Information Infrastructure II

INFO I211 – Spring 2014 – Sections 18530 & 22519

Lecture 5 – 2014.01.29-2014.01.30

Instructor:
Mitja Hmeljak,
http://mypage.iu.edu/~mitja
mitja@indiana.edu

Towards a Distributed Application: Connecting to the Web

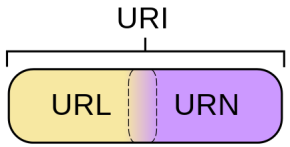
Python can do many things over the network:

- Emails
- FTP, SSH
  - Transferring files
- HTTP
  - Retrieving web pages

We'll be focusing on the HTTP part first!

Connecting to the Web: Uniform Resource Locators

URL – Uniform Resource Locator



Web address (a special case of a U.R.Identifier)

Ex: http://www.python.org

Ex: ftp://ftp.ncbi.nlm.nih.gov/

urllib – *a Python 2.x library*

For opening a connection to a URL & reading contents

Web contents are just like file contents!

Connecting to the Web from Python

Try this code – open pico *on* silo.soic.indiana.edu, *in your* ~/cgi-pub/ *directory –*

```
import urllib

web_page = urllib.urlopen("http://www.google.com/")

# it supports all of the read methods for files:
lines = web_page.readlines()
print lines

web_page.close()   # don't forget: close the connection
```

Connecting to the Web from Python

Let's write it out as a file that we can open in our browser:

```
import urllib
web_page = urllib.urlopen("http://www.google.com/")

# it supports all of the read methods for files:
lines = web_page.readlines()

f = open("page.html", "w")
for line in lines:
    line = line.decode("utf-8")  # renders the lines in a compatible encoding
    f.write(line)

f.close()
print "All done. Open page.html in your browser."
web_page.close()      # don't forget: close the connection
```

File from Web (Group Work)

Write a Python function called *getContent* which takes one argument, *url* and outputs the *content* of the page at *url* into a *file with the same name*. Save the output file in the same directory as the .py program you're running.

Helpful – to obtain a valid *name* for the file to write out, we need to extract the URL's *base name* thus:

```
import os
filename = os.path.basename(url)
```

e.g.

if url is "http://cgi.soic.indiana.edu/~mitja/hello.html"
then the resulting base name is "hello.html"

File from Web (Solution)

```
import urllib, os

def getContent(url):
    web_page = urllib.urlopen(url)
    lines = web_page.readlines()

    filename = os.path.basename(url)

    f = open(filename, "w")

    for line in lines:
        line = line.decode("utf-8")
        f.write(line)

    f.close()

    web_page.close()

getContent("http://cgi.soic.indiana.edu/~mitja/hello.html")
```