

Scrapy

- 웹사이트에서 데이터 수집을 위한 오픈소스 파이썬 프레임워크
- 멀티스레딩으로 데이터 수집
- gmarket 상품데이터 수집

```
In [1]: # install scrapy
        #!pip install scrapy
```

1. make project

```
In [2]: !scrapy startproject news
```

Error: scrapy.cfg already exists in /Users/rada/Desktop/20240919_KT_Aivle/2024_0924_AI_WebCrawling_day3/notebooks_full/news

```
In [3]: !tree news
```

```
news
├── news
│   ├── __init__.py
│   ├── __pycache__
│   │   ├── __init__.cpython-311.pyc
│   │   ├── items.cpython-311.pyc
│   │   └── settings.cpython-311.pyc
│   ├── items.py
│   ├── middlewares.py
│   ├── pipelines.py
│   ├── settings.py
│   └── spiders
│       ├── __init__.py
│       ├── __pycache__
│       │   ├── __init__.cpython-311.pyc
│       │   └── spider.cpython-311.pyc
│       └── spider.py
├── news.csv
└── scrapy.cfg
```

5 directories, 14 files

- scrapy structure
 - items : 데이터의 모양 정의
 - middlewares : 수집할때 header 정보와 같은 내용을 설정
 - pipelines : 데이터를 수집한 후에 코드를 실행
 - settings : robots.txt 규칙, 크롤링 시간 텀등을 설정
 - spiders : 크롤링 절차를 정의

2. xpath

- link, contents

```
In [4]: import scrapy, requests
        from scrapy.http import TextResponse
```

```
In [5]: # 링크 데이터
```

```
In [6]: request = requests.get('https://news.daum.net')
        response = TextResponse(request.url, body=request.text, encoding='utf-8')
        response
```

```
Out[6]: <200 https://news.daum.net/>
```

```
In [7]: selector = '/html/body/div[2]/main/section/div/div[1]/div[1]/ul/li'
        selector += '/div/div/strong/a/@href'
        links = response.xpath(selector).extract()
        len(links), links[:2]
```

```
Out[7]: (20,
        ['https://v.daum.net/v/20240912195017442',
         'https://v.daum.net/v/20240912190722611'])
```

```
In [8]: # 상세 데이터
```

```
In [9]: link = links[19]
        request = requests.get(link)
        response = TextResponse(request.url, body=request.text, encoding='utf-8')
        response
```

```
Out[9]: <200 https://v.daum.net/v/20240912170721505>
```

```
In [10]: title = response.xpath('//*[@id="mArticle"]/div[1]/h3/text()')[0].extract()
        content = response.xpath('//section//p/text()').extract()
        content = ' '.join(content).replace('\xa0', ' ').replace("\'", ' ')
        title, content[:100]
```

```
Out[10]: ('[단독] 고교평준화 폐지→학교다양화로 둔갑? 국교위 ‘짬짜미’ 의혹 덮고 ‘사학 퍼주기’ 밀어붙이나',
        '이 글자크기로 변경됩니다. (예시) 가장 빠른 뉴스가 있고 다양한 정보, 쌍방향 소통이 숨쉬는 다음뉴스를 만나보세요. 다음뉴스는 국내외 주요이슈와 실시간 속보, 문화생활 및 다양한')
```

3. items.py

- Data Model

```
In [11]: %%writefile news/news/items.py
import scrapy

class NewsContents(scrapy.Item):
    title = scrapy.Field()
    content = scrapy.Field()
    link = scrapy.Field()
```

Overwriting news/news/items.py

4. spider.py

- write crawling process

```
In [12]: %%writefile news/news/spiders/spider.py
import scrapy
```

```

from news.items import NewsContents

class NewsSpider(scrapy.Spider):
    name = 'news'
    allow_domain = ['daum.net']
    start_urls = ['https://news.daum.net']

    def parse(self, response):
        selector = '/html/body/div[2]/main/section/div/div[1]/div[1]/ul/li'
        selector += '/div/div/strong/a/@href'
        links = response.xpath(selector).extract()
        for link in links:
            yield scrapy.Request(link, callback=self.parse_content)

    def parse_content(self, response):
        item = NewsContents()
        item['title'] = response.xpath(
            '//*[@id="mArticle"]/div[1]/h3/text()')[0].extract()
        item['link'] = response.url
        content = response.xpath('//section//p/text()').extract()
        content = ' '.join(content).replace('\xa0', ' ').replace("\'", "'")
        item['content'] = content
        yield item

```

Overwriting news/news/spiders/spider.py

5. run scrapy

- news 디렉토리에서 아래의 커맨드 실행
- scrapy crawl news -o news.csv

```

In [13]: import pandas as pd
pd.read_csv("news/news.csv")[['title', 'link', 'content']].tail(2)

```

```

Out[13]:

```

	title	link	content
18	박주민 "한 총리, 응급실 뺑뺑이 사망이 가짜뉴스?...어느 나라 살고있나"	https://v.daum.net/v/20240912194014270	더불어민주당 의료대란 대책특위 위원장인 박주민 의원이 한덕수 국무총리를 향해 "의료...
19	中 외교부 "손준호, 죄 인정했다...법정서 참회"	https://v.daum.net/v/20240912180151665	중국촉구협회가 손준호에 대한 영구 제명 조치를 내린 것과 관련해 손준호측이 적극 반...