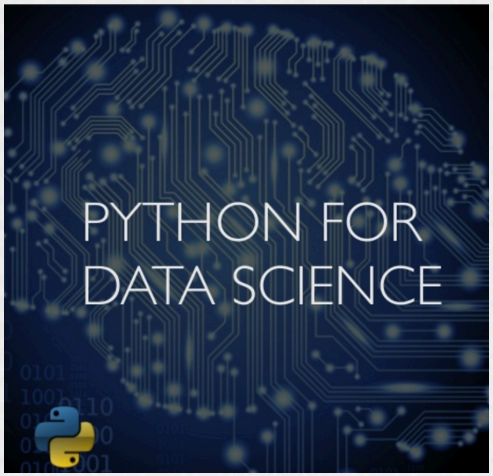
UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS



Python for Data Science

Lecture 16 (05/09, 05/011): Final review

Decision, Operations & Information Technologies
Robert H. Smith School of Business
Spring, 2016

String operations

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- greeting='hello world'
  - print greeting[1]
    - e
  - print greeting[1:5]
    - ello
- String concatenation
- String multiplies integer
- String methods: join(), replace(), split(), upper(), lower(), etc.

Files and modules

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- File open and write
  - Three ways to read a file
    - read()
    - readlines()
    - for line in fh: (suppose: fh=open(filename,'r'))
  - Write string into a file
- import modules
  - import fibo
  - from fibo import fib1,fib2
- Why .pyc (3 advantages)?

Data crawling - BeautifulSoup

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

```
soup = BeautifulSoup('<p class="body strikeout"></p>')
print soup.p['class']
["body", "strikeout"]

soup = BeautifulSoup('<p id="body strikeout"></p>')
print soup.p['id']
body strikeout
```

Data crawling - BeautifulSoup

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

```
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>
"""

from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc)

print soup.head
<head><title>The Dormouse's story</title></head>

print soup.title
<title>The Dormouse's story</title>
```

Data crawling - BeautifulSoup

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

```
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>
"""

from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc)
```

```
print soup.body.b
<b>The Dormouse's story</b>
```

Data crawling - BeautifulSoup

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

```
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>
"""
```

```
print soup.a
<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>
print soup.a.get('href')
 http://example.com/elsie

print 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>]
```

Data crawling - BeautifulSoup

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

```
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>
"""

from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc)
```

```
print soup.body.b.parent
<p class="title"><b>The Dormouse's story</b></p>

print soup.next_sibling.next_sibling
<a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>
```

Data crawling - json


```
[
    {
        "id": 2,
        "name": "An ice sculpture",
        "price": 12.50,
        "tags": ["cold", "ice"],
        "dimensions": {
            "length": 7.0,
            "width": 12.0,
            "height": 9.5
        },
        "warehouseLocation": {
            "latitude": -78.75,
            "longitude": 20.4
        }
    },
    {
        "id": 3,
        "name": "A blue mouse",
        "price": 25.50,
        "dimensions": {
            "length": 3.1,
            "width": 1.0,
            "height": 1.0
        },
        "warehouseLocation": {
            "latitude": 54.4,
            "longitude": -32.7
        }
    }
]
```


json_obj[0][‘name’] is “An ice sculpture”
json_obj[0][‘warehouseLocation’][‘latitude’] is “-78.75”

Regular expression

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Patterns
- Difference between search and match
- group(), start(), end(), span()
- findall()
- split()
- sub()
- Escape characters: \

TF-IDF

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

$$\text{tf}(\text{example}, d_2) = 3$$

$$\text{idf}(\text{example}, D) = \log \frac{2}{1} \approx 0.3010$$

$$\text{tfidf}(\text{example}, d_2) = \text{tf}(\text{example}, d_2) \times \text{idf}(\text{example}, D) = 3 \times 0.3010 \approx 0.9030$$

NLTK

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Word tokenize
  - word_tokenize()
- POS tagging
  - Input parameter: a word list
  - Output: a list of tuples, each tuple has two elements: word itself and the corresponding tag

matplotlib

- subplot