A Web Scraping based approach for data research through social media: An Instagram case

Ismael Camargo-Henríquez [1] and Yarisel Núñez-Bernal [2]
[1, 2:] Centro de Investigación, Desarrollo e Innovación en Tecnologías de la Información y las Comunicaciones, Universidad Tecnológica de Panamá
[1:] ismael.camargo@utp.ac.pa – [2:] yarisel.nunez@utp.ac.pa

2022 V Congreso Internacional en Inteligencia Ambiental, Ingeniería de Software y Salud Electrónica y Móvil (AmITIC) | 978-1-6654-8586-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/AmITIC55733.2022.9941290

Abstract* — *This paper presents a Web Scraping based approach to solve the data extraction problem in Instagram. Data from Instagram can be helpful in different research contexts. We describe the theoretical characterization of our proposal and its implementation in a real-case scenario. Also, we provide preliminary results and potential implementation challenges. The main contribution of our work is an instrument for research applying social media data analytics. Thus, researchers can take advantage of Instagram as a rich and dynamic data source.

Keywords* — *web data extraction, web scraping, web crawler, social media, big data, software engineering

I. INTRODUCTION

Every second, millions of valuable contents are created on social media. However, an emergent problem for data scientists is: *how to transform this content into data to generate useful knowledge?* Particularly, knowledge about social interactions between individuals and their contexts, either in society or organizations.

In this sense, Web Scraping is presented as a valuable technique to provide insights into massive amounts of data generated because of users' interactions on the Web. This technique may be useful in social media, where the data and information produced are difficult to extract and analyze through manual methods due to their volume and constant changes.

Particularly, government institutions with a presence on social media such as Instagram, offer first-hand and openly, official information related to different issues of public interest associated, for example: health, economy, education, etc.

Thus, this paper addresses, an approach focused on the data extraction from Instagram. We believe that it is possible using Web Scraping, to explore a public social media account related to an official government institution and obtain a dataset that can support, a data analysis, regarding situations of social, political, or economic interest, such as the COVID-19 pandemic. So, useful knowledge may be generated to promote correct decision-making, based on the trends provided by the data obtained and social perception.

Therefore, the main contribution of this paper is a description of a Web Scraping based approach to obtaining key data from a social media such as Instagram, it's outlining in structured data entities for their storage and retrieval, as well as the potential implementation using Software Engineering techniques. Additionally, it presents an approximation of results obtained with the proposed approach, to point out the feasibility in a real research scenario and encourage reflections about other potential uses and applications for this proposal.

II. WEB SCRAPING: CONCEPT

In many existing social media such as Instagram, most users allow access to the information produced by the content they generate. However, in these and other associated cases, it is impossible to directly access, download, extract or preserve the content, making it difficult to analyze it later [1]. Although this data retrieval could be done manually, it would be a highly time-consuming and excessively inefficient task, given the speed and volume at which this data is produced.

Faced with this scenario, a new approach is needed to automate and speed up this process. Thus, *"Web Scraping"*, also called *"Web Harvesting"* [2], *"Web Data Extraction"* [3][4][5], or even *"Web Data Mining"* [6], has emerged. In a precise definition, Web Scraping can be described as *"the construction of an agent to download, analyze and organize data from the Web in an automated way"* [7]. In this sense, Web Scraping is a technique that transfers and automates by software, the tasks of extracting, copying, and storing data from its source, executing them much faster, in real-time, and more correctly than a human effort.

The development of a Web Scraping process involves associating different technical components and specific knowledge along with a series of at least, three phases [8][6]. These phases and the set of knowledge and components needed are shown in *Fig. 1*.

In the *Website Analysis* phase (see *Fig. 1*), it is necessary to examine the underlying structure behind the content offered, to understand the form and presentation of the data, usually delivered in HTML. Thus, is required a general understanding of the DOM architecture and the structure of markup languages such as HTML, CSS, XML, etc.

The tasks to be automated for data extraction are developed during the *Website Crawling* phase (*Fig. 1*). These tasks are designed to perform the navigation of the content under analysis. Generally, these tasks can be programmed using programming languages such as Python, R, or Java due to the popularity of these languages in Data Science and the availability of libraries for certain reading tasks in complex HTML or XML structures (e.g., RVEST, Beautiful Soup).

In the Data Organization phase, the obtained data is processed. Here, it is necessary to perform cleaning, processing, and structuring tasks for subsequent storage and analysis. Considering the volume of data, its complexity, and structure, activities to clean and transfer it to a standard storage format such as CSV, JSON, TXT, can be carried out by using programming languages (e.g., R, Java, Python). Likewise, these tasks can be performed using libraries based on Natural Language Processing. Additionally, knowledge of Databases (e.g., MySQL, PostgreSQL) could be used to provide queries, filters, updates, etc.

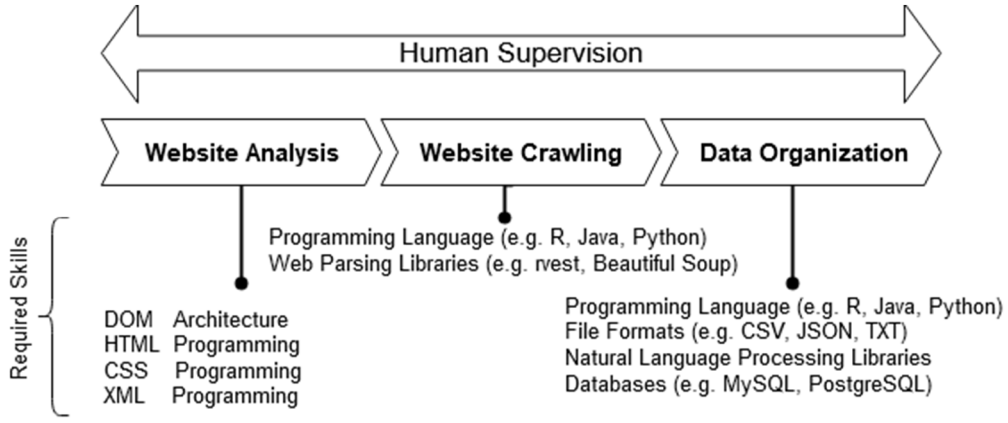

Fig. 1. Phases, components, and knowledge required to develop a Web Scraping process.

It should be noticed that, throughout the process, *Human Supervision* is required to manage the phases (see *Fig. 1*) although the automation of the tasks is done by software, the human element will be necessary to follow up and evaluate aspects such as the testing of the programmed tasks, intermediate data, as well as the storage, cleaning, and analysis of it.

A. Some Web Scraping models

Although Web Scraping is a relatively new technique, there are different implementation models for a variety of scenarios. Some of these models are reviewed in detail in the works of [5] and [1]. However, *Table. 1* summarizes the most common categories in which Web Scraping is used, with greater or lesser variations.

Table 1. Web Scraping models.

	Model	Description
A	Mimicry	This Web Scraping model works by using predefined rules from a template that recreates or mimics DOM selectors [9] from the original content to be analyzed. However, its effectiveness is reduced by changes in the template structure, requiring reprogramming tasks to know how to find the necessary data. Tools such as Import.io [10] or Mozenda [11] use this model.
B	Weight Measurement	This model uses a generic algorithm that parses the DOM tree [9] and measures the weight of words in each branch. The algorithm chooses a node as the starting point and extracts the text from all child nodes. The main advantage of this mechanism is that it can adapt to structural changes. However, the results are usually quite noisy.
C	Differential	This model assumes that two DOM documents on the same website will only differ in content [1]. The algorithm overlaps both documents removing only the differences, which are the required data between them.
D	Machine Learning	The main principle of this Web Scraping model is to train an algorithm using a large sample of content analyzed manually. Using machine learning [12] a statistical measurement is made of where the main text block is located compared to the other text blocks. The algorithm deduces by itself where the text is usually located. Thus, the larger the sampling, the more accurate the algorithm will be.

III. PROPOSED APPROACH

The proposed approach uses model *B* described in *Table. 1*, which starts by reading the DOM tree and exploring specific texts within the structure.

This model was chosen due to its ease of development and several technical advantages for the context of the study. Aspects such as the flexibility for adapting changes and a relative speed in performing searches within the DOM tree were considered.

Although this Web Scraping model can generate *"noise"* in the data gathered, this issue can be solved in the Data Organization phase, as previously described in Fig. 1.

In our approach, CSS classes identified in the *Website Analysis* phase of *Fig. 1* are used for this exploration. In general, the proposed approach as a solution performs Web Scraping according to the guidelines outlined in *Algorithm-1.*

Algorithm-1. Web Scraping Approach for Instagram Data Extraction.

```
Input:   - USR: Valid username
         - PSW: Valid password
         - URL: HTTPS address to Instagram
              e.g., 'https://www.instagram.com'
         - TPA: Target public account on Instagram
              e.g., 'https://www.instagram.com/minsapma'
 1  procedure WebScraping ()
 2     HTMLDoc ← NavigateTo(URL)
 3     LOGIN ← SetLogin(HTMLDoc, USR, PSW)
 4    if LOGIN = True then
 5      HTMLDoc ← NavigateTo(URL+'/'+TPA)
 6      POSTS ← HTMLDoc.getTag_Name('a')
 7      for each POST in POSTS do
 8         ID_POST ← POST.getTag_Name('/p/')
 9         POST_CAPTION ← POST.getClass_Name('MOdxS')
10         POST_DATE ← POST.getTag_Name('time')
11         COMMENTS ← HTMLDoc.getClass_Name('C4VMK')
12         for each COMMENT in COMMENTS do
13          COMMENT_TEXT ← COMMENT.getAttribute('text')
14          SaveComment(ID_POST, COMMENT_TEXT)
15         end for
16         SavePost(ID_POST, POST_CAPTION, POST_DATE)
17      end for
18    end if
19  end procedure
```

In a more detailed description of this algorithm, it starts by reading as inputs, a user (*USR*) and password (*PSW*) valid for Instagram access context. Likewise, the use of a target account with public access (*TPA*) is also considered. As described above, this work intends to be able to extract official data from a government institution with an account on Instagram.

For the study, the account of the *Ministerio de Salud de Panamá*, whose acronym is *MINSA*, was evaluated, being its official Instagram account *minsapma*. Selenium library [13] was used to browse and automate Web Scraping tasks. Python 3.10.4 was used as the programming language for this.

Thus, as described in lines 2 and 3 of *Algorithm-1*, the navigation to the Instagram Web (*URL*) is automated, the registered account is accessed and then the *login* to it is validated (see line 4). Once a successful login (True) is achieved, the target account profile (*minsapma*) is navigated to, and the generated HTML structure is acquired. Here the algorithm performs a specialization on the process as follows: (a) It explores and extracts the data associated with a post (see line 7) and (b) It continues to examine and retrieve the comments related to that post (see line 12).

In both cases, using the *Weight Measurement Approach* (see *Table. 1*), the notations or tags corresponding to the different elements required (see lines 8 to 10) are extracted from the HTML structure by performing the *Website Crawling* phase (see *Fig. 1*).

For example, for each *POST* in a list of *POSTS*, it will extract key data, such as the identifier of that post (*ID_POST*), publishing date (*POST_DATE*), and the descriptive text related to the post (*POST_CAPTION*).

The notations found (*'a'*, *'p'*, *'MOdxS'*, *'time'*, *'C4VMK'*, etc.), correspond to CSS tags defined within the HTML structure, these were obtained in the *Website Analysis* phase (see *Fig. 1*). The Web Scraping performed with our algorithm does not extract any values associated with multimedia elements of the posts, such as images or videos, as well as, for example, the identity of those who comment on the posts. This follows the guidelines discussed in [14][15] and [16] for a legal and ethical Web Scraping project.

Finally, each post and its comments are saved in storage format (TXT, JSON, CSV) for later retrieval, cleaning, or analysis (see lines 14 and 16). About this, the design decision chose a JSON file-based intermediate structure to optimize the processing of the collected data (texts) and reduce concurrency over a database, because, in preliminary tests, there were limited computational resources. This JSON structure corresponds to the format shown in *Source-1*.

Source-1. JSON data storage structure.

```
1   {
2       "posts": [
3           {
4               "idPost": "",
5               "postDate": "",
6               "postText": "",
7               "comments": [
8                       {"commentText": ""},
9                       {"commentText": ""},
10                      {"commentText": ""}
11              ]
12          }
13      ]
14  }
```

After intermediate storage, collected data were processed using cleaning algorithms that allowed removing *"noise"* by filtering out hashtags, mentions, emojis, etc. Finally, cleaned data were transferred to a database, whose data entity design is represented in *Fig. 2*, closing the *Data Organization* phase suggested in *Fig. 1*.

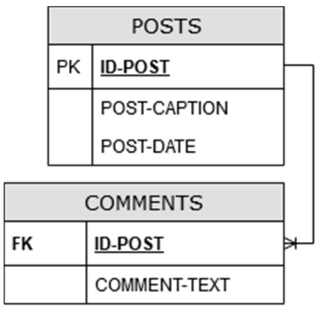

Fig. 2. Data entity design for final storage.

IV. RESULTS AND DISCUSSION

The proposed approach allowed getting so far, a partial sample of 1142 posts and 7309 comments. These results represent approximately 13% of the expected total. The dataset, once completed, will be used in a sentiment analysis study regarding social perception about official reports on the COVID-19 pandemic in Panamá.

Although our approach has achieved satisfactory results in preliminary tests as shown in *Fig. 3*, with better computational capabilities (e.g., cloud computing) it would be possible to improve performance, in terms of time, as well as underlying data cleaning and storage tasks.

	ID_POST	COMMENT_TEXT
1	CaYe6ZuOX_-	Eres muy útil, receptivo en las comunicaciones, conocedor y profesion...
2	CaYe6ZuOX_-	En el centro de salud de alcalde Díaz no quieren atender y solo dan cu...
3	CaYe6ZuOX_-	Quiero saber si tiene vigencia o se puede hacer la denuncia. Por lo me...
4	CaYe4O6uVyn	Dios los tenga en a gloria los que perdieron la Vida..
5	CaYe4O6uVyn	¿¿¿11 muertos??? Ahora no vengan con que ninguno estaba vacunado...
6	CaYe4O6uVyn	gracias por los datos
7	CaYe4O6uVyn	EL VIRUS BAJANDO Y ENTONCES ??POR QUÉ SEMEJANTE CANTIDAD D...
8	CaYet-ROxqq	Esas estadísticas están todas convenientemente analizadas
9	CaYet-ROxqq	no he podido vacunarne porque cierran muy temprano
10	CaYet-ROxqq	yo no he visto a ningún niño morirse y casi ninguno tiene. Vacuna
11	CaYet-ROxqq	Esos horarios TODOS estan mal!!! En todos los puntos a las 3pm dejan ...
12	CaYSn2vLw5t	Saben que contienen el polietinelglicol?
13	CaYSn2vLw5t	Que extraño los mismos efectos secundarios que da el polietinelglicol ...
14	CaYSn2vLw5t	Ahora yo no puedo reunirme con mi familia porque ya es algo malo . S...
15	CaYSn2vLw5t	Solo falta la D para que diga poldietinelglicol
16	CaYSn2vLw5t	Qie es el polietinelglicol?
17	CaYSRQhLmSx	Cuándo hay otra vacunación?
18	CaYSRQhLmSx	porke maña no asen una jornada de tersera dosi en la escuela porfirio ...
19	CaYR90UraD2	Buen dia aun estan vacunando en metro mall?
20	CaYR90UraD2	Hasta qué día estarán aplicando dosis pediátrica?
21	CaYR90UraD2	Aquí también aplican la 3ra dosis ???!
22	CaYR90UraD2	Hoy están vacunando ?
23	CaYR90UraD2	Publicidad engañosa. Ni westland mall. Ni la Licas Barcenas ni la escuel...
24	CaYR90UraD2	Estuve a la 1 de la tarde en Albrook para que dijeran que ya no estaba...
25	CaYR90UraD2	En Albrook no aceptan el código QR de Panamá digital por no tener las...
26	CaYR90UraD2	Hasta cuando vacunaran en los centros comerciales?
27	CaYR90UraD2	Vacunan este fin de semana en los mall?
28	CaYR90UraD2	Este domingo estarán aplicando las dosis en Albrook??
29	CaYR90UraD2	Donde estarán vacunando el sábado 26
30	CaYR90UraD2	Sean serios con los horarios en metro mall dicen hasta la 5 cuando van...

Fig. 3. Data sample for comments collected with the Web Scraping based approach.

Nevertheless, the feasibility of the proposed approach has been proven up to this point, as well as its potential use in other areas of data research applied to social media.

For instance, a direct benefit of implementing this proposal is that it could support similar domains such as market research, opinion mining on products or services, measurements of customers' satisfaction or dissatisfaction, etc.

This, by considering the widespread use of social media as a tool for marketing goods and services, notably driven during the COVID-19 pandemic.

Additionally, a further advantage offered by this approach is the capability to obtain a large amount of data in an extremely short time, reducing the effort in this task and transferring it towards other activities such as data analysis or data cleaning.

Another interesting property is that the proposed approach is adaptable. In technical terms, a review of *Algorithm-1* supports the idea that, although its implementation was oriented for Instagram, the principles discussed could be technically transferred with relatively few changes to other areas of use.

V. POTENTIAL CHALLENGES AND BIASES

The following is a brief description of the potential restrictions and challenges when replicating the proposed approach described in this work.

1. *Changes in the HTML document structure*: Design modifications to the HTML document could affect key sections or tags needed to extract the required data. However, the algorithm can be adjusted to deal with these changes.
2. *Data volume and computational capabilities*: The proposed approach may require high processing and storage capabilities. It is possible that, under certain hardware or network conditions, *"lags"* may be created both in extracting and storing data.
3. *Excessive data requests and user banning*: A constant data request to the social media server could be considered an external attack, leading to the user's banning.
4. *Extraction of short-lived posts*: Stories with a short duration are not considered by the approach. Although they generate interactions, do not produce an evaluable volume of data.
5. *Data extraction limits*: Due to its design, our approach does not include start or end points for stopping data collecting (e.g., dates). However, with some changes to *Algorithm-1*, this capability could be added.
6. *Private account data access*: The proposed approach does not provide any access to data in private accounts, because it would represent a privacy violation.

VI. CONCLUSIONS

This paper has proposed and described a Web Scraping based approach for key data extracting from social media, such as Instagram. In this regard, the main aspects to consider for the realization of a development process using this technique and how it can be approached theoretically and technically have been shown.

In this sense, we show the algorithmic strategy used by our approach to extract and collect data to generate a partial sample of comments and posts as results. These data were obtained with the Instagram account of the *Ministerio de Salud de Panamá*. Likewise, the mechanisms used for the storage of these data were illustrated.

On the other hand, the potential technical challenges that could hinder the replication of the approach in other implementation contexts were outlined.

An interesting contribution offered by this work was proving the viability of the approach. That contributes to the research and study of data mined from social media and how these may be used for potential inferences of knowledge in a way that is consistent with current social and technological realities.

Thus, from the perspective of the problem of data extraction in social media, the approach proposed provides an attractive solution space that may allow the extension and development of other mechanisms and tools to support research in this area.

ACKNOWLEDGEMENTS

This work was supported by the Universidad Tecnológica de Panamá (UTP) through the Centro de Investigación, Desarrollo e Innovación en Tecnologías de la Información y las Comunicaciones (CIDITIC).

REFERENCES

[1] R. Diouf, E. N. Sarr, O. Sall, B. Birregah, M. Bousso, and S. N. Mbaye, "Web Scraping: State-of-the-Art and Areas of Application," in *2019 IEEE International Conference on Big Data*, 2019, pp. 6040–6042. doi: 10.1109/BigData47090.2019.9005594.

[2] P. A. Johnson, R. E. Sieber, N. Magnien, and J. Ariwi, "Automated web harvesting to collect and analyse user-generated content for tourism," *Curr. Issues Tour.*, vol. 15, no. 3, pp. 293–299, 2012, doi: 10.1080/13683500.2011.555528.

[3] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, "Web data extraction, applications and techniques: A survey," *Knowledge-Based Syst.*, vol. 70, pp. 301–323, 2014, doi: 10.1016/j.knosys.2014.07.007.

[4] J. Myllymaki, "Effective Web Data Extraction with Standard XML Technologies," in *Proceedings of the 10th International Conference on World Wide Web*, 2001, pp. 689–696. doi: 10.1145/371920.372183.

[5] M. A. Bin Mohd Azir and K. B. Ahmad, "Wrapper approaches for web data extraction : A review," in *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*, 2017, pp. 1–6. doi: 10.1109/ICEEI.2017.8312458.

[6] B. Singh and H. K. Singh, "Web Data Mining research: A survey," in *2010 IEEE International Conference on Computational Intelligence and Computing Research*, 2010, pp. 1–10. doi: 10.1109/ICCIC.2010.5705856.

[7] S. vanden Broucke and B. Baesens, "Introduction," in *Practical Web Scraping for Data Science: Best Practices and Examples with Python*, Apress, 2018, pp. 3–23. doi: 10.1007/978-1-4842-3582-9_1.

[8] V. Krotov and M. Tennyson, "Research Note: Scraping Financial Data from the Web Using the R Language," *J. Emerg. Technol. Account.*, vol. 15, no. 1, pp. 169–181, 2018, doi: 10.2308/jeta-52063.

[9] S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm, "DOM-Based Content Extraction of HTML Documents," in *Proceedings of the 12th International Conference on World Wide Web*, 2003, pp. 207–214. doi: 10.1145/775152.775182.

[10] Import.IO, "Enterprise scale eCommerce data to drive growt - Import.io." 2022. [Online]. Available: https://www.import.io/

[11] U. Baskaran and K. Ramanujam, "Automated scraping of structured data records from health discussion forums using semantic analysis," *Informatics Med. Unlocked*, vol. 10, pp. 149–158, 2018, doi: 10.1016/j.imu.2018.01.003.

[12] E. Vargiu and M. Urru, "Exploiting web scraping in a collaborative filtering-based approach to web advertising.," *Artif. Intell. Res.*, vol. 2, no. 1, pp. 44–54, 2013, doi: 10.5430/air.v2n1p44.

[13] Selenium, "Selenium automates browsers. That's it!" Mar. 2022. [Online]. Available: https://www.selenium.dev/

[14] J. K. Hirschey, "Symbiotic relationships: Pragmatic acceptance of data scraping," *Berkeley Technol. Law J.*, vol. 29, pp. 897–927, 2014, doi: 10.2139/ssrn.2419167.

[15] R. O. Mason, "Four Ethical Issues of the Information Age," *MIS Q.*, vol. 10, no. 1, pp. 5–12, 1986, doi: 10.2307/248873.

[16] B. Ives and V. Krotov, "Anything you search can be used against you in a court of law: Data mining in search archives," *Commun. Assoc. Inf. Syst.*, vol. 18, no. 1, p. 29, 2006, doi: 10.17705/1CAIS.01829.