# Scraping the Web with Scrapy

# Hello, there!

- I'm **Matt Lebrun**
- PyClub Ring Leader at **Save22, Inc.**
- Trustee of **Python.PH**
- Been in love with **Python** since 2012
- ***#peanutbutterlife***

IF SOMEONE EVER TELLS YOU THAT YOU'RE PUTTING TOO MUCH PEANUT BUTTER ON YOUR BREAD,

STOP TALKING TO THEM. YOU DON'T NEED THAT KIND OF NEGATIVITY IN YOUR LIFE.

# So you want to get data from the web?

- Content extraction
- Page navigation

# Content extraction (Scraping)

# Content extraction (Scraping)

- request
- beautifulsoup
- lxml

# Page navigation (Crawling)

# Page navigation (Crawling)

- custom script

# What is Scrapy?

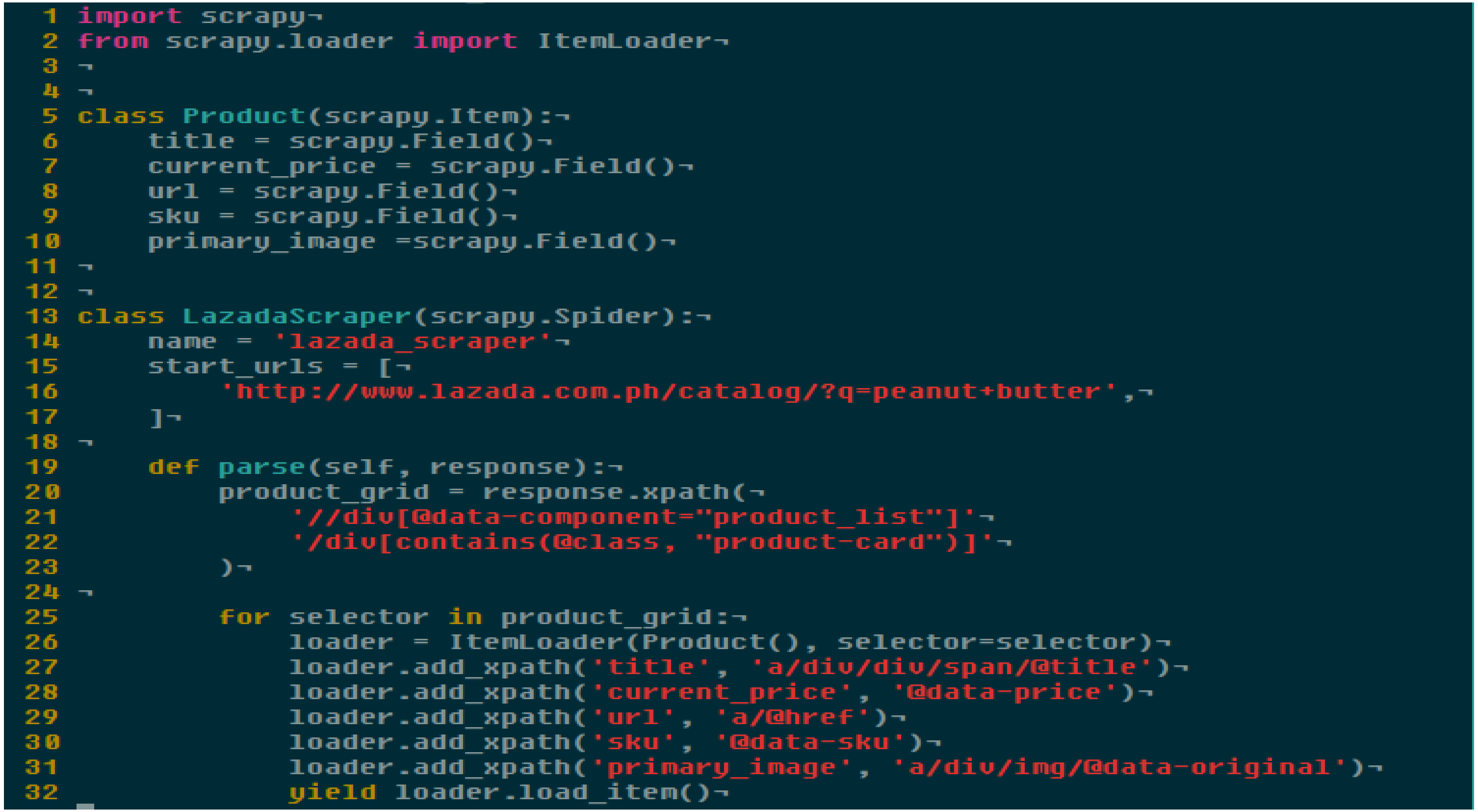
- Web scraping framework
- Handles scraping *and* crawling
- Has an interactive shell for testing **CSS** and **XPath** expressions
- Support for generating multiple feed formats (CSV, JSON, XML) and storing them in multiple backends (FTP, S3, local filesystem)
- Encoding support and auto detection
- Highly extensible via different signal hooks and middlewares
- Python3 support!
- (more)

# Install Scrapy

```
$ pip install scrapy
```

*Note: There might be caveats depending on you OS*

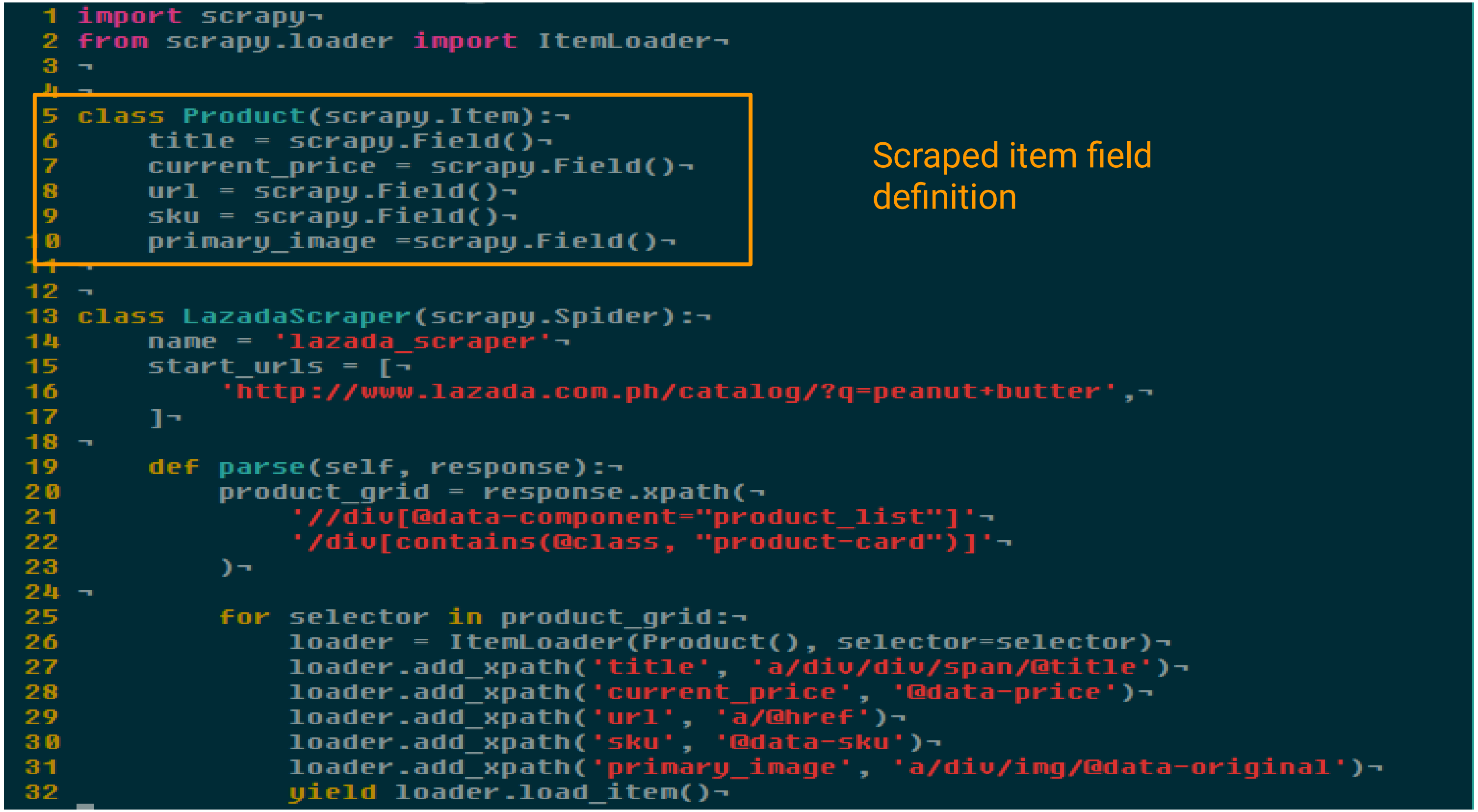
# Scrapy script



```python
import scrapy
from scrapy.loader import ItemLoader


class Product(scrapy.Item):
    title = scrapy.Field()
    current_price = scrapy.Field()
    url = scrapy.Field()
    sku = scrapy.Field()
    primary_image =scrapy.Field()


class LazadaScraper(scrapy.Spider):
    name = 'lazada_scraper'
    start_urls = [
        'http://www.lazada.com.ph/catalog/?q=peanut+butter',
    ]

    def parse(self, response):
        product_grid = response.xpath(
            '//div[@data-component="product_list"]'
            '/div[contains(@class, "product-card")]'
        )

        for selector in product_grid:
            loader = ItemLoader(Product(), selector=selector)
            loader.add_xpath('title', 'a/div/div/span/@title')
            loader.add_xpath('current_price', '@data-price')
            loader.add_xpath('url', 'a/@href')
            loader.add_xpath('sku', '@data-sku')
            loader.add_xpath('primary_image', 'a/div/img/@data-original')
            yield loader.load_item()
```

lazada_scraper.py

# Scrapy script

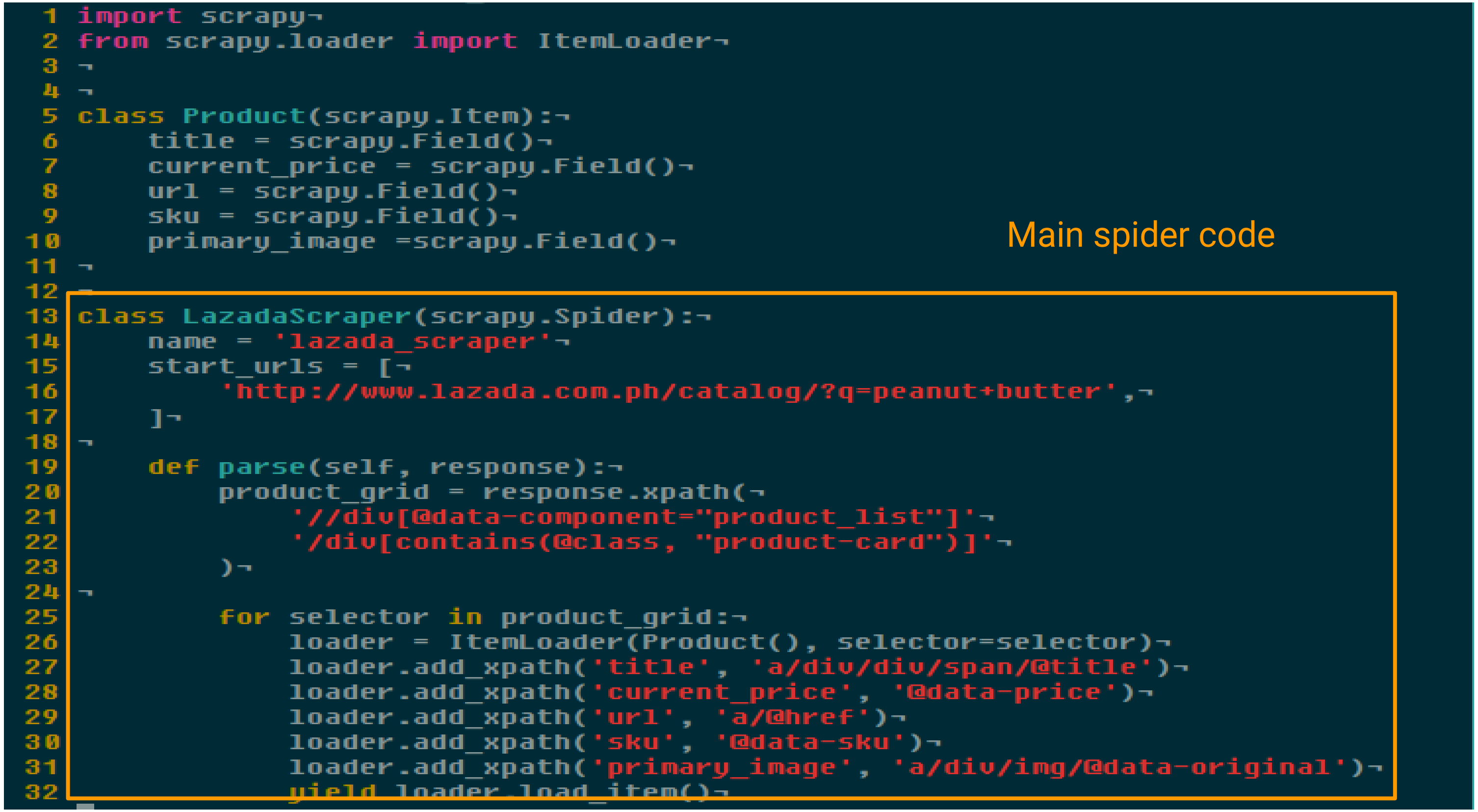


```
import scrapy¬
from scrapy.loader import ItemLoader¬
¬
¬
class Product(scrapy.Item):¬
    title = scrapy.Field()¬
    current_price = scrapy.Field()¬
    url = scrapy.Field()¬
    sku = scrapy.Field()¬
    primary_image =scrapy.Field()¬
¬
¬
class LazadaScraper(scrapy.Spider):¬
    name = 'lazada_scraper'¬
    start_urls = [¬
        'http://www.lazada.com.ph/catalog/?q=peanut+butter',¬
    ]¬
¬
    def parse(self, response):¬
        product_grid = response.xpath(¬
            '//div[@data-component="product_list"]'¬
            '/div[contains(@class, "product-card")]'¬
        )¬
¬
        for selector in product_grid:¬
            loader = ItemLoader(Product(), selector=selector)¬
            loader.add_xpath('title', 'a/div/div/span/@title')¬
            loader.add_xpath('current_price', '@data-price')¬
            loader.add_xpath('url', 'a/@href')¬
            loader.add_xpath('sku', '@data-sku')¬
            loader.add_xpath('primary_image', 'a/div/img/@data-original')¬
            yield loader.load_item()¬
```

Scraped item field definition

lazada_scraper.py

# Scrapy script

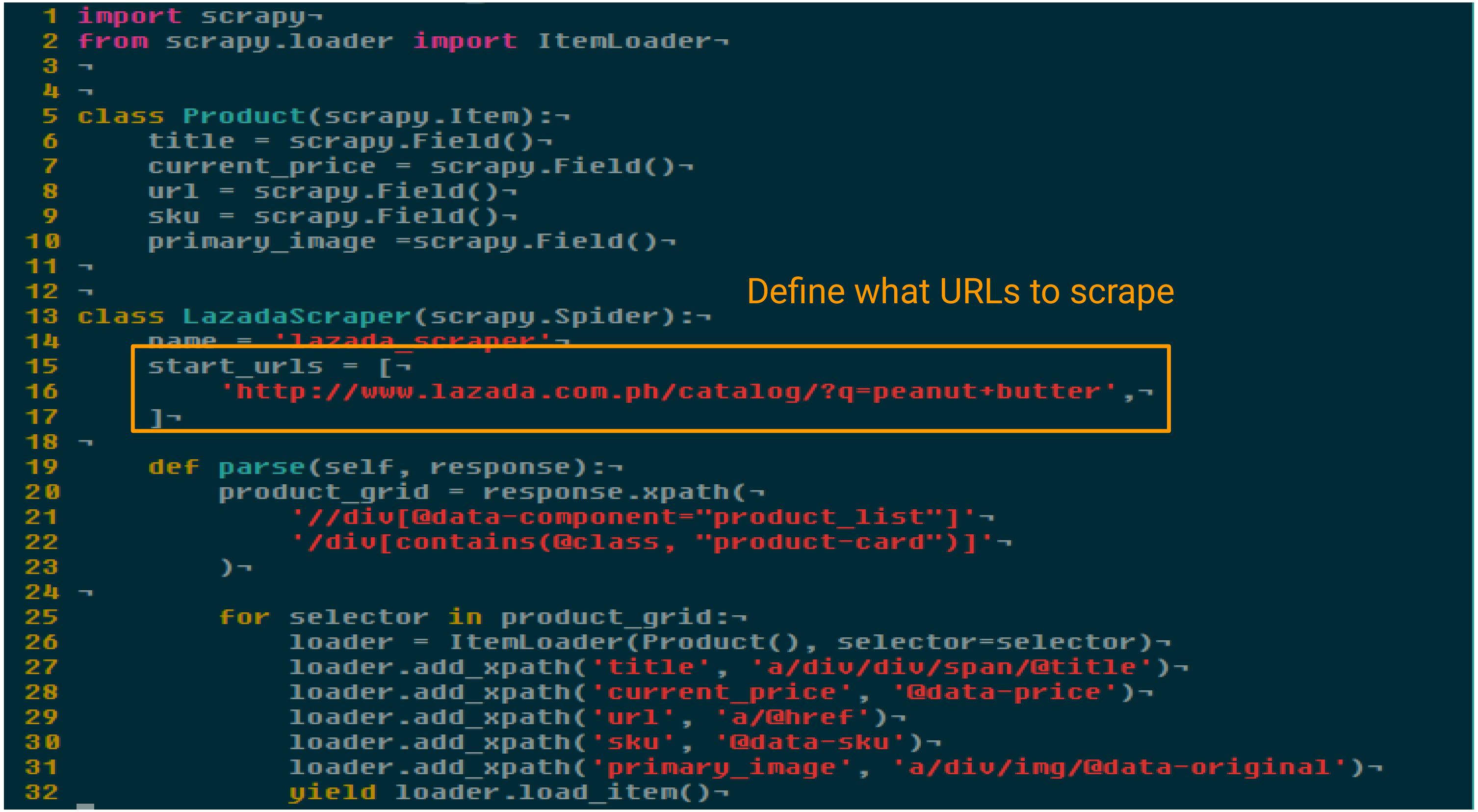


```
import scrapy¬
from scrapy.loader import ItemLoader¬
¬
¬
class Product(scrapy.Item):¬
    title = scrapy.Field()¬
    current_price = scrapy.Field()¬
    url = scrapy.Field()¬
    sku = scrapy.Field()¬
    primary_image =scrapy.Field()¬
¬
¬
class LazadaScraper(scrapy.Spider):¬
    name = 'lazada_scraper'¬
    start_urls = [¬
        'http://www.lazada.com.ph/catalog/?q=peanut+butter',¬
    ]¬
¬
    def parse(self, response):¬
        product_grid = response.xpath(¬
            '//div[@data-component="product_list"]'¬
            '/div[contains(@class, "product-card")]'¬
        )¬
¬
        for selector in product_grid:¬
            loader = ItemLoader(Product(), selector=selector)¬
            loader.add_xpath('title', 'a/div/div/span/@title')¬
            loader.add_xpath('current_price', '@data-price')¬
            loader.add_xpath('url', 'a/@href')¬
            loader.add_xpath('sku', '@data-sku')¬
            loader.add_xpath('primary_image', 'a/div/img/@data-original')¬
            yield loader.load_item()¬
```

Main spider code

lazada_scraper.py

# Scrapy script

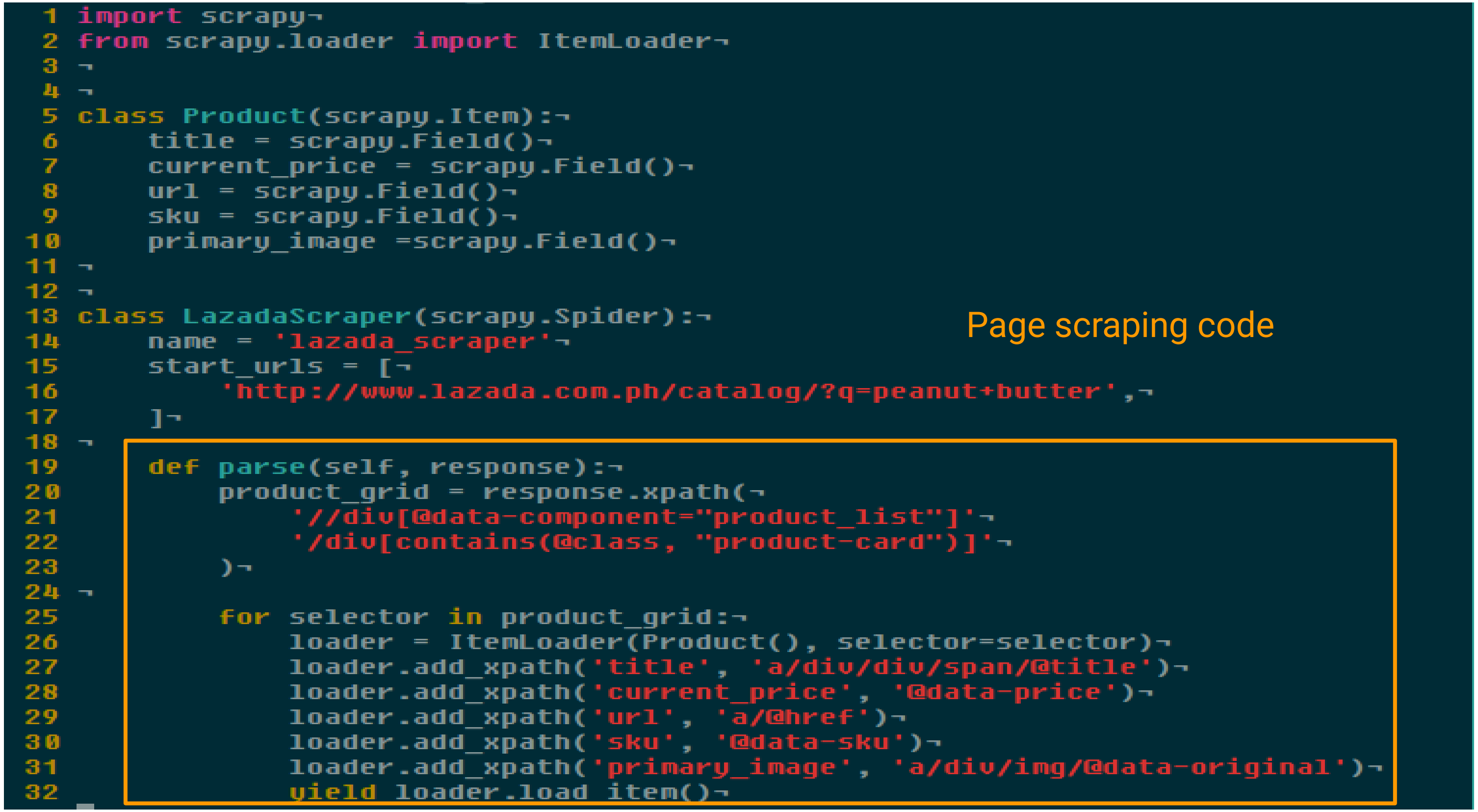


```
import scrapy¬
from scrapy.loader import ItemLoader¬
¬
¬
class Product(scrapy.Item):¬
    title = scrapy.Field()¬
    current_price = scrapy.Field()¬
    url = scrapy.Field()¬
    sku = scrapy.Field()¬
    primary_image =scrapy.Field()¬
¬
¬
class LazadaScraper(scrapy.Spider):¬
    name = 'lazada_scraper'¬
    start_urls = [¬
        'http://www.lazada.com.ph/catalog/?q=peanut+butter',¬
    ]¬
¬
    def parse(self, response):¬
        product_grid = response.xpath(¬
            '//div[@data-component="product_list"]'¬
            '/div[contains(@class, "product-card")]'¬
        )¬
¬
        for selector in product_grid:¬
            loader = ItemLoader(Product(), selector=selector)¬
            loader.add_xpath('title', 'a/div/div/span/@title')¬
            loader.add_xpath('current_price', '@data-price')¬
            loader.add_xpath('url', 'a/@href')¬
            loader.add_xpath('sku', '@data-sku')¬
            loader.add_xpath('primary_image', 'a/div/img/@data-original')¬
            yield loader.load_item()¬
```

Define what URLs to scrape

lazada_scraper.py

# Scrapy script

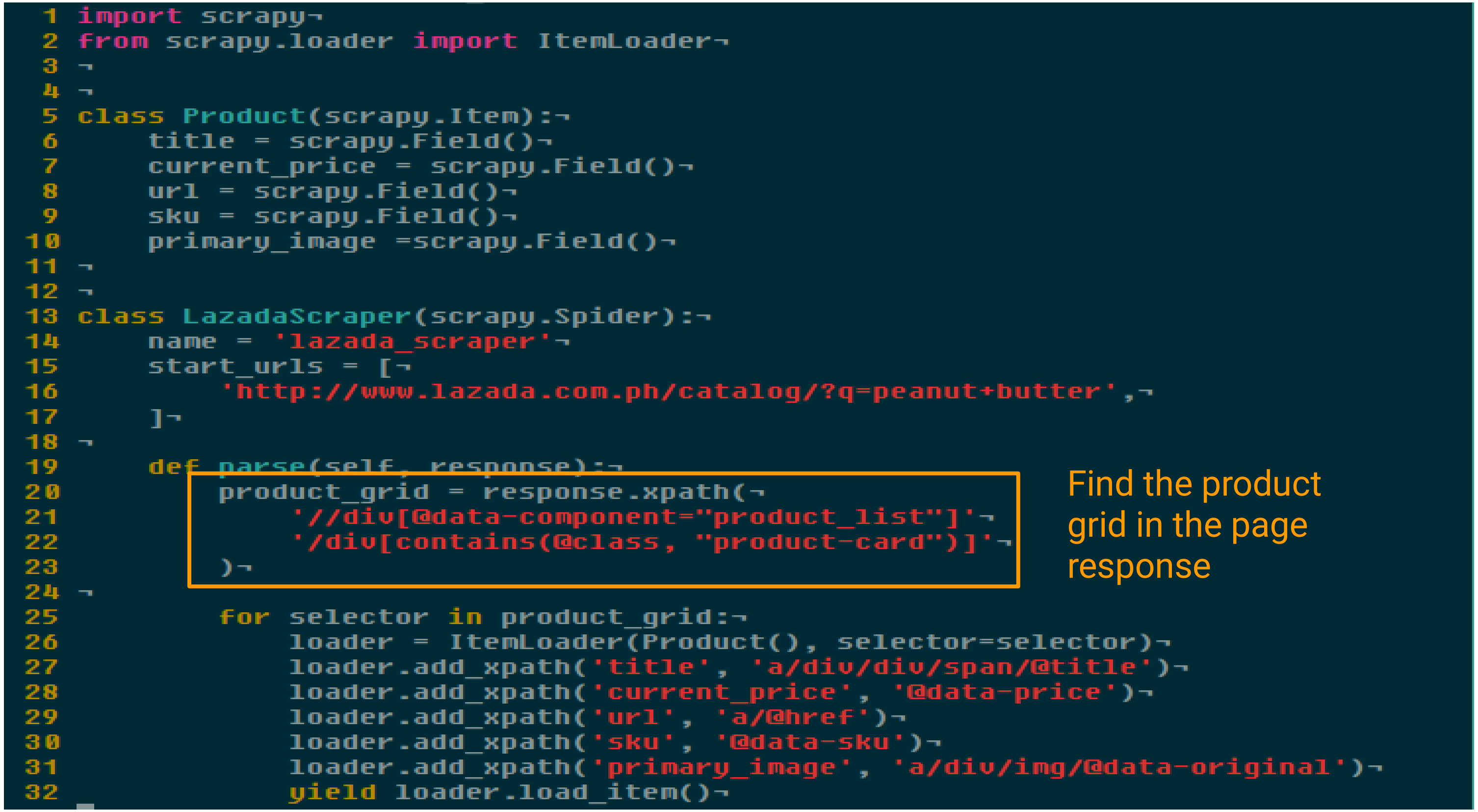


```python
import scrapy
from scrapy.loader import ItemLoader


class Product(scrapy.Item):
    title = scrapy.Field()
    current_price = scrapy.Field()
    url = scrapy.Field()
    sku = scrapy.Field()
    primary_image =scrapy.Field()


class LazadaScraper(scrapy.Spider):
    name = 'lazada_scraper'
    start_urls = [
        'http://www.lazada.com.ph/catalog/?q=peanut+butter',
    ]

    def parse(self, response):
        product_grid = response.xpath(
            '//div[@data-component="product_list"]'
            '/div[contains(@class, "product-card")]'
        )

        for selector in product_grid:
            loader = ItemLoader(Product(), selector=selector)
            loader.add_xpath('title', 'a/div/div/span/@title')
            loader.add_xpath('current_price', '@data-price')
            loader.add_xpath('url', 'a/@href')
            loader.add_xpath('sku', '@data-sku')
            loader.add_xpath('primary_image', 'a/div/img/@data-original')
            yield loader.load_item()
```

Page scraping code

lazada_scraper.py

# Scrapy script

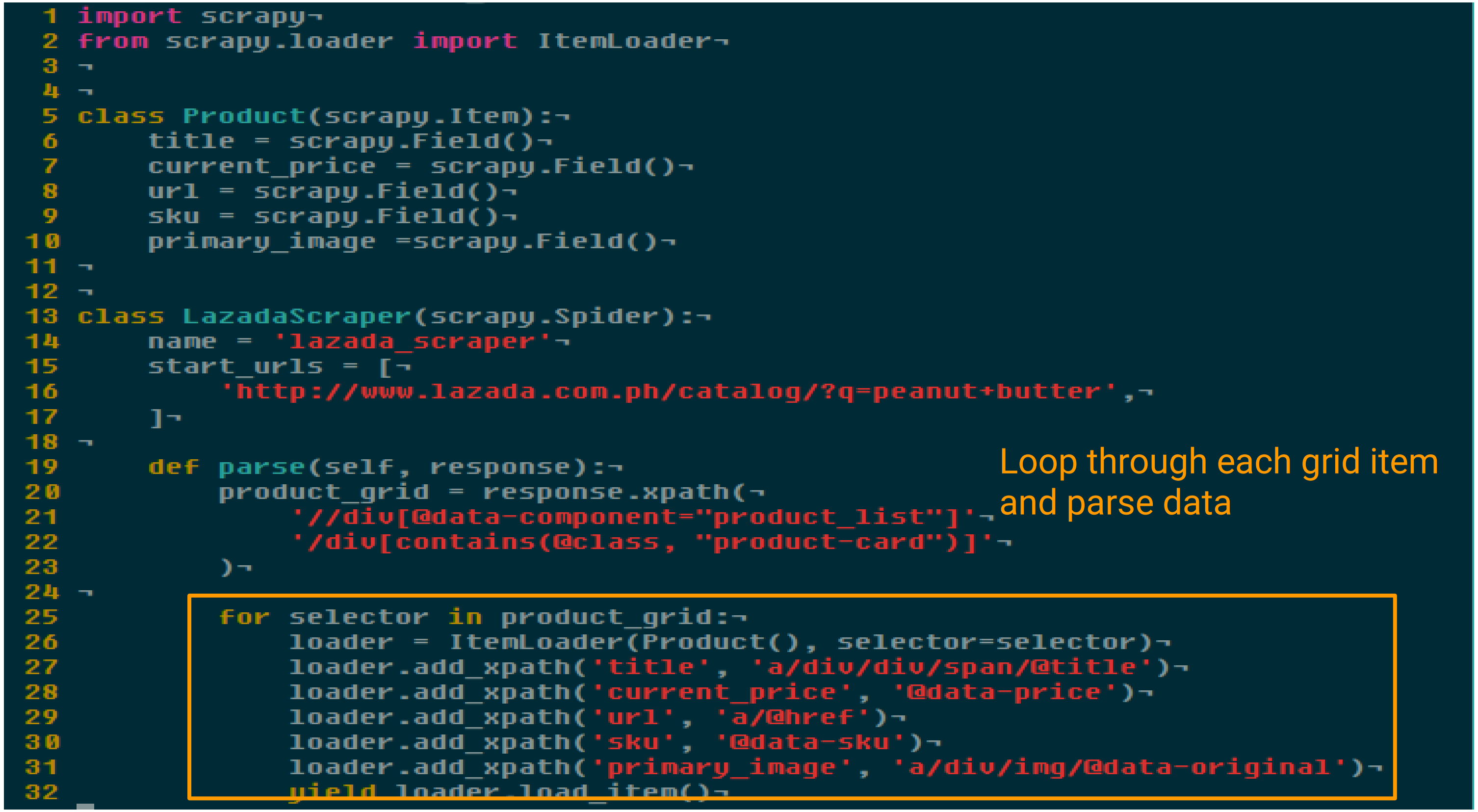


```
import scrapy¬
from scrapy.loader import ItemLoader¬
¬
¬
class Product(scrapy.Item):¬
    title = scrapy.Field()¬
    current_price = scrapy.Field()¬
    url = scrapy.Field()¬
    sku = scrapy.Field()¬
    primary_image =scrapy.Field()¬
¬
¬
class LazadaScraper(scrapy.Spider):¬
    name = 'lazada_scraper'¬
    start_urls = [¬
        'http://www.lazada.com.ph/catalog/?q=peanut+butter',¬
    ]¬
¬
    def parse(self, response):¬
        product_grid = response.xpath(¬
            '//div[@data-component="product_list"]'¬
            '/div[contains(@class, "product-card")]'¬
        )¬
¬
        for selector in product_grid:¬
            loader = ItemLoader(Product(), selector=selector)¬
            loader.add_xpath('title', 'a/div/div/span/@title')¬
            loader.add_xpath('current_price', '@data-price')¬
            loader.add_xpath('url', 'a/@href')¬
            loader.add_xpath('sku', '@data-sku')¬
            loader.add_xpath('primary_image', 'a/div/img/@data-original')¬
            yield loader.load_item()¬
```

Find the product grid in the page response

lazada_scraper.py

# Scrapy script



```
import scrapy¬
from scrapy.loader import ItemLoader¬
¬
¬
class Product(scrapy.Item):¬
    title = scrapy.Field()¬
    current_price = scrapy.Field()¬
    url = scrapy.Field()¬
    sku = scrapy.Field()¬
    primary_image =scrapy.Field()¬
¬
¬
class LazadaScraper(scrapy.Spider):¬
    name = 'lazada_scraper'¬
    start_urls = [¬
        'http://www.lazada.com.ph/catalog/?q=peanut+butter',¬
    ]¬
¬
    def parse(self, response):¬
        product_grid = response.xpath(¬
            '//div[@data-component="product_list"]'¬
            '/div[contains(@class, "product-card")]'¬
        )¬
¬
        for selector in product_grid:¬
            loader = ItemLoader(Product(), selector=selector)¬
            loader.add_xpath('title', 'a/div/div/span/@title')¬
            loader.add_xpath('current_price', '@data-price')¬
            loader.add_xpath('url', 'a/@href')¬
            loader.add_xpath('sku', '@data-sku')¬
            loader.add_xpath('primary_image', 'a/div/img/@data-original')¬
            yield loader.load_item()¬
```

Loop through each grid item and parse data

lazada_scraper.py

# Run the Scrapy spider

```
$ scrapy runspider lazada_scraper.py -t csv -o
lazada_scraper.csv
```

# Item data extraction tips

- Use scrapy shell
- Test and study how to extract each field (CSS, XPath, Beautifulsoup, RegEx, etc)
- Use the python debugger (pdb)

# Scrapy ecosystem

- Scrapyd
- Splash
- ScrapingHub

# Further reading

- **How to Install Scrapy in Windows**
  `https://scraper24x7.wordpress.com/2016/03/19/how-to-install-scrapy-in-windows/`
- **Official Scrapy Documentation**
  `https://doc.scrapy.org/en/1.2/`

# Q & A

# You guys are awesome!

Let's keep in touch ;)

email: **matt@lebrun.org**
social: **cr8ivecodesmith**

Looking for internship?
**`http://goo.gl/forms/mOQ39cm3Gy`**