Networked Programs
Chapter 12

Python for Informatics: Exploring Information
www.py4inf.com

Unless otherwise noted, the content of this course material is licensed under a Creative Commons Attribution 3.0 License.
http://creativecommons.org/licenses/by/3.0/.
Copyright 2009- Charles Severance, Jim Eng

Client  Server

Internet

Wikipedia

Internet

HTML  JavaScript  CSS

HTTP  Response  GET

Request

socket

POST

Data Store

memcache

Python

Templates

Data Store

memcache
Network Architecture....

Transport Control Protocol (TCP)

• Built on top of IP (Internet Protocol)
• Assumes IP might lose some data - stores and retransmits data if it seems to be lost
• Handles “flow control” using a transmit window
• Provides a nice reliable pipe


TCP Connections / Sockets

"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."

http://en.wikipedia.org/wiki/Internet_socket

http://www.flickr.com/photos/kitcowan/2103850699/
TCP Port Numbers

- A port is an application-specific or process-specific software communications endpoint
- It allows multiple networked applications to coexist on the same server.
- There is a list of well-known TCP port numbers

http://en.wikipedia.org/wiki/TCP_and_UDP_port

Common TCP Ports

- Telnet (23) - Login
- SSH (22) - Secure Login
- HTTP (80)
- HTTPS (443) - Secure
- SMTP (25) (Mail)
- IMAP (143/220/993) - Mail Retrieval
- POP (109/110) - Mail Retrieval
- DNS (53) - Domain Name
- FTP (21) - File Transfer


Sometimes we see the port number in the URL if the web server is running on a "non-standard" port.
Sockets in Python

- Python has built-in support for TCP Sockets

```python
import socket
mysock = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
mysock.connect( ('www.py4inf.com', 80) )
```

http://docs.python.org/library/socket.html

Application Protocol

- Since TCP (and Python) gives us a reliable socket, what to we want to do with the socket? What problem do we want to solve?
- Application Protocols
  - Mail
  - World Wide Web

HTTP - Hypertext Transport Protocol

- The dominant Application Layer Protocol on the Internet
- Invented for the Web - to Retrieve HTML, Images, Documents etc
- Extended to be data in addition to documents - RSS, Web Services, etc..
- Basic Concept - Make a Connection - Request a document - Retrieve the Document - Close the Connection

http://en.wikipedia.org/wiki/Http
HTTP

- The HyperText Transport Protocol is the set of rules to allow browsers to retrieve web documents from servers over the Internet

What is a Protocol?

- A set of rules that all parties follow so we can predict each other’s behavior
- And not bump into each other
- On two-way roads in USA, drive on the right hand side of the road
- On two-way roads in the UK, drive on the left hand side of the road

Getting Data From The Server

- Each the user clicks on an anchor tag with an href= value to switch to a new page, the browser makes a connection to the web server and issues a “GET” request - to GET the content of the page at the specified URL
- The server returns the HTML document to the Browser which formats and displays the document to the user.
Making an HTTP request

- Connect to the server like www.dr-chuck.com
  - a "hand shake"
- Request a document (or the default document)
  - GET http://www.dr-chuck.com/page1.htm
  - GET http://www.mlive.com/ann-arbor/
  - GET http://www.facebook.com
Let's Write a Web Browser!
Internet Standards

- The standards for all of the Internet protocols (inner workings) are developed by an organization
- Internet Engineering Task Force (IETF)
  - www.ietf.org
- Standards are called “RFCs” - “Request for Comments”


5 Request

A request message from a client to a server includes, within the first line of that message, the method to be applied to the resource, the identifier of the resource, and the protocol version in use.

```
Request = Request-Line ; Section 5.1
   *(general-header ; Section 4.5
      | request-header ; Section 5.3
      | entity-header ) CRLF ; Section 7.1
   CRLF
   [message-body] ; Section 4.3
```

5.1 Request-Line

The Request-Line begins with a method token, followed by the Request-URI and the protocol version, and ending with CRLF. The elements are separated by SP characters. No CR or LF is allowed except in the final CRLF sequence.

```
Request-Line = Method SP Request-URI SP HTTP-Version CRLF
```
Making an HTTP request

- Connect to the server like www.dr-chuck.com
  - a "hand shake"
- Request a document (or the default document)
  - GET http://www.dr-chuck.com/page1.htm
  - GET http://www.mlive.com/ann-arbor/
  - GET http://www.facebook.com

“Hacking” HTTP

$ telnet www.dr-chuck.com 80
Trying 74.208.28.177...
Escape character is '^]'.
GET http://www.dr-chuck.com/page1.htm
<h1>The First Page</h1>
<p>Choose a link below to try out the</p>
</p>

Accurate Hacking in the Movies

- Matrix Reloaded
- Bourne Ultimatum
- Die Hard 4
- ...

http://nmap.org/movies.html
http://www.youtube.com/watch?v=Zy5_gYu_isg

Port 80 is the non-encrypted HTTP port
Hmmm - This looks kind of complex.. Lots of GET commands

```
si-csev-mbp:tex csev$ telnet www.umich.edu 80
Trying 141.211.144.190...
Connected to www.umich.edu.
Escape character is '^]'.
GET /  
```
Firebug reveals the detail...

- If you haven’t already installed the Firebug FireFox extension you need it now
- It can help explore the HTTP request-response cycle
- Some simple-looking pages involve lots of requests:
  - HTML page(s)
  - Image files
  - CSS Style Sheets
  - Javascript files

An HTTP Request in Python

```python
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

while True:
    data = mysock.recv(512)
    if (len(data) < 1):
        break
    print data
mysock.close()
```
Uses urllib in Python

- Since HTTP is so common, we have a library that does all the socket work for us and makes web pages look like a file

```python
import urllib
fhand = urllib.urlopen('http://www.py4inf.com/code/romeo.txt')
for line in fhand:
    print line.strip()
```

Making HTTP Easier With urllib

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

HTTP Header

```
HTTP/1.1 200 OK
Date: Sun, 14 Mar 2010 23:52:41 GMT
Server: Apache
Last-Modified: Tue, 29 Dec 2009 01:31:22 GMT
ETag: "143c1b33-a7-4b395bea"
Accept-Ranges: bytes
Content-Length: 167
Connection: close
Content-Type: text/plain
```

HTTP Body

```
But soft what light through yonder window breaks
It is the east and Juliet is the sun
Arise fair sun and kill the envious moon
Who is already sick and pale with grief
```

Import urllib

```
import urllib
fhand = urllib.urlopen('http://www.py4inf.com/code/romeo.txt')
for line in fhand:
    print line.strip()
```

http://docs.python.org/library/urllib.html

urllib1.py

http://docs.python.org/library/urllib.html

urllib1.py
Like a file...

```python
import urllib
fhand = urllib.urlopen('http://www.py4inf.com/code/romeo.txt')
counts = dict()
for line in fhand:
    words = line.split()
    for word in words:
        counts[word] = counts.get(word, 0) + 1
print counts
```

urlwords.py

---

Reading Web Pages

```python
import urllib
fhand = urllib.urlopen('http://www.dr-chuck.com/page1.htm')
for line in fhand:
    print line.strip()
```

urllib1.py

---

Going from one page to another...

```python
import urllib
fhand = urllib.urlopen('http://www.dr-chuck.com/page1.htm')
for line in fhand:
    print line.strip()
```

---

Google

```python
import urllib
fhand = urllib.urlopen('http://www.dr-chuck.com/page1.htm')
for line in fhand:
    print line.strip()
```
Parsing HTML
(a.k.a Web Scraping)

What is Web Scraping?

- When a program or script pretends to be a browser and retrieves web pages, looks at those web pages, extracts information and then looks at more web pages.
- Search engines scrape web pages - we call this “spidering the web” or “web crawling”

http://en.wikipedia.org/wiki/Web_scraping
http://en.wikipedia.org/wiki/Web_crawler

Why Scrape?

- Pull data - particularly social data - who links to who?
- Get your own data back out of some system that has no “export capability”
- Monitor a site for new information
- Spider the web to make a database for a search engine
Scraping Web Pages

- There is some controversy about web page scraping and some sites are a bit snippy about it.
- Google: facebook scraping block
- Republishing copyrighted information is not allowed
- Violating terms of service is not allowed

The Easy Way - Beautiful Soup

- You could do string searches the hard way
- Or use the free software called BeautifulSoup from www.crummy.com

http://www.crummy.com/software/BeautifulSoup/

Place the BeautifulSoup.py file in the same folder as your Python code...

```python
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)
```

urllinks.py
Summary

- The TCP/IP gives us pipes / sockets between applications
- We designed application protocols to make use of these pipes
- HyperText Transport Protocol (HTTP) is a simple yet powerful protocol
- Python has good support for sockets, HTTP, and HTML parsing