크롤링 패턴 코드

> 다음 코드를 우선 그대로 쓰고, 두 부분만 수정!

```
import requests
from bs4 import BeautifulSoup

res = requests.get('http://v.media.daum.net/v/20170615203441266')

soup = BeautifulSoup(res.content, 'html.parser')

mydata = soup.find('title')

print(mydata.get_text())
```

① 크롤링할 페이지 주소 넣기

② 필요한 데이터 추출하는 코드 넣기

추출한 데이터를 변수에 넣은 후, 원하는 프로그래밍

HTML 언어 이해를 기반으로 크롤링해보기

```
from bs4 import BeautifulSoup
html = "<html> \
            <body> \
                <h1 id='title'>[1]크롤링이란?</h1> \
                <p class='cssstyle'>웹페이지에서 필요한 데이터를 추출하는 것</p> \
                <p id='body' align='center'>파이썬을 중심으로 다양한 웹크롤링 기술 발달</p>
            </body> \
        </html>"

soup = BeautifulSoup(html, "html.parser")
# 태그로 검색 방법
data = soup.find('h1')
print(data)
print(data.string)
print(data.get_text())
```

HTML 언어 이해를 기반으로 크롤링해보기

- 데이터 가져오기: 코드

```
print(data)
print(data.string)
print(data.get_text())
```

- 데이터 가져오기: 출력

```
<p class="cssstyle">웹페이지에서 필요한 데이터를 추출하는 것</p>
웹페이지에서 필요한 데이터를 추출하는 것
웹페이지에서 필요한 데이터를 추출하는 것
```

HTML 언어 이해를 기반으로 크롤링해보기

> p 태그 문장이 두 개인데 이 중에 하나를 선택하려면?

1. data = soup.find('p', class_='cssstyle')
2. data = soup.find('p', 'cssstyle')
3. data = soup.find('p', attrs = {'align': 'center'})
4. data = soup.find(id='body')

HTML 언어 이해를 기반으로 크롤링해보기

> p 태그 문장을 모두 가져오려면?

```
from bs4 import BeautifulSoup
html = "<html> \
            <body> \
                <h1 id='title'>[1]크롤링이란?</h1> \
                <p class='cssstyle'>웹페이지에서 필요한 데이터를 추출하는 것</p> \
                <p id='body' align='center'>파이썬을 중심으로 다양한 웹크롤링 기술 발달</p>
            </body> \
        </html>"
```

- find_all() 함수 사용하기

```
paragraph_data = soup.find_all('p')

for paragraph in paragraph_data:
    print(paragraph.get_text())
```

HTML 언어 이해를 기반으로 크롤링해보기

> 속성중에 가장 원하는 데이터들만 선택하기 수월한 속성: 클래스

```
from bs4 import BeautifulSoup
html = "<html> \
            <body> \
                <h1 id='title'>[1]크롤링이란?</h1> \
                <p class='cssstyle'>웹페이지에서 필요한 데이터를 추출하는 것</p> \
                <p id='body' align='center'>파이썬을 중심으로 다양한 웹크롤링 기술 발달</p>
            </body> \
        </html>"
```

> CSS 언어를 이해할 필요가 있습니다.