Scrapy
beyond the first steps

Eugenio Lacuesta

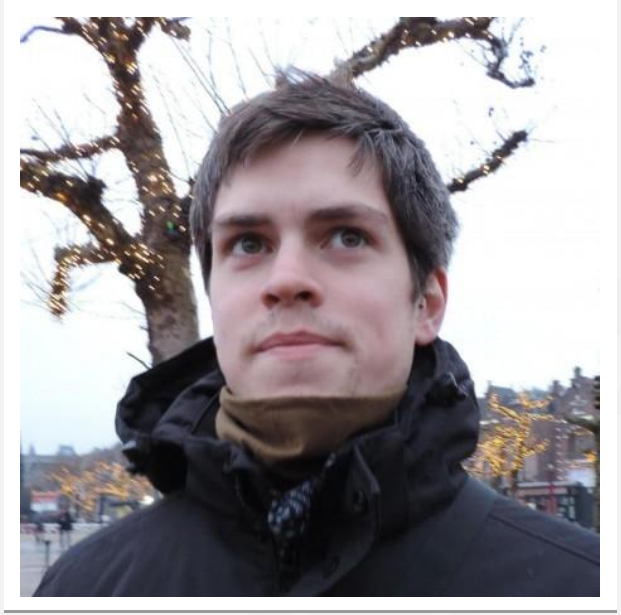
Python Brasil, october 2018

scrapinghub

Hello everyone!

About me:

- Python developer from Montevideo, Uruguay
- Scrapinghubber since 2015
- Scrapy contributor



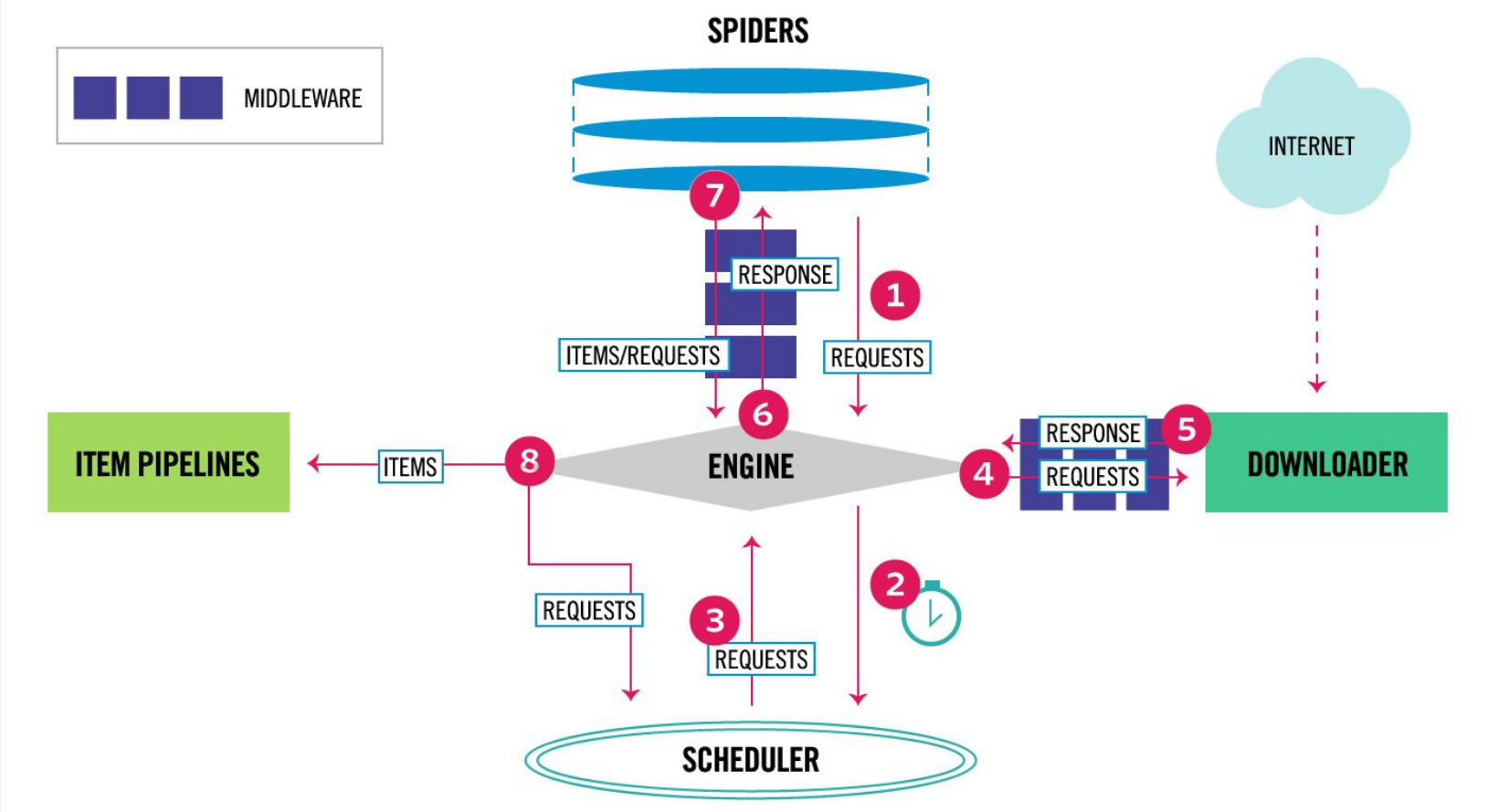
About this talk:

- Only basic Scrapy knowledge needed (spiders, requests)
- Scrapy project at https://github.com/elacuesta/scrapy-beyond-first-steps

Talk summary

- Quick architecture overview
- Basic spider
- Item exporters
- Spider middleware
- Item pipeline
- Signals
- from_crawler factory method
- Extensions
- Downloader middleware

Architecture



Basic spider - http://books.toscrape.com

```
from scrapy import Spider

class BooksSpider(Spider):
    name = 'books'
    start_urls = ['http://books.toscrape.com']

    def parse(self, response):
        for book_link in response.css('article.product_pod h3 a::attr(href)').getall():
            yield response.follow(book_link, callback=self.parse_book)

    def parse_book(self, response):
        return {
            'url': response.url,
            'title': response.css('h1::text').get(),
            'price': float(response.css('p.price_color::text').re_first(r'(\d+.?\d*)')),
        }
```

Basic spider - http://books.toscrape.com

```
from scrapy import Spider

class BooksSpider(Spider):
    name = 'books'
    start_urls = ['http://books.toscrape.com']

    def parse(self, response):
        for book_link in response.css('article.product_pod h3 a::attr(href)').getall():
            yield response.follow(book_link, callback=self.parse_book)

    def parse_book(self, response):
        return {
            'url': response.url,
            'title': response.css('h1::text').get(),
            'price': float(response.css('p.price_color::text').re_first(r'(\d+.?\d*)')),
        }
```

First URL to crawl

Basic spider - http://books.toscrape.com

```
from scrapy import Spider

class BooksSpider(Spider):
    name = 'books'
    start_urls = ['http://books.toscrape.com']

    def parse(self, response):
        for book_link in response.css('article.product_pod h3 a::attr(href)').getall():
            yield response.follow(book_link, callback=self.parse_book)

    def parse_book(self, response):
        return {
            'url': response.url,
            'title': response.css('h1::text').get(),
            'price': float(response.css('p.price_color::text').re_first(r'(\d+.?\d*)')),
        }
```

Iterate over the links and produce requests to the specific books

Basic spider - http://books.toscrape.com

```
from scrapy import Spider

class BooksSpider(Spider):
    name = 'books'
    start_urls = ['http://books.toscrape.com']

    def parse(self, response):
        for book_link in response.css('article.product_pod h3 a::attr(href)').getall():
            yield response.follow(book_link, callback=self.parse_book)

    def parse_book(self, response):
        return {
            'url': response.url,
            'title': response.css('h1::text').get(),
            'price': float(response.css('p.price_color::text').re_first(r'(\d+.?\d*)')),
        }
```

Extract information from each book page

```
$ scrapy crawl books -o books.json
2018-10-08 14:36:31 [scrapy.utils.log] INFO: Scrapy 1.5.0 started (bot: pybr2018)
(...)
2018-10-08 14:36:34 [scrapy.core.engine] INFO: Spider closed (finished)

$ cat books.json | jq .
[
  {
      "url": "http://books.toscrape.com/catalogue/a-light-in-the-attic_1000/index.html",
      "title": "A Light in the Attic",
      "price": 51.77,
  },
  {
      "url":
"http://books.toscrape.com/catalogue/scott-pilgrims-precious-little-life-scott-pilgrim-1_987/index.html",
      "title": "Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)",
      "price": 52.29,
  },
...

$
```

Item exporters

scrapinghub

Item exporters

In the previous example, the `-o books.json` part indicated Scrapy that we wanted the output in JSON format.

Scrapy comes with a few built-in exporters (JSON, XML, CSV)

Custom item exporters are supported though the `FEED_EXPORTERS` setting:

```
FEED_EXPORTERS = {
    'yaml': 'pybr2018.exporters.YAMLItemExporter',
}
```

Project's settings.py file

YAML Item exporter

```python
from ruamel.yaml import YAML
from scrapy.exporters import BaseItemExporter


class YAMLItemExporter(BaseItemExporter):

    def __init__(self, file, *args, **kwargs):
        super().__init__(*args, **kwargs)
        self.file = file
        self.yaml = YAML()
        self.yaml.encoding = self.encoding

    def export_item(self, item):
        self.yaml.dump([dict(item)], self.file)
```

YAML Item exporter

```
from ruamel.yaml import YAML
from scrapy.exporters import BaseItemExporter


class YAMLItemExporter(BaseItemExporter):

    def __init__(self, file, *args, **kwargs):
        super().__init__(*args, **kwargs)
        self.file = file
        self.yaml = YAML()
        self.yaml.encoding = self.encoding

    def export_item(self, item):
        self.yaml.dump([dict(item)], self.file)
```

Inherit from Scrapy's Base Item Exporter

YAML Item exporter

```python
from ruamel.yaml import YAML
from scrapy.exporters import BaseItemExporter


class YAMLItemExporter(BaseItemExporter):

    def __init__(self, file, *args, **kwargs):
        super().__init__(*args, **kwargs)
        self.file = file
        self.yaml = YAML()
        self.yaml.encoding = self.encoding

    def export_item(self, item):
        self.yaml.dump([dict(item)], self.file)
```

Initialize the base exporter.
Store the output file pointer and
create the YAML exporter

YAML Item exporter

```
from ruamel.yaml import YAML
from scrapy.exporters import BaseItemExporter


class YAMLItemExporter(BaseItemExporter):

    def __init__(self, file, *args, **kwargs):
        super().__init__(*args, **kwargs)
        self.file = file
        self.yaml = YAML()
        self.yaml.encoding = self.encoding

    def export_item(self, item):
        self.yaml.dump([dict(item)], self.file)
```

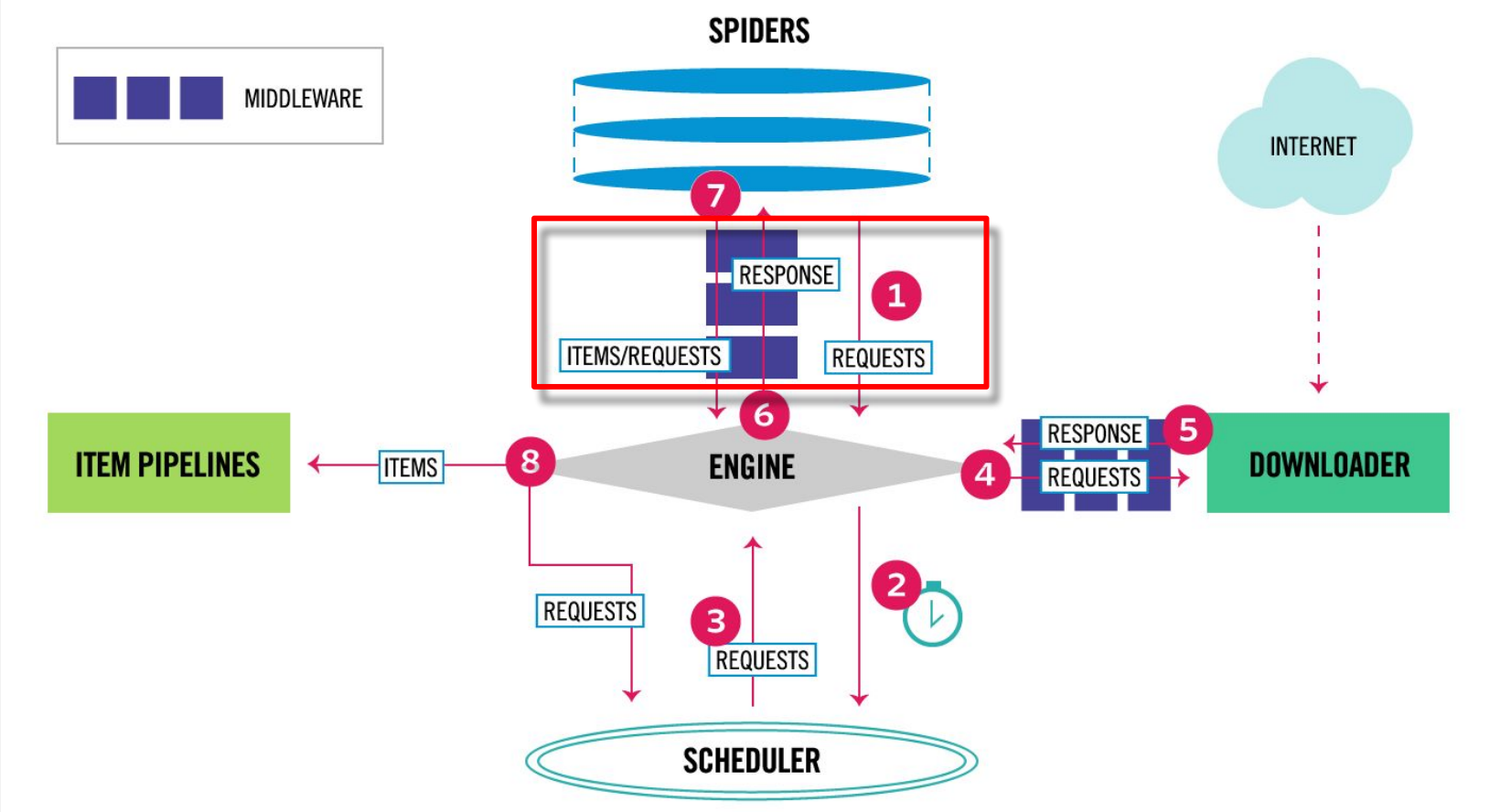
Serialize each item

```
$ scrapy crawl books -o books.yaml
2018-10-08 15:43:01 [scrapy.utils.log] INFO: Scrapy 1.5.0 started (bot: pybr2018)
(...several log lines...)
2018-10-08 15:43:04 [scrapy.core.engine] INFO: Spider closed (finished)

$ head -n 12 books.yaml
- url: http://books.toscrape.com/catalogue/a-light-in-the-attic_1000/index.html
  title: A Light in the Attic
  price: 51.77
- url: http://books.toscrape.com/catalogue/rip-it-up-and-start-again_986/index.html
  title: Rip it Up and Start Again
  price: 35.02
- url: http://books.toscrape.com/catalogue/set-me-free_988/index.html
  title: Set Me Free
  price: 17.46

$
```

Spider middleware

scrapinghub

Spider middleware



Spider middleware

The spider middleware is a framework of hooks into Scrapy's spider processing mechanism where you can plug custom functionality to process the responses that are sent to spiders and to process the requests and items that are generated from spiders.

From the docs

More at https://doc.scrapy.org/en/latest/topics/spider-middleware.html

Spider middleware - Example

In this example we will check each generated Book request and replace it by an already generated item if we have it cached, to avoid making unnecessary requests.

We enable the middleware in the spider class:

```
class BooksSpider(Spider):
    custom_settings = {
        SPIDER_MIDDLEWARE: {'pybr2018.middlewares.books.BookCacheSpiderMiddleware: 543}
    }
```

Spider middleware - Example

```
import os
import json

import scrapy


class BookCacheSpiderMiddleware:
    def __init__(self, *args, **kwargs):
        with open('{}/../data/books.cache'.format(os.path.dirname(__file__)), 'r') as f:
            self.books = {item['url']: item for item in json.load(f)}

    def process_spider_output(self, response, result, spider):
        for elem in result:
            if isinstance(elem, scrapy.Request) and elem.url in self.books:
                yield self.books[elem.url]
            else:
                yield elem
```

Spider middleware - Example

```
import os
import json

import scrapy


class BookCacheSpiderMiddleware:
    def __init__(self, *args, **kwargs):
        with open('{}/../data/books.cache'.format(os.path.dirname(__file__)), 'r') as f:
            self.books = {item['url']: item for item in json.load(f)}

    def process_spider_output(self, response, result, spider):
        for elem in result:
            if isinstance(elem, scrapy.Request) and elem.url in self.books:
                yield self.books[elem.url]
            else:
                yield elem
```

Get the Book cache (from a disk file in this example)

Spider middleware - Example

```
import os
import json

import scrapy


class BookCacheSpiderMiddleware:
    def __init__(self, *args, **kwargs):
        with open('{}/../data/books.cache'.format(os.path.dirname(__file__)), 'r') as f:
            self.books = {item['url']: item for item in json.load(f)}

    def process_spider_output(self, response, result, spider):
        for elem in result:
            if isinstance(elem, scrapy.Request) and elem.url in self.books:
                yield self.books[elem.url]
            else:
                yield elem
```

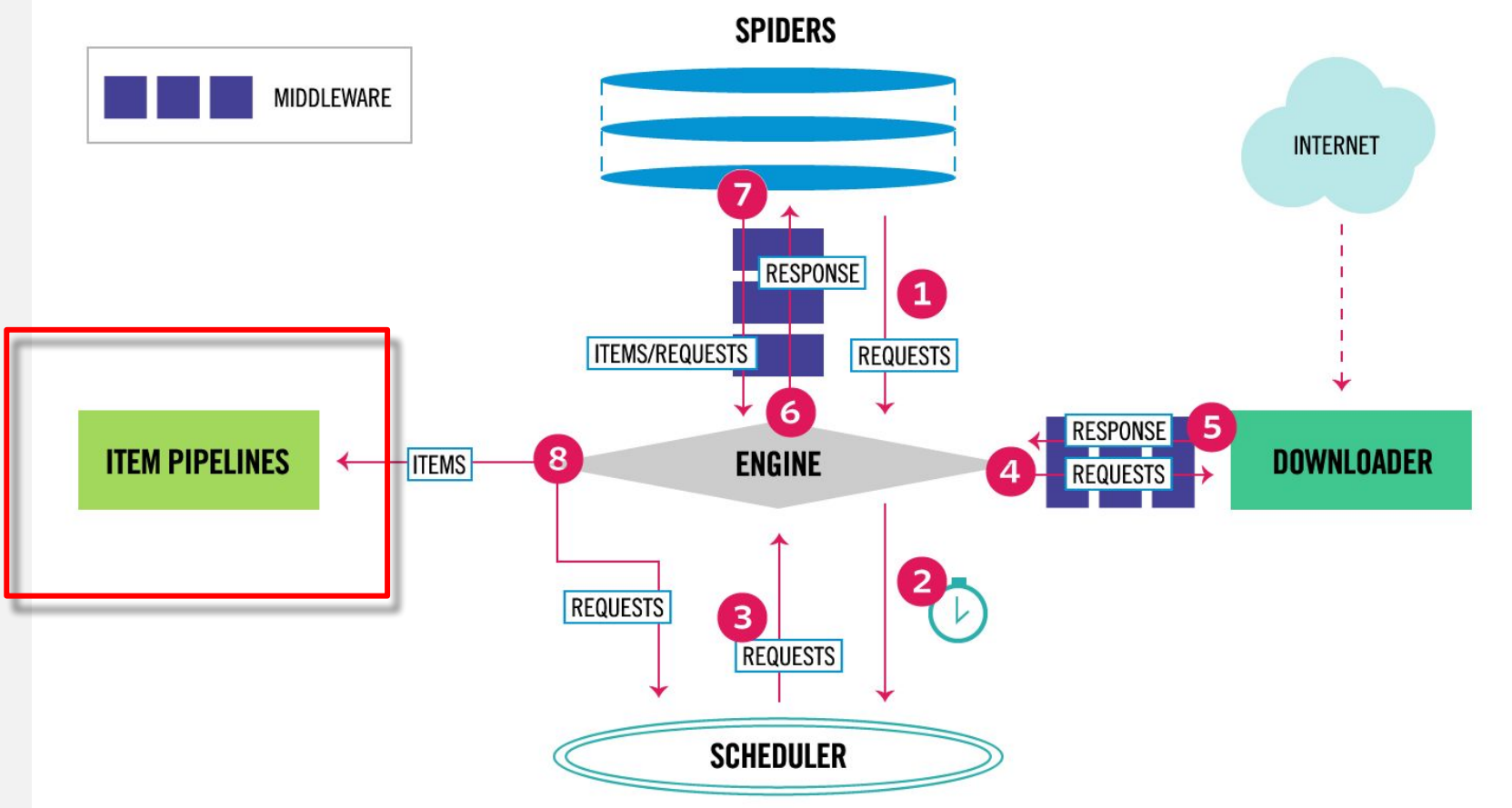
Process each item/request. Yield items unaltered, replace a request with an item if we have the URL in the cache

```
$ scrapy crawl books
2018-10-08 17:13:29 [scrapy.utils.log] INFO: Scrapy 1.5.0 started (bot: pybr2018)
(...)
2018-10-08 17:13:30 [BookCacheSpiderMiddleware] INFO:
[http://books.toscrape.com/catalogue/a-light-in-the-attic_1000/index.html] Book found in cache,
not making request
2018-10-08 17:13:30 [BookCacheSpiderMiddleware] INFO:
[http://books.toscrape.com/catalogue/sharp-objects_997/index.html] Book found in cache, not making
request
2018-10-08 17:13:30 [BookCacheSpiderMiddleware] INFO:
[http://books.toscrape.com/catalogue/sapiens-a-brief-history-of-humankind_996/index.html] Book
found in cache, not making request
(...)
2018-10-08 17:13:31 [scrapy.core.engine] INFO: Spider closed (finished)

$
```

Item pipeline

scrapinghub

Item pipeline



Item pipeline

After an item is scraped, it is sent through the Item Pipeline.

Each Pipeline stage is just a Python class that implements a `process_item` method, which receives the item and the spider object that produced it

This method can perform any action on the item, and it can also decide if the item should be sent to the next stage (by `return`ing it) or discarded (by raising the `DropItem` exception)

Item pipeline - Validate books

```
class BooksSpider(Spider):
    custom_settings = {
        'ITEM_PIPELINES': {'pybr2018.pipelines.ValidateBookPipeline': 100}
    }
```

```
from scrapy.exceptions import DropItem
import jsonschema

class ValidateBookPipeline:
    schema = {...}  # a valid JSON schema

    def process_item(self, item, spider):
        try:
            jsonschema.validate(dict(item), self.schema)
        except jsonschema.ValidationError as ex:
            raise DropItem(ex.message)
        else:
            return item
```

Item pipeline - Validate books

```
class BooksSpider(Spider):
      custom_settings = {
            'ITEM_PIPELINES': {'pybr2018.pipelines.ValidateBookPipeline': 100}
      }
```

Enable the Pipeline in the spider class

```
from scrapy.exceptions import DropItem
import jsonschema

class ValidateBookPipeline:
      schema = {...}  # a valid JSON schema

      def process_item(self, item, spider):
            try:
                  jsonschema.validate(dict(item), self.schema)
            except jsonschema.ValidationError as ex:
                  raise DropItem(ex.message)
            else:
                  return item
```

Item pipeline - Validate books

```
class BooksSpider(Spider):
    custom_settings = {
        'ITEM_PIPELINES': {'pybr2018.pipelines.ValidateBookPipeline': 100}
    }
```

```
from scrapy.exceptions import DropItem
import jsonschema

class ValidateBookPipeline:
    schema = {...}  # a valid JSON schema

    def process_item(self, item, spider):
        try:
            jsonschema.validate(dict(item), self.schema)
        except jsonschema.ValidationError as ex:
            raise DropItem(ex.message)
        else:
            return item
```

An actual JSON schema (missing here because it's too long)

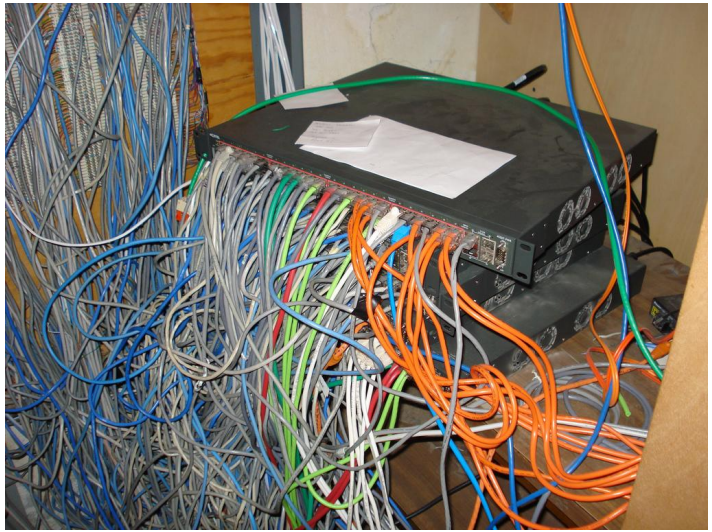
If the validation fails raise a DropItem exception to indicate Scrapy the item should be discarded

```
$ scrapy crawl books
2018-10-08 17:18:06 [scrapy.utils.log] INFO: Scrapy 1.5.0 started (bot: pybr2018)
(...)
2018-10-08 17:18:08 [scrapy.core.scraper] WARNING: Dropped: 52.15 is greater than the maximum of
50
{'url': 'http://books.toscrape.com/catalogue/the-black-maria_991/index.html', 'title': 'The Black
Maria', 'price': 52.15}
2018-10-08 17:18:08 [scrapy.core.scraper] WARNING: Dropped: 54.23 is greater than the maximum of
50
{'url': 'http://books.toscrape.com/catalogue/sapiens-a-brief-history-of-humankind_996/index.html',
'title': 'Sapiens: A Brief History of Humankind', 'price': 54.23}
2018-10-08 17:18:08 [scrapy.core.scraper] WARNING: Dropped: 'The Coming Woman: A Novel Based on
the Life of the Infamous Feminist, Victoria Woodhull' is too long
{'url':
'http://books.toscrape.com/catalogue/the-coming-woman-a-novel-based-on-the-life-of-the-infamous-fe
minist-victoria-woodhull_993/index.html', 'title': 'The Coming Woman: A Novel Based on the Life of
the Infamous Feminist, Victoria Woodhull', 'price': 17.93}
(...)
2018-10-08 17:18:09 [scrapy.core.engine] INFO: Spider closed (finished)

$
```

Signals

scrapinghub

Problem - Stateful server



https://www.flickr.com/photos/dcmorton/2446443463/

Now imagine the site stores information about the followed links, and returns responses based on the user's navigation patterns.
Scheduling parallel requests is not an option anymore.

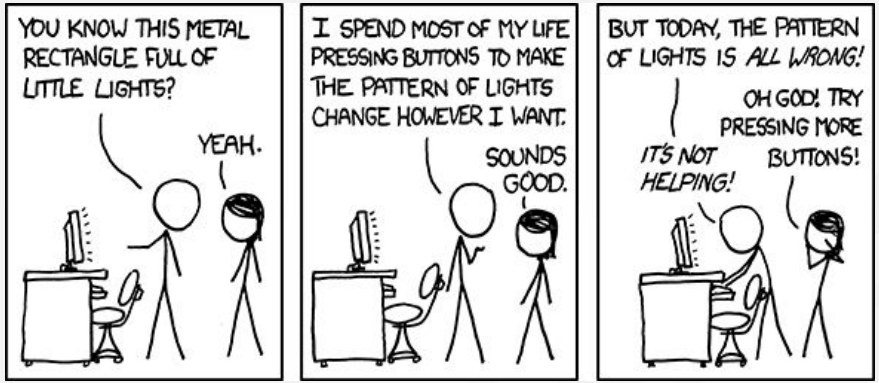
Possible solutions:

- Sequential crawl
- Parallel sessions

(One) solution - Sequential crawling

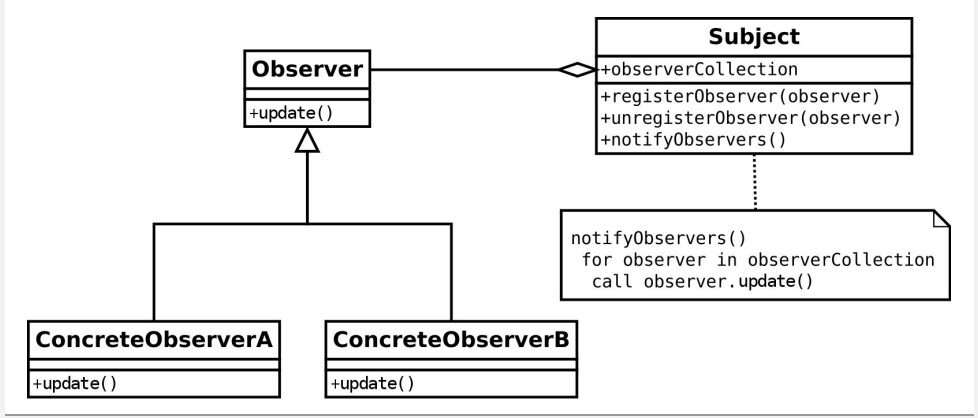
We'll take advantage of two design patterns employed on Scrapy:

- Observer pattern (signals)
- Factory pattern (from_crawler class method)



https://www.xkcd.com/722/

Signals



https://commons.wikimedia.org/wiki/File:Observer_w_update.svg

Observer design pattern: attach handlers to act on system events.

Several available events to handle:

- Engine is started
- Response is downloaded
- Item is scraped
- Spider is idle
- etc

`spider_idle` signal

From the Documentation:

Fired when the spider has no further requests scheduled/waiting to be downloaded or items being processed by the Item Pipeline

Provides *one* approach to sequential crawling

`from_crawler` class method

scrapinghub

`from_crawler` class method

The main entry point to Scrapy API is the *Crawler* object, passed (...) through the `from_crawler` class method. This object provides access to all Scrapy core components, and it's the only way (...) to access them and hook their functionality into Scrapy.

```
class MySpider(scrapy.Spider):
    @classmethod
    def from_crawler(cls, crawler, *args, **kwargs):
        spider = super().from_crawler(crawler, *args, **kwargs)
        crawler.signals.connect(spider.schedule_request, signal=scrapy.signals.spider_idle)
        return spider
```

`from_crawler` class method

The main entry point to Scrapy API is the *Crawler* object, passed (...) through the `from_crawler` class method. This object provides access to all Scrapy core components, and it's the only way (...) to access them and hook their functionality into Scrapy.

```
class MySpider(scrapy.Spider):
    @classmethod
    def from_crawler(cls, crawler, *args, **kwargs):
        spider = super().from_crawler(crawler, *args, **kwargs)
        crawler.signals.connect(spider.schedule_request, signal=scrapy.signals.spider_idle)
        return spider
```

Factory design pattern

Connect a handler to the signal

Putting it all together

```
class SequentialBooksSpider(BooksSpider):
	name = 'books-sequential'
	pending = collections.deque()

	@classmethod
	def from_crawler(cls, crawler, *args, **kwargs):
		spider = super().from_crawler(crawler, *args, **kwargs)
		crawler.signals.connect(spider.schedule_request, signal=scrapy.signals.spider_idle)
		return spider

	def schedule_request(self):
		if self.pending:
			request = self.pending.popleft()
			self.crawler.engine.crawl(request, self)

	def parse(self, response):
		for book_link in response.css('article.product_pod h3 a::attr(href)').getall():
			self.pending.append(response.follow(book_link, callback=self.parse_book))
			self.pending.append(response.request.replace(dont_filter=True, callback=self.parse_dummy))

	def parse_dummy(self, response):
		self.logger.info('Back at the main page')
```

Putting it all together

```
class SequentialBooksSpider(BooksSpider):
    name = 'books-sequential'
    pending = collections.deque()

    @classmethod
    def from_crawler(cls, crawler, *args, **kwargs):
        spider = super().from_crawler(crawler, *args, **kwargs)
        crawler.signals.connect(spider.schedule_request, signal=scrapy.signals.spider_idle)
        return spider

    def schedule_request(self):
        if self.pending:
            request = self.pending.popleft()
            self.crawler.engine.crawl(request, self)

    def parse(self, response):
        for book_link in response.css('article.product_pod h3 a::attr(href)').getall():
            self.pending.append(response.follow(book_link, callback=self.parse_book))
            self.pending.append(response.request.replace(dont_filter=True, callback=self.parse_dummy))

    def parse_dummy(self, response):
        self.logger.info('Back at the main page')
```

Inherit from our previous spider to reuse the parse_book method

Container to store pending requests

Putting it all together

```
class SequentialBooksSpider(BooksSpider):
    name = 'books-sequential'
    pending = collections.deque()

    @classmethod
    def from_crawler(cls, crawler, *args, **kwargs):
        spider = super().from_crawler(crawler, *args, **kwargs)
        crawler.signals.connect(spider.schedule_request, signal=scrapy.signals.spider_idle)
        return spider

    def schedule_request(self):
        if self.pending:
            request = self.pending.popleft()
            self.crawler.engine.crawl(request, self)

    def parse(self, response):
        for book_link in response.css('article.product_pod h3 a::attr(href)').getall():
            self.pending.append(response.follow(book_link, callback=self.parse_book))
            self.pending.append(response.request.replace(dont_filter=True, callback=self.parse_dummy))

    def parse_dummy(self, response):
        self.logger.info('Back at the main page')
```

Factory method: attach signal handler

Signal handler: crawl next request when the spider is idle

Putting it all together

```
class SequentialBooksSpider(BooksSpider):
    name = 'books-sequential'
    pending = collections.deque()

    @classmethod
    def from_crawler(cls, crawler, *args, **kwargs):
        spider = super().from_crawler(crawler, *args, **kwargs)
        crawler.signals.connect(spider.schedule_request, signal=scrapy.signals.spider_idle)
        return spider

    def schedule_request(self):
        if self.pending:
            request = self.pending.popleft()
            self.crawler.engine.crawl(request, self)

    def parse(self, response):
        for book_link in response.css('article.product_pod h3 a::attr(href)').getall():
            self.pending.append(response.follow(book_link, callback=self.parse_book))
            self.pending.append(response.request.replace(dont_filter=True, callback=self.parse_dummy))

    def parse_dummy(self, response):
        self.logger.info('Back at the main page')
```

Store book requests instead of scheduling directly

Simulate going back to the main page

```
$ scrapy crawl books-sequential
2018-10-09 12:54:19 [scrapy.utils.log] INFO: Scrapy 1.5.0 started (bot: pybr2018)
(...)
2018-10-09 12:54:20 [books-sequential] INFO: Scheduling:
http://books.toscrape.com/catalogue/a-light-in-the-attic_1000/index.html
2018-10-09 12:54:21 [books-sequential] INFO: Extracting:
http://books.toscrape.com/catalogue/a-light-in-the-attic_1000/index.html
2018-10-09 12:54:21 [books-sequential] INFO: Scheduling: http://books.toscrape.com
2018-10-09 12:54:21 [books-sequential] INFO: Back at the main page
2018-10-09 12:54:21 [books-sequential] INFO: Scheduling:
http://books.toscrape.com/catalogue/tipping-the-velvet_999/index.html
2018-10-09 12:54:22 [books-sequential] INFO: Extracting:
http://books.toscrape.com/catalogue/tipping-the-velvet_999/index.html
2018-10-09 12:54:22 [books-sequential] INFO: Scheduling: http://books.toscrape.com
2018-10-09 12:54:22 [books-sequential] INFO: Back at the main page
(...)
2018-10-09 12:54:30 [scrapy.core.engine] INFO: Spider closed (finished)

$
```

Extensions

scrapinghub

Extensions

The extensions framework provides a mechanism for inserting your own custom functionality into Scrapy.

Extensions are just regular classes that are instantiated at Scrapy startup.

Extensions - Example

Extensions can use signal handlers to perform custom actions.

In the next slide we will listen to the `item_scraped` signal to send items to a remote storage service (a local file in the example :wink:)

```
EXTENSIONS = {
      'pybr2018.extensions.RemoteStorageExtension': 550,
}
```

Enable the Extension in the project's settings.py file

```python
from twisted.internet.threads import deferToThread
from scrapy import signals

class RemoteStorageExtension:
	@classmethod
	def from_crawler(cls, crawler):
		ext = cls()
		crawler.signals.connect(ext.spider_opened, signal=signals.spider_opened)
		crawler.signals.connect(ext.item_scraped, signal=signals.item_scraped)
		crawler.signals.connect(ext.spider_closed, signal=signals.spider_closed)
		return ext

	def spider_opened(self, spider):
		self.file = open('{}_items.txt'.format(spider.name), 'w')

	def item_scraped(self, item, response, spider):
		return deferToThread(self._write_item, item)

	def _write_item(self, item):
		row = ', '.join([
			'{}: {}'.format(key, value) for key, value in
			sorted(dict(item).items(), key=lambda elem: elem[0])
		])
		self.file.write(row + '\n')
		return item

	def spider_closed(self, spider, reason):
		self.file.close()
```

```python
from twisted.internet.threads import deferToThread
from scrapy import signals

class RemoteStorageExtension:
    @classmethod
    def from_crawler(cls, crawler):
        ext = cls()
        crawler.signals.connect(ext.spider_opened, signal=signals.spider_opened)
        crawler.signals.connect(ext.item_scraped, signal=signals.item_scraped)
        crawler.signals.connect(ext.spider_closed, signal=signals.spider_closed)
        return ext

    def spider_opened(self, spider):
        self.file = open('{}_items.txt'.format(spider.name), 'w')

    def item_scraped(self, item, response, spider):
        return deferToThread(self._write_item, item)

    def _write_item(self, item):
        row = ', '.join([
            '{}: {}'.format(key, value) for key, value in
            sorted(dict(item).items(), key=lambda elem: elem[0])
        ])
        self.file.write(row + '\n')
        return item

    def spider_closed(self, spider, reason):
        self.file.close()
```

Attach signal handlers

```python
from twisted.internet.threads import deferToThread
from scrapy import signals

class RemoteStorageExtension:
    @classmethod
    def from_crawler(cls, crawler):
        ext = cls()
        crawler.signals.connect(ext.spider_opened, signal=signals.spider_opened)
        crawler.signals.connect(ext.item_scraped, signal=signals.item_scraped)
        crawler.signals.connect(ext.spider_closed, signal=signals.spider_closed)
        return ext

    def spider_opened(self, spider):
        self.file = open('{}_items.txt'.format(spider.name), 'w')

    def item_scraped(self, item, response, spider):
        return deferToThread(self._write_item, item)

    def _write_item(self, item):
        row = ', '.join([
            '{}: {}'.format(key, value) for key, value in
            sorted(dict(item).items(), key=lambda elem: elem[0])
        ])
        self.file.write(row + '\n')
        return item

    def spider_closed(self, spider, reason):
        self.file.close()
```

Signal handler: open connection

Signal handler: close connection

```python
from twisted.internet.threads import deferToThread
from scrapy import signals

class RemoteStorageExtension:
    @classmethod
    def from_crawler(cls, crawler):
        ext = cls()
        crawler.signals.connect(ext.spider_opened, signal=signals.spider_opened)
        crawler.signals.connect(ext.item_scraped, signal=signals.item_scraped)
        crawler.signals.connect(ext.spider_closed, signal=signals.spider_closed)
        return ext

    def spider_opened(self, spider):
        self.file = open('{}_items.txt'.format(spider.name), 'w')

    def item_scraped(self, item, response, spider):
        return deferToThread(self._write_item, item)

    def _write_item(self, item):
        row = ', '.join([
            '{}: {}'.format(key, value) for key, value in
            sorted(dict(item).items(), key=lambda elem: elem[0])
        ])
        self.file.write(row + '\n')
        return item

    def spider_closed(self, spider, reason):
        self.file.close()
```

Signal handler:
Return a Twisted Deferred to handle the blocking operation.
(Just an example, writing to the same file from multiple threads can lead to unexpected results!)

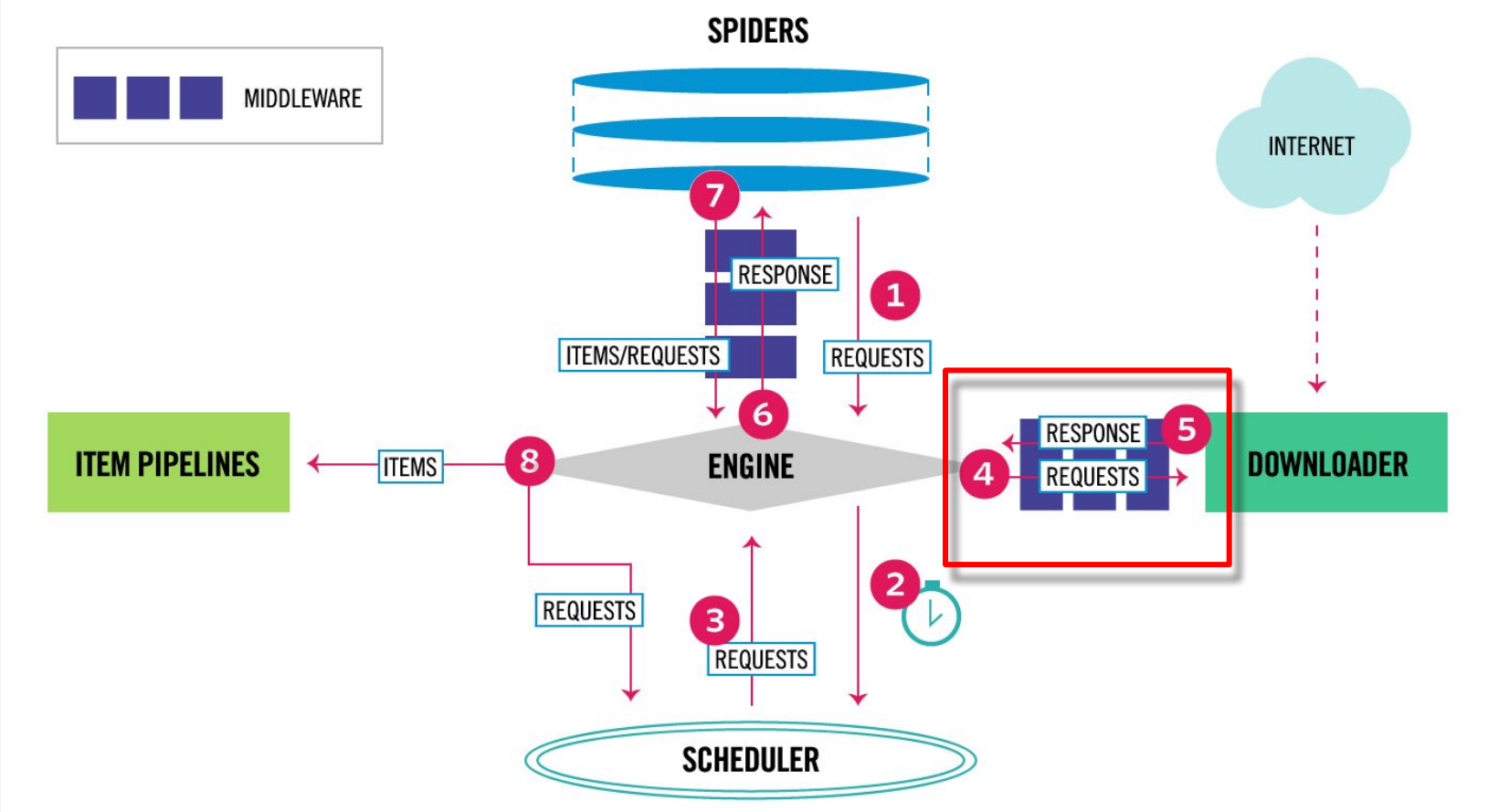
Serialize and write the item

```
$ scrapy crawl books
2018-10-09 14:11:36 [scrapy.utils.log] INFO: Scrapy 1.5.0 started (bot: pybr2018)
(...)
2018-10-09 14:11:36 [RemoteStorageExtension] INFO: Writing items to a remote location
(...)
2018-10-09 14:12:56 [scrapy.core.engine] INFO: Spider closed (finished)

$ head -n 5 books_items.txt
price: 51.77, title: A Light in the Attic, url:
http://books.toscrape.com/catalogue/a-light-in-the-attic_1000/index.html
price: 51.33, title: Libertarianism for Beginners, url:
http://books.toscrape.com/catalogue/libertarianism-for-beginners_982/index.html
price: 17.46, title: Set Me Free, url:
http://books.toscrape.com/catalogue/set-me-free_988/index.html
price: 35.02, title: Rip it Up and Start Again, url:
http://books.toscrape.com/catalogue/rip-it-up-and-start-again_986/index.html
price: 45.17, title: It's Only the Himalayas, url:
http://books.toscrape.com/catalogue/its-only-the-himalayas_981/index.html

$
```

Downloader middleware

scrapinghub

Downloader middleware



Downloader middleware

The downloader middleware is a framework of hooks into Scrapy's request/response processing. It's a light, low-level system for globally altering Scrapy's requests and responses.

From the docs

More at https://doc.scrapy.org/en/latest/topics/downloader-middleware.html

Downloader middleware - SOAP example

In the following example we will use Zeep (https://python-zeep.readthedocs.io) to communicate with a SOAP web service that converts between Celsius and Fahrenheit.

The SOAP-related heavy lifting will be handled by a downloader middleware, making the process relatively transparent to the spider.

```python
class TemperatureConversionMiddleware:

    def __init__(self):
        self.client = zeep.Client('https://www.w3schools.com/xml/tempconvert.asmx?WSDL')

    def process_request(self, request, spider):
        # ...
        body = self.client.create_message(request, ...)
        return Request(
            url=request.url,
            method='POST',
            body=lxml.etree.tostring(body),
            # ...
        )

    def process_response(self, request, response, spider):
        # ...
        request.meta['result'] = process_reply(response, ...)
        return response
```

```python
class TemperatureConversionMiddleware:

    def __init__(self):
        self.client = zeep.Client('https://www.w3schools.com/xml/tempconvert.asmx?WSDL')

    def process_request(self, request, spider):
        # ...
        body = self.client.create_message(request, ...)
        return Request(
            url=request.url,
            method='POST',
            body=lxml.etree.tostring(body),
            # ...
        )

    def process_response(self, request, response, spider):
        # ...
        request.meta['result'] = process_reply(response, ...)
        return response
```

The actual middleware is longer, lots of stuff omitted in this slide!

```python
class TemperatureConversionMiddleware:

    def __init__(self):
        self.client = zeep.Client('https://www.w3schools.com/xml/tempconvert.asmx?WSDL')

    def process_request(self, request, spider):
        # ...
        body = self.client.create_message(request, ...)
        return Request(
            url=request.url,
            method='POST',
            body=lxml.etree.tostring(body),
            # ...
        )

    def process_response(self, request, response, spider):
        # ...
        request.meta['result'] = process_reply(response, ...)
        return response
```

Create Zeep client

```python
class TemperatureConversionMiddleware:

    def __init__(self):
        self.client = zeep.Client('https://www.w3schools.com/xml/tempconvert.asmx?WSDL')

    def process_request(self, request, spider):
        # ...
        body = self.client.create_message(request, ...)
        return Request(
            url=request.url,
            method='POST',
            body=lxml.etree.tostring(body),
            # ...
        )

    def process_response(self, request, response, spider):
        # ...
        request.meta['result'] = process_reply(response, ...)
        return response
```

Process requests to the SOAP service.

Use Zeep to create XML request bodies for Scrapy to send

```
class TemperatureConversionMiddleware:

    def __init__(self):
        self.client = zeep.Client('https://www.w3schools.com/xml/tempconvert.asmx?WSDL')

    def process_request(self, request, spider):
        # ...
        body = self.client.create_message(request, ...)
        return Request(
            url=request.url,
            method='POST',
            body=lxml.etree.tostring(body),
            # ...
        )

    def process_response(self, request, response, spider):
        # ...
        request.meta['result'] = process_reply(response, ...)
        return response
```

Process responses coming back from the SOAP service.

Use Zeep to parse the XML bodies

Spider - Temperature converter

```
class TemperatureSpider(scrapy.Spider):
    name = 'temperature'
    url = 'https://www.w3schools.com/xml/tempconvert.asmx'
    custom_settings = {
        'DOWNLOADER_MIDDLEWARES': {
            'pybr2018.middlewares.temperature.TemperatureConversionMiddleware': 543,
        }
    }

    def start_requests(self):
        for operation in ('CelsiusToFahrenheit', 'FahrenheitToCelsius'):
            for _ in range(5):
                meta = {'operation_name': operation, 'source_value': random.uniform(0, 50)}
                yield scrapy.Request(self.url, dont_filter=True, meta=meta)

    def parse(self, response):
        source_unit, destination_unit = response.meta['operation_name'].split('To')
        return {
            'source': '{} {}'.format(response.meta['source_value'], source_unit),
            'destination': '{} {}'.format(response.meta['result'], destination_unit),
        }
```

Spider - Temperature converter

```
class TemperatureSpider(scrapy.Spider):
    name = 'temperature'
    url = 'https://www.w3schools.com/xml/tempconvert.asmx'
    custom_settings = {
        'DOWNLOADER_MIDDLEWARES': {
            'pybr2018.middlewares.temperature.TemperatureConversionMiddleware': 543,
        }
    }

    def start_requests(self):
        for operation in ('CelsiusToFahrenheit', 'FahrenheitToCelsius'):
            for _ in range(5):
                meta = {'operation_name': operation, 'source_value': random.uniform(0, 50)}
                yield scrapy.Request(self.url, dont_filter=True, meta=meta)

    def parse(self, response):
        source_unit, destination_unit = response.meta['operation_name'].split('To')
        return {
            'source': '{} {}'.format(response.meta['source_value'], source_unit),
            'destination': '{} {}'.format(response.meta['result'], destination_unit),
        }
```

SOAP service base URL

Enable the downloader middleware

Spider - Temperature converter

```
class TemperatureSpider(scrapy.Spider):
    name = 'temperature'
    url = 'https://www.w3schools.com/xml/tempconvert.asmx'
    custom_settings = {
        'DOWNLOADER_MIDDLEWARES': {
            'pybr2018.middlewares.temperature.TemperatureConversionMiddleware': 543,
        }
    }

    def start_requests(self):
        for operation in ('CelsiusToFahrenheit', 'FahrenheitToCelsius'):
            for _ in range(5):
                meta = {'operation_name': operation, 'source_value': random.uniform(0, 50)}
                yield scrapy.Request(self.url, dont_filter=True, meta=meta)

    def parse(self, response):
        source_unit, destination_unit = response.meta['operation_name'].split('To')
        return {
            'source': '{} {}'.format(response.meta['source_value'], source_unit),
            'destination': '{} {}'.format(response.meta['result'], destination_unit),
        }
```

Produce requests with random temperature values for the SOAP service to convert

Spider - Temperature converter

```
class TemperatureSpider(scrapy.Spider):
    name = 'temperature'
    url = 'https://www.w3schools.com/xml/tempconvert.asmx'
    custom_settings = {
        'DOWNLOADER_MIDDLEWARES': {
            'pybr2018.middlewares.temperature.TemperatureConversionMiddleware': 543,
        }
    }

    def start_requests(self):
        for operation in ('CelsiusToFahrenheit', 'FahrenheitToCelsius'):
            for _ in range(5):
                meta = {'operation_name': operation, 'source_value': random.uniform(0, 50)}
                yield scrapy.Request(self.url, dont_filter=True, meta=meta)

    def parse(self, response):
        source_unit, destination_unit = response.meta['operation_name'].split('To')
        return {
            'source': '{} {}'.format(response.meta['source_value'], source_unit),
            'destination': '{} {}'.format(response.meta['result'], destination_unit),
        }
```

Produce items with the converted value, processed by the middleware

```
$ scrapy crawl temperature -o temperature.json
2018-10-10 10:38:41 [scrapy.utils.log] INFO: Scrapy 1.5.0 started (bot: pybr2018)
(...)
2018-10-10 10:38:42 [TemperatureConversionMiddleware] INFO: Creating request for "CelsiusToFahrenheit" operation
(...)
2018-10-10 10:38:43 [TemperatureConversionMiddleware] INFO: Processing response for "CelsiusToFahrenheit" operation
(...)
2018-10-10 10:38:43 [scrapy.core.engine] INFO: Spider closed (finished)

$ cat temperature.json | jq .
[
  {
      "source": "44.562162995713486 Celsius",
      "destination": "112.211893392284 Fahrenheit"
  },
  {
      "source": "4.497578874865576 Celsius",
      "destination": "40.095641974758 Fahrenheit"
  },
(...)

$
```

Questions?

scrapinghub

Scrapinghub is hiring!

We're a globally distributed team of over 140 scrapinghubbers who are passionate about scraping, web crawling, and data science

Come join the team: https://scrapinghub.com/jobs

scrapinghub

Thanks!

Eugenio Lacuesta
lacuesta@scrapinghub.com

scrapinghub