BeautifulSoup

COMS10012 / COMSM0085

Software Tools

BeautifulSoup

Requirements

Python requirements:

sudo apt install python3-bs4 这句命令才是在exercise上用的

```
$ sudo apk add python3 py3-pip
```

BeautifulSoup itself:

```
$ pip install bs4
```

Test using interpreter.

Loading a page

BeautifulSoup will read from a file pointer

```
>>> file = "cattax/index.html"
>>> soup = BeautifulSoup(open(file, 'r'))
```

(Can also pass a string directly if needed).

Printing page text

Common desire in scraping: get the visible text on a page.

```
>>> text = soup.get_text()
>>> print(text)
```

Navigating page elements

Select a page element by tag name:

```
>>> soup.title
```

Navigate the element heirarchy

```
>>> soup.body.main
```

Parent and child (+ ‘sibling’) relationships.

Finding page elements

Avoid navigation, have BeautifulSoup find elements by a specification.

```
>>> soup.find('strong')
```

Find more than the first match:

```
>>> soup.find_all('strong')
```

Can also refine search by 'attrs=' – see documentation!