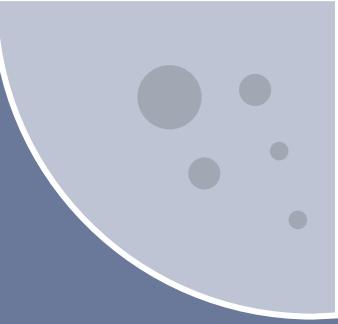




Scrapy 사용해서

Crawling하기



1. 실제 크롤링할 스파이더(spider, scrapy 기반 크롤링 프로그램) 생성
2. 크롤링할 사이트(시작점)와 크롤링할 아이템(item)에 대한 selector 설정
3. 크롤러 실행

크롤링 프로젝트 생성

scrapy startproject <프로젝트이름> 으로 생성

```
scrapy startproject ecommerce
```

Scrapy로 작성된 프로젝트

ecommerce 폴더 내의 파일 구조 확인

```
scrapy.cfg          # deploy configuration file  
ecommerce/          # project's Python module, you'll import your code from here  
    __init__.py  
    items.py        # project items definition file  
    pipelines.py    # project pipelines file  
    settings.py     # project settings file  
    spiders/         # a directory where you'll later put your spiders  
        __init__.py
```

크롤러(spider) 작성

터미널에서 ecommerce/ecommerce 폴더에서 다음 명령으로 작성 가능

- 크롤러이름: 크롤링 프로젝트 내에, 여러 가지 크롤러(scrapy에서는 spider라고 함) 있을 수 있으므로, 각 크롤러의 이름을 지정
- 크롤링페이지주소: 각 크롤러가 크롤링을 시작할 페이지 주소를 지정

```
scrapy genspider <크롤러이름> <크롤링페이지주소>
```

```
scrapy genspider gmarket www.gmarket.co.kr
```

크롤러(spider) 작성

ecommerce/ecommerce/spiders 디렉토리에 [gmarket.py](#) 파일(템플릿)이 생김
직접 scrapy genspider 명령을 사용하지 않고, 만들어도 됨

크롤러(spider) 실행

- 터미널 환경에서, ecommerce 디렉토리에서 scrapy crawl gmarket 명령

```
scrapy startproject ecommerce
```

```
scrapy genspider gmarket www.gmarket.co.kr
```

```
scrapy crawl gmarket
```

크롤러(spider) 작성

- 클래스 이름은 마음대로 정하면 됨, 단 scrapy.Spider 를 상속받아야 함
- name이 크롤러(spider)의 이름
- allowed_domains는 옵션 (삭제해도 무방함)
 - 별도 상세 설정으로 허용된 주소 이외의 주소는 크롤링 못하게끔 하는 기능을 위한 변수

```
# -*- coding: utf-8 -*-
import scrapy

class GmarketSpider(scrapy.Spider):
    name = 'gmarket'
    allowed_domains = ['www.gmarket.co.kr']
    start_urls = ['http://www.gmarket.co.kr/']

    def parse(self, response):
        pass
```

크롤러(spider) 작성

- start_urls 가 중요함. 크롤링할 페이지 주소를 나타냄.
- parse 함수는 클래스의 메서드로 response를 반드시 인자로 받아야 함
 - response에 start_urls 에 기록된 주소의 크롤링 결과가 담아져오기 때문임

```
# -*- coding: utf-8 -*-
import scrapy

class GmarketSpider(scrapy.Spider):
    name = 'gmarket'
    allowed_domains = ['www.gmarket.co.kr']
    start_urls = ['http://www.gmarket.co.kr/']

    def parse(self, response):
        pass
```

크롤러(spider) 작성

- start_urls는 리스트로 크롤링할 주소를 여러개 써도 됨
- 동작 방식
 - i. start_urls에서 주소를 하나씩 가져와서 크롤링한 후,
 - ii. response에 넣고, parse 함수를 호출함
 - iii. parse 함수에 response에 담아져있는 크롤링 결과를 원하는 대로 처리하면 됨

```
# 예
start_urls = ['https://corners.gmarket.co.kr/Bestsellers/', 'https://sports.v.daum.net/schedule']
```

크롤러(spider)로 작성해보기

- 아이템 데이터 처리하기

| 지금부터 scrapy의 강점이 나타난다.

- 다양한 데이터 포맷으로 아이템 저장하기

- csv, xml, json 포맷

- 터미널 환경에서, ecommerce 폴더에서 다음 명령

```
scrapy crawl 크롤러명 -o 저장할파일명 -t 저장포맷  
# 예  
scrapy crawl gmarket -o gmarket.csv -t csv  
scrapy crawl gmarket -o gmarket.xml -t xml
```

| json 파일을 확인하면, 한글문자가 깨져나옴

```
scrapy crawl gmarket -o gmarket.json -t json
```

크롤러(spider)로 작성해보기

- [settings.py](#) 수정

```
FEED_EXPORT_ENCODING = 'utf-8'
```

```
scrapy crawl gmarket -o gmarket.json -t json
```

```
In [9]: response.css('head > title::text').get()
```

```
Out[9]: 'G마켓 - G마켓 베스트'
```

```
}
```

```
In [10]: response.css('div.best-list li a::text').getall()
```

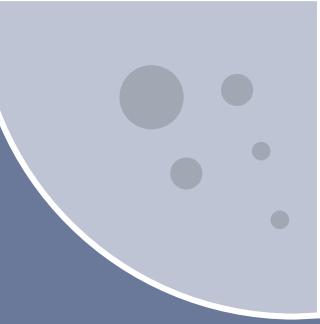
```
● ● ● 1. scrapy shell 'http://corners.gmarket.co.kr/Bestsellers' (python3.7)
['[Chicline]여름신상/빅사이즈/블라우스/남방/니트/레이스/하객룩',
 '동크앤팅크 70g 프리미엄 물티슈 리필 72매 10팩 b',
 '[GS25](GS25) 모바일 2만원권',
 '보가 G 클로티 402 여성 샌들 힐 스트랩 가보시힐',
 '힐링 크릴 100 (60캡슐 x1000mg) 남극크릴새우오일',
 '어죽탕 600g 남한강물고기 푸드 5개 이상 무료배송',
 '14kg 일반세탁기 DWF-14GAWR 전국무료설치배송',
 '약산성 트리플에스 탈모샴푸 2000ml+250ml (무료샘플)',
 '허그 여름 레깅스/치마레깅스/반바지레깅스/치마바지',
 '마블링 굿 LA갈비 1KG / 짬갈비 초특가 (기한임박)',
 '1+1애견패드 14종/강남패드/요요쉬패드/강아지패드',
 'V130 최대 할인 30만 4415U/4G/NVMe128G/15.6',
 '갓지은 구워먹는 치즈절편 앙꼬앙금절편 치즈떡 찰떡',
 '아로니아 10kg 냉동생과 우량품 40%새일 창고정리',
 '주말특가 SALE 로고패턴추가 44종 여성가방/숄더백',
 '엠보 미끄럼방지 욕실화 실내화 /푹신한 EVA소재',
 '사과5kg/사과9~10kg 어린이날선물세트 부사 간식 tkrhk',
 '5개 1천 원 할인 /여름특가/상하복/원피스/티셔츠/팬츠',
 '~160호 여아/수영복/아동/래쉬가드/주니어/쥬니어',
 'BEREUKAN 편광 선글라스 등산/낚시/스포츠 BER-2719',
 '아동복/티셔츠/레깅스/바지/민소매/치마/반팔/원피스',
 '프라임 국산정품 건전지 AA AAA/망간 C/D/9V국산꼭확인',
```

```
.../ecommerce/ecommerce ➤ ls
__init__.py           gmarket_products.csv middlewares.py      settings.py
__pycache__          items.py           pipelines.py       spiders
.../ecommerce ➤ scrapy crawl gmarket_best -o gmarket_products.xml -t csv
```

```
<?xml version="1.0" encoding="utf-8"?>
<items>
<item><title>(100원 응모될 ) 카카오 배그데이 에어드랍 인형 </title><price>100원 </price></item>
<item><title>복을 부르는 인테리어 소품 코끼리조각 2P 세트 </title><price>12,900원 </price></item>
<item><title>11kg 가정용 꼬마 성주 꿀맛 참외 (포장제 포함무게 )</title><price>10,900원 </price></item>
<item><title>[백설 ]포도씨유 900ml 보다 사용하기 편리한 500mlx6개 </title><price>17,500원 </price></item>
<item><title>[아이스샌드 ](신세계경기점)아동 래쉬가드 플패키지 (상의+하의+플랩캡+아쿠아슈즈 ) 14종 </title><price>22,730원 </price></item>
<item><title>무안군자연식품국내산 100%빨간양파즙 100팩 (팩당 100ml)</title><price>19,000원 </price></item>
<item><title>[GFresh]수안보농협 햇 대학찰옥수수 15개 특 예약 순차출고 </title><price>8,900원 </price></item>
<item><title>[탑 모델 ]9900균 일가 /조리 /쪼리 /스트랩 /샌들 /웨지힐 /슬리퍼 </title><price>9,900원 </price></item>
<item><title>(3+1) 빅 사이즈 냉장고바지 상하세트 </title><price>5,900원 </price></item>
<item><title>[미스타셰프 ]미스타셰프 육개장 600g 4팩 /즉석탕 /즉석국 /특가 </title><price>9,900원 </price></item>
<item><title>[프로스펙스 ]풋볼 트레이닝 3/4 팬츠 (블랙 /네이비 ) 19 신상 </title><price>12,900원 </price></item>
<item><title>[프로스펙스 ]풋볼 트레이닝 쇼트 반바지 (블랙 /네이비 ) 19 신상 </title><price>9,900원 </price></item>
<item><title>푸마바디웨어 남성 /여성 드로즈 /브라 /팬티 </title><price>7,900원 </price></item>
<item><title>[칼카니 ]마블 X칼카니 성인 /아동 반팔티 나시티 반바지 100종 </title><price>6,900원 </price></item>
<item><title>개국특집 /역대최다구성 창억떡 세트_호박인절미 1팩 더 (총 110개 )</title><price>54,900원 </price></item>
<item><title>모던브라운 코끼리조각 2P 세트 </title><price>12,900원 </price></item>
<item><title>티메이 블라우스 원피스 티셔츠 여름신상 반팔티 </title><price>9,900원 </price></item>
<item><title>동배순삭 재구매 레전드 순면 동배커버 복대팬티 </title><price>3,900원 </price></item>
<item><title>[웅진 ]빅토리아 탄산수 /음료 500mlx40pet 11종 중 택 2</title><price>18,900원 </price></item>
<item><title>[코카콜라 ]코카콜라 190mlx30캔 </title><price>13,900원 </price></item>
```



Q & A



Thank you

