SMARTINTERNZ

STANLEY BUILD-A-THON

(ROBOTIC PROCESS AUTOMATION)

TEAM : DATA PIRATES

PROJECT TITLE: *DATA SCRAPING*

TEAM MEMBERS :

- *Khansa Nazeer*
- *Umaima Adeena Waheed*
- *Rumandla Shreya*
- *Zainab Unisa*
- *Badugula Mounika*
- *Vadla Spandana*

DATA SCRAPING

Data Scraping is the process of importing information from the website into a spreadsheet or local file saved on a computer. It is an efficient way to get data from the web or to channel that data to another website.

Table of Contents

OVERVIEW 4
AUTOMATED DATA SCRAPING WITH TOOLS 5
Data Scraper (Chrome Plugin) 5
WebHarvy 5
import.io 6
PURPOSE 6
Blue Prism 6
EXISTING SYSTEM 7
Human copy-and-paste 7
Text pattern matching 7
HTTP programming 7
HTML parsing 7
Vertical aggregation 8
DOM parsing 8
Semantic annotation recognizing 8
Computer vision web-page analysis 8
PROPOSED SYSTEM 9
REQUIREMENTS 10
Software Requirements: 10
Pre-requirements for Blue Prism 10
FLOW CHART 11
OUTPUT SCREEN 12
ADVANTAGES OF DATA SCRAPING 13
DISADVANTAGES OF DATA SCRAPING 14
APPLICATIONS 15
OTHER USES 15
CONCLUSION 16
FUTURE SCOPE 16
REFERENCES 16

OVERVIEW:

Data scraping, also known as web scraping, is the process of importing information from a website into a spreadsheet or local file saved on a computer. It is one of the most efficient ways to get data from the web, and in some cases, to channel that data to another website.

Popular uses of data scraping include:

- Research for web content/business intelligence
- Pricing for travel booking sites/price comparison sites
- Finding sales leads/conducting market research by crawling public data sources (e.g., Yell and Twitter)
- Sending product data from an e-commerce site to another online vendor (e.g., Google Shopping)

AUTOMATED DATA SCRAPING WITH TOOLS

Getting to grips with using dynamic web queries in Excel is a useful way to understand data scraping. A dedicated data scraping tool is more effective if data scraping needs to be used regularly.

Here are a few of the most popular data scraping tools on the market:

Data Scraper (Chrome Plugin)

Data Scraper slots straight into our Chrome browser extensions, allowing us to choose from a range of ready-made data scraping "recipes" to extract data from whichever web page is loaded in our browser.

WebHarvy

WebHarvy is a point-and-click data scraper with a free trial version. Its biggest selling point is its flexibility – we can use the tool's in-built web browser to navigate to the data we would like to import, and can then create our own mining specifications to extract exactly what we need from the source website.

import.io

Import.io is a feature-rich data mining tool suite that does much of the hard work for us. It has some interesting features, including "What's changed?" reports that can notify us of updates to specified websites – ideal for in-depth competitor analysis.

PURPOSE

Blue Prism

Blue Prism provides browsers, windows, java-based automation and more. The business layer is encapsulated in an object. This means all application interactions are done within the object level. The object contains a component called an Application Modeler (AM). This component helps model the application and reveals the DOM elements exposed by the application in question.

Through the AM, we can specify the type of application, in this case a browser-based application. This takes us through a series of configured screens that help us model our application. Once the configuration is complete, a default element called Element1 is created. The Element1 has two buttons: identify and highlight. Clicking Identify reveals the available spy modes. Depending on the application, we might see HTML mode, Active Accessibility (AA) mode, Region mode, Win32 or UIA. It's worthy to note, UIA is available in newer versions of Blue Prism (version 5 and above).

EXISTING SYSTEM

Human copy-and-paste

The simplest form of web scraping is manually copying and pasting data from a web page into a text file or spreadsheet. Sometimes even the best web-scraping technology cannot replace a human's 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.

Text pattern matching

A simple yet powerful approach to extracting information from web pages can be based on the UNIX grep command or regular expression-matching facilities of programming languages (for instance Perl or Python).

HTTP programming

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

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. 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.

Vertical aggregation

There are several companies that have developed vertical specific harvesting platforms. These platforms create and monitor a multitude of "bots" for specific verticals with no "man in the loop" (no direct human involvement), and no work related to a specific target site. The preparation involves establishing the knowledge base for the entire vertical and then the platform creates the bots automatically. The platform's robustness is measured by the quality of the information it retrieves (usually number of fields) and its scalability (how quick it can scale up to hundreds or thousands of sites). This scalability is mostly used to target the Long Tail of sites that common aggregators find complicated or too labor-intensive to harvest content from.

DOM parsing

By embedding a full-fledged web browser, such as the Internet Explorer or the Mozilla browser control, programs can retrieve the dynamic content generated by client-side scripts. These browser controls also parse web pages into a DOM tree, based on which programs can retrieve parts of the pages. Languages such as X path can be used to parse the resulting DOM tree.

Semantic annotation recognizing

The pages being scraped may embrace metadata or semantic markups and annotations, which can be used to locate specific data snippets. If the annotations are embedded in the pages, as Microformat does, this technique can be viewed as a special case of DOM parsing. In another case, the annotations, organized into a semantic layer, are stored and managed separately from the web pages, so the scrapers can retrieve data schema and instructions from this layer before scraping the pages.

Computer vision web-page analysis

There are efforts using machine learning and computer vision that attempt to identify and extract information from web pages by interpreting pages visually as a human being might.

PROPOSED SYSTEM

Data Scraping is achieved using RPA Blue Prism Web automation. Robotics Process Automation is more cost-effective and easier to use than Application Programming Interface (API). Like people RPA uses user interface, thereby requiring nominal technical know-how and lesser dependence on IT, as many important components of business logic is implemented in the user interface, thus, it is beneficial than using API. Blue Prism RPA is probably one of the greatest technological boons of the 21st century. With the enormous volume of data generated and higher customer demands, efficiency and the quantity of work has become integral in the workforce. And the solution to this mammoth task at hand is- RPA Blue Prism. Blue Prism is a set of tools and libraries for RPA. It has two basic components- business objects, which interact with the user interface and the logic that runs the robot, called process.

REQUIREMENTS

Software Requirements:

- Blue Prism Software (6.10)
- Browser (latest version)
- Active Internet Connection

Hardware Requirements

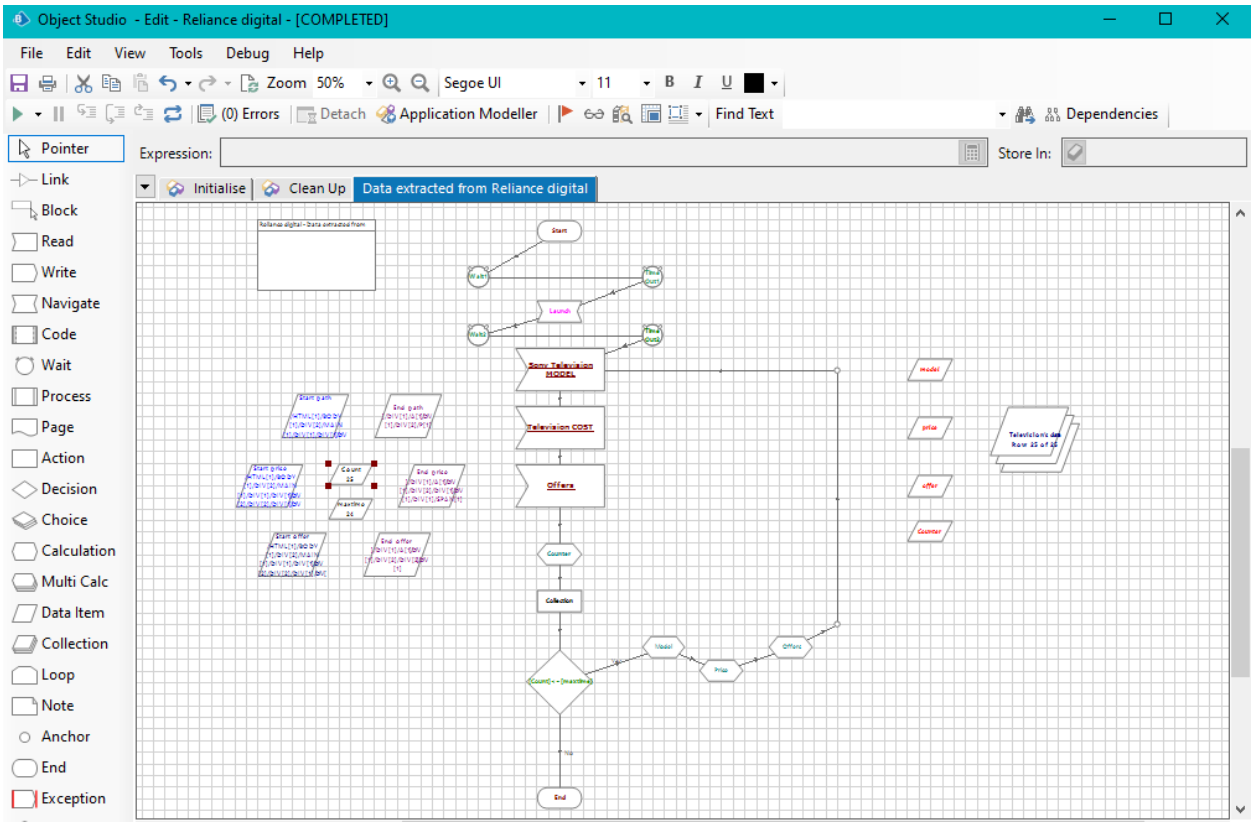
- Processor: INTEL CORE i5 Generation-10 (latest model)
- Processor Speed: 2.11 GHz
- RAM: 12GB

Pre-requirements for Blue Prism

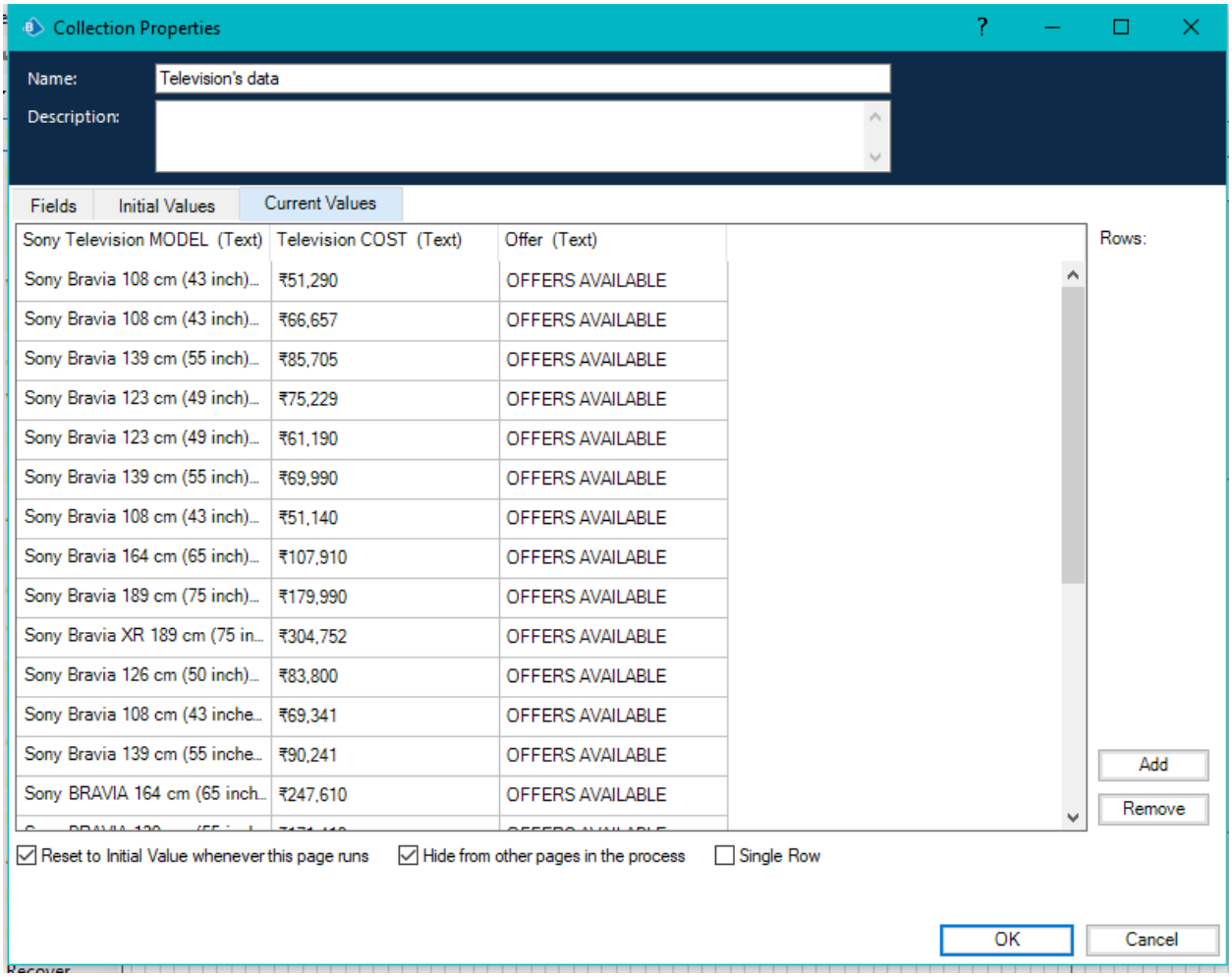
The following are the prerequisites for Blue Prism. Is the only software which:

1. Creates and supports a digital workforce of industrial strength and enterprise-scale.
2. Does not require IT skills to implement.
3. Can be implemented in sprints of 4 to 8 weeks (start to finish).
4. Is meager cost compared to the TCO of alternative solutions.
5. Provides tremendous payback with self-funding returns and an ROI that has been as high as 80%.
6. Can be managed within IT infrastructure and processes.

FLOW CHART



OUTPUT SCREEN



ADVANTAGES OF DATA SCRAPING

- Eliminates the need for manual data entry

- Reduces costly human errors
- No need for custom processing and filtering algorithms
- Easy drag-and-drop features eliminate the need to write scripts

DISADVANTAGES OF DATA SCRAPING

- Learning curve
- The structure of websites changes frequently
- It is not easy to handle complex websites
- To extract data on a large scale is way harder
- A web scraping tool is not omnipotent
- Your IP may get banned by the target website
- There are even some legal issues involved

APPLICATIONS

1. Lead Generation for Marketing

2. Price Comparison & Competition Monitoring
3. E-Commerce
4. Real Estate
5. Data Analysis
6. Academic Research
7. Training and Testing Data for Machine Learning Projects
8. Sports Betting Odds Analysis

OTHER USES

1. Scrape hotel/restaurant ratings and reviews from websites like TripAdvisor
2. Scrape hotel room prices and details from websites like Booking.com and Hotels.com
3. Scrape tweets related to an account or hashtag from Twitter
4. Scrape profile data from social networks like Facebook, LinkedIn etc. for tracking online reputation.
5. Scrape hospital/clinic websites to build a catalog of physicians including their contact details
6. Scrape images and profile data from Instagram
7. Crawl forums and communities to extract data from posts and authors
8. Scrape articles from various article/PR websites
9. Scrape data from various Government websites, most of which do not provide an easy way to download the data which they display

CONCLUSION

We have shown that Robotic Process Automation with a platform such as Blue Prism

helps in Data Scraping. Web scraping is an application of Robotic Process Automation which is use most industries. Either it is stock trading websites, e-commerce websites, commodities trading websites, etc, you can scrape the data from any of them based on your interest. Now, the problem with performing web scraping manually is that it is quite prone to errors and takes time. Also, the data present on the websites is never static. It gets a substancial amount of updated very frequently. So, the data that is stored at a point instance might not be accurat. So, industries can simply automate this task.

FUTURE SCOPE

The Internet is large, complex, and ever-evolving.90% of all the data in the world has been generated over the last two years. In this vast ocean of data, how does one get to the relevant piece of information? This is where web scraping takes over.

Web scrapers attach themselves to this beast and ride the waves by extracting information from websites at will. Granted, "scraping" doesn't have a lot of positive connotations, yet it happens to be the only way to access data or content from a website without RSS or an open API.

Web scraping faces testing times ahead. We outline why there may be some serious challenges to its future.

1. With the rise in data, redundancies in web scraping are rising. No more is a web scraping a domain of the coders; in fact, companies now offer customized scraping tools to clients which they can use to get the data they want. The outcome of everyone equipped to crawl, crawl, and extract is an unnecessary waste of precious manpower. Collaborative scraping could well heal this injury.

Here, where one web crawler does a broad scraping, the others crawl data off an API. An extension of the problem is that text retrieval attracts more attention than multimedia; and with websites becoming more complex, this enforces limited scraping capacity.

2. Easily, the biggest challenge to web scraping technology is privacy concerns. With data freely available (most of it voluntary, much of it involuntary), the call for stricter legislation rings loudest.

REFERENCES

- https://www.whizlabs.com/blog/rpa-blue-prism-introduction/
- https://en.wikipedia.org/wiki/Web_scraping
- https://www.helpsystems.com/blog/rpa-action-automated-data-scraping
- https://www.targetinternet.com/what-is-data-scraping-and-how-can-you-use-it/
- https://www.octoparse.com/blog/web-scraping-limitations
- https://www.webharvy.com/articles/web-scraper-use-cases.html