WEB DATA EXTRACTION WITH ROBOT PROCESS AUTOMATION. STUDY ON LINKEDIN WEB SCRAPING USING UIPATH STUDIO

Marcu Florentina, *The Bucharest University of Economic Studies, Bucharest, ROMANIA*

ABSTRACT:

In 2015 „a full 90 percent of all the data in the world has been generated over the last two years".[4] From that time to the present, the data available on the internet is increasing rapidly, and along with these, the web scraping tools have significantly evolved. This paper presents the RPA technology concept, the industries in which it applies, and its application in web scraping, a technique employed to extract a huge amount of data from websites. After this brief overview, will be presented an example of data extraction from LinkedIn job vacancies and a way to automatically save generated information in the CSV format data file using UiPath automation tool. This research can be a general guideline for anyone who wants to extract data easily, in a short time, with zero costs and from any site that allows or disallows the downloads of information.

KEY WORDS: Robotic Process Automation, UiPath, Web Scraping, LinkedIn

1. INTRODUCTION

According to "The social economy: Unlocking value and productivity through social technologies" study, published by McKinsey Global Institute, only 39% of employees time is spent for role-specific tasks, 28 percent is spent reading and answering e-mail, 19 percent for searching and gathering information and the remaining 14% being used for communicating and collaborating internally.[11] All these tasks, which accounts 61 percent of the total working time, can be automated using the Robotic Process Automation technology. A similar situation is encountered when a researcher looking for data on websites and arises the problem of downloading them, data can only be extracted manually and the time spent doing this can be quite long, from hours to days. In response to this need, web scraping tools have appeared and allow the extraction of data (numerical, categorical, text data and so on) from any site using methods created by RPA software vendors. In addition, using all the advantages offered by RPA technology, complex applications can be made to combine web scraping tools with automatic saving of the generated file in the format desired by the user.

2. ROBOTIC PROCESS AUTOMATION (RPA)

A literature research identified several definition for RPA, all of this provides the main purpose of this technology: human repetitive tasks automation using robots.

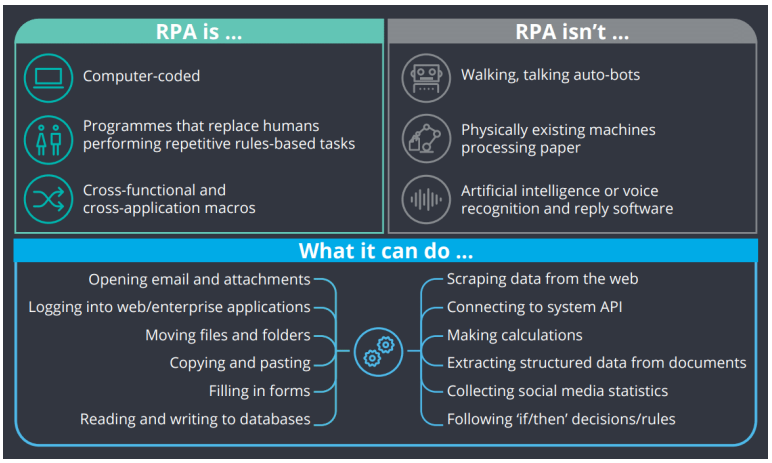

Figure 1. Definition of Robotic Process Automation
Source: https://research.aimultiple.com/what-is-robotic-process-automation/

RPA „is one of the most advanced technologies"[3] and is used in many industry, such as:

- telecommunications
- banking (call center operations: can help the employees with information retrieved from all of bank systems about the clients, the onboarding process, help desk, credit card applications)
- public administration (healthcare and education)
- retail (e-commerce, Supply Chain Management: update the status of orders, identification of the products in storage and removal of them)
- insurance (claims processing, manual data entry, form registration)
- human resources (automatic search of candidates according to job requirements, onboarding employees).

The benefits of tasks automation are costs reduce, lower operational risk (data entry errors), improved internal processes and execution time, increase productivity and data quality, reduced workload, employee satisfaction increase, the solutions of RPA could be used by non-IT people. So, the benefits cannot be denied and there are many tools that can be used to put into practice any repetitive and monotonous tasks such as Blue Prism, UiPath, Automation Anywhere, HelpSystems, Pega, OpenConnect, G1ANT, WorkFusion and so on.

3. THE PROCESS OF WEB DATA EXTRACTION

Nowadays, there are many information that can be easily retrieved from internet and widely used in research, development and forecast. Web scraping is a technique used to extract data from websites (extract product details and product prices from e-commerce websites job requirements and the benefits offered by companies from job ads, details about companies – name, location, website, phone, financial results of all banks from Romania, customer contact details and so on), to collect information and then save into a format desired by the user (csv, excel, json).

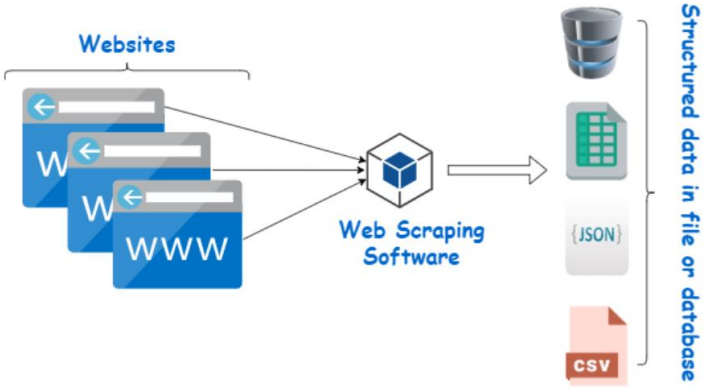

Figure 2. The process of web data extraction
Source: https://www.webharvy.com/articles/what-is-web-scraping.html

In fact, not infrequently, website do not allow to save a copy of the dataset for personal use. This can get complicated when the only option is to extract data manually and everything becomes a waste of time. Here comes the web scraping process that will perform this tasks in a few moments.

4. LinkedIn Web Scraping using UiPath Studio

The UiPath Studio is an application that can be downloaded for free (community version) from the internet and has advanced automation tools, including web scraping tools. "UiPath can extract literally anything you can see in a web browser. This includes statistics, finance and stock info, real-estate data, product catalogs, search-engine results, job listings, social networks feeds, customer opinions, and competitive pricing. Within a company, you can find even an even larger variety of data formats that UiPath can handle: reports, dashboards, customers, employees, finance, and medical data that you need to transform and migrate."[10]

In this section, will be presented the possibility of extracting information about jobs posted on the LinkedIn platform. This process includes the following steps:

Step 1: **Open** the application UiPath Studio, **create** a new project with the name "WebScrapingLinkedIn" and then click on the button Create.

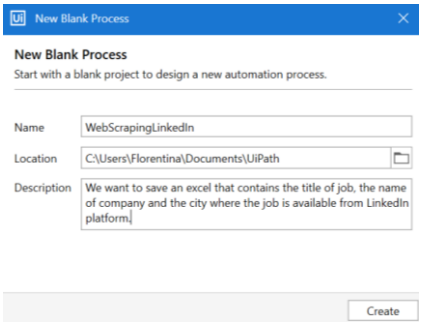

Figure 3. Create a new project in UiPath Studio

Step 2: Open LinkedIn and search for a job (for example IT jobs)

Step 3: Click on the Data Scraping from the Design tab

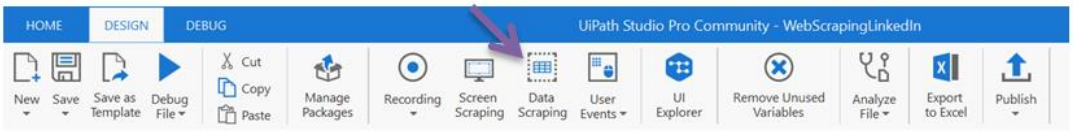

Figure 4. The Design tab in UiPath Studio

Step 4: A pop-up called "Select Element" will open, then click Next. The first column of the dataset to be extracted is the job titles, so it will be selected the first job title found on the recruitment platform (IT Support Specialist- Bucuresti). A pop-up called "Select Second Element" will open, click on the button Next and select the second job title (IT Trainee), so that UiPath Studio will create a pattern of this information.

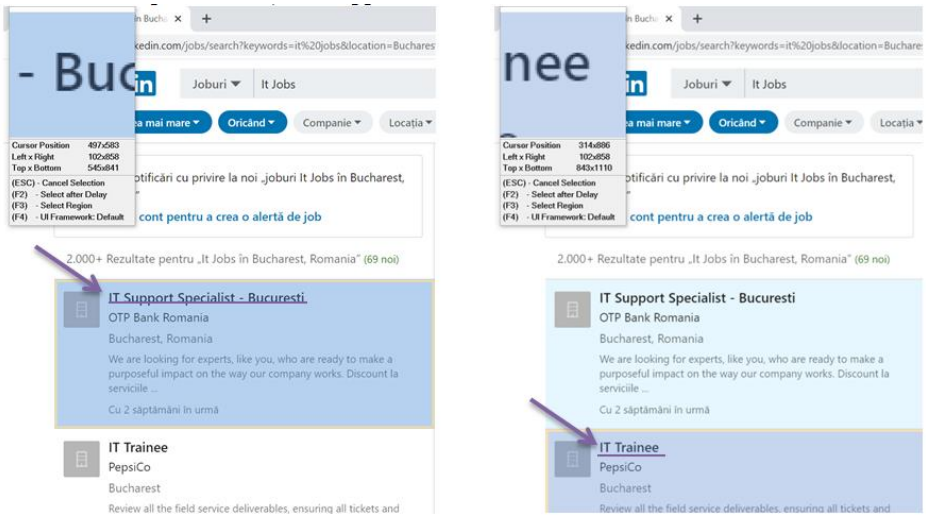

Figure 5. Create a pattern for extracting job titles

Step 5: Modify the column headers and choose to extract URLs, then click Next. The data set will be provided after that.

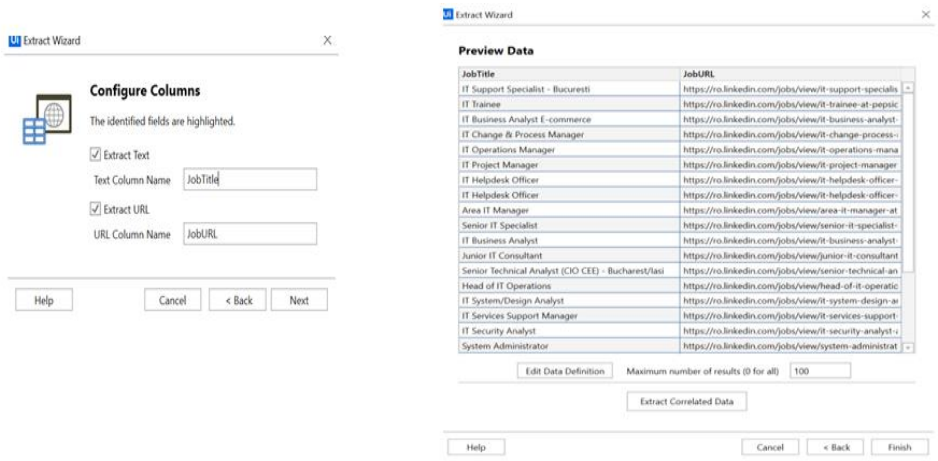

Figure 6. Modify the column headers and the dataset generated (with JobTitle and JobURL columns)

In this section, can be edited the maximum number of results (the default value is 100). To extract additional information (company name - information positioned under the job title - or the city), select Extract Correleted Data and then return to the Step 4. After all columns have been selected, click on Finish.

Step 6: The project is complete now and a sequence is generated in Uipath Studio.
To save the document in excel format, click on the Write CSV from Activities Panel and drag it to the work space and connects this activity with Attach Browser window, write the file name and use the method ExtractDataTable, like in the figure below:

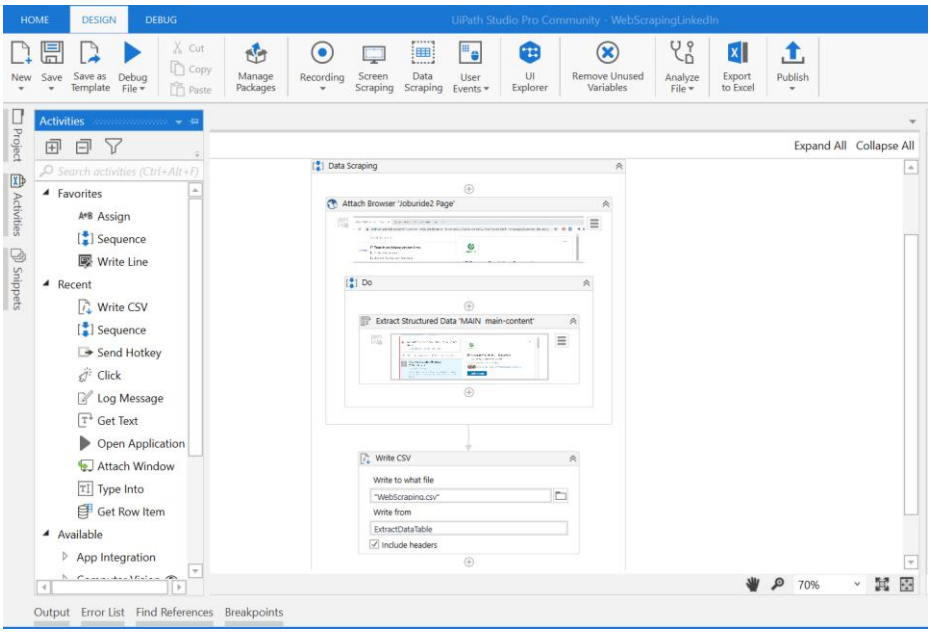

Figure 7. The final project created in UiPath

Click on the Debug File from the Design tab and Run the project with the LinkedIn page open. The file will be saved in the location specified in Step 1 in a CSV format and will contain a number of rows equal to the one selected in Step 5.

	A	B	C	D	E
1	JobTitle	JobURL	CompanyName	City	
2	IT Support Specialist - Bucuresti	https://ro.linkedin.c	OTP Bank Romania	Bucharesc, Romania	
3	IT Trainee	https://ro.linkedin.c	PepsiCo	Bucharest	
4	IT Business Analyst E-commerce	https://ro.linkedin.c	Auchan Retail România	Bucharest	
5	IT Change & Process Manager	https://ro.linkedin.c	ING Romania	Bucharest	
6	IT Operations Manager	https://ro.linkedin.c	Ubisoft	Bucharest	
7	IT Project Manager	https://ro.linkedin.c	Enel România	Bucharest	
8	IT Helpdesk Officer	https://ro.linkedin.c	BCR	Bucharest	
9	IT Helpdesk Officer	https://ro.linkedin.c	EveryMatrix Ltd	Bucharest	
10	Area IT Manager	https://ro.linkedin.c	Confidential	Bucharest	
11	Senior IT Specialist	https://ro.linkedin.c	Deutsche Telekom Services E	Bucharest	
12	IT Business Analyst	https://ro.linkedin.c	ENGIE Romania	Bucharest	
13	Junior IT Consultant	https://ro.linkedin.c	PRAS Consulting	Bucharest	
14	Senior Technical Analyst (CIO CEE) - B	https://ro.linkedin.c	UniCredit Services	Bucharesc, Romania	
15	Head of IT Operations	https://ro.linkedin.c	Bitdefender	Bucharesc, Romania	
16	IT System/Design Analyst	https://ro.linkedin.c	Vodafone	Bucharesc, Romania	
17	IT Services Support Manager	https://ro.linkedin.c	Amazon	Bucharesc, Romania	
18	IT Security Analyst	https://ro.linkedin.c	ALTEN Romania	Bucharesc, Romania	
19	System Administrator	https://ro.linkedin.c	BCR	Bucharesc, Romania	
20	Information Technology Specialist - B	https://ro.linkedin.c	Kalypso	Bucharesc, Romania	
21	IT SW Project Manager	https://ro.linkedin.c	SII Romania	Bucharesc, Romania	
22	IT Organization Change and Commun	https://ro.linkedin.c	NXP Semiconductors	Bucharest	

WebScraping

Figure 8. The data set extracted from the jobs posted on the LinkedIn platform in excel format (with JobTitle, JobURL, CompanyName and City columns)

5. CONCLUSION

In this study I presented the data extraction process using Robotic Process Automation
and an example of extracting data from the LinkedIn site with UiPath Studio. From this data can be extracted information about the percentage of IT jobs in the total jobs posted or an overview of labor market supply and demand. The data can be included in comparative analyze between IT jobs and financial-accounting jobs, constructions, engineering and so on. So, the example presented can be used as a model in extracting data from any website: the eMAG products, Amazon products (such as price, specifications and so on), information from Wikipedia, booking sites, home appliance sites, recruitment platforms and so on.

This study can be a starting point for the deepening of RPA technology and the Uipath tool, developing data mining applications or a simple way to extract data from websites with the help of which any interested person can create a graph, analyze data from various sources or make predictions. For academicians, data scientists or data engineers, students and researchers this paper represent a good way to implement web scarping in the process of obtaining data for their future research.

REFERENCES

[1] Lusiana C., Dewi M., Alvin C., Social Media Web Scraping using Social Media Developers API and Regex, 4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 2019 https://www.sciencedirect.com/science/article/pii/S1877050919311561

[2] Osman C. C., Robotic Process Automation: Lessons Learned from Case Studies, Bucharest, 2019 http://revistaie.ase.ro/content/92/06%20-%20osman.pdf

[3] Somayya M., Rajesh M. H., Durgesh K.J., The Future Digital Work Force: Robotic Process Automation (RPA), TECSI Laboratório de Tecnologia e Sistemas de Informação - FEA/USP, 2019 https://www.redalyc.org/jatsRepo/2032/203261541001/html/index.html

[4] ScienceDaily, Big Data, for better or worst: 90% of world's data generated over last two years, 2013 https://www.sciencedaily.com/releases/2013/05/130522085217.htm

[5]***https://docs.uipath.com/studio/docs/about-data-scraping

[6]***https://docs.uipath.com/studio/docs/example-of-using-data-scraping

[7]***https://docs.uipath.com/activities/docs/write-csv-file

[8]***https://www2.deloitte.com/content/dam/Deloitte/fr/Documents/services-financiers/publications/deloitte_global-robotics-survey-2018-full-report.pdf

[9]***https://ro.linkedin.com/jobs/search?keywords=IT&location=Bucharest%2C%20Bucharest%2C%20Romania&trk=homepage-jobseeker_jobs-search-bar_search-submit&redirect=false&position=1&pageNum=0

[10]**https://www.uipath.com/developers/video-tutorials/web-data-extraction-automation

[11]**https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-social-economy