PRICE SCRAPE

INTELLIGENT SOFTWARE AGENTS PROJECT REPORT

ISA – GROUP MEMBERS
KAH GHI LIM (A0100172A)
CHNG YAN HAO (A0024023A)

Contents

1. Executive Summary 2

2. Problem Description 3

2.1 Project Scope and Objective 4

3. Software Agent System 5

3.1 Knowledge Acquisition 5

3.2 Knowledge Model 5

3.3 Existing Solutions 6

4. Solution 7

4.1 System Architecture / System Flow 7

4.2 System Scope 9

4.3 System Assumptions 9

4.4 System Features 10

4.5 Limitations 10

5. Conclusion 11

6. Bibliography 12

7. Appendices 13

1. Executive Summary

Digital transformation has accelerated in the last year, namely due to the COVID-19 pandemic, changing our daily lives from work to personal. COVID-19 has driven many processes into the digital world – from working at offices to working from home, from having physical meetings with our family and friends to having virtual ones, and from having physical stores to them setting up e-commerce websites.

With more e-commerce websites available online, prices of products can now be easily found online with a quick click of a button on one's preferred search engine. Consumers are also getting more tech-savvy and can easily compare prices from various sources to buy the most cost-competitive product. For businesses to set an informed pricing strategy to attract more customers and avoid losing their current market share, efforts would have to be made to monitor competitor's prices frequently.

In this project, we would like to propose our solution, named Price Scrape (PS). PS can automate monitoring of competitor prices using software robots, thereby allowing for up-to-date prices for the company's pricing strategy while allowing the employee to spend their worktime on more value-added tasks. In this report, we will focus on grocery shops for our use case scenario. We will also cover some of the limitations of our solutions and suggest possible features for future work.

2. Problem Description

In recent years, digital transformation has gained traction in Singapore – through the launching of the Smart Nation initiative locally by Prime Minister Lee Hsien Loong back in 2014. Some strategic national projects under the Smart Nation initiative include sharing of data sets across agencies, setting up GoBusiness and establishing e-payment services (Smart Nation Singapore, 2021). With more processes made digital and online, information such as product specification, model and price can be easily obtained from the web.

The transformation was further accelerated by the COVID-19 pandemic, as indicated by 26 percent of businesses and enterprises surveyed in Singapore in November 2020 (Hirschmann, 2021). Businesses as well as national and global economies were badly affected due to lockdowns and safe distancing measures, where many companies and industries turned to digital transformation to help shorten these gaps. In a global survey conducted by McKinsey (LaBerge et al, 2020), COVID-19 has accelerated the digitization of customer interactions by 4 years in the Asia-Pacific region as shown in Figure 1.

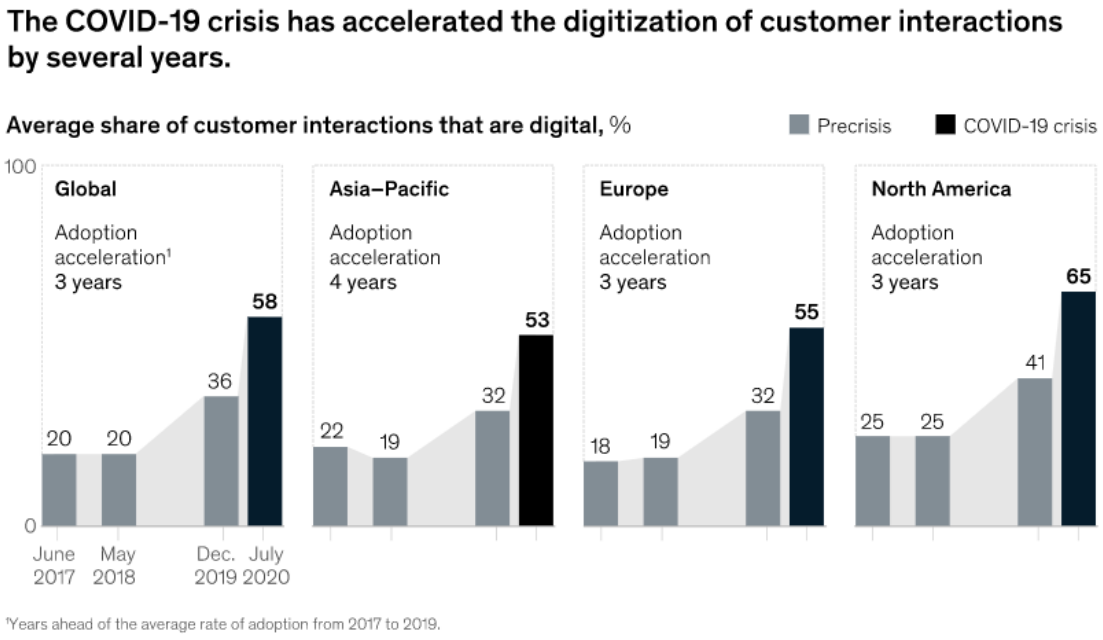


Figure 1 COVID-19 accelerated digitization of customer interactions

Consumers have moved towards online channels to do their shopping, perhaps due to the greater convenience compared to complying with contact tracing and safe distancing measures. In Jan 2019, only 5.5% of retail trade was online and 7.6% of supermarkets and hypermarkets were online. One year on, the proportion has increased to 10.3% and 11.6% for retail trade and

supermarkets respectively (Department of Statistics Singapore, 2021). Experts predict that this trend will set to continue, even as safe distancing measures ease (Ng, H. & Chen, S., 2020; Tan, S., 2020).

With so many products listed online, information can now be easily obtained such as product description and price. Consumers now have the option of choosing from more than one store to buy their desired items. Some may find it a hassle to compare between different products and prices across different websites and ultimately decide to stick to one store, thereby losing potential cost savings they may have enjoyed from another store.

From a business point of view, the company's sustainability is largely tied to its revenue and profit generated in a year. This means that businesses would need to regularly update their price strategy, to retain existing customers while attracting potential customers with special offers and deals. To ensure that their prices are up to date and cost-competitive, businesses would need to monitor and track competitor prices regularly. This is a significant undertaking for a human to execute, especially if there are many products available in their store.

These pain points are what we hope to resolve with our Price Scrape (PS) tool, with primary focus on grocery stores in Singapore such NTUC FairPrice, RedMart, Cold Storage and Sheng Siong.

2.1 Project Scope and Objective

PS is a software agent that takes in a user's list of items, goes to various websites, finds and records the actual item name and its price with just a simple push of a button. PS then runs a similarity score based on the input and the item name from the website and sorts the item with the highest score at the top. This allows the user to be able to validate the item scraped efficiently and utilise the information effectively.

PS is set in a way where users can freely adjust the items that they wish to scrape by providing their own list of items. It can also be expanded in scope to include other websites should the user have some understanding of XPath selectors. PS can also be scheduled to run at scheduled times, even when the user is sleeping, provided the computer is on. PS is written with a simple user interface and with it, we hope to save users valuable time during working hours, freeing up their time for more value-added tasks and increasing their efficiency and productivity.

3. Software Agent System

Software agent refers to “a component of software and/or hardware that is capable of acting exactingly in order to accomplish tasks on behalf of its user” (Nwana, 1998). Simply put, a software agent works like a personal assistant and interacts with software as how a human would, including keystrokes and clicks. Some advantages of a software agent is that it can do repetitive tasks without feeling tired or bored, have low error rates and can be scheduled to run automatically with minimal human intervention.

This is the backbone of our PS tool, where TagUI is utilised to scrape prices off websites with a simple click of a button. The system architecture and detailed features will be covered in Section 4.

3.1 Knowledge Acquisition

For users to compare prices and determine if their current price strategy is competitive, prices from various different competitors would have to be scraped, forming their database for comparison. Having a human do this manually would not only be tedious, but also error-prone.

Prominent supermarkets in Singapore, namely NTUC FairPrice, RedMart, Cold Storage and Sheng Siong, were selected to be the main websites to scrape from. This is due to their market prominence and as such, their websites should be more developed and stable, where we can then utilise XPath to specify the elements to be scraped and stored into the data base.

3.2 Knowledge Model

The item name scraped from the website would be compared against the actual input based on cosine similarity score, to confirm that the item scraped is indeed the item of interest. The score is used to sort the entire list of entire scraped such that the items deemed to be of higher relevance would be at the top of the list. This makes it easier for the user to verify the items and focus on items that may need to have price adjustment as part of their pricing strategy.

3.3 Existing Solutions

Aside from PS, there are other web scraping tools available online. Some features of the top 3 solutions listed by guru99 (Guru99, 2021) such as ScrapingBee, Octoparse and Scraping-bot are compared against our PS solution in Table 1.

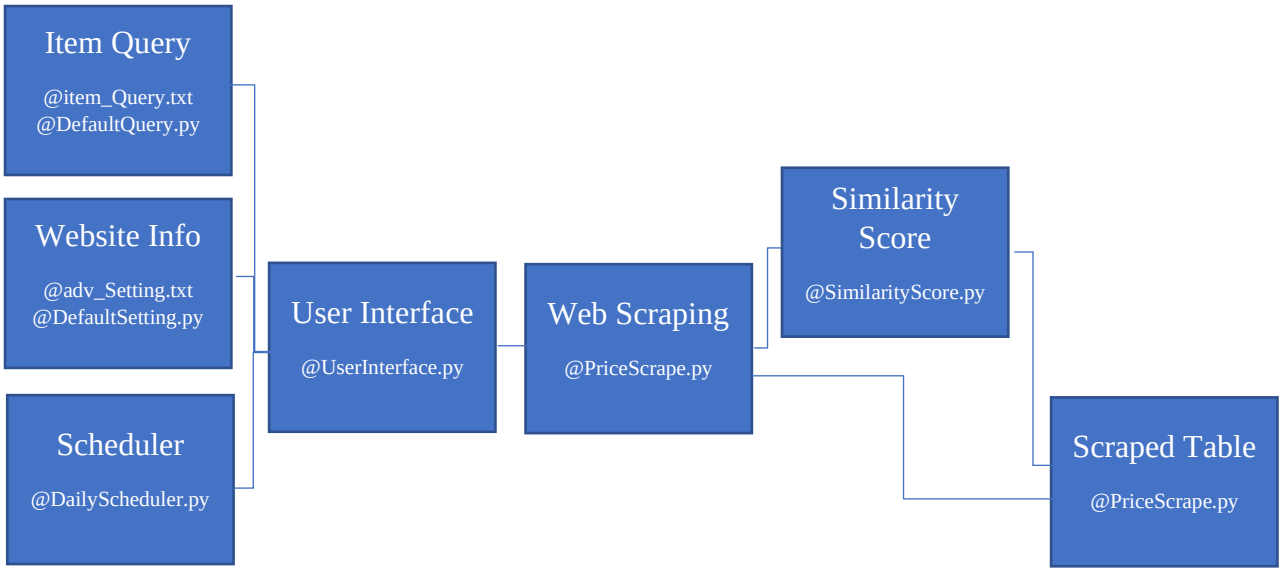
Table 1: Comparison between Price Scrape (PS) and other existing web scraping tools

	Price Scrape (PS)	**Scrapingbee**	**Octoparse**	**Scraping-bot**
Price (cheapest plan)	Free	1000 Free Credits + Paid Plan @ $29/mo	Free Trial + Paid Plan @ $75/mo	100 Free Credits + Paid Plan @ ~$72/mo
Coding required	No to minimal	Minimal	No	Minimal
Bot detection	Possible to get blocked by websites that monitor for bots	Rotating proxies	Automatic IP rotation	Able to avoid captcha and blocking
Output	.csv file containing specific information relevant to user	Entire HTML webpage	.csv, excel or API call	.json or raw HTML content of entire webpage

Although PS may encounter issues with websites that detect bots, these were not encountered for the four grocery stores used for this project. The information downloaded is also directly applicable to the user, instead of being in HTML code where non-tech users may have difficulty understanding the language. As such, we believe that PS has its unique role to play in the web-scrape space to help users in their work and improve their efficiency.

4. Solution

4.1 System Architecture / System Flow



i. Item Query

a) item_Query.txt: Stores user variable item_Query

b) DefaultQuery.py: Writes a default item_Query.txt

(use when item_Query.txt is missing / during set default)

ii. Website Info

a) adv_Setting: Stores all stores' links, selectors and time/item thresholds

b) DefaultSetting.py: Writes a default adv_Setting.txt

(use when adv_Setting.txt is missing / during set default)

iii. Scheduler

Takes in hh:mm from user to run scheduler daily.

iv. User Interface

A GUI with 5 buttons:

a) <Run Scrape Now>

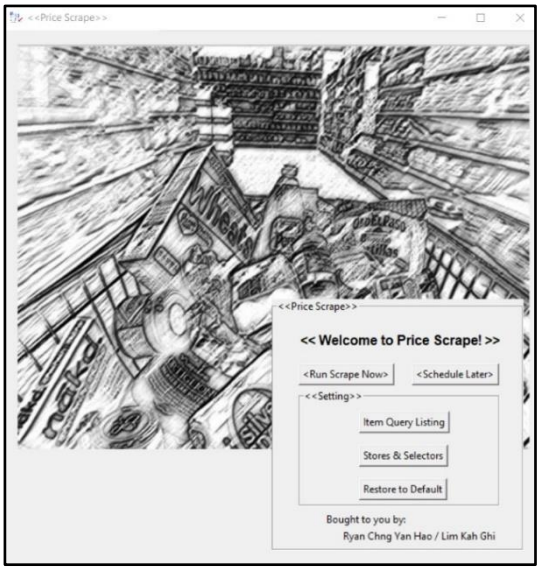
b) <Schedule Later>

c) Item Query Listing

d) Store & Selectors

e) Restore to Default

Use of buttons covered in QuickUserGuide.pdf



v. Web Scraping

Steps:

a) Read store information.
b) Read item query list.
c) Open browser.
d) Locate search box.
e) Enter “item 1” from query list.
f) Find first result name.
g) Find first result price, unit, availability using first result name as anchor.
h) Run cosine similarity between “item 1” and “first result name”.
i) Data scraped stored in python dictionary, append to a table.
j) Go to (b) for next item until end of list
k) Go to (a) for next store information until end of list
l) Appended table gets sorted by Item Query No. (ascending), Item Availability (ascending), Item Similarity (descending), Item Price (ascending) & Item Name.
m) Table is timestamped and populated as .csv file.

vi. Similarity Score

Use in web scraping step (h).

Steps:

a) Tokenize “query name” and “result name”.
b) Remove their English stop words.
c) Stemming.
d) Take union of “query name set” and “result name set”
e) Vectorize and count items in sets. i.e. “query name set”, “result name set” & “union”
f) Calculate score based on cosine similarity formula.

vii. Scraped Table

Item Query No.	Item Query	Store Name	Item Name	Item Price	Item Unit	Item Availability	Item Similarity
1	Philippines Banana	Sheng Siong					
1	Philippines Banana	RedMart					
1	Philippines Banana	ColdStorage	COLD STORAGE Banana Cavendish Aloha Philippines	$2.50 / 1000g	NA	Available	0.577
1	Philippines Banana	ColdStorage	COLD STORAGE Banana Cavendish Aloha Philippines	$2.50 / 1000g	NA	Available	0.577
1	Philippines Banana	NTUC	Just Fruit Just Banana	$5.95	30g	Available	0.408
1	Philippines Banana	NTUC	Orell's Glazed Banana Thins 200g	$6.80	200 G	Available	0.316
1	Philippines Banana	NTUC	Just Fruit Mix (Banana, Mango, Pineapple)	$5.95	30g	Available	0.289
1	Philippines Banana	NTUC	Sumifru Philippines Banana	$2.35	700g	Not Available	0.816
1	Philippines Banana	NTUC	Sumifru Sweet Mountain Banana	$3.95	820g	Not Available	0.354
1	Philippines Banana	NTUC	Orell's Glazed Banana Thins	$4.50	100g	Not Available	0.354
1	Philippines Banana	NTUC	Sumifru Poongmi Wang Banana	$4.95	820g	Not Available	0.354
1	Philippines Banana	NTUC	UFC Banana Sauce Regular 320g	$1.60	320 G	Not Available	0.316
1	Philippines Banana	NTUC	UFC Banana Sauce Regular 550g	$2.80	550 G	Not Available	0.316
1	Philippines Banana	NTUC	Orell's Glazed Banana Thins - Chocolate Flavour	$4.50	100g	Not Available	0.289
1	Philippines Banana	NTUC	Orell's Glazed Banana Thins - Cinnamon Flavour	$4.50	100g	Not Available	0.289
1	Philippines Banana	NTUC	Oh So Healthy! Guava Purple Yam Banana Fruit Crisps	$1.95	20g	Not Available	0.236
1	Philippines Banana	NTUC	Oh So Healthy! Mango Sweet Potato Banana Fruit Crisps	$1.95	20g	Not Available	0.236
1	Philippines Banana	NTUC	Oh So Healthy! Purple Yam Banana Coconut Fruit Crisps	$1.95	20g	Not Available	0.236
1	Philippines Banana	NTUC	Oh So Healthy! Purple Yam Banana Coconut Fruit Crisps	$1.95	20g	Not Available	0.236

Item Query No: Item position in the list

Item Query: Name of query, keyed into search bar

Item Store: User definable name for website & its information

Item Name: Name search result

Item Price: Price of item

Item Unit: Unit of item

Item Availability: If "Out of Stock", item is labelled not available

Item Similarity: Generated from web scraping step (h) and similarity score

Blank entries indicate that item is not found in the store.

4.2 System Scope

Price Scrape default setting provides website selectors for:

i. Sheng Siong
ii. NTUC
iii. RedMart
iv. ColdStorage

4.3 System Assumptions

i. Steps in web scraping are the same as stated in 4.1 Part v.
ii. Websites does not change their selectors.
iii. Important search results are unlikely to be found beyond first page.

4.4 System Features

Knowledge & Intelligence

i.	Rule-based web scraping:	Saves man-hours from scraping manually.
ii.	Similarity score:	Compares queries versus respective results based on cosine similarity; interprets how close is a query against its match without human inferring.
iii.	Smart sorting:	With most desired information at the top. i.e. User query order (first on top), similarity (highest on top) and price (lowest, non-budget price on top)

Technical Features

Ease of access:	Available in both attended (direct trigger) and unattended (scheduler) versions. Controls only involved 5 buttons.
Scalability:	Possibility to input larger sets of data. Allows complete list of products within the company. Advance setting available for power users who understand selectors and add more websites.
Compatibility:	Threshold setting for time / number of items. Compatible with slow/fast internet connections.

4.5 Limitations

The tool is at the mercy of the websites it is scraping from. Despite power users can edit the selectors, most common users will have difficulty editing the settings. The tool only scrapes the first loaded page with the assumption of relevant searches should be upfront. Tool is rigidly set to perform steps as per 4.1 Part v.

5. Conclusion

With COVID-19 accelerating digital transformation and changing our daily routines from physical to virtual and online, there will be many opportunities for our solution to be used. Our solution is flexible to cater for both non-tech savvy and tech savvy users, where the former can easily utilise PS to scrape for item information subjected to that four websites while the latter have the option to alter and expand the scope of PS to other websites and other item information as well.

With our solution, we hope to enable people to be more efficient and productive in this rapidly changing world.

6. Bibliography

Department of Statistics, Government of Singapore. 2021, Feb 1. *M601861 - Online Retail Sales Proportion (Out Of The Respective Industry's Total Sales), Monthly*. Retrieved from https://www.tablebuilder.singstat.gov.sg/publicfacing/createDataTable.action?refId=17033

Guru99. (2021). 15 Best Web Scraping Tools for Data Extraction in 2021 [Blog Post]. Retrieved from 15 Best Web Scraping Tools for Data Extraction in 2021 (guru99.com)

Hirschmann, R. 2021, Feb 10. *COVID-19 Impact on Business Digital Transformation in Singapore 2020*. Retrieved from https://www.statista.com/statistics/1200997/singapore-covid-19-impact-on-business-digital-transformation/

LaBerge, L., O'Toole, C., Schneider, J. & Smaje, K. 2020, Oct 5. *How COVID-19 has pushed companies over the technology tipping point—and transformed business forever*. Retrieved https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/how-covid-19-has-pushed-companies-over-the-technology-tipping-point-and-transformed-business-forever#

Ng, H. & Chen, S. 2020, May 25. *How Covid-19 is changing what Singaporeans shop for online.* Retrieved from https://www.straitstimes.com/singapore/covid-shopping

Nwana H.S., Ndumu D.T. (1998) A Brief Introduction to Software Agent Technology. In: Jennings N.R., Wooldridge M.J. (eds) *Agent Technology*. Springer, Berlin, Heidelberg. Retrieved from https://doi.org/10.1007/978-3-662-03678-5_2

Smart Nation Singapore, Government of Singapore. 2021, Mar 29. *Transforming Singapore Through Technology*. Retrieved from https://www.smartnation.gov.sg/why-Smart-Nation/transforming-singapore

Tan, S. A. 2020, May 25. *Online shopping trend set to stay after curbs ease, say analysts*. Retrieved from https://www.straitstimes.com/business/economy/online-shopping-trend-set-to-stay-after-curbs-ease-say-analysts

7. Appendices

1. Mapped System Functionalities

No.	System Function	Modular Course	Knowledge/Technique/Skills used
1	Web Scraping	RISM & SRBP	RPA via TagUI / XPath Selectors / Anchor Base logic
2	Similarity Score	IPA	Natural Language Processing - mimic human inferential skill on identifying similarity in product names
3	Scheduler	SRBP	Allows unattended bot use
4	Coding Phases (found in miscellaneous)	RISM	Launching/improving workable bots by phases via agile methodology

2. Project Proposal (uploaded into GitHub)

3. Installation and User Guide (uploaded into GitHub)

4. Individual project report (uploaded into GitHub)