Exotic README

English | □□□□ | □□□□

Exotic (representing Exotic Star) is a professional version of PulsarR, which contains a upgraded PulsarR server, a set of top e-commerce site scraping examples, and a applet for auto extraction supported by advanced AI.

Never write another web scraper. Exotic learns from the website, automatically generates all the extract rules, queries the Web as a database, and delivers web data completely and accurately at scale:

1. STEP1: automatically extract every field in webpages using advanced AI and generate extract SQLs
2. STEP2: test the SQLs and improve them to match frontend business requirements if necessary
3. STEP3: create crawl rules in the web console to run extract SQLs continuously and download all the web data to drive your business forward

There are already dozens of scraping cases for the most popular websites, we are constantly adding more cases.

Features

- Web spider: browser rendering, ajax data crawling
- High performance: highly optimized, rendering hundreds of pages in parallel on a single machine without be blocked
- Low cost: scraping 100,000 browser rendered e-comm webpages, or n * 10,000,000 data point each day, only 8 core CPU/32G memory are required
- Web UI: a very simple yet powerful web UI to manage spiders and download data
- Machine learning: automatically extract every field in webpages using unsupervised machine learning and generate extract rules and SQLs
- Data quantity assurance: smart retry, accurate scheduling, web data lifecycle management
- Large scale: fully distributed, designed for large scale crawling
- Simple API: single line of code to scrape, or single SQL to turn a website into a table
- X-SQL: extended SQL to manage web data: Web crawling, scraping, Web content mining, Web BI
- Bot stealth: IP rotation, web driver stealth, never get banned
- RPA: simulating human behaviors, SPA crawling, or do something else awesome
- Big data: various backend storage support: MongoDB/HBase/Gora
- Logs & metrics: monitored closely and every event is recorded

Requirements

- Memory 4G+
- The latest version of the Java 11 JDK
- Java and jar on the PATH
- Google Chrome 90+

Download

Download the latest executable jar:

```
wget http://static.platonic.fun/repo/ai/platon/exotic/exotic-standalone.jar
```

Build from source

Add the following lines to your .m2/settings.xml.

```
<mirrors>
    <mirror>
        <id>maven-default-http-blocker</id>
        <mirrorOf>dummy</mirrorOf>
        <name>Dummy mirror to override default blocking mirror that blocks http</name>
        <url>http://0.0.0.0/</url>
    </mirror>
</mirrors>
```

```
git clone https://github.com/platonai/exotic.git
cd exotic
mvn clean && mvn
cd exotic-standalone/target/
```

Run the standalone server and open web console

```
# Linux:
java -jar exotic-standalone*.jar serve

# Windows:
java -jar exotic-standalone[-the-actual-version].jar serve
```

Note: if you are using CMD or PowerShell on Windows, you may need to remove the wildcard * and use the full name of the jar.

If Exotic is running in GUI mode, the web console should open within a few seconds, or you can open it manually:

http://localhost:2718/exotic/crawl/

Run Auto Extract

We can use the `harvest` command to leans from a set of item pages using unsupervised machine learning.

```
java -jar exotic-standalone*.jar harvest https://shopee.sg/Computers-Peripherals-cat.11013247 -diagnose -refresh
```

The URL in the command above should be an portal URL, such as the URL of the product listing page.

Exotic visits the portal URL, finds out the best link set for item pages, fetches item pages and then learn from them.

Here is a snapshot of the result of auto extract using unsupervised machine learning for an e-comm site.

3. eXtracted **25** fields from page area **.product-briefing.flex.card.-Esc+w**

		T1C2	T1C3	T1C4	T1C5	T1C6	T1C7	T1C8	T1C9	T1C10
1	[SG Seller] USB 3.0 4 Ports High Speed 5Gb…	4.9	3k	ratings	8.3k	Sold	$6.90	$0.8 OFF		Bundle Deals
2	Easylabeltech USB /Bluetooth /WiFi Thermal…	4.9	1.5k	ratings	4.3k	Sold	$27.50 - $179.90	$10 OFF	Free shipping	$45.00 - $211.40
3	Apple 13 inch MacBook Air Laptop (Apple M…	4.9	301	ratings			$1,399.00		Free shipping	$1,449.00
4	[SG LOCAL SELLER] Displayport DP to HD…	4.9	920	ratings	3.6k	Sold	$5.79 - $6.29		Free shipping	$6.80 - $7.50
5	[✅SG Ready Stock] LED Star Night Light M…	4.9	612	ratings	1.7k	Sold	$22.50 - $32.50	$2 OFF	Free shipping	$50.00 - $69.90
6	CANON ORIGINAL INK 745 745XL 746XL B…	4.9	836	ratings	2.4k	Sold		$20.99 - $68.39	Free shipping	
7	ATZ High Speed HDMI v2.0 4K (1m / 1.5m / …	5.0	1.9k	ratings	7.6k	Sold	$6.11 - $8.81	$3 OFF		$6.79 - $9.79
8	ATZ USB 2.0 A-Male to B-Male Printer/ Scan…	4.9	177	ratings	606	Sold		$3 OFF		Shop Vouchers
9	(Local Stock) (GEBIZ&ACRA REG) PLA 3D …	4.9	373	ratings	2.6k	Sold	Buy (11 -..1) $14.80	$16.40	Free shipping	wholesale
10	(Local Stock) PLA 3D Printer Filament Stand…	4.9	360	ratings	1.7k	Sold		$20.00	Free shipping	
11	ATZ HDMI Cable 4K Ver 2.0 (0.5m / 1m / 2m…	4.9	866	ratings	3k	Sold		$3 OFF		Shop Vouchers
12	[SG Local Seller] HDMI to VGA adapter 1080P	4.9	2k	ratings	7.1k	Sold	$5.29 - $7.88		Free shipping	$6.30 - $8.80
13	100*150mm 350PCS Waterproof AWB Ther…	5.0	243	ratings	3.2k	Sold	$5.90	$10 OFF	Free shipping	Shop Vou Free
14	COMFAST 1300Mbps WiFi Adapter Dual Ba…	4.9	926	ratings	2.5k	Sold	$9.98	$0.3 OFF		Shop Vouchers
15	[21x25] Leather Mousepad Waterproof Mous…	4.9	1.7k	ratings	5.4k	Sold	$2.80 - $3.15	25*21		$8.00
16	ARCTIC P12 P14 PWM/PST ARGB RGB 3-P…	5.0	148	ratings	939	Sold		$1 OFF	Free shipping	Shop Vouchers
17	[✅SG Ready Stock] Laptop Stand Portable …	4.9	10.4k	ratings	29.7k	Sold	$9.90 - $21.99	$2 OFF	Free shipping	$19.90 - $39.90
18	[SG] Macbk PU Leather Laptop Case (11.6/1…	4.9	1.3k	ratings	4.1k	Sold	$9.99 - $16.99	$2 OFF	Free shipping	$23.99 - $29.99
19	[SG] LionShield Webcam Ultra Thin Privacy …	4.9	2k	ratings	5.4k	Sold	$3.99 - $7.99	$2 OFF		$12.99 - $19.99
20	[Free Brush] Switch Lubricant Mechanical Ke…	5.0	1.8k	ratings	6.4k	Sold	$3.99 - $9.49		Free shipping	$11.49 - $20.00
tp	0	20	20	20	20	19	19	14	26	13
fp	0	0	0	0	0	0	0	0	4	0
fn	0	0	0	0	0	12	7	2	11	0
tn	0	0	0	0	0	1	1	0	4	0
precision	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.87	1.00
recall	0.00	1.00	1.00	1.00	1.00	0.61	0.73	0.87	0.70	1.00
f1	0.00	1.00	1.00	1.00	1.00	0.76	0.84	0.93	0.78	1.00

The best CSS selectors for each field are generated automatically, you can use these rules for web scraping in the old-fashioned way:

Generated CSS Paths

T1C2	div.-Esc+w.card.product-briefing div.HLQqkk div.flex-column.imEX5V span
T1C3	div.HLQqkk div.flex-column.imEX5V div.W2tD8- div.MrYJVA.Ga-lTj
T1C4	div.HLQqkk div.flex-column.imEX5V div.W2tD8- div.MrYJVA
T1C5	div.HLQqkk div.flex-column.imEX5V div.W2tD8- div.Wz7RdC
T1C6	div.HLQqkk div.flex-column.imEX5V div.W2tD8- div._45NQT5
T1C7	div.HLQqkk div.flex-column.imEX5V div.W2tD8- div.Cv8D6q
T1C8	div.-Esc+w.card.product-briefing div.HLQqkk div.imEX5V div.pmmxKx
T1C9	div.-Esc+w.card.product-briefing div.HLQqkk div.imEX5V div.mini-vouchers__label
T1C10	div.imEX5V div.PMuAq5 div.flex-no-overflow span.voucher-promo-value.voucher-promo-absolute-value
T1C11	div.HLQqkk div.imEX5V div.PMuAq5 label._0b8hHE

And also the generated SQL:

Generated X-SQL

```
select
    dom_first_text(dom, 'div.-Esc+w.card.product-briefing div.HLQqkk div.flex-column.imEX5V span') as `T1C2`,
    dom_first_text(dom, 'div.HLQqkk div.flex-column.imEX5V div.W2tD8- div.MrYJVA.Ga-lTj') as `T1C3`,
    dom_first_text(dom, 'div.HLQqkk div.flex-column.imEX5V div.W2tD8- div.MrYJVA') as `T1C4`,
    dom_first_text(dom, 'div.HLQqkk div.flex-column.imEX5V div.W2tD8- div.Wz7RdC') as `T1C5`,
    dom_first_text(dom, 'div.HLQqkk div.flex-column.imEX5V div.W2tD8- div._45NQT5') as `T1C6`,
    dom_first_text(dom, 'div.HLQqkk div.flex-column.imEX5V div.W2tD8- div.Cv8D6q') as `T1C7`,
    dom_first_text(dom, 'div.-Esc+w.card.product-briefing div.HLQqkk div.imEX5V div.pmmxKx') as `T1C8`,
    dom_first_text(dom, 'div.-Esc+w.card.product-briefing div.HLQqkk div.imEX5V div.mini-vouchers__label') as `T1C9`,
    dom_first_text(dom, 'div.imEX5V div.PMuAq5 div.flex-no-overflow span.voucher-promo-value.voucher-promo-value--absolute-valu
    dom_first_text(dom, 'div.HLQqkk div.imEX5V div.PMuAq5 label._0b8hHE') as `T1C11`,
    dom_first_text(dom, 'div.PMuAq5 div.MGNOw3.hInOdW div.dHS5e4.xIMb1R div.LgUWja') as `T1C12`,
    dom_first_text(dom, 'div.PMuAq5 div.MGNOw3.hInOdW div.dHS5e4.xIMb1R div.Nd79Ux') as `T1C13`,
    dom_first_text(dom, 'div.MGNOw3.hInOdW div.dHS5e4.xIMb1R div.flex-row div.NPdOlf') as `T1C14`,
    dom_first_text(dom, 'div.imEX5V div.PMuAq5 div.-+gikn.hInOdW label._0b8hHE') as `T1C15`,
    dom_first_text(dom, 'div.PMuAq5 div.-+gikn.hInOdW div.items-center button.product-variation') as `T1C16`,
    dom_first_text(dom, 'div.PMuAq5 div.-+gikn.hInOdW div.items-center button.product-variation') as `T1C17`,
    dom_first_text(dom, 'div.imEX5V div.PMuAq5 div.-+gikn.hInOdW div._0b8hHE') as `T1C18`,
    dom_first_text(dom, 'div.PMuAq5 div.-+gikn.hInOdW div.G2C2rT.items-center div') as `T1C19`,
    dom_first_text(dom, 'div.flex-column.imEX5V div.vdf0Mi div.OozJX2 span') as `T1C20`,
    dom_first_text(dom, 'div.HLQqkk div.flex-column.imEX5V div.vdf0Mi button.btn.btn-solid-primary.btn--l.GfiOwy') as `T1C21`,
    dom_first_text(dom, 'div.-Esc+w.card.product-briefing div.HLQqkk div.flex-column.imEX5V span.zevbuo') as `T1C22`,
    dom_first_text(dom, 'div.-Esc+w.card.product-briefing div.HLQqkk div.flex-column.imEX5V span') as `T1C23`
from load_and_select('https://shopee.sg/(Local-Stock)-(GEBIZ-ACRA-REG)-PLA-3D-Printer-Filament-Standard-Colours-Series-1.75mm-1
```

Note that the website in this demo uses CSS obfuscation techniques, so the CSS selectors are hard to read and changes frequently. There is no other effective technology to solve this problem other than machine learning based solutions.

The complete code can be found here.

Scrape pages using the generated SQLs

The `harvest` command extracts fields automatically using unsupervised machine learning, and also generates the best css selectors for all possible fields and the extract SQLs. We can execute the SQLs using `sql` command.

```
# Note: remove the wildcard `*` and use the full name of the jar on Windows
java -jar exotic-standalone*.jar sql "
select
    dom_first_text(dom, 'div.-Esc+w.card.product-briefing div.HLQqkk div.flex-
column.imEX5V span') as T1C2,
    dom_first_text(dom, 'div.HLQqkk div.flex-column.imEX5V div.W2tD8- div.MrYJVA.Ga-
lTj') as T1C3,
    dom_first_text(dom, 'div.HLQqkk div.flex-column.imEX5V div.W2tD8- div.MrYJVA') as
T1C4,
    dom_first_text(dom, 'div.HLQqkk div.flex-column.imEX5V div.W2tD8- div.Wz7RdC') as
T1C5,
    dom_first_text(dom, 'div.HLQqkk div.flex-column.imEX5V div.W2tD8- div._45NQT5') as
T1C6,
    dom_first_text(dom, 'div.HLQqkk div.flex-column.imEX5V div.W2tD8- div.Cv8D6q') as
T1C7,
    dom_first_text(dom, 'div.-Esc+w.card.product-briefing div.HLQqkk div.imEX5V
div.pmmxKx') as T1C8,
    dom_first_text(dom, 'div.-Esc+w.card.product-briefing div.HLQqkk div.imEX5V
div.mini-vouchers__label') as T1C9,
    dom_first_text(dom, 'div.imEX5V div.PMuAq5 div.flex-no-overflow span.voucher-
promo-value.voucher-promo-value--absolute-value') as T1C10,
    dom_first_text(dom, 'div.HLQqkk div.imEX5V div.PMuAq5 label._0b8hHE') as T1C11,
    dom_first_text(dom, 'div.PMuAq5 div.MGNOw3.hInOdW div.dHS5e4.xIMb1R div.LgUWja')
```

```
as T1C12,
    dom_first_text(dom, 'div.PMuAq5 div.MGNOw3.hInOdW div.dHS5e4.xIMb1R div.Nd79Ux')
as T1C13,
    dom_first_text(dom, 'div.MGNOw3.hInOdW div.dHS5e4.xIMb1R div.flex-row div.NPdOlf')
as T1C14,
    dom_first_text(dom, 'div.imEX5V div.PMuAq5 div.-+gikn.hInOdW label._0b8hHE') as
T1C15,
    dom_first_text(dom, 'div.PMuAq5 div.-+gikn.hInOdW div.items-center button.product-
variation') as T1C16,
    dom_first_text(dom, 'div.PMuAq5 div.-+gikn.hInOdW div.items-center button.product-
variation') as T1C17,
    dom_first_text(dom, 'div.imEX5V div.PMuAq5 div.-+gikn.hInOdW div._0b8hHE') as
T1C18,
    dom_first_text(dom, 'div.PMuAq5 div.-+gikn.hInOdW div.G2C2rT.items-center div') as
T1C19,
    dom_first_text(dom, 'div.flex-column.imEX5V div.vdf0Mi div.OozJX2 span') as T1C20,
    dom_first_text(dom, 'div.HLQqkk div.flex-column.imEX5V div.vdf0Mi button.btn.btn-
solid-primary.btn--l.GfiOwy') as T1C21,
    dom_first_text(dom, 'div.-Esc+w.card.product-briefing div.HLQqkk div.flex-
column.imEX5V span.zevbuo') as T1C22,
    dom_first_text(dom, 'div.-Esc+w.card.product-briefing div.HLQqkk div.flex-
column.imEX5V span') as T1C23
from load_and_select('https://shopee.sg/(Local-Stock)-(GEBIZ-ACRA-REG)-PLA-3D-Printer-
Filament-Standard-Colours-Series-1.75mm-1kg-i.182524985.8326053759?sp_atk=3afa9679-
22cb-4c30-a1db-9d271e15b7a2&xptdk=3afa9679-22cb-4c30-a1db-9d271e15b7a2', 'div.page-
product');
"
```

Explore the Exotic executable jar

Run the executable jar directly for help to explore more power provided:

```
# Note: remove the wildcard `*` and use the full name of the jar on Windows
java -jar exotic-standalone*.jar
```

This command will print the help message and most useful examples.

Q & A

Q: How to use proxies?

A: Follow this guide for proxy rotation.