



*42 SME INNOVATION 2022*

**TEAM CHANDRAN**

*Wong Ding Hang*

*Nazrin Shah*

*Lok Eu Lee*

*Hoo Yun Zhe*

TEAM MEMBERS



# PRICESHOP CHALLENGE

Develop a tool that scrapes  
information from the web, matches  
them automatically with our  
product database and generates a  
formatted CSV file

# SUMMARY

We developed a program that extract data with user provided list of URLs and generate a output for database updating that split into 3 component

## SCRAPER

Data scrapped with Scrapy and save in JSON format

## MATCHER

Read and match the saved data with database

## OUTPUT

Append and update information on matching product



## SCRAPY

Scrapy iterates through the provided list of URLs, and retrieves the HTML which is later converted into JSON, which we then use to query the required information



## CHALLENGES

- Since Scrapy only gets the HTML we needed a way to find the information that was hidden behind clickable elements. To get around this, we had to siphon through information that was hidden in the script tags
- To get around IP blocking we had to add throttling to the scraper, to make it seem less like a bot. We also rotated user agents so it seemed like we were making requests from different browsers

# HOW OUR SCRAPER WORK

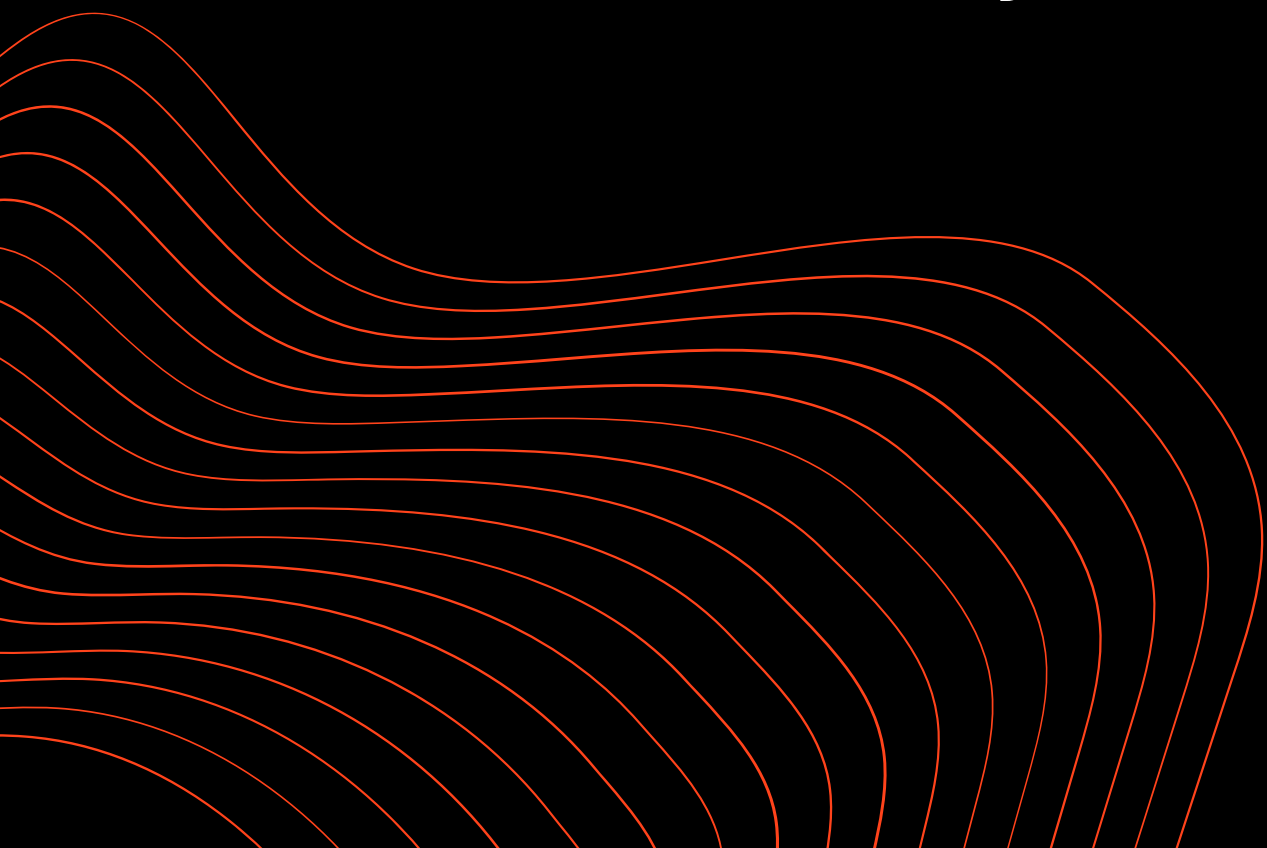
*—Scraping data with a single click*





# OUR SCRAPER ACHIEVEMENT

We managed to scrap all 66 URLs given  
by Priceshop within seconds



# TIME TO SCRAPE

with throttle

```
'start_time': datetime.datetime(2022, 6, 30, 21, 24, 7, 406161)}  
2022-07-01 05:24:24 [scrapy.core.engine] INFO: Spider closed (finished)  
scrapy crawl lazada -o lazada.json 1.76s user 0.13s system 10% cpu 17.873 total
```

without throttle

```
'start_time': datetime.datetime(2022, 6, 30, 21, 24, 45, 201605)}  
2022-07-01 05:24:46 [scrapy.core.engine] INFO: Spider closed (finished)  
scrapy crawl lazada -o lazada.json 1.47s user 0.12s system 74% cpu 2.136 total
```

# HOW DOES IT WORK?

01



## Tokenization

Tokenization of the database provided by splitting them into smaller chunk for pattern recognition on natural language processing

02



## Labeling

The token that has been collected during tokenization stage is analyzed and given the appropriate label and stored as a database for future pattern recognition

03



## Extracting

The scrapped data will be treated by natural language processing with the created pattern database and extract all the key information out with their corresponding label

04



## Matching

The received key information will be matched with the product name in database one at a time, from brand, to model, then specification to narrow down and determine the correct match

05



## Tabulating

The vendor information will be updated and appended on the matched product column and exported as a CSV file for backend storage



# TOOLS WE USED

- **Spacy** for natural language processing
- **Regex** for general tokenization
- **Pandas** for dataframe and data processing

# EXAMPLE OF OUR MATCHED DATA

Apple iPad 10.2-inch 9th Gen Wi-Fi + Cellular (2021)	256GB Silver	Apple iPad 10.2 (2021) (256GB) Wi-Fi + Cellular
Apple iPad 10.2-inch 9th Gen Wi-Fi + Cellular (2021)	64GB Silver	Apple iPad 10.2 (2021) (64GB) Wi-Fi + Cellular
Apple iPad 10.2-inch 9th Gen Wi-Fi + Cellular (2021)	64GB Space Grey	Apple iPad 10.2 (2021) (64GB) Wi-Fi + Cellular
Apple iPad 10.2-inch 9th Gen Wi-Fi + Cellular (2021)	256GB Space Grey	Apple iPad 10.2 (2021) (256GB) Wi-Fi + Cellular
Apple 10.9-inch iPad Air 5th Gen Wi-Fi [2022]	256GB Pink	Apple iPad Air (2022) (256GB) Wi-Fi
Apple 10.9-inch iPad Air 5th Gen Wi-Fi [2022]	256GB Purple	Apple iPad Air (2022) (256GB) Wi-Fi
Apple 10.9-inch iPad Air 5th Gen Wi-Fi [2022]	64GB Purple	Apple iPad Air (2022) (64GB) Wi-Fi
Apple 10.9-inch iPad Air 5th Gen Wi-Fi [2022]	256GB Space Grey	Apple iPad Air (2022) (256GB) Wi-Fi
Apple 10.9-inch iPad Air 5th Gen Wi-Fi [2022]	64GB Pink	Apple iPad Air (2022) (64GB) Wi-Fi
Apple 10.9-inch iPad Air 5th Gen Wi-Fi [2022]	256GB Starlight	Apple iPad Air (2022) (256GB) Wi-Fi
Apple 10.9-inch iPad Air 5th Gen Wi-Fi [2022]	256GB Blue	Apple iPad Air (2022) (256GB) Wi-Fi
Apple 10.9-inch iPad Air 5th Gen Wi-Fi [2022]	64GB Space Grey	Apple iPad Air (2022) (64GB) Wi-Fi
Apple 10.9-inch iPad Air 5th Gen Wi-Fi [2022]	64GB Blue	Apple iPad Air (2022) (64GB) Wi-Fi
Apple 10.9-inch iPad Air 5th Gen Wi-Fi [2022]	64GB Starlight	Apple iPad Air (2022) (64GB) Wi-Fi
Apple 10.9-inch iPad Air 5th Gen Wi-Fi + Cellular [2022]	64GB Space Grey	Apple iPad Air (2022) (64GB) Wi-Fi + Cellular
Apple 10.9-inch iPad Air 5th Gen Wi-Fi + Cellular [2022]	256GB Starlight	Apple iPad Air (2022) (256GB) Wi-Fi + Cellular
Apple 10.9-inch iPad Air 5th Gen Wi-Fi + Cellular [2022]	64GB Purple	Apple iPad Air (2022) (64GB) Wi-Fi + Cellular
Apple 10.9-inch iPad Air 5th Gen Wi-Fi + Cellular [2022]	256GB Space Grey	Apple iPad Air (2022) (256GB) Wi-Fi + Cellular
Apple 10.9-inch iPad Air 5th Gen Wi-Fi + Cellular [2022]	256GB Blue	Apple iPad Air (2022) (256GB) Wi-Fi + Cellular
Apple 10.9-inch iPad Air 5th Gen Wi-Fi + Cellular [2022]	256GB Pink	Apple iPad Air (2022) (256GB) Wi-Fi + Cellular
Apple 10.9-inch iPad Air 5th Gen Wi-Fi + Cellular [2022]	64GB Pink	Apple iPad Air (2022) (64GB) Wi-Fi + Cellular
Apple 10.9-inch iPad Air 5th Gen Wi-Fi + Cellular [2022]	64GB Blue	Apple iPad Air (2022) (64GB) Wi-Fi + Cellular
Apple 10.9-inch iPad Air 5th Gen Wi-Fi + Cellular [2022]	256GB Purple	Apple iPad Air (2022) (256GB) Wi-Fi + Cellular
Apple 10.9-inch iPad Air 5th Gen Wi-Fi + Cellular [2022]	64GB Starlight	Apple iPad Air (2022) (64GB) Wi-Fi + Cellular
Apple 10.9 Inch 4th Gen iPad Air Wi-Fi + Cellular [2020]	64GB Silver	Apple iPad Air (2020) (64GB) Wi-Fi + Cellular
Apple 10.9 Inch 4th Gen iPad Air Wi-Fi + Cellular [2020]	256GB Rose Gold	Apple iPad Air (2020) (256GB) Wi-Fi + Cellular

# With NLP We Achieved

91%








## OVERALL ACCURACY

Our matcher achieved a matching rate of 98% when only given single product pages, 75% when given multiple product pages and an overall 91% matching rate



# Why Use Natural Language Processing

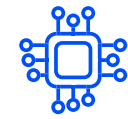
-  Dynamically extract tokens in different formats
-  Classification of tokens through pattern recognition
-  Scalability with machine learning
-  Selective matching with labelled data for enhanced precision
-  User friendly

A conceptual image featuring a hand typing on a keyboard. The entire scene is overlaid with a semi-transparent blue filter. In the upper left corner, several thin, white, wavy lines flow across the frame. The text 'THE FUTURE' is centered in a bold, white, sans-serif font.

THE FUTURE



# THINGS TO IMPROVE ON



## MACHINE LEARNING

With more data we can train a model to get more precise token matching. With a bigger data set machine learning is definitely the way to go as it scales with the size of database



## THREADING

When scraping a lot of websites, we can try to speed things up by threading. We can allocate threads to handle a few URLs at a time



## ROUTING

Routing is definitely a necessity when things scale up. Currently we're only able to access free proxies which are slow and unstable, but a company can easily afford premium proxies with consistent uptime and low latency



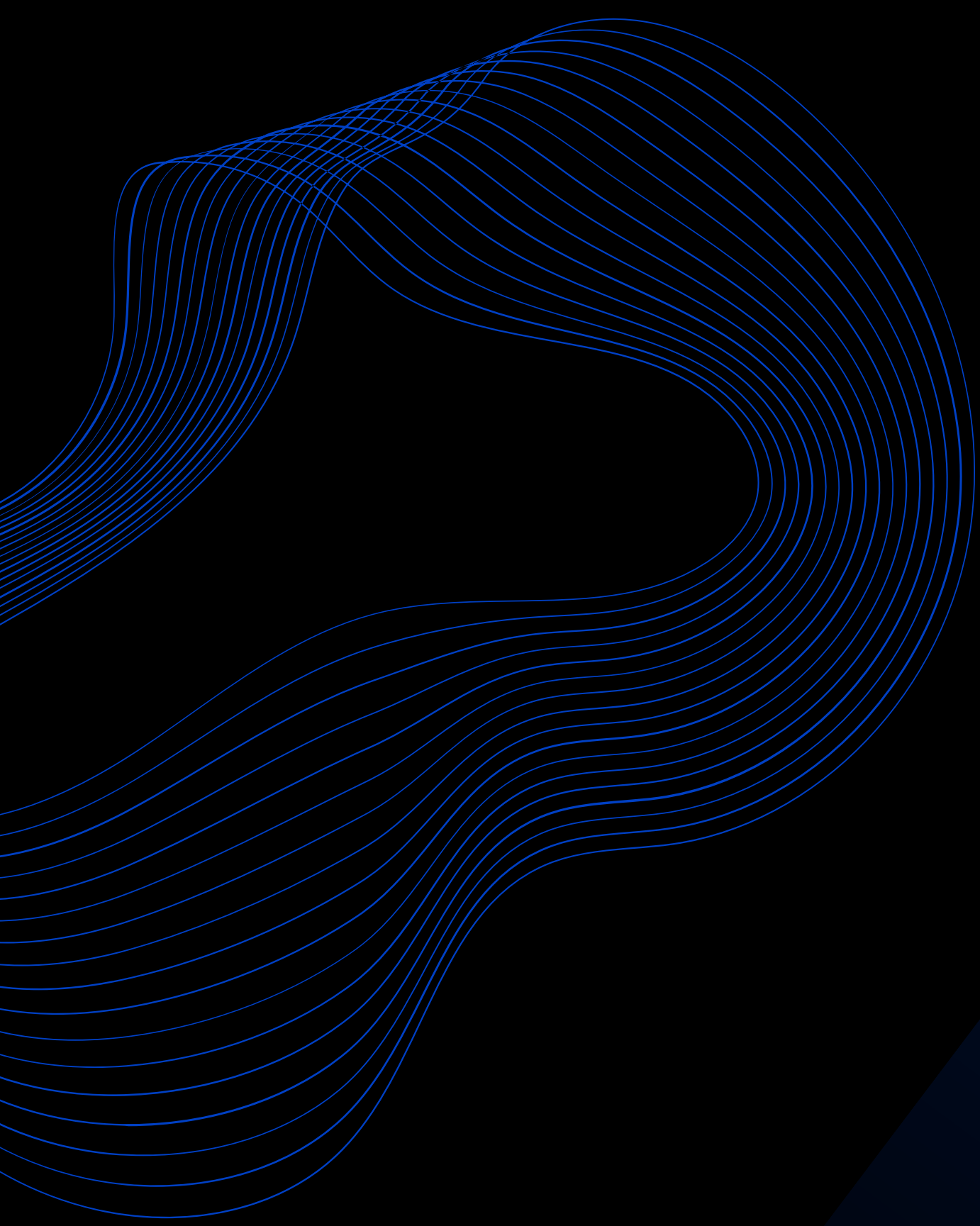
## PATTERN

With a trained model, we can generate a better pattern to generate a better dataset



## FRONT END UI

A front end UI will improve the user experience, instead of using a bash script, the user can use our tool through a beautifully designed UI



# THANK YOU

Thank you for listening!