Week 2

Regular expression

Quick guide

^	Matches the beginning of a line
$	Matches the end of the line
.	Matches any character
\s	Matches whitespace
\S	Matches any non-whitespace character
*****	Repeats a character zero or more times
***?**	Repeats a character zero or more times (non-greedy)
+	Repeats a character one or more times
+?	Repeats a character one or more times (non-greedy)
[aeiou]	Matches a single character in the listed set
[^XYZ]	Matches a single character *not* in the listed set
[a-z0-9]	The set of characters can include a range
(	Indicates where string extraction is to start
)	Indicates where string extraction is to end

Quiz: chapter

1. Which of the following best describes "Regular Expressions"?

- (●) A small programming language unto itself
- ( ) A way to solve Algebra formulas for the unknown value
- ( ) A way to calculate mathematical values paying attention to operator precedence
- ( ) The way Python handles and recovers from errors that would otherwise cause a traceback

1. Which of the following regular expressions would extract 'uct.ac.za' from this string using re.findall?

```
From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
```

- ( ) ..@\S+..
- ( ) @\S+
- ( ) F.+:
- (●) @(\S+)

2. Which of the following is the way we match the "start of a line" in a regular expression?

- (●) ^
- ( ) str.startswith()
- ( ) \linestart
- ( ) String.startsWith()
- ( ) variable[0:1]

2. What will the '\$' regular expression match?

- ( ) The beginning of a line
- ( ) The end of a line
- (●) An empty line
- ( ) A dollar sign
- ( ) A new line at the end of a line

\$ = $

3. What would the following mean in a regular expression? [a-z0-9]

- (●) Match a lowercase letter or a digit
- ( ) Match an entire line as long as it is lowercase letters or digits
- ( ) Match any text that is surrounded by square braces
- ( ) Match any number of lowercase letters followed by any number of digits
- ( ) Match anything but a lowercase letter or digit

4. What is the type of the return value of the re.findall() method?

- ( ) A single character
- ( ) An integer
- ( ) A string
- (●) A list of strings
- ( ) A boolean

5. What is the "wild card" character in a regular expression (i.e., the character that matches any character)?

- [ ] $
- [x] .
- [ ] +
- [ ] ?
- [ ] ^
- [ ] *

6. What is the difference between the "+" and "*" character in regular expressions?

- [x] The "+" matches at least one character and the "*" matches zero or more characters
- [ ] The "+" matches upper case characters and the "*" matches lowercase characters
- [ ] The "+" matches the beginning of a line and the "*" matches the end of a line
- [ ] The "+" matches the actual plus character and the "*" matches any character
- [ ] The "+" indicates "start of extraction" and the "*" indicates the "end of extraction"

7. What does the "[0-9]+" match in a regular expression?

- [x] One or more digits
- [ ] Any mathematical expression
- [ ] Several digits followed by a plus sign
- [ ] Zero or more digits
- [ ] Any number of digits at the beginning of a line

8. What does the following Python sequence print out?

```
x = 'From: Using the : character'
y = re.findall('^F.+:', x)
print y
```

- From:
- ^F.+:
- ['From: Using the :'] (selected)
- :
- ['From:']

9. What character do you add to the "+" or "*" to indicate that the match is to be done in a non-greedy manner?

- ++
- \g
- ^
- ? (selected)
- **
- $

10. Given the following line of text:

```
From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
```

What would the regular expression '\S+?@\S+' match?

- From
- d@u
- stephen.marquard@uct.ac.za (selected)
- \@\
- marquard@uct

Week 3

Notes

```
# Using python to access web data

#### Socket
# In computer networking, an internet socket or network socket is an
# endpoint of a bidirectional inter-process communication flow across an
# internet protocol-based computer network, such as the internet.

#### Socket library
import socket
#%%

#### Write a browser

import socket
mysock = socket.socket(socket.AF_INEF, socket.SOCK_STREAM)
mysock.connect(('www.py4inf.com', 80))

mysock.send('GET http://www.py4inf.com/code/romeo.txt HTTP/1.0\n\n')

while True:
    data = mysock.recv(512)
    if (len(data) < 1):
        break
    print (data)
mysock.close()
#%%

#### Make it easier by another library
import urllib

fhand = urllib.request('http://www.baidu.com')

for line in fhand:
    print (line.strip())
#%%
```

Quiz: networks and sockets

1. What do we call it when a browser uses the HTTP protocol to load a file or page from a server and display it in the browser?

- ○ SMTP
- ○ DECNET
- ○ Internet Protocol (IP)
- ◉ The Request/Response Cycle
- ○ IMAP

2. What separates the HTTP headers from the body of the HTTP document?

- ( ) X-End-Header: true
- ( ) Four dashes
- (•) A blank line
- ( ) A less-than sign indicating the start of an HTML tag

2. Which of the following is most similar to a TCP port number?

- (•) A telephone extension
- ( ) A telephone number
- ( ) A street number in an address
- ( ) The distance between two locations
- ( ) The GPS coordinates of a building

3. What must you do in Python before opening a socket?

- ( ) import tcp-socket
- ( ) open socket
- ( ) import tcp
- ( ) _socket = true
- (•) import socket

4. Which of the following TCP sockets is most commonly used for the web protocol (HTTP)?

- (•) 80
- ( ) 119
- ( ) 22
- ( ) 25
- ( ) 23

4. In a client-server application on the web using sockets, which must come up first?

- [x] server
- [ ] client
- [ ] it does not matter

5. Which of the following is most like an open socket in an application?

- [x] An "in-progress" phone conversation
- [ ] Fiber optic cables
- [ ] The wheels on an automobile
- [ ] The chain on a bicycle
- [ ] The ringer on a telephone

6. What does the "H" of HTTP stand for?

- [ ] Hyperspeed
- [ ] Manual
- [ ] Simple
- [ ] wHolsitic
- [x] HyperText

HTTP: Hypertext Transfer Protocol

7. What is an important aspect of an Application Layer protocol like HTTP?

- [x] Which application talks first? The client or server?
- [ ] What is the IP address for a domain like www.dr-chuck.com?
- [ ] How much memory does the server need to serve requests?
- [ ] How long do we wait before packets are retransmitted?

8. What are the three parts of this URL (Uniform Resource Locator)?

```
http://www.dr-chuck.com/page1.htm
```

- ○ Protocol, document, and offset
- ○ Host, offset, and page
- ○ Document, page, and protocol
- ◉ Protocol, host, and document
- ○ Page, offset, and count

9. When you click on an anchor tag in a web page like below, what HTTP request is sent to the server?

```
<p>Please click <a href="page1.htm">here</a>.</p>
```

- ◉ GET
- ○ POST
- ○ PUT
- ○ DELETE
- ○ INFO

10. Which organization publishes Internet Protocol Standards?

- ○ SIFA
- ○ SCORM
- ○ IMS
- ○ LDAP
- ◉ IETF

Week 4

```
import urllib
from BeautifulSoup import *

url = raw_input('Enter - ')

html = urllib.urlopen(url).read()
soup = BeautifulSoup(html)

# Retrieve a list of the anchor tags
# Each tag is like a dictionary of HTML attributes

tags = soup('a')

for tag in tags:
   print tag.get('href', None)
```

1. Which of the following Python data structures is most similar to the value returned in this line of Python:

```
x = urllib.urlopen('http://www.py4inf.com/code/romeo.txt')
```

- ( ) dictionary
- (•) file handle
- ( ) list
- ( ) regular expression
- ( ) socket

2. In this Python code, which line actually reads the data?

```
import socket
mysock = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
mysock.connect(('www.py4inf.com', 80))
mysock.send('GET http://www.py4inf.com/code/romeo.txt HTTP/1.0\n\n')
while True:
    data = mysock.recv(512)
    if ( len(data) < 1 ) :
        break
    print data
mysock.close()
```

- (●) mysock.recv()
- ( ) socket.socket()
- ( ) mysock.close()
- ( ) mysock.connect()
- ( ) mysock.send()

3. Which of the following regular expressions would extract the URL from this line of HTML:

```
<p>Please click <a href="http://www.dr-chuck.com">here</a></p>
```

- (●) href="(.+)"
- ( ) href=".+"
- ( ) http://.*
- ( ) <.*>

4. In this Python code, which line is most like the open() call to read a file:

```
import socket
mysock = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
mysock.connect(('www.py4inf.com', 80))
mysock.send('GET http://www.py4inf.com/code/romeo.txt HTTP/1.0\n\n')
while True:
    data = mysock.recv(512)
    if ( len(data) < 1 ) :
        break
    print data
mysock.close()
```

- (●) mysock.connect()
- ( ) import socket
- ( ) mysock.recv()
- ( ) mysock.send()
- ( ) socket.socket()

5. Which HTTP header tells the browser the kind of document that is being returned?

- ( ) Document-Type:
- ( ) ETag:
- (●) Content-Type:
- ( ) Metadata:
- ( ) HTML-Document:

6. What should you check before scraping a web site?

- (●) That the web site allows scraping
- ( ) That the web site only has links within the same site
- ( ) That the web site returns HTML for all pages
- ( ) That the web site supports the HTTP GET command

7. What is the purpose of the BeautifulSoup Python library?

- ○ It builds word clouds from web pages
- ◉ It repairs and parses HTML to make it easier for a program to understand
- ○ It allows a web site to choose an attractive skin
- ○ It animates web operations to make them more attractive
- ○ It optimizes files that are retrieved many times

8. What ends up in the "x" variable in the following code:

```
html = urllib.urlopen(url).read()
soup = BeautifulSoup(html)
x = soup('a')
```

- ◉ A list of all the anchor tags (<a..) in the HTML from the URL
- ○ True if there were any anchor tags in the HTML from the URL
- ○ All of the externally linked CSS files in the HTML from the URL
- ○ All of the paragraphs of the HTML from the URL