Thực hành Data Mining lab 2

Phạm Ngọc Hải, Lê Thị Minh Anh, Lương Đức Anh, Cao Diệu Ly

February 28, 2024

1 Tìm hiểu và làm ví dụ về Scrapy

1.1 Đáp ứng điều kiện cài đặt

[ ]:
```
# Xác định thông tin nền tảng

# Settings for notebook
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
# Show Python version
import platform
platform.python_version()
```

[ ]:
```
'3.11.7'
```

1.2 Cài đặt và import

[ ]:
```
# Cài đặt và import scrapy

try:
    import scrapy
except:
    !pip install scrapy
    import scrapy
from scrapy.crawler import CrawlerProcess
```

1.3 Thiết lập một đường ống pipeline

Lớp này tạo một đường dẫn đơn giản ghi tất cả các mục tìm thấy vào một tệp JSON, trong đó mỗi dòng chứa một phần tử JSON.

[ ]:
```
import json

class JsonWriterPipeline(object):

    def open_spider(self, spider):
        self.file = open('quoteresult.jl', 'w')
```

```
    def close_spider(self, spider):
        self.file.close()

    def process_item(self, item, spider):
        line = json.dumps(dict(item)) + "\n"
        self.file.write(line)
        return item
```

1.4 Define the spider

Lớp QuoteSpider xác định URL nào sẽ bắt đầu thu thập dữ liệu và giá trị nào cần truy xuất. Tôi đặt mức ghi nhật ký của trình thu thập thông tin thành cảnh báo, nếu không sổ ghi chép sẽ bị quá tải với các thông báo GỠ LỖI về dữ liệu được truy xuất.

[ ]:
```
import logging

class QuotesSpider(scrapy.Spider):
    name = "quotes"
    start_urls = [
        'http://quotes.toscrape.com/page/1/',
        'http://quotes.toscrape.com/page/2/',
    ]
    custom_settings = {
        'LOG_LEVEL': logging.WARNING,
        'ITEM_PIPELINES': {'__main__.JsonWriterPipeline': 1}, # Used for␣
 ↪pipeline 1
        'FEED_FORMAT':'json',                                 # Used for␣
 ↪pipeline 2
        'FEED_URI': 'quoteresult.json'                        # Used for␣
 ↪pipeline 2
    }

    def parse(self, response):
        for quote in response.css('div.quote'):
            yield {
                'text': quote.css('span.text::text').extract_first(),
                'author': quote.css('span small::text').extract_first(),
                'tags': quote.css('div.tags a.tag::text').extract(),
            }
```

1.5 Start the crawler

[ ]:
```
process = CrawlerProcess({
    'USER_AGENT': 'Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)'
})

process.crawl(QuotesSpider)
```

```
process.start()
```

```
2024-02-28 15:55:31 [scrapy.utils.log] INFO: Scrapy 2.11.1 started (bot:
scrapybot)
2024-02-28 15:55:31 [scrapy.utils.log] INFO: Versions: lxml 5.1.0.0, libxml2
2.12.3, cssselect 1.2.0, parsel 1.8.1, w3lib 2.1.2, Twisted 23.10.0, Python
3.11.7 (main, Jan 29 2024, 16:03:57) [GCC 13.2.1 20230801], pyOpenSSL 24.0.0
(OpenSSL 3.1.4 24 Oct 2023), cryptography 41.0.7, Platform
Linux-6.7.4-zen1-1-zen-x86_64-with-glibc2.39
2024-02-28 15:55:31 [py.warnings] WARNING:
/home/harito/venv/py/lib/python3.11/site-packages/scrapy/utils/request.py:254:
ScrapyDeprecationWarning: '2.6' is a deprecated value for the
'REQUEST_FINGERPRINTER_IMPLEMENTATION' setting.

It is also the default value. In other words, it is normal to get this warning
if you have not defined a value for the 'REQUEST_FINGERPRINTER_IMPLEMENTATION'
setting. This is so for backward compatibility reasons, but it will change in a
future version of Scrapy.

See the documentation of the 'REQUEST_FINGERPRINTER_IMPLEMENTATION' setting for
information on how to handle this deprecation.
  return cls(crawler)

2024-02-28 15:55:31 [py.warnings] WARNING:
/home/harito/venv/py/lib/python3.11/site-
packages/scrapy/extensions/feedexport.py:406: ScrapyDeprecationWarning: The
`FEED_URI` and `FEED_FORMAT` settings have been deprecated in favor of the
`FEEDS` setting. Please see the `FEEDS` setting docs for more details
  exporter = cls(crawler)
```

[ ]:
```
<Deferred at 0x715f3c2b3d10>
```

1.6 Kiểm tra và hiện ra kết quả thu được

[ ]:
```
ll quoteresult.*
```

```
-rwxrwxrwx 1 root 5551 Feb 28 15:55 quoteresult.jl*
-rwxrwxrwx 1 root 5573 Feb 28 15:55 quoteresult.json*
```

[ ]:
```
!tail -n 2 quoteresult.jl
```

```
{"text": "\u201cGood friends, good books, and a sleepy conscience: this is the
ideal life.\u201d", "author": "Mark Twain", "tags": ["books", "contentment",
"friends", "friendship", "life"]}
{"text": "\u201cLife is what happens to us while we are making other
plans.\u201d", "author": "Allen Saunders", "tags": ["fate", "life",
"misattributed-john-lennon", "planning", "plans"]}
```

[ ]:
```
!tail -n 2 quoteresult.json
```

```
{"text": "\u201cLife is what happens to us while we are making other
plans.\u201d", "author": "Allen Saunders", "tags": ["fate", "life",
"misattributed-john-lennon", "planning", "plans"]}
]
```

[ ]:
```
import pandas as pd
dfjson = pd.read_json('quoteresult.json')
dfjson
```

[ ]:

	text	author	\
0	"The world as we have created it is a process …	Albert Einstein	
1	"It is our choices, Harry, that show what we t…	J.K. Rowling	
2	"There are only two ways to live your life. On…	Albert Einstein	
3	"The person, be it gentleman or lady, who has …	Jane Austen	
4	"Imperfection is beauty, madness is genius and…	Marilyn Monroe	
5	"Try not to become a man of success. Rather be…	Albert Einstein	
6	"It is better to be hated for what you are tha…	André Gide	
7	"I have not failed. I've just found 10,000 way…	Thomas A. Edison	
8	"A woman is like a tea bag; you never know how…	Eleanor Roosevelt	
9	"A day without sunshine is like, you know, nig…	Steve Martin	
10	"This life is what you make it. No matter what…	Marilyn Monroe	
11	"It takes a great deal of bravery to stand up …	J.K. Rowling	
12	"If you can't explain it to a six year old, yo…	Albert Einstein	
13	"You may not be her first, her last, or her on…	Bob Marley	
14	"I like nonsense, it wakes up the brain cells…	Dr. Seuss	
15	"I may not have gone where I intended to go, b…	Douglas Adams	
16	"The opposite of love is not hate, it's indiff…	Elie Wiesel	
17	"It is not a lack of love, but a lack of frien…	Friedrich Nietzsche	
18	"Good friends, good books, and a sleepy consci…	Mark Twain	
19	"Life is what happens to us while we are makin…	Allen Saunders	

	tags
0	[change, deep-thoughts, thinking, world]
1	[abilities, choices]
2	[inspirational, life, live, miracle, miracles]
3	[aliteracy, books, classic, humor]
4	[be-yourself, inspirational]
5	[adulthood, success, value]
6	[life, love]
7	[edison, failure, inspirational, paraphrased]
8	[misattributed-eleanor-roosevelt]
9	[humor, obvious, simile]
10	[friends, heartbreak, inspirational, life, lov…
11	[courage, friends]
12	[simplicity, understand]
13	[love]

```
14                                          [fantasy]
15                                 [life, navigation]
16  [activism, apathy, hate, indifference, inspira…
17  [friendship, lack-of-friendship, lack-of-love,…
18    [books, contentment, friends, friendship, life]
19  [fate, life, misattributed-john-lennon, planni…
```

[ ]:
```
dfjl = pd.read_json('quoteresult.jl', lines=True)
dfjl
```

[ ]:
```
                                                text               author  \
0   "The world as we have created it is a process …      Albert Einstein
1   "It is our choices, Harry, that show what we t…         J.K. Rowling
2   "There are only two ways to live your life. On…      Albert Einstein
3   "The person, be it gentleman or lady, who has …          Jane Austen
4   "Imperfection is beauty, madness is genius and…       Marilyn Monroe
5   "Try not to become a man of success. Rather be…      Albert Einstein
6   "It is better to be hated for what you are tha…           André Gide
7   "I have not failed. I've just found 10,000 way…     Thomas A. Edison
8   "A woman is like a tea bag; you never know how…    Eleanor Roosevelt
9   "A day without sunshine is like, you know, nig…         Steve Martin
10  "This life is what you make it. No matter what…       Marilyn Monroe
11  "It takes a great deal of bravery to stand up …         J.K. Rowling
12  "If you can't explain it to a six year old, yo…      Albert Einstein
13  "You may not be her first, her last, or her on…           Bob Marley
14  "I like nonsense, it wakes up the brain cells…            Dr. Seuss
15  "I may not have gone where I intended to go, b…        Douglas Adams
16  "The opposite of love is not hate, it's indiff…          Elie Wiesel
17  "It is not a lack of love, but a lack of frien…  Friedrich Nietzsche
18  "Good friends, good books, and a sleepy consci…           Mark Twain
19  "Life is what happens to us while we are makin…      Allen Saunders

                                                tags
0             [change, deep-thoughts, thinking, world]
1                                  [abilities, choices]
2      [inspirational, life, live, miracle, miracles]
3                   [aliteracy, books, classic, humor]
4                        [be-yourself, inspirational]
5                         [adulthood, success, value]
6                                       [life, love]
7       [edison, failure, inspirational, paraphrased]
8                    [misattributed-eleanor-roosevelt]
9                              [humor, obvious, simile]
10  [friends, heartbreak, inspirational, life, lov…
11                                 [courage, friends]
12                           [simplicity, understand]
13                                             [love]
```

```
14                                          [fantasy]
15                                [life, navigation]
16  [activism, apathy, hate, indifference, inspira…
17  [friendship, lack-of-friendship, lack-of-love,…
18    [books, contentment, friends, friendship, life]
19  [fate, life, misattributed-john-lennon, planni…
```

[ ]:
```
dfjson.to_pickle('quotejson.pickle')
dfjl.to_pickle('quotejl.pickle')
```

[ ]:
```
ll *pickle
```

```
-rwxrwxrwx 1 root 5454 Feb 28 15:57 quotejl.pickle*
-rwxrwxrwx 1 root 5454 Feb 28 15:57 quotejson.pickle*
```

2 Thử nghiệm việc sử dụng Scraper để crawl data làm project nhóm

Do nhóm em đã có nguồn dữ liệu tải được từ trước nên định hướng sẽ không sử dụng Scraper. Tuy nhiên trong khuôn khổ nội dung bài thực hành, nhóm em sẽ sử dụng crawl ảnh (nguồn dữ liệu dự án nhóm em sẽ sử dụng).

2.0.1 Cách 1.

[ ]:
```
!pip install ImageScraper
```

```
Requirement already satisfied: ImageScraper in
/home/harito/venv/py/lib/python3.11/site-packages (2.0.7)
Requirement already satisfied: lxml>=3.2.3 in
/home/harito/venv/py/lib/python3.11/site-packages (from ImageScraper) (5.1.0)
Requirement already satisfied: requests>=2.1.0 in
/home/harito/venv/py/lib/python3.11/site-packages (from ImageScraper) (2.31.0)
Requirement already satisfied: future>=0.14.3 in
/home/harito/venv/py/lib/python3.11/site-packages (from ImageScraper) (1.0.0)
Requirement already satisfied: setproctitle>=1.1.8 in
/home/harito/venv/py/lib/python3.11/site-packages (from ImageScraper) (1.3.3)
Requirement already satisfied: SimplePool in
/home/harito/venv/py/lib/python3.11/site-packages (from ImageScraper) (0.1)
Requirement already satisfied: charset-normalizer<4,>=2 in
/home/harito/venv/py/lib/python3.11/site-packages (from
requests>=2.1.0->ImageScraper) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in
/home/harito/venv/py/lib/python3.11/site-packages (from
requests>=2.1.0->ImageScraper) (2.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/home/harito/venv/py/lib/python3.11/site-packages (from
requests>=2.1.0->ImageScraper) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
```

```
/home/harito/venv/py/lib/python3.11/site-packages (from
requests>=2.1.0->ImageScraper) (2023.7.22)
```

```
[ ]: !image-scraper --max-images 10 'https://vnexpress.net/'
```

```
ImageScraper
============
Requesting page…

Found 2 images:
Progress: 100% |||||||||||||||||||||||||||||||||||||||| Time: 00:00:00   1.82 K/s

Done!
Downloaded 2 images
Failed: 0
```

Vậy là đã download được 2 ảnh có từ web

2.0.2 Cách 2. Vấn dùng Scraper

```
[ ]: !scrapy startproject oral_cancer_images
```

```
New Scrapy project 'oral_cancer_images', using template directory
'/home/harito/venv/py/lib/python3.11/site-packages/scrapy/templates/project',
created in:
    /mnt/DataK/Univer/UniSubject/_3th_year/_2nd_term/3ii_DM/Lec_Ass/oral_cancer_
images

You can start your first spider with:
    cd oral_cancer_images
    scrapy genspider example example.com
```

```
[ ]: %cd oral_cancer_images/oral_cancer_images/spiders/
```

```
2024-02-28 16:27:08 [py.warnings] WARNING:
/home/harito/venv/py/lib/python3.11/site-
packages/IPython/core/magics/osm.py:417: UserWarning: using dhist requires you
to install the `pickleshare` library.
  self.shell.db['dhist'] = compress_dhist(dhist)[-100:]


/mnt/DataK/Univer/UniSubject/_3th_year/_2nd_term/3ii_DM/Lec_Ass/oral_cancer_imag
es/oral_cancer_images/spiders
```

```
[ ]: !scrapy genspider images_spider "https://oralcancerfoundation.org/dental/
     ↪oral-cancer-images"
```

```
Created spider 'images_spider' using template 'basic' in module:
  oral_cancer_images.spiders.images_spider
```

[ ]:
```
!scrapy crawl images_spider -o output.json
```

```
2024-02-28 16:27:51 [scrapy.utils.log] INFO: Scrapy 2.11.1 started (bot:
oral_cancer_images)
2024-02-28 16:27:51 [scrapy.utils.log] INFO: Versions: lxml 5.1.0.0, libxml2
2.12.3, cssselect 1.2.0, parsel 1.8.1, w3lib 2.1.2, Twisted 23.10.0, Python
3.11.7 (main, Jan 29 2024, 16:03:57) [GCC 13.2.1 20230801], pyOpenSSL 24.0.0
(OpenSSL 3.1.4 24 Oct 2023), cryptography 41.0.7, Platform
Linux-6.7.4-zen1-1-zen-x86_64-with-glibc2.39
2024-02-28 16:27:51 [scrapy.addons] INFO: Enabled addons:
[]
2024-02-28 16:27:51 [asyncio] DEBUG: Using selector: EpollSelector
2024-02-28 16:27:51 [scrapy.utils.log] DEBUG: Using reactor:
twisted.internet.asyncioreactor.AsyncioSelectorReactor
2024-02-28 16:27:51 [scrapy.utils.log] DEBUG: Using asyncio event loop:
asyncio.unix_events._UnixSelectorEventLoop
2024-02-28 16:27:51 [scrapy.extensions.telnet] INFO: Telnet Password:
d51179d5211b3bbd
2024-02-28 16:27:52 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.feedexport.FeedExporter',
 'scrapy.extensions.logstats.LogStats']
2024-02-28 16:27:52 [scrapy.crawler] INFO: Overridden settings:
{'BOT_NAME': 'oral_cancer_images',
 'FEED_EXPORT_ENCODING': 'utf-8',
 'NEWSPIDER_MODULE': 'oral_cancer_images.spiders',
 'REQUEST_FINGERPRINTER_IMPLEMENTATION': '2.7',
 'ROBOTSTXT_OBEY': True,
 'SPIDER_MODULES': ['oral_cancer_images.spiders'],
 'TWISTED_REACTOR': 'twisted.internet.asyncioreactor.AsyncioSelectorReactor'}
2024-02-28 16:27:52 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
 'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2024-02-28 16:27:52 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
```

```
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2024-02-28 16:27:52 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2024-02-28 16:27:52 [scrapy.core.engine] INFO: Spider opened
2024-02-28 16:27:52 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0
pages/min), scraped 0 items (at 0 items/min)
2024-02-28 16:27:52 [scrapy.extensions.telnet] INFO: Telnet console listening on
127.0.0.1:6023
2024-02-28 16:27:53 [scrapy.core.engine] DEBUG: Crawled (200) <GET
https://oralcancerfoundation.org/robots.txt> (referer: None)
2024-02-28 16:27:53 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting
(301) to <GET https://oralcancerfoundation.org/dental/oral-cancer-images/> from
<GET https://oralcancerfoundation.org/dental/oral-cancer-images>
2024-02-28 16:27:55 [scrapy.core.engine] DEBUG: Crawled (200) <GET
https://oralcancerfoundation.org/dental/oral-cancer-images/> (referer: None)
2024-02-28 16:27:55 [scrapy.core.engine] INFO: Closing spider (finished)
2024-02-28 16:27:55 [scrapy.extensions.feedexport] INFO: Stored json feed (0
items) in: output.json
2024-02-28 16:27:55 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 748,
 'downloader/request_count': 3,
 'downloader/request_method_count/GET': 3,
 'downloader/response_bytes': 159070,
 'downloader/response_count': 3,
 'downloader/response_status_count/200': 2,
 'downloader/response_status_count/301': 1,
 'elapsed_time_seconds': 3.296637,
 'feedexport/success_count/FileFeedStorage': 1,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2024, 2, 28, 9, 27, 55, 593668,
tzinfo=datetime.timezone.utc),
 'httpcompression/response_bytes': 1109061,
 'httpcompression/response_count': 2,
 'log_count/DEBUG': 6,
 'log_count/INFO': 11,
 'memusage/max': 141742080,
 'memusage/startup': 141742080,
 'response_received_count': 2,
 'robotstxt/request_count': 1,
 'robotstxt/response_count': 1,
 'robotstxt/response_status_count/200': 1,
 'scheduler/dequeued': 2,
 'scheduler/dequeued/memory': 2,
 'scheduler/enqueued': 2,
 'scheduler/enqueued/memory': 2,
 'start_time': datetime.datetime(2024, 2, 28, 9, 27, 52, 297031,
```

```
tzinfo=datetime.timezone.utc)}
2024-02-28 16:27:55 [scrapy.core.engine] INFO: Spider closed (finished)
```