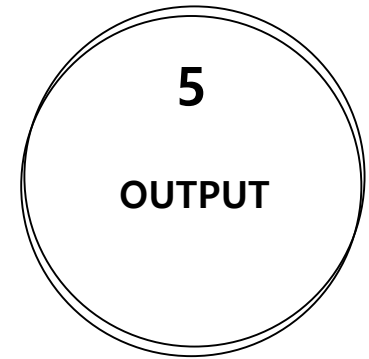
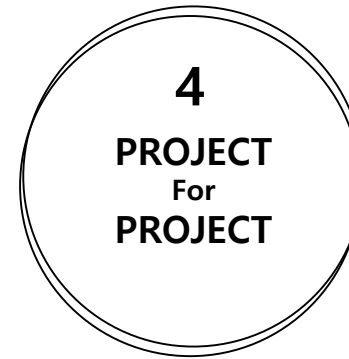
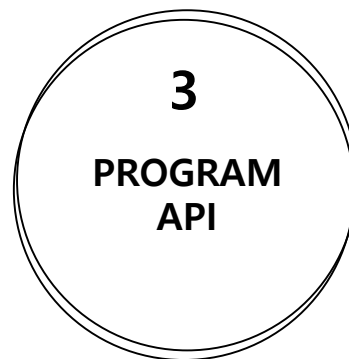
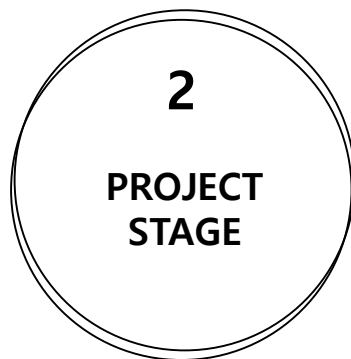




PYTHON PROJECT

- 시각장애인들을 위한 핵심 데이터 추출 -

INDEX

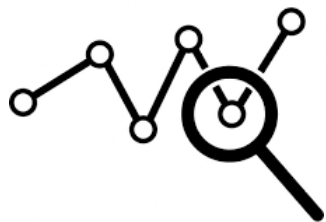


BACKGROUND

1. 시각장애인들을 위한 음성 알림 기술 = 매우 미흡.
2. 단지, TEXT를 읽어주고 안내하는 역할.
3. 원하는 정보를 찾기까지 많은 시간이 걸림.



CONCEPT



크롤링을 이용해 핵심 데이터 추출하여 해당 데이터를 음성으로 출력.

시각장애인들을 위한 쓸모 있는 데이터를 추출해보자.

PROJECT STAGE

STEP 1



크롤링을 이용해 해당 URL
웹 페이지 전체 데이터 추출.

STEP 2



원하는 데이터 필터링.

STEP 1



Google TTS(text-to-speech)로
해당 데이터 음성 출력.

PROGRAM & API



ubuntu 18.04 LTS



{JSON}

PROCESS FOR PROJECT-1

```
$sudo apt-get update
$sudo apt-get install build-essential
$sudo apt-get install python (version 2.7)
$sudo apt-get install python-pip (파이썬에서 사용하는 pip(package manager) 설치)
```

```
$sudo pip install virtualenv virtualenvwrapper (파이썬 가상환경 설치)
$vi .bashrc (profile 설정 – 맨 마지막 줄에 write하고 save)
```

```
export WORKON_HOME=$HOME/.virtualenvs
source /usr/local/bin/virtualenvwrapper.sh
:wq!
```

\$mkvirtualenv [가상환경이름]

```
solteee@solteee-VirtualBox:~$ mkvirtualenv hyunji
New python executable in /home/solteee/.virtualenvs/hyunji/bin/python
Installing setuptools, pip, wheel...done.
virtualenvwrapper.user_scripts creating /home/solteee/.virtualenvs/hyunji/bin/predeactivate
virtualenvwrapper.user_scripts creating /home/solteee/.virtualenvs/hyunji/bin/postdeactivate
virtualenvwrapper.user_scripts creating /home/solteee/.virtualenvs/hyunji/bin/preactivate
virtualenvwrapper.user_scripts creating /home/solteee/.virtualenvs/hyunji/bin/postactivate
virtualenvwrapper.user_scripts creating /home/solteee/.virtualenvs/hyunji/bin/get_env_details
(hyunji) solteee@solteee-VirtualBox:~$
```

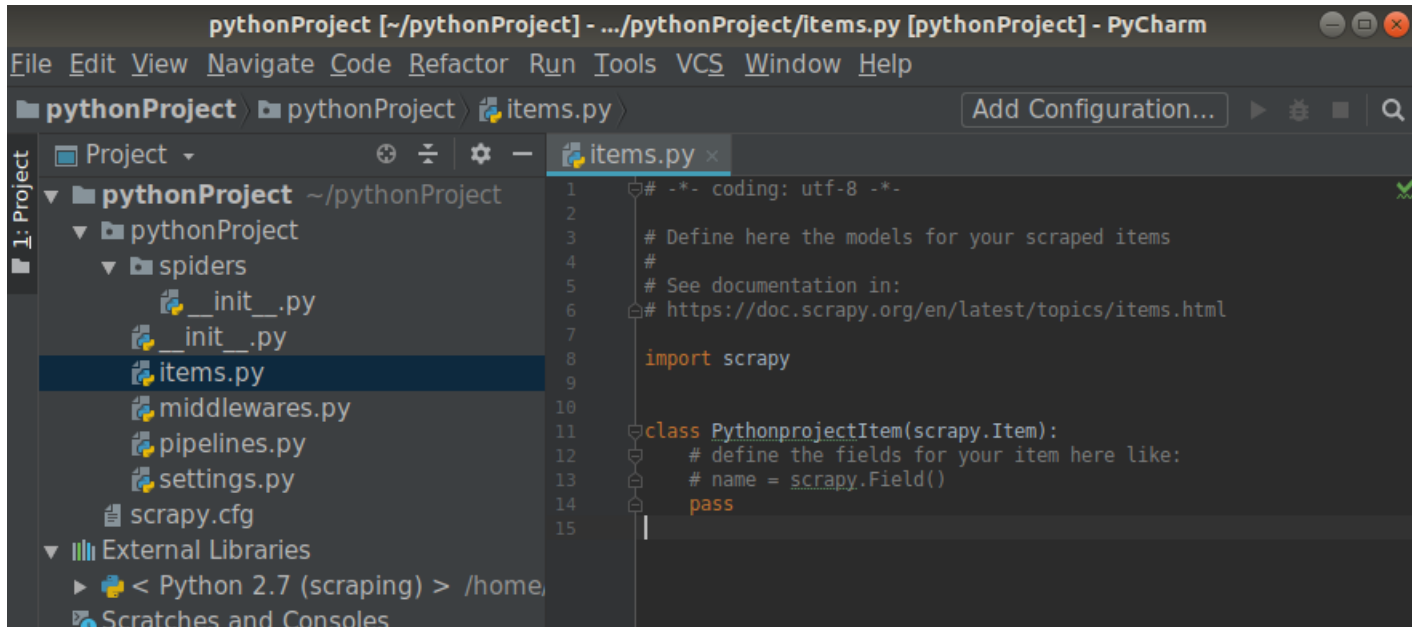
(hyunji) \$deactivate (가상환경 종료)
\$workon [가상환경이름] (가상환경 실행)

-----가상환경에서 패키지 설치!!-----

```
(hyunji) $sudo apt-get install libxml2-dev libxslt-dev python-dev zlib1g-dev
(hyunji) $sudo apt-get install python-lxml
(hyunji) $pip install lxml
(hyunji) $pip install beautifulsoup4
(hyunji) $sudo apt-get install libffi-dev libssl-dev
```

PROCESS FOR PROJECT-2

(hyunji) \$scrapy startproject [pythonProject]



[Items.py]
웹 환경에서 title, link 만 가져오고 싶을 때,
items에 그 두개만 지정)

[Pipelines.py]
Scrapy를 통해 데이터를 들고 와서 데이터를
처리하고자 할 때 사용.

[Settings.py]
Scrapy, Spider에 대한 설정.

[Spiders 폴더]
실제적으로 Scrapy할 내용들을 프로그래밍.

PROCESS FOR PROJECT-3

```
items.py x yes24.py x
1
2  #-*- coding: utf-8 -*-
3
4  # Define here the models for your scraped items
5  #
6  # See documentation in:
7  # https://doc.scrapy.org/en/latest/topics/items.ht
8
9  import scrapy
10
11
12 class bookItem(scrapy.Item):
13     # define the fields for your item here like:
14     # name = scrapy.Field()
15     title = scrapy.Field()
16     link = scrapy.Field()
17
```

[items.py]

- Yes24홈페이지에서 책의 title과 link 가져올 것!

[yes24.py]

- 크롤링할 url 설정.

- 추출할 데이터 xpath 이용하여 각각의 리스트 변수에 대입.

```
Project x pythonProject x items.py x yes24.py x
1  __author__ = 'solteee'
2
3  import scrapy
4  from pythonProject.items import bookItem
5
6  class YesSpider(scrapy.Spider):
7      name = "yes24"
8      allowed_domains = ["www.yes24.com"]
9      start_urls = [
10         "http://www.yes24.com/searchcorner/Search?keywordAd=&keyword=&domain=ALL&qdomain=%C0%FC%C3%BC&wcode=001_005&query=crawling="
11     ]
12
13     def parse(self, response):
14         for sel in response.xpath('//td[re:test(@class, "goods_infogr")] /p[re:test(@class, "goods_name goods_icon")]/a'):
15             item = bookItem()
16             item['title'] = sel.xpath('strong/text()').extract()
17             item['link'] = sel.xpath('@href').extract()
18             yield item
19
20
21
```


PROCESS FOR PROJECT-4

The screenshot shows a web browser window with the URL www.yes24.com/searchcorner/Search?keywordAd=&keyword=&domain=ALL&qdomain=%C0%FC%3BC&Wcode=001_005&que. The search results are for the keyword "crawling".

Four items are highlighted with red boxes and labeled as Item[0] through Item[3]:

- Item[0] TITLE & LINK**: [직수입양서] A Snail Crawling on a Twig Just Above the Water: Blank 150 Page Lined Journal for Your Thoughts, Ideas, and Inspiration (Paperback)
- Item[1] TITLE & LINK**: [직수입양서] A Snail Crawling Across a Leaf with Morning Dew Drops: Blank 150 Page Lined Journal for Your Thoughts, Ideas, and Inspiration (Paperback)
- Item[2] TITLE & LINK**: [직수입양서] The Crawling Buttress (Paperback)
- Item[3] TITLE & LINK**: [직수입양서] Crawling Into the Light: From Tragedy to Triumph and Beyond (Paperback)

Each item includes a price, a discount, and a shipping date. The items are listed in a grid format with a sidebar on the left containing navigation links and a right sidebar with promotional banners.

PROCESS FOR PROJECT-5

YES24 | 대한민국 대표 인터넷 서점

www.yes24.com/searchcorner/Search?keywordAd=&keyword=&domain=ALL&qdomain=%C0%FC%33%BC&Wcode=001_005&que

통합검색: "crawling=" 검색결과 1-20 / 143건

인기도 | 정확도 | 신상품 | 최저가 | 최고가 | 평균

20개 | 옵션선택 | 품질포함

1. [직수입양서] A Snail Crawling on a Twig Just Above the Water: Blank 150 Page Lined Journal for Your Thoughts, Ideas, and Inspiration (Paperback)
Journal, Unique | Createspace Independent Publishing Platform | 2016년 10월
13,660원 → 11,200원(18% 할인) | YES포인트 560원(5%지급)
출고 예상일: 7 일 이내

2. [직수입양서] A Snail Crawling Across a Leaf with Morning Dew Drops: Blank 150 Page Lined Journal for Your Thoughts, Ideas, and Inspiration (Paperback)
Journal, Unique | Createspace Independent Publishing Platform | 2016년 10월
13,660원 → 11,200원(18% 할인) | YES포인트 560원(5%지급)
출고 예상일: 7 일 이내

HTML 검색

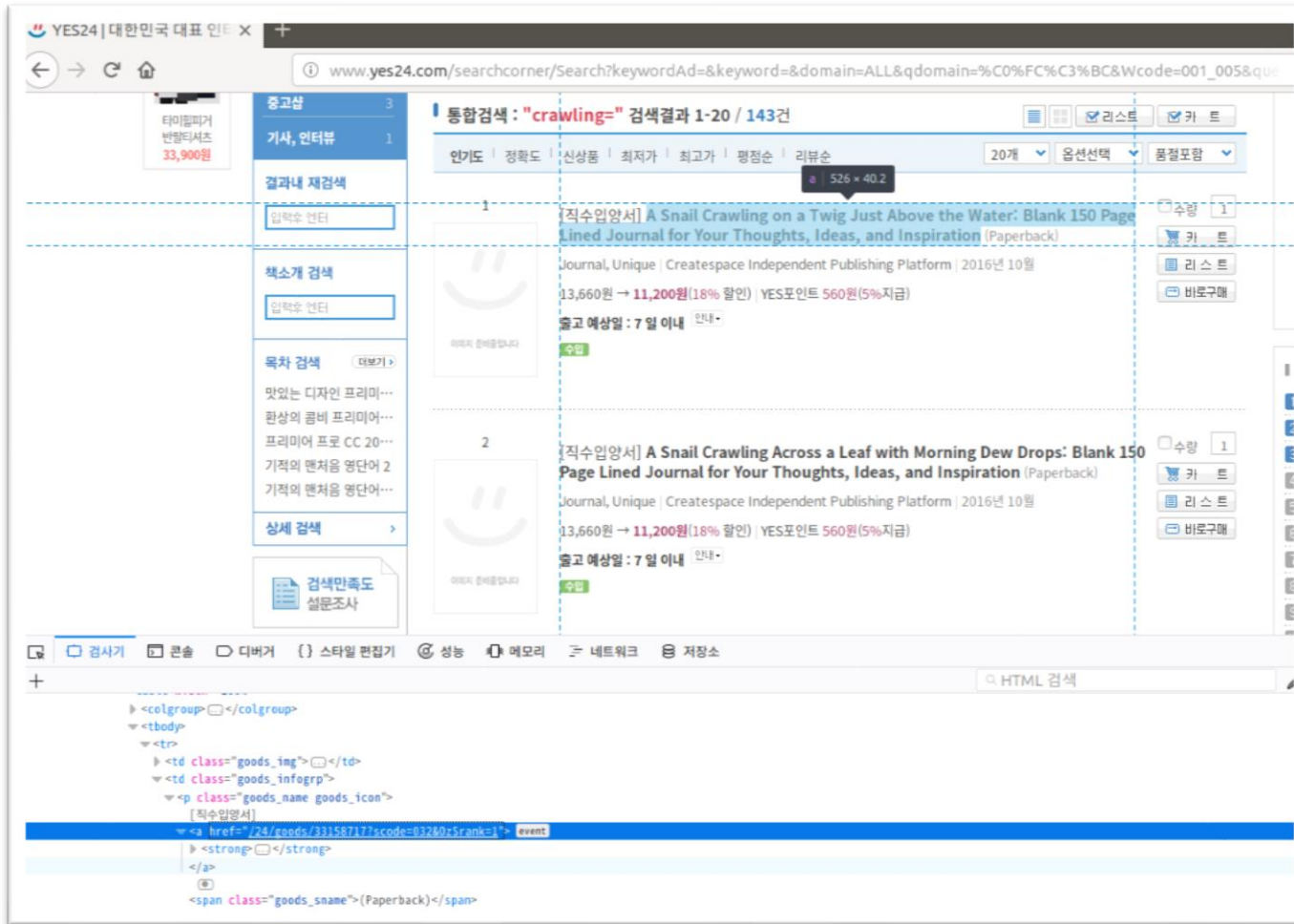
```

<celgroup>
  <tbody>
    <tr>
      <td class="goods_img">
      <td class="goods_infogr">
        <p class="goods_name goods_icen">
          [직수입양서]
          <a href="/24/goods/231582177?code=03240+5rank=1">event
            <strong>
          </a>
          <span class="goods_sname">(Paperback)</span>
        </td>
    </tr>
  </tbody>
</celgroup>

```

<td class="goods_infogrp">
이 코드를 중심으로 item의 index 증가!!

PROCESS FOR PROJECT-6



```
<colgroup></colgroup>
<tbody>
  <tr>
    <td class="goods_img"></td>
    <td class="goods_infogr">
      <p class="goods_name goods_icon">
        [직수입양서]
        <a href="/24/goods/33158717?scode=03260z&rank=1">event
          <strong>
            A Snail Crawling on a Twig Just Above the Water: Blank 150 Page Lined Journal for Your Thoughts, Ideas, and Inspiration
          </strong>
        </a>
      <span class="goods_sname">(Paperback)</span>
    </td>
  </tr>
  <tr>
    <td class="goods_img"></td>
    <td class="goods_infogr">
      <p class="goods_name goods_icon">
        [직수입양서]
        <a href="/24/goods/33158717?scode=03260z&rank=1">event
          <strong>
            A Snail Crawling Across a Leaf with Morning Dew Drops: Blank 150 Page Lined Journal for Your Thoughts, Ideas, and Inspiration
          </strong>
        </a>
      <span class="goods_sname">(Paperback)</span>
    </td>
  </tr>
</tbody>
</table>
```

```
<tbody>
  <tr>
    <td>
      <p>
        <a href= "LINK" />
          <strong> TITLE </strong>
      </p>
    </td>
  </tr>
</tbody>
```

```
def parse(self, response):
    for sel in response.xpath('//td[re:test(@class, "goods_infogr")]/p[re:test(@class, "goods_name goods_icon")]/a'):
        item = bookItem()
        item['title'] = sel.xpath('strong/text()').extract()[0]
        item['link'] = sel.xpath('@href').extract()[0]
        yield item
```

- LINK : <a>태그의 href 속성값
Item['link'] = sel.xpath('@href').extract()

- TITLE : <a>태그안의 태그의 text값
Item['title'] = sel.xpath('strong/text()').extract()

PROCESS FOR PROJECT-7

```
(scraping) solteee@solteee-VirtualBox:~$ cd pythonProject/
(scraping) solteee@solteee-VirtualBox:~/pythonProject$ scrapy crawl yes24
2018-07-31 01:20:42 [scrapy.utils.log] INFO: Scrapy 1.5.1 started (bot: pythonProject)
2018-07-31 01:20:42 [scrapy.utils.log] INFO: Versions: lxml 4.2.3.0, libxml2 2.9.8,
lt, Apr 15 2018, 21:51:34) - [GCC 7.3.0], pyOpenSSL 18.0.0 (OpenSSL 1.1.0h 27 Mar 2018)
2018-07-31 01:20:42 [scrapy.crawler] INFO: Overridden settings: {'NEWSPIDER_MODULE': 'pythonProject.spiders', 'BOT_NAME': 'pythonProject'}
2018-07-31 01:20:42 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.memusage.MemoryUsage',
'scrapy.extensions.logstats.LogStats',
'scrapy.extensions.telnet.TelnetConsole',
'scrapy.extensions.corestats.CoreStats']
2018-07-31 01:20:42 [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']
2018-07-31 01:20:42 [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']
2018-07-31 01:20:42 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2018-07-31 01:20:42 [scrapy.core.engine] INFO: Spider opened
2018-07-31 01:20:42 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min)
2018-07-31 01:20:42 [scrapy.extensions.telnet] DEBUG: Telnet console listening on 127.0.0.1
2018-07-31 01:20:42 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://www.yes24.com/search/1_005&query=crawling=> (referer: None)
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/33158717?scode=032&0zSrank=1'],
'title': ['u'A Snail Crawling on a Twig Just Above the Water: Blank 150 Page Lined Journal for 2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/33158716?scode=032&0zSrank=2'],
'title': ['u'A Snail Crawling Across a Leaf with Morning Dew Drops: Blank 150 Page Lined Journal for 2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/30717822?scode=032&0zSrank=3'],
'title': ['u'The Crawling Buttress']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/30717047?scode=032&0zSrank=4'],
'title': ['u'Crawling Into the Light: From Tragedy to Triumph and Beyond']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/30402414?scode=032&0zSrank=5'],
'title': ['u'Query Selection in Deep Web Crawling']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/29043147?scode=032&0zSrank=6'],
'title': ['u'Crawling to My Death and Other Poems: A Poetry Anthology']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/27864982?scode=032&0zSrank=7'],
'title': ['u"Crawling Out of Hell: The True Story of a British Sniper's Greatest Battle"']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/27118732?scode=032&0zSrank=8'],
'title': ['u'The Crawling Flesh']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/25977053?scode=032&0zSrank=9'],
'title': ['u'Crawling Darkness']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/25490525?scode=032&0zSrank=10'],
'title': ['u'Crawling in the Dark: A Novel of Suspense and Thriller']]
```

```
{'link': ['u'/24/goods/33158717?scode=032&0zSrank=1'],
'title': ['u'A Snail Crawling on a Twig Just Above the Water: Blank 150 Page Lined Journal for 2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/33158716?scode=032&0zSrank=2'],
'title': ['u'A Snail Crawling Across a Leaf with Morning Dew Drops: Blank 150 Page Lined Journal for 2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/30717822?scode=032&0zSrank=3'],
'title': ['u'The Crawling Buttress']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/30717047?scode=032&0zSrank=4'],
'title': ['u'Crawling Into the Light: From Tragedy to Triumph and Beyond']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/30402414?scode=032&0zSrank=5'],
'title': ['u'Query Selection in Deep Web Crawling']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/29043147?scode=032&0zSrank=6'],
'title': ['u'Crawling to My Death and Other Poems: A Poetry Anthology']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/27864982?scode=032&0zSrank=7'],
'title': ['u"Crawling Out of Hell: The True Story of a British Sniper's Greatest Battle"']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/27118732?scode=032&0zSrank=8'],
'title': ['u'The Crawling Flesh']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/25977053?scode=032&0zSrank=9'],
'title': ['u'Crawling Darkness']]
2018-07-31 01:20:43 [scrapy.core.scraper] DEBUG: Scraped from <200 http://www.yes24.com/search/1_005&query=crawling=>
{'link': ['u'/24/goods/25490525?scode=032&0zSrank=10'],
'title': ['u'Crawling in the Dark: A Novel of Suspense and Thriller']]
```

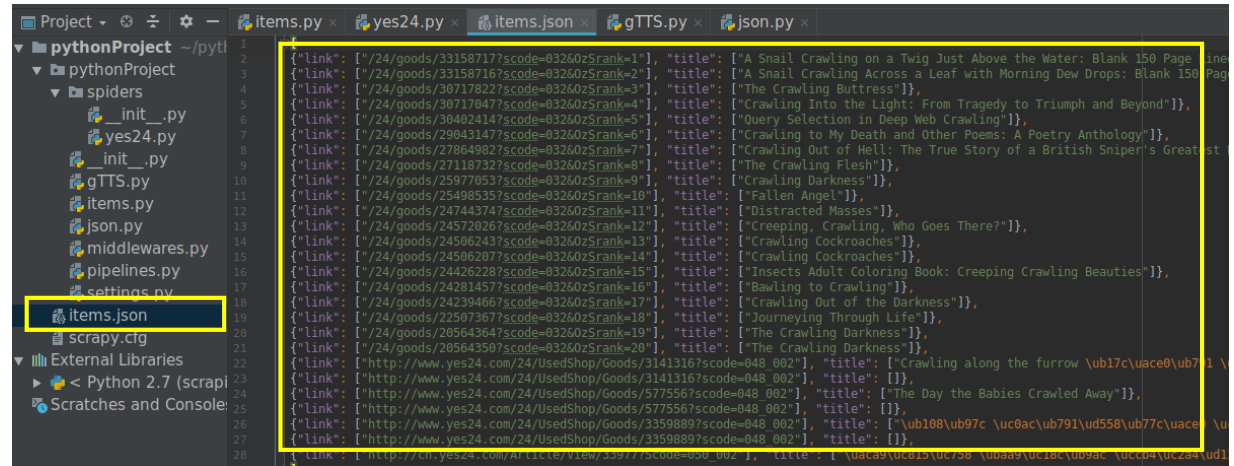
Scrapy가 설치된 가상환경에서 이전에 만든 project 폴더로 이동하여,
\$scrapy crawl [name] 적으면, crawling 과 동시에 scrapy 됨.

▶ cmd창에서 'link'와 'title'이 추출된 것을 볼 수 있다.

PROCESS FOR PROJECT-8

(scraping) \$scrapy crawl [name] -o items.json (Scrapy 한 데이터를 json 파일형식으로 저장!!)

```
(scraping) solteee@solteee-VirtualBox:~/pythonProject$ scrapy crawl yes24 -o items.json
2018-07-31 01:29:46 [scrapy.utils.log] INFO: scrapy 1.5.1 started (bot: pythonProject)
2018-07-31 01:29:46 [scrapy.utils.log] INFO: Versions: lxml 4.2.3.0, libxml2 2.9.8, cssselect 1.0.3, parsel 1.5.0, Twisted 18.0.0, pyOpenSSL 18.0.0 (OpenSSL 1.1.0h 27 Mar 2018), cryptography 2.3.1
2018-07-31 01:29:46 [scrapy.crawler] INFO: Overridden settings: {'NEWSPIDER_MODULE': 'pythonProject.spider', 'BOT_NAME': 'pythonProject', 'ROBOTSTXT_OBEY': True, 'FEED_FORMAT': 'json'}
2018-07-31 01:29:46 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.feedexport.FeedExporter',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.logstats.LogStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.corestats.CoreStats']
2018-07-31 01:29:46 [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']
2018-07-31 01:29:46 [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']
2018-07-31 01:29:46 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2018-07-31 01:29:46 [scrapy.core.engine] INFO: Spider opened
2018-07-31 01:29:46 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (0 items/min)
2018-07-31 01:29:46 [scrapy.extensions.telnet] DEBUG: Telnet console listening on 127.0.0.1:6023
2018-07-31 01:29:46 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://www.yes24.com/robots.txt> (referer: http://www.yes24.com/searchcorner/Search)
2018-07-31 01:29:46 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://www.yes24.com/searchcorner/Search>
```



Items.json 파일 생성!!

Json파일 open 해보면, link와 title이 저장되어 있는 것을 볼 수 있다.

PROCESS FOR PROJECT-9

- Json 파일에 있는 데이터를 python으로 가져온다.

```

1 import json
2 from pprint import pprint
3
4 class json_item():
5     with open('/home/soltee/pythonProject/items.json') as data_file:
6         item = json.load(data_file)
7
8     pprint(item)
9

```

```
{u'link': [u'/24/goods/29043147?scode=032&0zSrank=6'],  
  u'title': [u'Crawling to My Death and Other Poems: A  
{u'link': [u'/24/goods/27864982?scode=032&0zSrank=7'],  
  u'title': [u"Crawling Out of Hell: The True Story of  
{u'link': [u'/24/goods/27118732?scode=032&0zSrank=8'],  
  u'title': [u'The Crawling Flesh']},  
{u'link': [u'/24/goods/25977053?scode=032&0zSrank=9'],  
  u'title': [u'Crawling Darkness']},
```

- gTTS API 이용하여 title 음성출력!!

```
items.py x yes24.py x items.json x gTTS.py x
1 from gtts import gTTS
2 from pythonProject.type import json_item
3
4 items = list(json_item().item)
5 title_item = []
6
7 for i in range(1,11):
8     title_item.append('number' + str(i))
9     title_item.append((items[i])['title'])
10    print(title_item)
11    tts_en = gTTS(str(title_item), lang='en')
12    tts_ko = gTTS(str(title_item), lang='ko')
13
14    tts_en.save('title(en).mp3')
15    tts_ko.save('title(ko).mp3')
```

The screenshot shows a Windows File Explorer window with the address bar displaying the path to a folder named 'pythonProject'. The left sidebar contains navigation icons and labels: '최근' (Recent), '출' (Home), '바탕 화면' (Desktop), '다운로드' (Downloads), '문서' (Documents), and '비디오' (Videos). The main pane shows a grid of files and folders. A red rectangle highlights two MP3 files: 'title(en).mp3' and 'title(ko).mp3'.

Title의 text만 10가지 추출!!

(en)버전과 (ko)버전의 음성파일 생성!

OUTPUT

1. Crawling하여 Scrapy한 데이터

```
[{"link": ["/24/goods/33158717?scode=032&0zSrank=1"], "title": ["A Snail Crawling on a Twig Just Above the Water: Blank 150 Page Lined Jo", "A Snail Crawling Across a Leaf with Morning Dew Drops: Blank 150 Page Li", "The Crawling Buttress"]}, {"link": ["/24/goods/30717822?scode=032&0zSrank=3"], "title": ["Crawling Into the Light: From Tragedy to Triumph and Beyond"]}, {"link": ["/24/goods/30402414?scode=032&0zSrank=5"], "title": ["Query Selection in Deep Web Crawling"]}, {"link": ["/24/goods/29043147?scode=032&0zSrank=6"], "title": ["Crawling to My Death and Other Poems: A Poetry Anthology"]}, {"link": ["/24/goods/27864982?scode=032&0zSrank=7"], "title": ["Crawling Out of Hell: The True Story of a British Sniper's Greatest Batt", "The Crawling Flesh"]}, {"link": ["/24/goods/25977053?scode=032&0zSrank=9"], "title": ["Crawling Darkness"]}, {"link": ["/24/goods/25498535?scode=032&0zSrank=10"], "title": ["Fallen Angel"]}, {"link": ["/24/goods/24744374?scode=032&0zSrank=11"], "title": ["Distracted Masses"]}, {"link": ["/24/goods/24572026?scode=032&0zSrank=12"], "title": ["Creeping, Crawling, Who Goes There?"]}, {"link": ["/24/goods/24506243?scode=032&0zSrank=13"], "title": ["Crawling Cockroaches"]}, {"link": ["/24/goods/24506207?scode=032&0zSrank=14"], "title": ["Crawling Cockroaches"]}, {"link": ["/24/goods/24426228?scode=032&0zSrank=15"], "title": ["Insects Adult Coloring Book: Creeping Crawling Beauties"]}, {"link": ["/24/goods/24281457?scode=032&0zSrank=16"], "title": ["Bawling to Crawling"]}, {"link": ["/24/goods/24239466?scode=032&0zSrank=17"], "title": ["Crawling Out of the Darkness"]}, {"link": ["/24/goods/22507367?scode=032&0zSrank=18"], "title": ["Journeying Through Life"]}, {"link": ["/24/goods/20564364?scode=032&0zSrank=19"], "title": ["The Crawling Darkness"]}, {"link": ["/24/goods/20564350?scode=032&0zSrank=20"], "title": ["The Crawling Darkness"]}, {"link": ["http://www.yes24.com/24/UsedShop/Goods/3141316?scode=048_002"], "title": ["Crawling along the furrow \ub17c\uace0\ub791 \uae3", "http://www.yes24.com/24/UsedShop/Goods/3141316?scode=048_002"], "title": []}, {"link": ["http://www.yes24.com/24/UsedShop/Goods/577556?scode=048_002"], "title": ["The Day the Babies Crawled Away"]}, {"link": ["http://www.yes24.com/24/UsedShop/Goods/577556?scode=048_002"], "title": []}, {"link": ["http://www.yes24.com/24/UsedShop/Goods/3359889?scode=048_002"], "title": ["\ub108\ub97c \uc0ac\ub791\ud558\ub77c\uace0 \uc544", "http://www.yes24.com/24/UsedShop/Goods/3359889?scode=048_002"], "title": []}, {"link": ["http://ch.yes24.com/Article/View/33977?Scode=050_002"], "title": ["\uaca9\uc815\uc758 \ubaa9\uc18c\ub9ac \ucbb4\uc2a4\ud130 \"/>
```

2. 추출된 데이터 mp3파일



Title(en).mp3



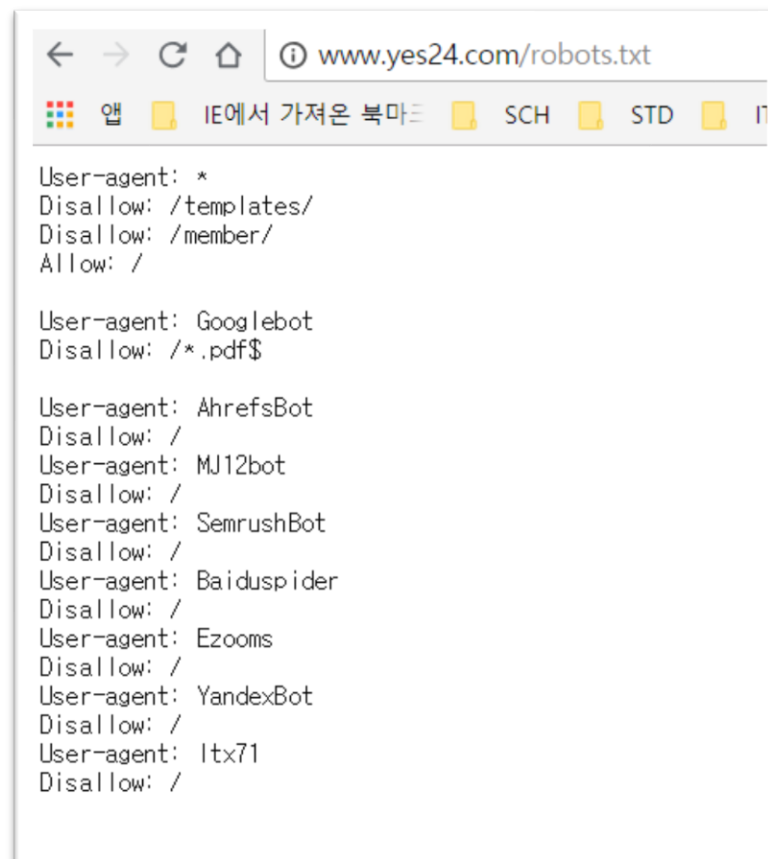
Title(ko).mp3

CRAWLING 가능여부

Main url + /robots.txt

주소 검색하면, User-agent: *
Allow와 Disallow가 나온다.

>> 해당 사이트는 Disallow에 명시된 부분을 제외하고는
크롤링 하는 것을 허용한다.



```
User-agent: *  
Disallow: /templates/  
Disallow: /member/  
Allow: /  
  
User-agent: Googlebot  
Disallow: /*.pdf$  
  
User-agent: AhrefsBot  
Disallow: /  
User-agent: MJ12bot  
Disallow: /  
User-agent: SemrushBot  
Disallow: /  
User-agent: Baiduspider  
Disallow: /  
User-agent: Ezooms  
Disallow: /  
User-agent: YandexBot  
Disallow: /  
User-agent: Itx71  
Disallow: /
```


THANK YOU

