scrapy

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

In [1]:

```
# install scrapy
#!pip install scrapy
```

In [2]:

```
# 1. 스크레피 프로젝트 생성
```

In [3]:

```
!scrapy startproject gmarket
```

```
Error: scrapy.cfg already exists in /Users/rada/Desktop/kt_abler/code/day3/gmarket
```

In [4]:

```
!tree gmarket
```

```
zsh:1: command not found: tree
```

In [5]:

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

In [6]:

```
# 2. xpath 찾기 : 링크, 상세 페이지
```

In [7]:

```
import scrapy, requests
from scrapy.http import TextResponse
```

In [8]:

```
# 링크 데이터
```

In [9]:

```
request = requests.get("http://corners.gmarket.co.kr/Bestsellers")
response = TextResponse(request.url, body=request.text, encoding="utf-8")
```

In [10]:

```
links = response.xpath('//*[@id="gBestWrap"]/div/div[3]/div/ul/li/a/@href').extract()
```

In [11]:

```
# 상세 데이터 : 상품명, 가격
```

In [12]:

```
link = links[0]
request = requests.get(link)
response = TextResponse(request.url, body=request.text, encoding="utf-8")
```

In [13]:

```
title = response.xpath('//*[@id="itemcase_basic"]/div[1]/h1/text()')[0].extract()
price = response.xpath('//*[@id="itemcase_basic"]/div[1]/p/span/strong/text()')[0].extract()
title, price
```

Out[13]:

```
('비비고  (CJ제일제당)  차돌된장찌개 460G 5봉 ', '21,710')
```

In [14]:

```
# 3. items.py : model
```

In [15]:

```
%%writefile gmarket/gmarket/items.py
import scrapy

class GmarketItem(scrapy.Item):
    title = scrapy.Field()
    price = scrapy.Field()
    link = scrapy.Field()
```

```
Overwriting gmarket/gmarket/items.py
```

In [16]:

```
# 4. spider.py : 크롤링 절차 정의
```

In [17]:

```
%%writefile gmarket/gmarket/spiders/spider.py
import scrapy
from gmarket.items import GmarketItem

class GMSpider(scrapy.Spider):
    name = "GMB"
    allow_domain = ["gmarket.co.kr"]
    start_urls = ["http://corners.gmarket.co.kr/Bestsellers"]

    def parse(self, response):
        links = response.xpath('//*[@id="gBestWrap"]/div/div[3]/div/ul/li/a/@href').extract()
        for link in links[:20]:
            yield scrapy.Request(link, callback=self.parse_content)

    def parse_content(self, response):
        item = GmarketItem()
        item["title"] = response.xpath('//*[@id="itemcase_basic"]/div[1]/h1/text()')[0].extract()
        item["price"] = response.xpath('//*[@id="itemcase_basic"]/div[1]/p/span/strong/text()')[0].ex
        item["link"] = response.url
        yield item
```

```
Overwriting gmarket/gmarket/spiders/spider.py
```

In [18]:

```
# 5. 스크래피 실행
# gmarket 디렉토리에서 아래의 커멘드 실행
# scrapy crawl GMB -o items.csv
```

In [19]:

```
%ls gmarket/
```

```
gmarket/    items.csv   scrapy.cfg
```

In [20]:

```
import pandas as pd
```

In [21]:

```
pd.read_csv("gmarket/items.csv")
```

Out[21]:

	link	price	title
0	http://item.gmarket.co.kr/Item?goodscode=13862...	9,400	성주 참외 가정용 랜덤 실중량 10kg 대한민국 최저가
1	http://item.gmarket.co.kr/Item?goodscode=17925...	45,000	(12%+18%쿠폰) 엘칸토 남여 쿨썸머 슈즈 BEST 모음전
2	http://item.gmarket.co.kr/Item?goodscode=22506...	12,900	(무료반품) 제이프랑 여름신상 티셔츠/팬츠/반팔/셋업
3	http://item.gmarket.co.kr/Item?goodscode=21454...	21,710	비비고 (CJ제일제당) 차돌된장찌개 460G 5봉
4	http://item.gmarket.co.kr/Item?goodscode=66793...	33,000	(최대50%+10%) 여름 뷰티케어는 이니스프리로 PICK
...	...	...	...
194	http://item.gmarket.co.kr/Item?goodscode=22361...	9,900	얼티미스틱 벨벳 립스틱 2개
195	http://item.gmarket.co.kr/Item?goodscode=15667...	17,900	청정우 차돌박이 (3초구이 샤브용) 250gX3팩
196	http://item.gmarket.co.kr/Item?goodscode=21848...	13,900	부광 얇은헤드 칫솔 초극세모 12입 X 2세트
197	http://item.gmarket.co.kr/Item?goodscode=16357...	8,900	1+1+1행사 모이스쳐 틴트 립밤
198	http://item.gmarket.co.kr/Item?goodscode=17764...	6,900	치실 국내생산 이편한 치실 80개입 x 5팩(400개)

199 rows × 3 columns