

Beautiful Soap

Beautiful Soup란 무엇인가?

- Beautiful Soup란 무엇인가?
- "The Fish-Footman began by producing from under his arm a great letter, nearly as large as himself."
- Beautiful Soup은 HTML 및 XML 파일에서 데이터를 가져 오는 Python 라이브러리임.
- Parser tree를 통해서 탐색, 검색, 수정등의 기능을 제공함.
- Python 2.7 또는 Python 3.2에 호환성을 제공함.
- 현재 버전은 Beautiful Soup 4.x임

Beautiful Soup 살펴보기

```
html_doc = """
<html><head><title>The Dormouse's story</title></head>
<body>
<p class="title"><b>The Dormouse's story</b></p>

<p class="story">Once upon a time there were three little sisters; and
their names were
<a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
<a href="http://example.com/lacie" class="sister" id="link2">Lacie</a>
and
<a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;
and they lived at the bottom of a well.</p>
<p class="story">...</p>
"""
```

Beautiful Soup 살펴보기

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc, 'html.parser')
print(soup.prettify())
soup.title
# <title>The Dormouse's story</title>
soup.title.name
# u'title'
soup.title.string
# u'The Dormouse's story'
soup.title.parent.name
# u'head'
soup.p
# <p class="title"><b>The Dormouse's story</b></p>
soup.p['class']
```

Beautiful Soup 살펴보기

```
# u'title'
```

```
soup.a
```

```
# <a class="sister" href="http://example.com/elsie"
id="link1">Elsie</a>
```

```
soup.find_all('a')
```

```
# [<a class="sister" href="http://example.com/elsie"
id="link1">Elsie</a>,
```

```
# <a class="sister" href="http://example.com/lacie"
id="link2">Lacie</a>,
```

```
# <a class="sister" href="http://example.com/tillie"
id="link3">Tillie</a>]
```

```
soup.find(id="link3")
```

```
# <a class="sister" href="http://example.com/tillie"
id="link3">Tillie</a>
```

Beautiful Soup 살펴보기

```
for link in soup.find_all('a'):
    print(link.get('href'))
# http://example.com/elsie
# http://example.com/lacie
# http://example.com/tillie
print(soup.get_text())
# The Dormouse's story
#
# The Dormouse's story
#
# Once upon a time there were three little sisters; and their names were
# Elsie,
# Lacie and
# Tillie;
# and they lived at the bottom of a well.
#
# ...
```

Beautiful Soap설치하기

- \$ apt-get install python-bs4 (for Python 2)
- \$ apt-get install python3-bs4 (for Python 3)
- \$ easy_install beautifulsoup4
- \$ pip install beautifulsoup4
- \$ python setup.py install

- parser 설치하기
 - \$ apt-get install python-lxml
 - \$ easy_install lxml
 - \$ pip install lxml
 - \$ apt-get install python-html5lib
 - \$ easy_install html5lib
 - \$ pip install html5lib

Parser library

Parser	Typical usage	Advantages	Disadvantages
Python's html.parser	BeautifulSoup(markup, "html.parser")	<ul style="list-style-type: none">•Batteries included•Decent speed•Lenient (as of Python 2.7.3 and 3.2.)	<ul style="list-style-type: none">•Not very lenient•(before Python 2.7.3 or 3.2.2)
lxml's HTML parser	BeautifulSoup(markup, "lxml")	<ul style="list-style-type: none">•Very fast•Lenient	<ul style="list-style-type: none">•External C dependency
lxml's XML parser	BeautifulSoup(markup, "lxml-xml") BeautifulSoup(markup, "xml")	<ul style="list-style-type: none">•Very fast•The only currently supported XML parser	<ul style="list-style-type: none">•External C dependency
html5lib	BeautifulSoup(markup, "html5lib")	<ul style="list-style-type: none">•Extremely lenient•Parses pages the same way a web browser does•Creates valid HTML5	<ul style="list-style-type: none">•Very slow•External Python dependency

Soup 생성 및 객체

- 문서를 파싱하려면 BeautifulSoup 생성자에 전달함 . 문자열이나 열려있는 파일 핸들을 전달할 수 있음.

```
From bs4 import BeautifulSoup
```

```
With open( " index.html " ) as fp:
```

```
    soup = BeautifulSoup(fp)
```

```
Soup = BeautifulSoup( " <html>data</html> " )
```

- Beautiful Soup은 HTML 문서를 Python 개체 트리로 변환함.
- Tag, NavigableString, BeautifulSoup, Comment 객체로 구분됨

Soup 객체

- Tag
 - 원본 문서의 XML 또는 HTML 태그에 해당
- Name
 - 모든 태그에는 다음과 같이 액세스 할 수있는 이름이 있음
- Attributes
 - 태그는 여러 속성을 가질 수 있음. 태그의 속성은 " id " 이며 값은 " boldest"임.
 - 사전처럼 태그를 처리하여 태그 속성에 액세스 할 수 있음.
- Multi-valued attributes
 - HTML 4는 여러 값을 가질 수있는 몇 가지 속성을 정의함.
 - HTML 5는 두 가지를 제거하지만 몇 가지를 더 정의함.
 - 가장 일반적인 다중 값 속성은 class태그가 둘 이상의 CSS 클래스를 가질 수 있다는 것임.
 - rel, rev, accept-charset, headers,와 accesskey. BeautifulSoup는 다중 값 속성의 값을 목록으로 표시함.
- Comments 및 다른 special strings