Introduction to Programming for Scientists

Lecture 8
HTML, XML and the Web

Prof. Steven Ludtke
N410, sludtke@bcm.edu

Homework Review

HTML

- Declarative language
- HTML is a type of XML, XHTML obeys XML rules more completely
- Python HTMLParser module
- 'commands' in HTML are denoted by
  <command option=value option=value>text</command>
- For example:

  <HTML>

  <HEAD><TITLE>My Page</TITLE></HEAD>

  <BODY>

  <H3>Hi Everyone</H3>

  <P>This is really just some test text to demonstrate how HTML works.
  I can do interesting things like italicize or make text bold,
  or even both together. ta da

  </BODY>

urllib2

- import urllib2
- f=urllib2.urlopen("http://blake.bcm.edu/dl/test.html)
- for i in f: print i

CSS

- Cascading Style Sheet
  - Used to present websites in a uniform way
  - <link rel="stylesheet" type="text/css" charset="utf-8" media="all" href="/moin_static185/modern/css/common.css">
  -
  -

XML Basics

- <?xml version="1.0" encoding="UTF-8" ?>
- Tags
  - <tag> content </tag> or <tag />
- Attributes
  - <tag attr1="value" attr2="value2"> </tag>
- Nesting
  - <tag 1>content <tag2>nested</tag2></tag1>
- Tags/Attributes case insensitive, content and values not.

XML Example

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<CATALOG>
    <PLANT>
        <COMMON>Bloodroot</COMMON>
        <BOTANICAL>Sanguinaria canadensis</BOTANICAL>
        <ZONE>4</ZONE>
        <LIGHT>Mostly Shady</LIGHT>
        <PRICE CURRENCY="dollar">2.44</PRICE>
        <AVAILABILITY>031599</AVAILABILITY>
    </PLANT>
    <PLANT>
        <COMMON>Columbine</COMMON>
        <BOTANICAL>Aquilegia canadensis</BOTANICAL>
        <ZONE>3</ZONE>
        <LIGHT>Mostly Shady</LIGHT>
        <PRICE CURRENCY="dollar" >9.37</PRICE>
        <AVAILABILITY>030699</AVAILABILITY>
    </PLANT>
</CATALOG>
```

XML in Python

- xml.dom
  - Document Object Model (W3C)
  - View XML as a single hierarchical document
- xml.sax
  - Simple API for XML (W3C)
  - Parse XML files sequentially, callbacks
- xml.etree
  - Python specific, similar to DOM
  - Easier to use !

Using ElementTree

- import xml.etree.cElementTree (or xml.etree.ElementTree)
- et=xml.etree.cElementTree.parse("xml_example.xml")
- e=et.getroot()
- e.getitems()
- e.getchildren()

XML Schemas

- Schemas Specifications
  - DTD
  - XML Schema
  - RELAX NG
- Specific Schemas/Ontologies
  - http://www.bioontology.org
  - http://en.wikipedia.org/wiki/List_of_XML_markup_languages

Real World Example

- www.pdb.org
  - 3 formats, which to use ?
  - XML always the best ?

Python Webservers

- Simple server built-in (SimpleHTTPServer, CGIHTTPServer)
- Zope (Full OODB with online management)
- Twisted (Full internet server framework)
- Many more...

Python Webserver

```
# This will serve files from the current directory and below
# we use port 8080 because port 80 is restricted on most platforms


print "http://%s:8080"%socket.gethostbyname(socket.gethostname())
from BaseHTTPServer import *
from SimpleHTTPServer import *
httpd=HTTPServer(("",8080),SimpleHTTPRequestHandler)
httpd.serve_forever()
```

Homework 8

- Write a program that retrieves something from the web and processes it in some useful fashion.