**Regular Expressions for Web Advertising Detection based on an Automatic Sliding Algorithm**

1*D. Riaño <donovan20@comunidad.unam.mx>*

1*R. Piñon <rodrigo\_pinon@comunidad.unam.mx>*

1*G. Molero-Castillo <gmoleroca@fi-b.unam.mx>*

1*E. Bárcenas <ebarcenas@unam.mx >*

1*A. Velázquez-Mena <mena@fi-b.unam.mx>*

1*Engineering Faculty, National Autonomous University of Mexico, UNAM, Mexico*

**Abstract**.This paper presents the automation of a Web advertising recognition algorithm, using regular expressions. Currently, the use of regular expressions, optical character recognition, Databases, and automation tests have been critical for multiple Software implementations. The tests were carried out in three Web browsers. As a result, the detection of advertisements in Spanish, that distract attention and that above all extract information from users was achieved. The main feature of the algorithm is that automatic and versatile execution does not require access to the code of the page in question and that in the future it can be an application with background operation. In addition, being supported by optical character recognition gives us acceptable efficiency in detecting advertising.

**Keywords:** Digital Marketing; Optical Character Recognition; Regular Expressions; Web Advertising.

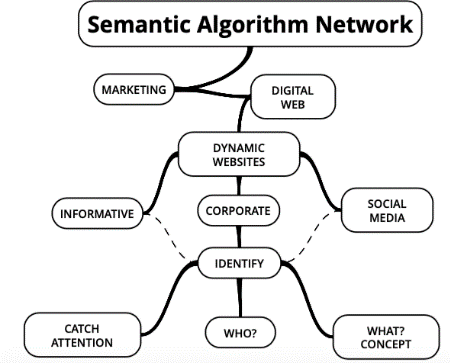
***1. Introduction***

In the past, marketing only exists offline and its main objective was to coordinate the media, make deals so that people or even other companies have positive opinions about the products that are advertised, or about the ideas that are planned, to be sold. But now, with all the tools that technology has given us this is over, today users use search engines to find what they want and not only that, but they can also access criticism and comments made by the community.

Within digital marketing, the main objective is the user [1], and therefore the marketing techniques changed their paradigm. Today a digital strategy must include all the relevant spaces for the 'target' to interact, looking for people who influence their opinions to enter the same network of users, and can give more strength to ideas or products. This can also focus on improving search engines and, according to the experience acquired, becoming increasingly invasive in their ways of entering the minds of users.

Therefore, a semantic network was designed to understand the current digital marketing trends used in dynamic Web pages. It is important to mention that a semantic network is a form of representation of knowledge through interrelationships in the form of a graph [2]. Figure 1 shows the interrelationships of the semantic network, in which advertisements on Web pages can be divided into three groups: a) sites that subsist on advertising; b) informational, corporate and online stores; and c) social networks or similar.

This semantic network relates: i) who offers, ii) what product or service does it offer, and iii) in what way does it attract attention (offer or promotion). These types of advertising are distinguished when browsing the Internet, which some of these were the object of analysis in this research work.



*Fig. 1. Semantic network of current Web advertising*

Websites that subsist from advertising offer a brief content of news, education, or entertainment and their ads are provided by Google. In these, the ads change constantly. In the case of informational, corporate, and online stores, these include sites such as MSN, Amazon, Sanborns, Adidas, and others, whose purpose is to offer current trends in products or services. While in the case of social networks and derivatives such as Facebook, Instagram, LinkedIn, Twitter, and others, they are sites that offer interaction with users and ads related to the user profile.

This paper presents the automation of a Web advertising recognition algorithm, using regular expressions. The tests were carried out in three browsers: Chrome, Firefox, and Safari. A feature of the algorithm is its automatic and versatile execution, since it does not require access to the code of the Web page in question and that it is an application that operates in the background.

***2. Background***

Given the potential and strategies of current digital marketing, this activity is used more frequently, as an important part of brand loyalty campaigns, since it is a communication channel that achieves the interaction between the potential client and brand. Among the actions focused on online marketing include [3]: customer loyalty; increase the image of the brand and its sales; generate promotions and product tests; encourage the repeat purchase of the product; and conduct a direct and personalized communication campaign.

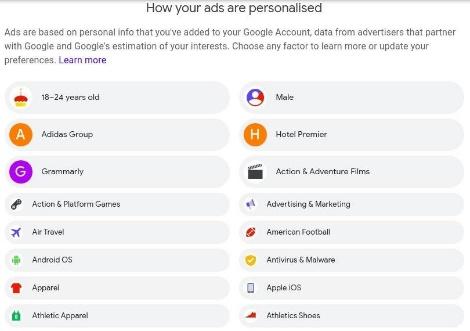
One way to deal with advertising and privacy in Web browsers is through regular expressions (RegExp, Regex, or RE). These expressions are a way of representing character strings that fit a certain pattern [4] [5]. Its applications are diverse, for example, validation of form fields, identification of text strings in social networks, search commands, among others [6] [7].

* 1. **User Privacy Management**

At present, information and privacy management are in a critical stage [8]. Companies such as Facebook, Twitter, Google, Amazon, among others, have included in their websites, applications, or search engines for information hoarding. These companies provide ‘free’ for the service, but use the information of the users, at convenience, for advertising purposes.

As advertising increases in web browsers, it puts the privacy of users at risk. That is, when searching the Web for certain types of services or products, today there is a saturation of information on the website where it is positioned. Another important aspect is that user searches generate a large amount of collected data. Therefore, these data, under strict treatment, could also provide useful information for better management of targeted advertising [9] [10].

As an example, Figure 2 shows the Google search engine for the automatic creation of user profiles based on Web browsing. For example, a male user is relevant, with an age range between 18 and 24 years, whose interest is to buy online, do outdoor activities, develop mobile applications, watch movies, and others. This user, through the collection of data through cookies, trackers, and information contained in the Gmail profile, can be used to group profiles with common characteristics and, therefore, receive advertising for various offers and promotions.



*Fig. 2. Google search engine for automatic creation of user profiles based on Web browsing*

Based on the above, therefore, when we constantly interact with Google services, it is advisable to review the data and privacy policies frequently about the adjustments made by Mountain View, in order to have greater knowledge about the security and use of our information.

* 1. **Ad-Blockers of Web Browsers**

In recent years, efforts have been increased to implement Web advertising blockers [11]. These were initially implemented for the Firefox browser, which over the years was improving to achieve a browser without ads. But the companies sought ways to continue sending advertising. Consequently, in a short time, they began to create controversies [1]. Controversies for significant losses of possible income [12].

Another term that is also currently used is Trackers in marketing, which are indicators and trackers of the effectiveness of targeted advertising campaigns. These tools store information through cookies, and these, in turn, provide the location of the searches performed. Thus, as demand grew and the possibilities of blocking advertising, new Ad-blockers were implemented in most Web browsers.

An important fact to consider is that, in recent years, Google controls 85% of the global search engine advertising business and about 50% of all online advertising [13]. People and society, in general, see Google as a service [14], but behind that, there is a technology that contains specific functions and includes exclusive and restrictive extensions.

* 1. **Selenium Automatization**

Selenium, also known as Selenium Webdriver, is a tool to automate processes in different Web browsers [15]. Its purpose is to improve support for detecting problems in any Web browser [16]. This tool allows testing any Web browser to obtain HTML code data, change, open and move between tabs of the browser windows, return or advance according to the test history, resize the windows, take screenshots, complete fields, clicks on parts of a website, among others.

These tasks are applicable to Java, Python, C#, Ruby, Perl, and JavaScript programming languages. The operating systems that Selenium supports are Windows, Mac OS, and Linux, each with their respective packages and Integrated Development Environments (IDEs) [17].

On the other hand, at present there is interest in using Optical Character Recognition (OCR) in the detection of Web advertising [14], however, more efforts are still needed to cover all digital marketing strategies.

* 1. **Related Works**

One of the works focused on the detection of Web advertising was [18], where an algorithm that performs Web crawling is presented. It consists of obtaining information from websites through tagging, the test Web page was MSN. The classification is made through keywords throughout the Web page, that is, at the beginning (B-beginning), in the middle (I-inside), at the end (l-last), unique (U), or outside (O).

[19] describes the contextual advertising analysis through PageSense, which aims to associate ads on Web pages. Through this platform, blank regions are detected and the non-intrusive area is selected for ad placement without breaking the original style of the Web page. Bayesian combinations and probabilities were used for the analysis, which reflects the percentages of advertisements for different types of products or services.

In another work, in [20], an analysis was made based on Euclidean distances. These distances were with respect to the way the ads are of interest to users, the search for products, and the adaptation of the objective profile, dividing it by sections, such as health, sports, business society, education, art, science, computer, among others.

On the other hand, [21] describes the analysis carried out on approximately 500 Web pages, in which tests aimed at detecting ad types, but not content was carried out. Among the types of ads analyzed, pop-ups, carousels, videos, gifs, games, stickers, or text stand out. The countries where the ads come from, the frequency, the size, and the origin of the URLs (Uniform Resource Locator) were also analyzed.

***3. Method***

The implementation of the algorithm was in Python. This algorithm has been assisted by the Tesseract [22], Pillow, and OpenCV [23] libraries, as well as the Tesseract-OCR package. For the implementation, we start from a point before the OCR process, that is, optical character recognition. In addition, we used the Selenium library to perform the Automatic sliding in the Web browser.

In the implementation, the height difference between the different sizes of monitors on the market was also taken into account, therefore, the algorithm performs a dynamic detection of the height of the windows, thus adjusting to any screen size. Therefore, the displacement of information in this work is vertical.

* 1. **Website Sliding**

Through Selenium, a new window opens for a browser compatible with this tool. The terminal asks the user for the URL field. Subsequently, the browser under test expands to full screen for a quick and complete scan of the website.

As part of the algorithm, and with the aim of stopping the process for a few moments, programming threads are used. The first pause is made to allow the page to load, since depending on the speed of the Internet, and even on the state of the page, it takes a loading time to perform the correct screenshot for subsequent analysis using the OCR.

In the next step, the window is maximized, before sliding to capture and store screens. This causes a second pause, which is 0.5 seconds. Subsequently, it starts the screen capture process until it finishes with all the vertical content of the Web page. Once the capture process is complete, the Web browser window where the query was made closes automatically.

* 1. **Optical Character Recognition**

Optical character recognition, OCR, allows extracting text from an image with alphabetical writing, regardless of the language, size, or color of the text, with high effectiveness. The effectiveness of the OCR ranges from 71 to 98%. This system is capable of reaching average values of 85.1% for handwritten text and 90.93% for printed or digital text [24].

A specific function was defined for the search of all the images of the screenshots saved with an established path. Then, a cycle was implemented where all the images are analyzed within the established route in the order in which they were captured.

When the desired image is accessed, the OCR results are saved in a variable, which consists of text that is then formatted, placing everything in capital letters, separating the words, eliminating spaces and characters not supported by the language SQL like “', \,. and &”, this to be able to compare with regular expressions.

* 1. **Regular Expressions**

Once the Web content information has been obtained and synthesized into text strings, the local server is accessed to validate the website content with a database, which houses around 600 different words based on regular expressions distributed in three tables, and whose main objective is to validate the following topics: i) words most used in digital marketing; ii) brands, considering their respective sub-brands in the products or services offered by the company; y iii) the type or what the product or service consists of.

Tables 1, 2, and 3 show an fragment of the regular expressions related to the most used words in digital marketing, some recognized brands, and some products that are published more frequently in Mexico; respectively.

Table. 1

Fragment of the most used words in digital marketing in Spanish

|  |  |  |  |
| --- | --- | --- | --- |
| Word | Plural | Accent | Character |
| Ahorro  Bajo | Ahorros  Bajos | Null  Null | Null  Null |
| Comprar Cotiza | Null  Null | Null  Null | Null  Cotizar |
| Descuento | Descuentos | Null | % |
| Dinero | Dineros | Null | Null |
| Especial  Gratuito  Hasta  Ilimitado  Interes  Internet  Oferta | Especiales  Gratuitos  Null  Ilimitados  Intereses  Null  Ofertas | Null  Null  Null  Null  Interés  Null  Null | Null  Gratis  Null  Null  Null  Web  Null |

Table. 2

Fragment of some of the best-selling and highest-paid brands in Mexico

|  |  |  |  |
| --- | --- | --- | --- |
| Brand | Sub-brand | Product | Acronym |
| Adidas  Adidas | Boost  NMD | Boost  NMD | Null  Null |
| Apple  Apple | iMac  iPhone | iMac  iPhone | Null  Null |
| Bancomer | BBVA | BBVA | BBVA |
| Banorte | Banorte | Banorte | Null |
| Levis  Levis  Mazda  Mazda  Microsoft  Microsoft  Nike | Trucker  Western  Mazda2  Mazda3  Azure  Office  Jordan | Trucker  Western  Mazda2  Mazda3  Azure  Office  Jordan | Null  Null  Null  Null  Null  Null  Null |

Table. 3

Fragment of some products advertised in Mexico

|  |  |  |
| --- | --- | --- |
| Key | Type | Concept |
| 5G  Americano | Internet  Deportes | Internet  Entretenimiento |
| Basquetbol  Béisbol | Deportes  Deportes | Entretenimiento Entretenimiento |
| Chamarra | Ropa | Vestimenta |
| Chico | Talla | Vestimenta |
| Compacto  Ellas  Ellos  Familia  Grande  Hatchback  Jeans  Laptops  Licuadora  Smartphones | Autos  Género  Género  Género  Talla  Autos  Ropa  Electrónicos  Electrónicos  Electrónicos | Automóvil  Social  Social  Social  Vestimenta  Automóvil  Vestimenta  Electrónica  Electrodomésticos  Electrónica |

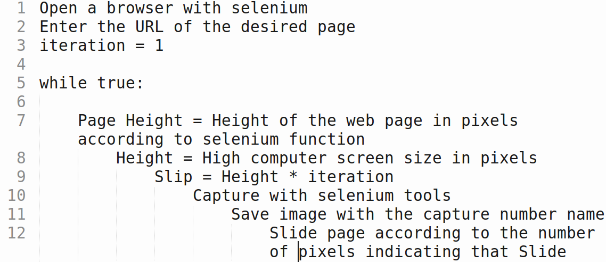
These words and symbols that make up regular expressions are those that are commonly used in advertisements on the Web [25].

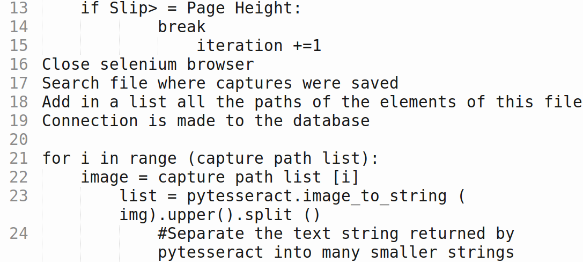
Furthermore, since advertising not only plays with the visuals but also with the letters, sizes, and styles, the search range with plurals, accents, and symbols referring to some keywords was extended over these words.

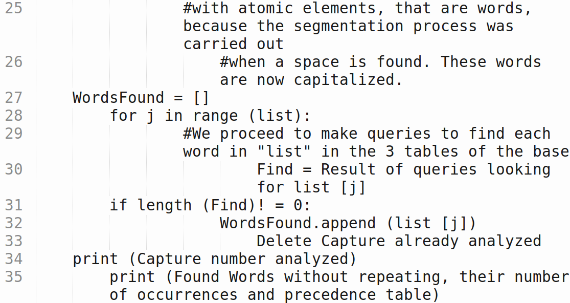
It is important to note that these regular expressions of brand and product are related to brand studies in Mexico and some studies in Latin America. For Mexico, the statistics for 2019 were searched in the databases of the National Institute of Statistics and Geography (INEGI, by its acronym in Spanish), which is an autonomous body of the Mexican government responsible for geographic statistics regarding resources, population, and the economy. Another source of data was the Economic Commission for Latin America and the Caribbean (ECLAC), which is an agency of the United Nations that allows access to information in some Latin American countries.

* 1. **Pseudocode**

Based on the above, the basic idea of the algorithm is to use automatic scrolling, so that the information contained in Web pages can be captured. Then these captured images are processed and transformed into text format to later identify the existing advertisements based on filters and matches with the word families used in digital marketing as regular expressions. Finally, the most advertised brands are identified and are trending in searches through Web browsers. This is useful for keeping track of these. Figure 3 presents the pseudocode of the algorithm developed.







*Fig. 3. Pseudocode of the implemented algorithm*

In general, as already described in previous sections, the algorithm has three main stages: i) you must obtain the URL of the page, and then take the screenshots through automatic scrolling; ii) the screenshots images are then processed, in the order in which they were taken, to convert them to text with the help of OCR; and iii) the text is analyzed, obtaining, as a result, a list of coincidences with the regular expressions, stored in tables, about the companies that advertise, the products and their strategies.

As a restriction, this work does not make Web Scraping, which is a process of automatic collection of data and information from the Internet, commonly on Web pages that use languages such as HTML, whose data is analyzed for certain needs and purposes [25]. Thus, no personal extensions added to the Web browser were included for the test, nor any linked account for synchronization with the devices. In addition, the ads with lateral extension were not considered, because the sliding is vertical, from top to bottom.

***4. Results***

For the analysis of Web advertising, three types of dynamic Web pages were considered, tested in three different browsers: Chrome by Google, Mozilla Firefox, and Safari by Apple Inc. These analyzed websites were:

* MSN: www.msn.com/es-mx
* Sanborns: www.sanborns.com.mx
* AhorraSeguros: https://ahorraseguros.mx

Table 4 summarizes the results obtained for each URL in each Web browser, the total number of words (regular expressions) that appear as advertising on each evaluated website, the number of screenshots, and the execution time since the opening of the Web browser until final comparison.

Table. 4

Results of the algorithm evaluation

|  |  |  |  |
| --- | --- | --- | --- |
| Web Browser | Advertising | Total Screenshots | Time (sec) |
| Chrome  Firefox  Safari  Chrome | URL 1  106  103  118  URL 2  56 | MSN  9  9  9  Sanborns  4 | 11.636  12.144  14.539  4.449 |
| Firefox  Safari | 62  68 | 4  4 | 4.036  4.209 |
|  | URL 3 | Ahorra Seguros |  |
| Chrome  Firefox  Safari | 133  149  153 | 9  9  9 | 13.547  14.547  14.556 |

Based on the results obtained, it was possible to identify that through the Safari browser the algorithm detected a greater number of advertisements compared to the other two Chrome and Firefox browsers. This better identification of advertisements is due to the fact that the algorithm makes a better adjustment of the content of the Web page, and therefore it takes more time to perform the tests. In the case of Chrome and Firefox, both also detected a significant amount of advertisements, but when scrolling the Web page, a small amount of information was lost.

To determine the efficiency of the algorithm, a visual review of the information contained in the evaluated Web pages was performed. This review consisted of counting the RegGex in the Web ads, which should match the total number of words detected by the algorithm. Table 5 summarizes the results obtained from the comparison of matches between the words detected by the algorithm and the total existing words as part of the advertising in the Web.

Table. 5

Words identified by the algorithm with respect to the total of words with advertising content

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| RegExp/Coincidence | Visual Count | Chrome | Firefox | Safari |
| Microsft  News  IOS  Android  MSN  Rebaja  **Total**  $ | URL 1  22  9  10  10  12  15  **124**  URL 2  27 | MSN  22  9  10  10  3  13  **106**  Sanborns  18 | 14  9  0  0  10  12  **103**  23 | 22  9  10  10  2  14  **118**  24 |
| Libros  Perfumes  Tecnología  Videojuegos  **Total** | 3  2  3  2  **75** | 2  2  3  2  **56** | 1  2  3  2  **62** | 3  2  3  2  **68** |
|  | URL 3 | Ahorra | Seguros |  |
| Seguros  Seguro  Beneficios  Precios  Servicios  **Total** | 57  22  7  5  5  **172** | 48  22  5  4  2  **133** | 54  22  7  5  1  **149** | 56  21  7  4  4  **153** |

In the case of www.msn.com/es-mx, URL 1, a remarkable performance of the algorithm in detecting Web advertising through the Safari browser was achieved, obtaining a 95.16% confidence. While Chrome and Firefox browsers also reached significant confidence percentages, whose values were 85.48 and 83.06%, respectively.

For URL 2, www.sanborns.com.mx, where Safari reached effectiveness of 90.66%, Chrome 74.66%, and Firefox 82.66%. These results are due to the striking visual design understandable to human beings but complicated to analyze because some words collided, as well as the presence of logos and words with different size and typography concatenated with each other.

In the case of https://ahorraseguros.mx, URL 3, significant results were also achieved, with an 88.95% confidence in Safari, 86.62% in Firefox, and 77.32% in Chrome. In this test, the main factor in not achieving a higher level of confidence was the number of logos of different brands with a variety of fonts and backgrounds. This was the main problem when doing optical character recognition.

The particularities of the results obtained in the tests are due to the cuts and screen adjustments in the automatic sliding, that is, the configuration in each Web browser is varied, changing the way in which the information on the website is organized, and this causes loss of content.

***6. Conclusions***

The remarkable advances in contemporary technology also bring negative consequences for the end-user, such as the invasion of advertisements on the Web. Advertisements that are directed based on searches, needs, and interests of users.

Web advertising does not only consist of words or sentences that capture the user's attention with promotions or offers. For advertisers, it is vital that the user knows who is promoting it, regardless of whether the client really plans to buy said product, but the most important thing is to get your attention and remember the brand for future purchases.

The use of regular expressions was useful, in addition, the implementation of the database facilitated the organization for the detection of Web advertising, covering more cases of use of advertisements.

Expected results were found for the tests performed, with a percentage of reliability from acceptable to high, ranging from 74.66% to 95.16%, and the highest reliability rate was given through MSN, due to the simple design, common typography, constant sizes, absence of concatenated words, logos and striking designs.

Undoubtedly, the effectiveness of the algorithm in Safari is remarkable due to its way of distributing information on the screens of the end-users.

As future work it is intended to include more regular expressions in the database and make an extension in the algorithm, that is, to include artificial intelligence algorithms capable of recognizing advertising based on color patterns, size, and location of banners, text in bold and typography among other features in today advertisers.

**References**

1. Marketing Digital, ¿Qué es el marketing digital?, www.mdmarketingdigital.com/que-es-el-marketing-digital.
2. Redes Semánticas, http://tesis.uson.mx/digital/tesis/docs/9049/Capitulo1.pdf.
3. Marketing Online: Potencial y Estrategias, 2019. www.cecarm.com/Guia\_Marketing\_Online\_Potencial\_y\_Estrategias\_-\_CECARM.pdf-6120
4. Pomol, R., González, C., González, S. Una herramienta didáctica para el aprendizaje interactivo de expresiones regulares. 2013. http://repositorio.uigv.edu.pe/handle/20.500.11818/804
5. Beltrán, R., El uso de expresiones regulares en la detección de errores escritos: implicaciones para el diseño de un corrector gramatical, 2008. https://dialnet.unirioja.es/servlet/articulo?codigo=4007478
6. Gallego, A., La jerarquía de Chomsky y la facultad del lenguaje: consecuencias para la variación y la evolución. *Teorema*, 2008, vol 27-2, pp. 47-60.
7. García, I., Herramienta para la corrección automática de autómatas finitos, 2017. https://riull.ull.es/xmlui/handle/915/5846
8. Sánchez, J., López, L., Martínez, J., Solución para garantizar la privacidad en el Internet de las Cosas, *El profesional de la información*, 2015, vol. 24, pp. 62-70.
9. Ortiz, M., Aguilar, L., Marín, L., Los desafíos del marketing en la era del big data, *e-Ciencias de la Información*, 2016, vol. 6, pp. 1-30.
10. Riaño, D., Molero-Castillo, G., Velázquez-Mena,  A., Bárcenas E., Expresiones regulares para el tratamiento de privacidad de navegadores Web, *Abstraction and Application*, 2019, vol. 25, pp.121-130.
11. Cerezo, P., Ad blocking: el modelo publicitario digital, a revisión, *Cuadernos de periodistas: revista de la Asociación de la Prensa de Madrid*, 2016, pp. 81-89.
12. Londaitz, A., Publicidad en los celulares: Publicidad invasiva vs. derecho a la privacidad, *Thesis*, Universidad del Salvador, 2011. https://racimo.usal.edu.ar/4312
13. Google, Bienvenido a Google, la mejor empresa para trabajar, 2013. www.expansion.com/2013/08/23/directivos/1377273795.html
14. Jarvis, J., Y Google, ¿cómo lo haría?, 2000. https://narrativabreve.com/2013/10/libro-google-jeff-harvis.html
15. Leotta, M., Clerissi, D., Ricca, F., Spadaro, C., Comparing the maintainability of selenium webdriver test suites employing different locators: A case study, In Proceedings of 1st International Workshop on Joining AcadeMiA and Industry Contributions to Testing Automation, 2013. https://dl.acm.org/doi/10.1145/2489280.2489284
16. Gojare, S., Joshi, R., Gaigaware, D., Analysis and Design of Selenium WebDriver Automation Testing Framework, *Procedia Computer Science*, 2015, vol. 50, pp. 341-346.
17. Selenium Webdriver, 2017. www.tutorialspoint.com/selenium/pdf/selenium\_webdriver.pdf
18. Yih, W., Goodman, J., Carvalho, V., Finding Advertising Keywords on Web Pages, In Proceedings of the 15th International Conference on World Wide Web, 2006. https://dl.acm.org/doi/pdf/10.1145/1135777.1135813
19. Mei, T., Li, L., Tian, X., Tao, D., Ngo, C., PageSense: Toward Stylewise Contextual Advertising via Visual Analysis of Web Pages, In IEEE Transactions on Circuits and Systems for Video Technology, 2018. dl.acm.org/doi/abs/10.1109/TCSVT.2016.2598702
20. Sánchez, D., Viejo, A., Privacy-preserving and advertising-friendly web surfing. *Computer Communications*, 2018, vol. 130, pp. 113-123.
21. Krammer, V., An Effective Defense against Intrusive Web Advertising, In Sixth Annual Conference on Privacy, Security and Trust, 2008. https://ieeexplore.ieee.org/document/4641268
22. Sajjad, K., Automatic license plate recognition using Python and Opencv, College of Engineering, 2010. https://pdfs.semanticscholar.org/bddf/1200eb17f239e4dce2a9cec938eb8cf305f5.pdf
23. Patel, C., Patel, A., Patel D., Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study, In International Journal of Computer Applications, 2012. https://research.ijcaonline.org/volume55/number10/pxc3882784.pdf
24. Vallez, M., Keyword Research: métodos y herramientas para identificar palabras clave, *BiD: textos universitaris de biblioteconomia i documentació*, 2011, vol. 27, pp. 1-14.
25. Slamet, C., Andrian, R., Maylawati, D., Darmalaksana, W., Ramdhani, M., Web Scraping and Naïve Bayes Classification for Job Search Engine, In the 2nd Annual Applied Science and Engineering Conference, 2018. https://iopscience.iop.org/article/10.1088/1757-899X/288/1/012038/pdf