1 – Piece of collected by the HTML-extractor data

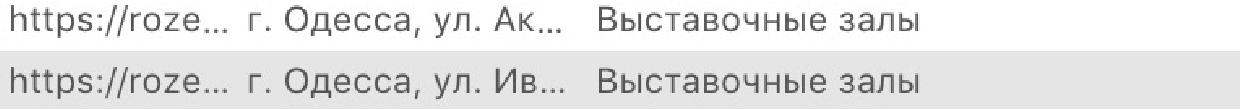
As seen in the image above, phone numbers in different formats were parsed successfully. But title parsing still needs to be modified to exclude useless data, though sometimes it can be useful (pic. 2).



https://deshev...	(067) 500-...	Отдел логистики (только Viber)
https://allo.ua/ru/	(056) 790 ...	По Днепру
https://allo.ua/ru/	(044) 459 ...	По Киеву

Picture 2 – Piece of correctly parsed phone number title

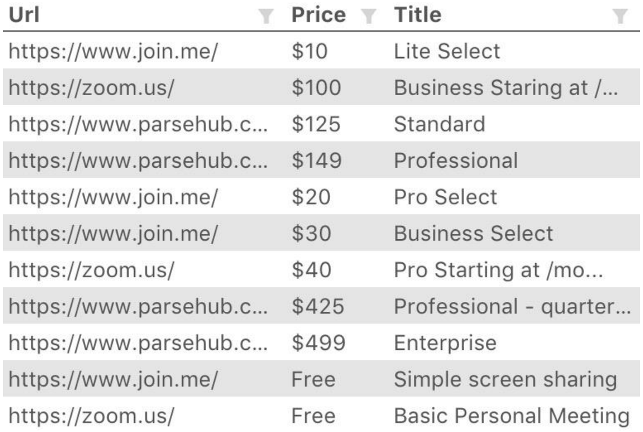
Collected address data is shown in pic. 3. Addresses are often extracted without errors, which is a good sign for the module scraping algorithms.



https://roze...	г. Одесса, ул. Ак...	Выставочные залы
https://roze...	г. Одесса, ул. Ив...	Выставочные залы

Picture 3 – Piece of correctly parsed addresses

The last but not least module to be tested is the pricing extractor. It has been fed with multiple website URLs with a subscription strategy (websites with this type of monetization strategy often have pricing page). The results are shown in the pic. 4.



Url	Price	Title
https://www.join.me/	$10	Lite Select
https://zoom.us/	$100	Business Staring at /...
https://www.parsehub.c...	$125	Standard
https://www.parsehub.c...	$149	Professional
https://www.join.me/	$20	Pro Select
https://www.join.me/	$30	Business Select
https://zoom.us/	$40	Pro Starting at /mo...
https://www.parsehub.c...	$425	Professional - quarter...
https://www.parsehub.c...	$499	Enterprise
https://www.join.me/	Free	Simple screen sharing
https://zoom.us/	Free	Basic Personal Meeting

Picture 4 – Piece of collected pricing information

Extracted from multiple websites pricing information can be used to decide pricing of our own project, because this information provides an average pricing number of different features on the market. The collected information can also be used to decide which monetization strategy suits the managed project the most.

So, automatic web-scraping can be useful for analyzing project analogs' characteristics, like their features, pricing information, services they use and so on. This information can later be used to make decisions at every stage of the project life cycle. It can help in choosing popular and trusted services and tools, analyzing analogs' features and, most importantly, make the developed project unique and distinct from them.

REFERENCES

1. Nadkarni P.M. An introduction to information retrieval: applications in genomics [Electronic resource]. — Access: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137130/

2. Teslenko P. 3-Level Approach to the Projects Planning / P. Teslenko, D. Bedrii, S. Antoshchuk, H. Lytvynchenko // XIII th International Scientific and Technical Conference «Computer science and information technologies» 11-14 September, 2018. — Lviv, 2018. — pp. 195-198.

3. Barska I. Algorithm of Distributing the Team Load for IT-Project / Barska I., Teslenko P., Fesenko T., Voznyi O. // Proceedings of the 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). — Warsaw : University of Technology, 2015. — p. 559 – 562.

4. Teslenko P. Increasing probability of successful projects complete / P. Teslenko, S. Antoshchuk V.Krylov // Proceedings of the International Research Conference at the Dortmund University of Applied Sciences and Arts took place on June 30th -July 1st 2017 for the seventh time. — 2017. — Dortmund : the Dortmund University. — P. 28-30