

Naver Café Crawling

CLab.

2014040007 박병조



*Computer
communication Lab.*



Index

◆ Setting

◆ Source Code

◆ Result

◆ Reference





Setting

- ◆ Tool
 - ❖ Pycharm

- ◆ Python version : 3.8

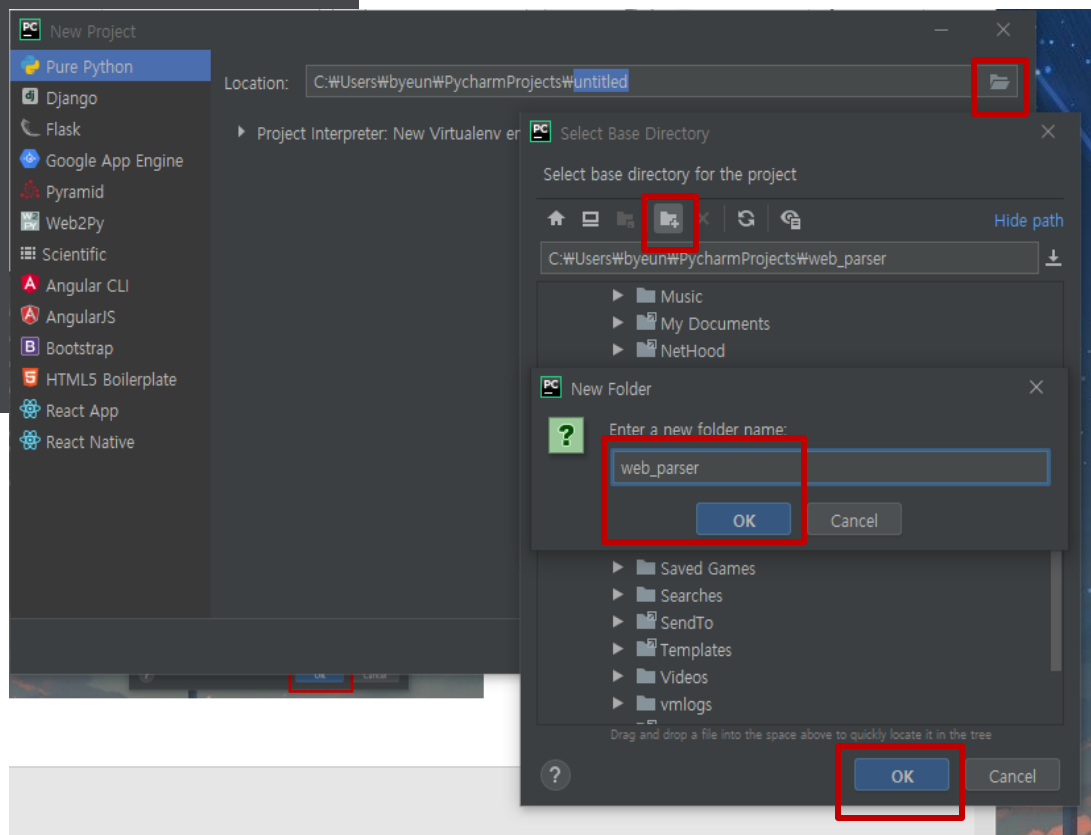
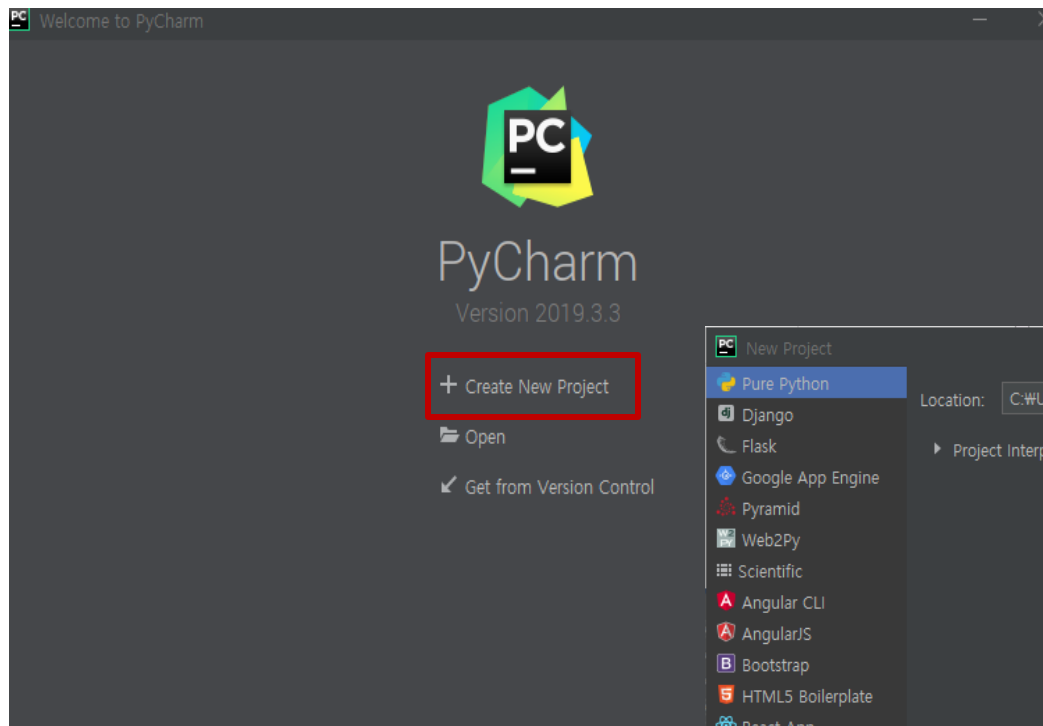
- ◆ Chromedriver
 - ❖ Chrome > Chrome://settings/help > version check
 - ❖ <https://chromedriver.chromium.org/downloads> > now version download

- ◆ Install Packages
 - ❖ Selenium
 - ❖ Django
 - ❖ Bs4





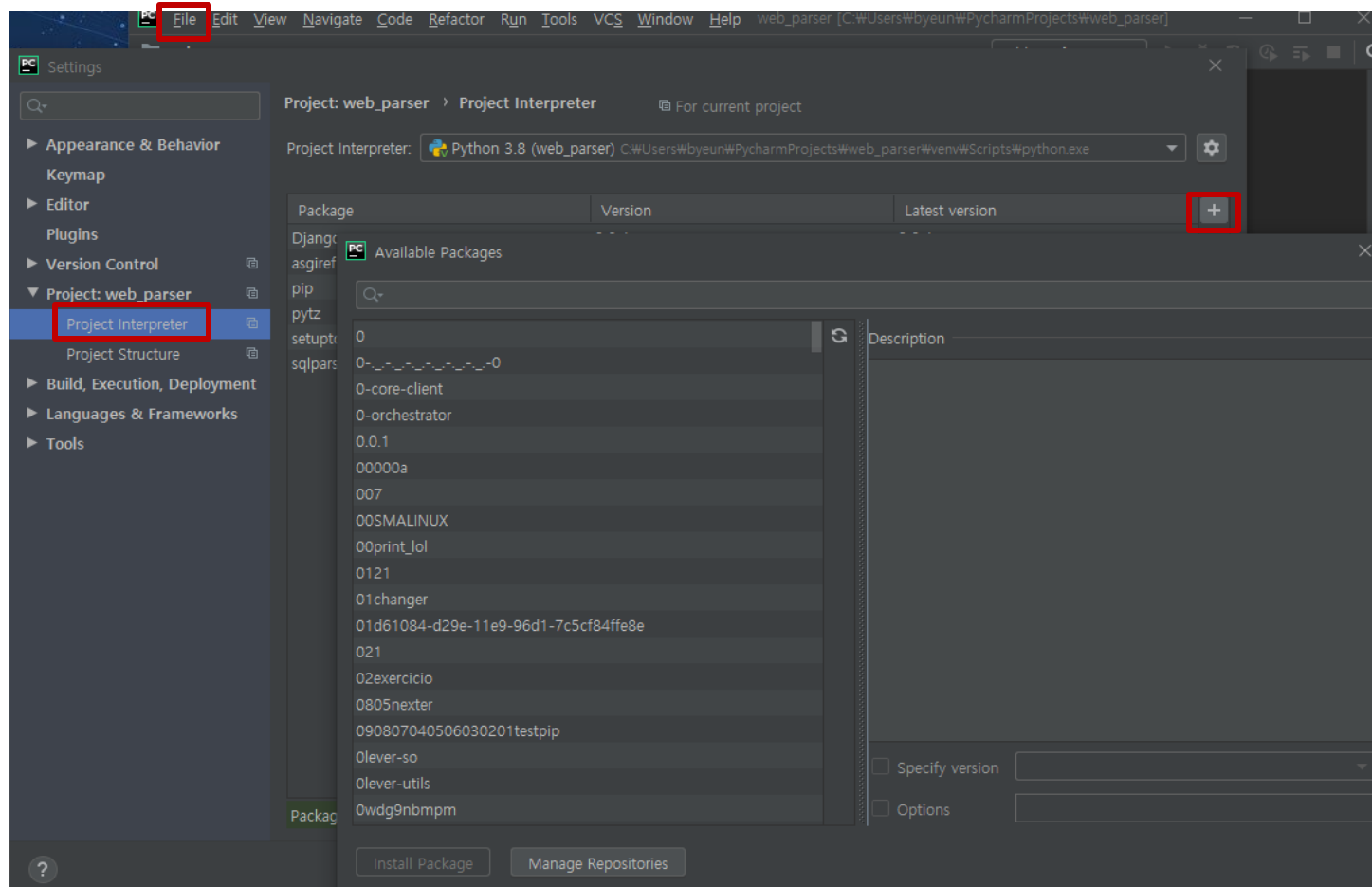
Pycharm Setting





Pycharm Setting

- ◆ File > Settings > Project:name > Project interpreter
- ◆ Install Packages : Django, selenium, bs4





Pycharm Setting

◆ Terminal setting

- ❖ Django-admin startproject name_1
- ❖ cd name_1
- ❖ python manage.py startapp name_2

◆ Setting.py

- ❖ Add name_2 to INSTALLED_APPS

The screenshot shows the PyCharm IDE interface. The left sidebar displays the project structure for 'web_parser', with 'settings.py' selected under the 'web_parser' directory. The main editor window shows the content of 'settings.py'. The 'INSTALLED_APPS' list is visible, and the entry 'parsed_data' is highlighted with a red box. Below the editor, a terminal window is open, showing the execution of Django commands: 'django-admin startproject web_parser', 'cd web_parser', and 'python manage.py startapp parsed_data'. The terminal output indicates that the application was created successfully.

```
ALLOWED_HOSTS = []

# Application definition

INSTALLED_APPS = [
    'django.contrib.admin',
    'django.contrib.auth',
    'django.contrib.contenttypes',
    'django.contrib.sessions',
    'django.contrib.messages',
    'django.contrib.staticfiles',
    'parsed_data',
]

MIDDLEWARE = [
    'django.middleware.security.SecurityMiddleware',
    'django.contrib.sessions.middleware.SessionMiddleware',
    'django.middleware.common.CommonMiddleware',
    'django.middleware.csrf.CsrfViewMiddleware',
    'django.contrib.auth.middleware.AuthenticationMiddleware',
    'django.contrib.messages.middleware.MessageMiddleware',
]
```

```
Terminal: Local x +
Microsoft Windows [Version 10.0.18363.720]
(c) 2019 Microsoft Corporation. All rights reserved.

(venv) C:\Users\byeun\PycharmProjects\web_parser>django-admin startproject web_parser

(venv) C:\Users\byeun\PycharmProjects\web_parser>cd web_parser

(venv) C:\Users\byeun\PycharmProjects\web_parser\web_parser>python manage.py startapp parsed_data

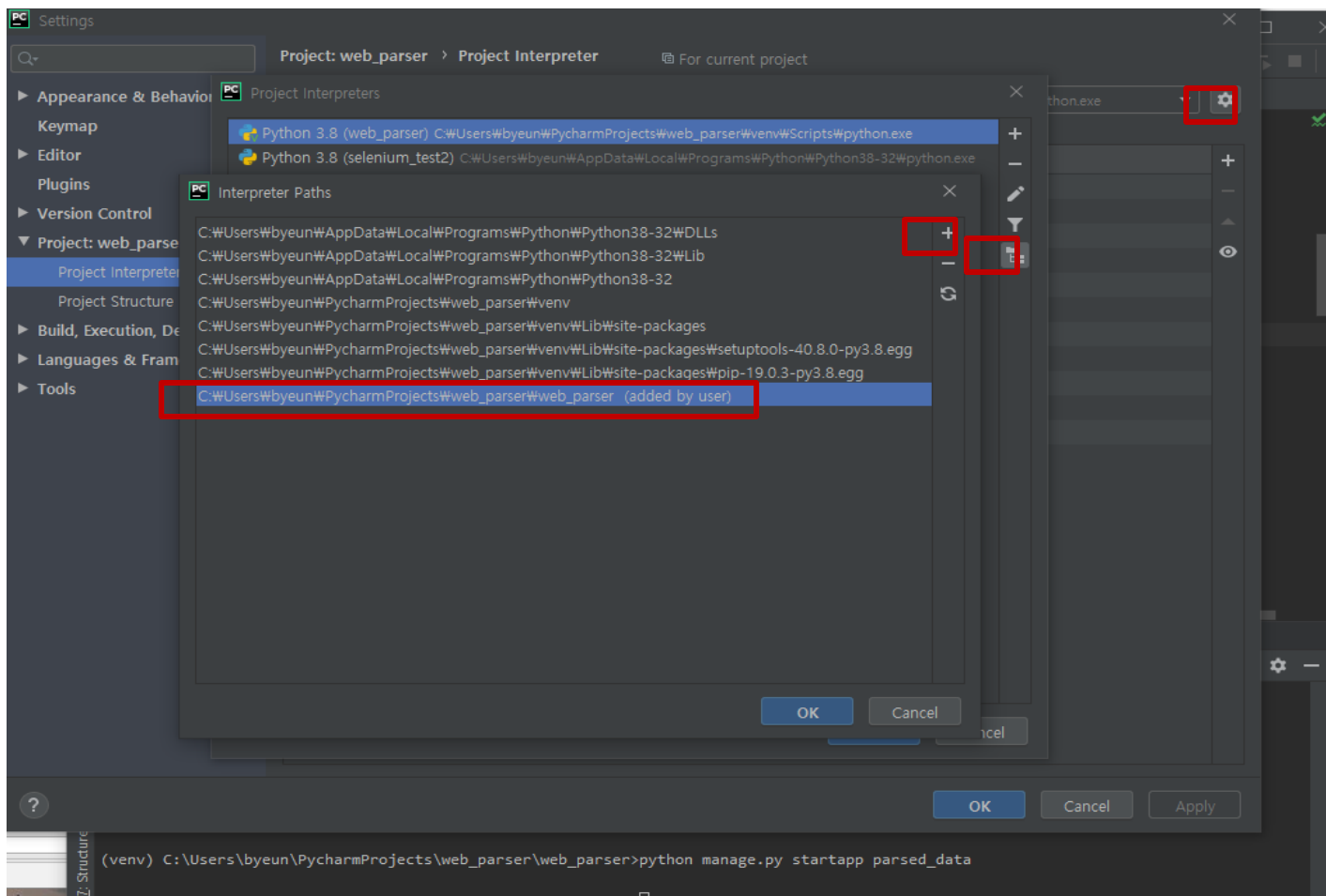
(venv) C:\Users\byeun\PycharmProjects\web_parser\web_parser>
```



Pycharm Setting

◆ Add Path

❖ File > Setting > Project:name > Project Interpreters





Pycharm Setting

◆ DB Setting

- ❖ models.py
- ❖ admin.py

The screenshot shows the PyCharm IDE interface. On the left, the 'Project' view displays the file structure of a project named 'web_parser'. The 'parsed_data' folder is expanded, showing files like 'migrations', 'init.py', 'admin.py', 'apps.py', 'models.py', 'tests.py', and 'views.py'. The 'models.py' file is selected and its content is displayed in the main editor. The code in 'models.py' includes an import from 'django.db import models', a comment '# Create your models here.', a class definition 'class CafeData(models.Model):', and two fields: 'title = models.CharField(max_length=200)' and 'link = models.URLField()'. There is also a 'time = models.TimeField()' line. At the bottom, there is a comment '#django admin page title overloading' and a method 'def __str__(self):' which returns 'self.title'.

The screenshot shows the 'admin.py' file in the PyCharm IDE. The code includes imports for 'from django.contrib import admin' and 'from .models import CafeData'. Below these, there is a comment '# Register your models here.' followed by the line 'admin.site.register(CafeData)'.

❖ DB create : Terminal

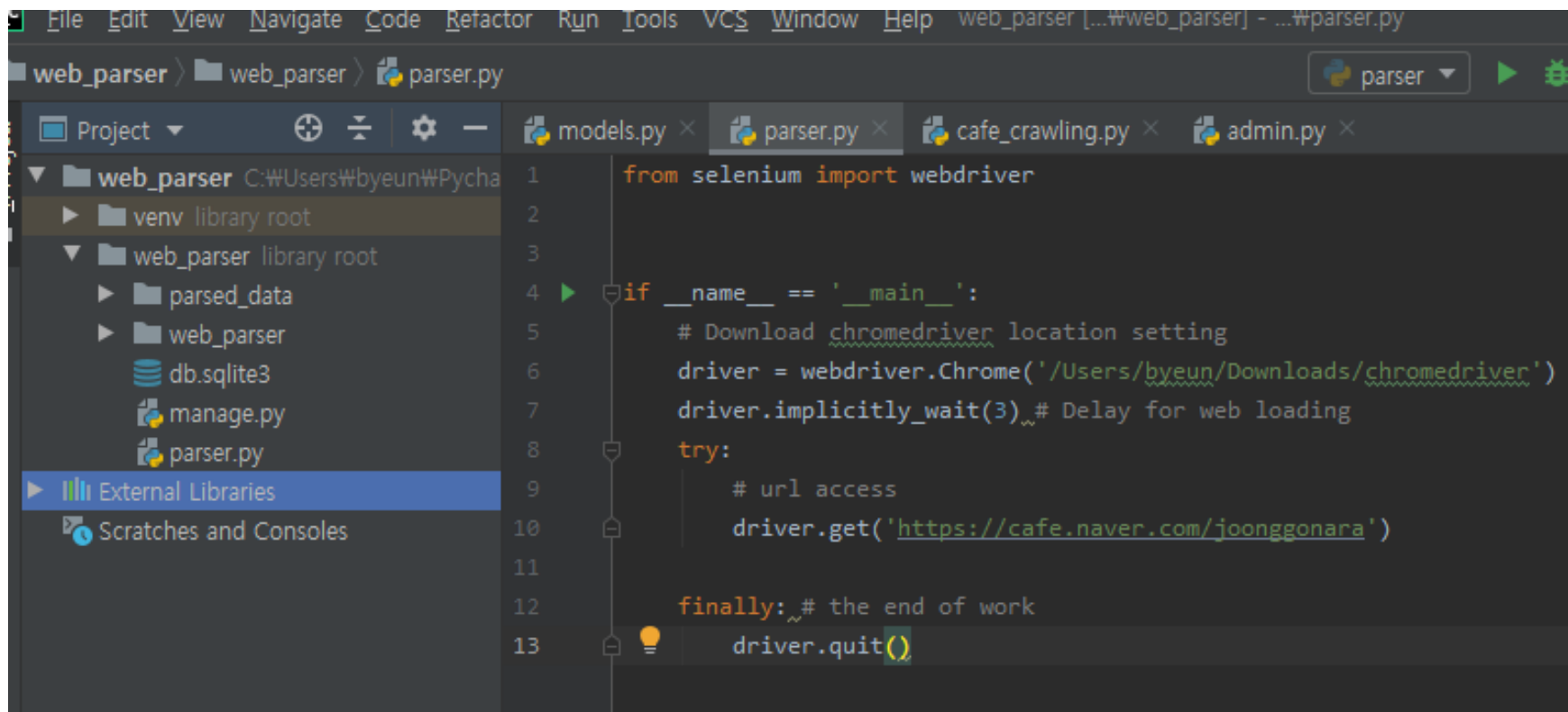
- ◆ python manage.py migrate
- ◆ python manage.py makemigrations parsed_data
- ◆ python manage.py migrate parsed_data

❖ Django Admin

- ◆ python manage.py createsuperuser
- ◆ Don not need to Email



◆ Create parser.py



```
File Edit View Navigate Code Refactor Run Tools VCS Window Help web_parser [...#web_parser] - ...#parser.py
web_parser > web_parser > parser.py
Project
web_parser C:\Users\byeun\Pycha
  venv library root
  web_parser library root
    parsed_data
    web_parser
    db.sqlite3
    manage.py
    parser.py
  External Libraries
  Scratches and Consoles
models.py x parser.py x cafe_crawling.py x admin.py x
1 from selenium import webdriver
2
3
4 if __name__ == '__main__':
5     # Download chromedriver location setting
6     driver = webdriver.Chrome('/Users/byeun/Downloads/chromedriver')
7     driver.implicitly_wait(3) # Delay for web loading
8     try:
9         # url access
10        driver.get('https://cafe.naver.com/joonggonara')
11
12    finally: # the end of work
13        driver.quit()
```



◆ Create modules > café_crawling.py

```
PC File Edit View Navigate Code Refactor Run Tools VCS Window Help web_parser [...#web_parser] - ...#café_crawling.py
web_parser > web_parser > parsing_data > modules > café_crawling.py
Project
web_parser C:\Users\byeun\Pychar
venv library root
web_parser library root
parsing_data
migrations
modules
café_crawling.py
__init__.py
admin.py
apps.py
models.py
tests.py
views.py
web_parser
db.sqlite3
manage.py
parser.py
External Libraries
Scratches and Consoles
1 from selenium import webdriver
2 from bs4 import BeautifulSoup
3 import time
4
5 # crawling data format
6 class café_data:...
11
12 # café-menu search and page change
13 def parse_café(driver, query, data_time):...
33
34 # change pages until find the data_time want
35 def page_café(driver, data, data_time):...
50
51 # page title and link and time crawling
52 def crawling_page(driver, data, data_time):...
66
67 # check until data_time
68 def time_check(café_time, data_time):...
```



◆ café_crawling.py

```
4
5 # crawling data format
6 class cafe_data:
7     def __init__(self, title, link, time):
8         self.title = title
9         self.link = link
10        self.time = time
11
12 # cafe-menu search and page change
13 def parse_cafe(driver, query, data_time):
14     # query is cafe-menu / data_time is the desired time
15     # url access
16     driver.get("https://cafe.naver.com/joonggonara")
17     time.sleep(1)
18     # find cafe-menu and click
19     driver.find_element_by_partial_link_text(query).click()
20     time.sleep(1)
21     driver.switch_to.frame('cafe_main')
22
23     data = []
24
25     t = True
26     while t:
27         # page crawl
28         t = page_cafe(driver, data, data_time)
29         if t != '다음':
30             t = False
31
32     return data
```



◆ café_crawling.py

```

33
34 # change pages until find the data_time want
35 def page_cafe(driver, data, data_time):
36     soup = BeautifulSoup(driver.page_source, 'html.parser')
37     hf = soup.select('div.prev-next > a')
38     count = 15
39     for t in hf:
40         if count < 14:
41             return count
42         elif t.text == '다음':
43             driver.find_element_by_link_text(t.text).click()
44             return t.text
45         elif t.text == '이전':
46             continue
47         else:
48             driver.find_element_by_link_text(t.text).click()
49             count = crawling_page(driver, data, data_time)
50
51 # page title and link and time crawling
52 def crawling_page(driver, data, data_time):
53     soup = BeautifulSoup(driver.page_source, 'html.parser')
54     cafe_title = soup.select('div:nth-child(6) > table > tbody > tr > td > div > div > a.article')
55     cafe_time = soup.select('div:nth-child(6) > table > tbody > tr > td.td_date')
56     count = 0
57     for t in cafe_title:
58         ti = time_check(cafe_time[count], data_time)
59         if ti is True:
60             break
61         else:
62             data.append(cafe_data(t.text.strip(), t.get('href'), cafe_time[count].text))
63             count += 1
64
65     return count
66

```





◆ café_crawling.py

```
web_parser > web_parser > parsed_data > modules > café_crawling.py

Project ▾ | + | - | ⚙ | — | models.py × | parser.py × | café_crawling.py × | admin.py ×

web_parser C:\Users\byeun\Pychar
├── venv library root
├── web_parser library root
│   ├── parsed_data
│   │   ├── migrations
│   │   └── modules
│   │       ├── café_crawling.py
│   │       ├── __init__.py
│   │       ├── admin.py
│   │       ├── apps.py
│   │       ├── models.py
│   │       ├── tests.py
│   │       └── views.py
│   └── web_parser
│       ├── db.sqlite3
│       ├── manage.py
│       └── parser.py
└── External Libraries
    └── Scratches and Consoles

1  from selenium import webdriver
2      from bs4 import BeautifulSoup
3  import time
4
5  # crawling data format
6  class cafe_data:...
11
12  # cafe-menu search and page change
13  def parse_cafe(driver, query, data_time):...
33
34  # change pages until find the data_time want
35  def page_cafe(driver, data, data_time):...
50
51  # page title and link and time crawling
52  def crawling_page(driver, data, data_time):...
66
67  # check until data_time
68  def time_check(cafe_time, data_time):
69      if cafe_time.text.find(':') == -1 or cafe_time.text < data_time:
70          t = True
71      else:
72          t = False
73      return t
```



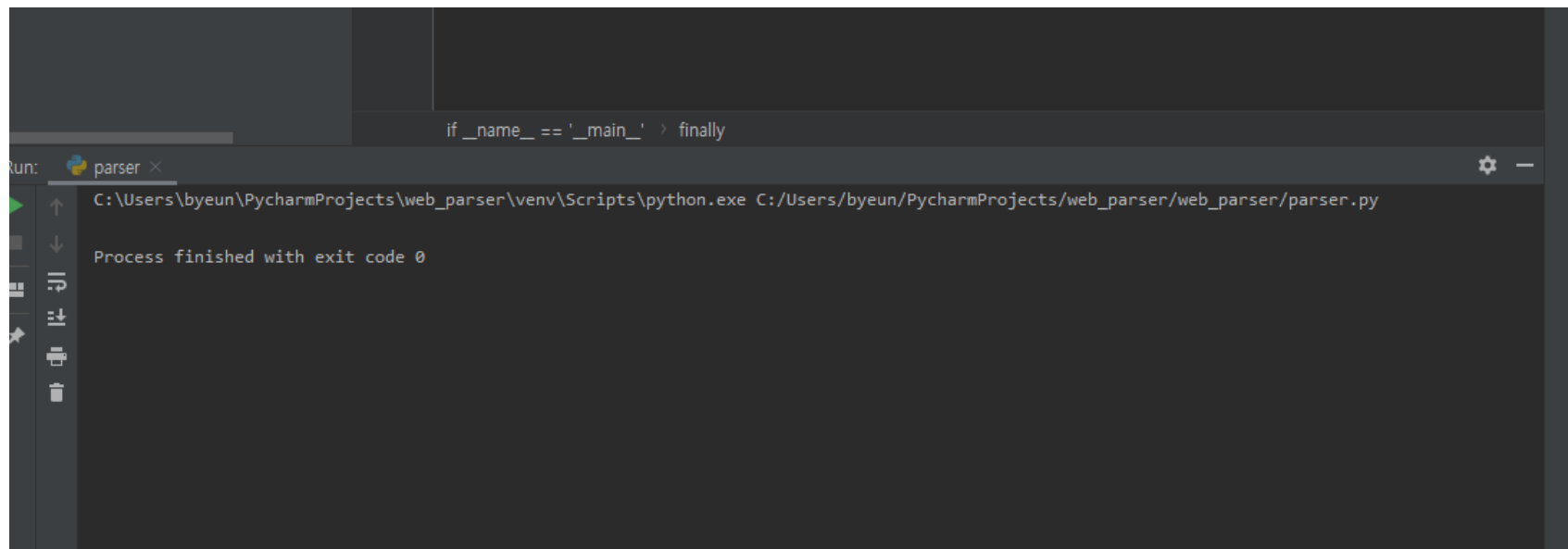
◆ parser.py

```
1 from selenium import webdriver
2 from parsed_data.modules.cafe_crawling import parse_cafe
3 import os
4 import django
5 os.environ.setdefault("DJANGO_SETTINGS_MODULE", "web_parser.settings")
6 django.setup()
7 from parsed_data.models import CafeData
8
9 if __name__ == '__main__':
10     # Download chromedriver location setting
11     driver = webdriver.Chrome('/Users/byeun/Downloads/chromedriver')
12     driver.implicitly_wait(3) # Delay for web loading
13     try:
14         # url access
15         driver.get('https://cafe.naver.com/joonggonara')
16         data = parse_cafe(driver, '컴퓨터', '07:00')
17
18         for t in data:
19             CafeData(title=t.title, link=t.link, time=t.time).save()
20
21     # the end of work
22     finally:
23         driver.quit()
```



Result

◆ Run parser.py



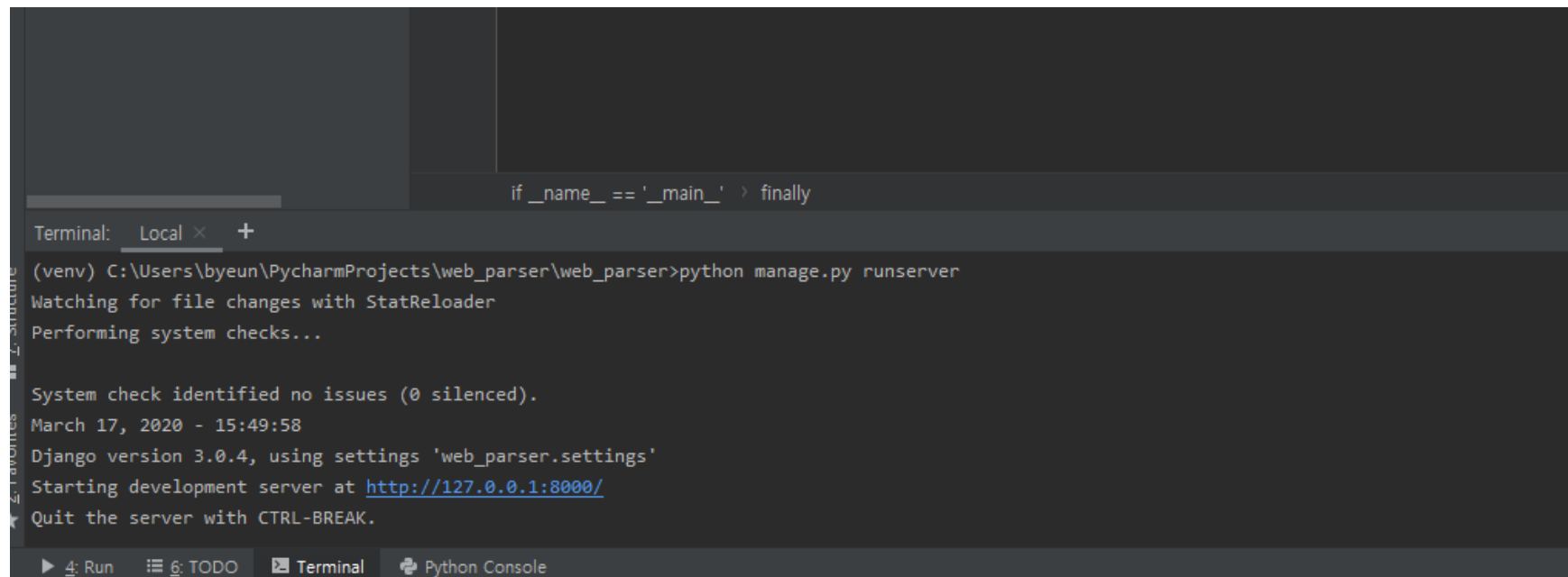
The image shows a PyCharm Run window for a file named 'parser.py'. The top bar indicates the file is open. The main area shows the command executed: `C:\Users\byeun\PycharmProjects\web_parser\venv\Scripts\python.exe C:/Users/byeun/PycharmProjects/web_parser/web_parser/parser.py`. Below the command, it states 'Process finished with exit code 0', indicating a successful execution. The left sidebar contains standard IDE icons for navigation and development.





Result

◆ Terminal > python manage.py runserver



```
if __name__ == '__main__': > finally

Terminal: Local x +
(venv) C:\Users\byeun\PycharmProjects\web_parser\web_parser>python manage.py runserver
Watching for file changes with StatReloader
Performing system checks...

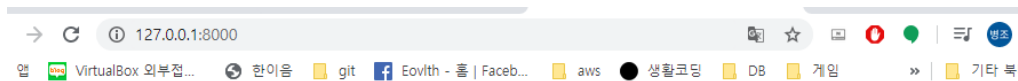
System check identified no issues (0 silenced).
March 17, 2020 - 15:49:58
Django version 3.0.4, using settings 'web_parser.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.
```





Result

◆ http://127.0.0.1:8000/



django

[View release notes for Django 3.0](#)



The install worked successfully! Congratulations!

You are seeing this page because `DEBUG=True` is in your settings file and you have not configured any URLs.





◆ <http://127.0.0.1:8000/admin>



Result

◆ http://127.0.0.1:8000/admin

127.0.0.1:8000/admin/parsed_data/cafedata/

VirtualBox 외부접... 한이음 git Eovith - 홈 | Faceb... aws 생활코딩 DB 게임 기타 북마크

Django administration

WELCOME, ADMIN. [VIEW SITE](#) / [CHANGE PASSWORD](#) / [LOG OUT](#)

Home > Parsed_Data > Cafe datas

Select cafe data to change

ADD CAFE DATA +

Action: Go 0 of 17 selected

☐ CAFE DATA

☐ 컴퓨터 팝니다 (개인) i7 7700k rtx2070

☐ 라온 모니터 RAON 24인치

☐ 서버컴 리니지 VPN 작업장 리니지2M DAVEN

☐ 미니컴 g3260

☐ (개인) 컴퓨터 본체 조립PC 견적★내용 필★

☐ 잠이 왔다 간만에 글 쓴다 이제좀 한가하니 매일 체크해주마~

☐ 이애텍이나 갤럭시 gtx1660 슈퍼 구매합니다

☐ ▶▶▶[윈도우10] [정품인증시디키] ms오피스2016/2019 즉시발송◀◀◀

☐ 선배님들 키보드 추천좀 해주세요!!!!

☐ 일체형 컴퓨터 DM500A2L-K12TV 팔아요. <상태A급><인천 남동구 직거래>

☐ 피식 17 0303 마지막 즐길거고 20 03 03 즐길거다 ㅋㅋㅋ

☐ 잠이 왔다 명 하면 지워진다 니 존재 조차조 ㅎㅎ

☐ 랜선cat7 팝니다

☐ 사기맞기 딱 좋은날씨네

☐ 게이밍 컴퓨터 풀셋트팝니다!!! 사기꾼이에요~~



Computer
Communication Lab.





Refence

- ◆ <https://docs.djangoproject.com/ko/3.0/intro/>
- ◆ <https://beomi.github.io/gb-crawling/>

