Introduction to Python and Webscraping

Robert Vesco

Yale School of Management / ILO

May 30, 2014

Class Objectives

- Programming is (can be) hard

`http://techcrunch.com/2014/05/24/dont-believe-anyone-who-tells-you-learning-to-code-is-easy/`

- Introduce basic python and webscraping
- Provide skills & knowledge not in online tutorials
- Tools that can be used with any programming language
- Provide some guidance for your personal projects

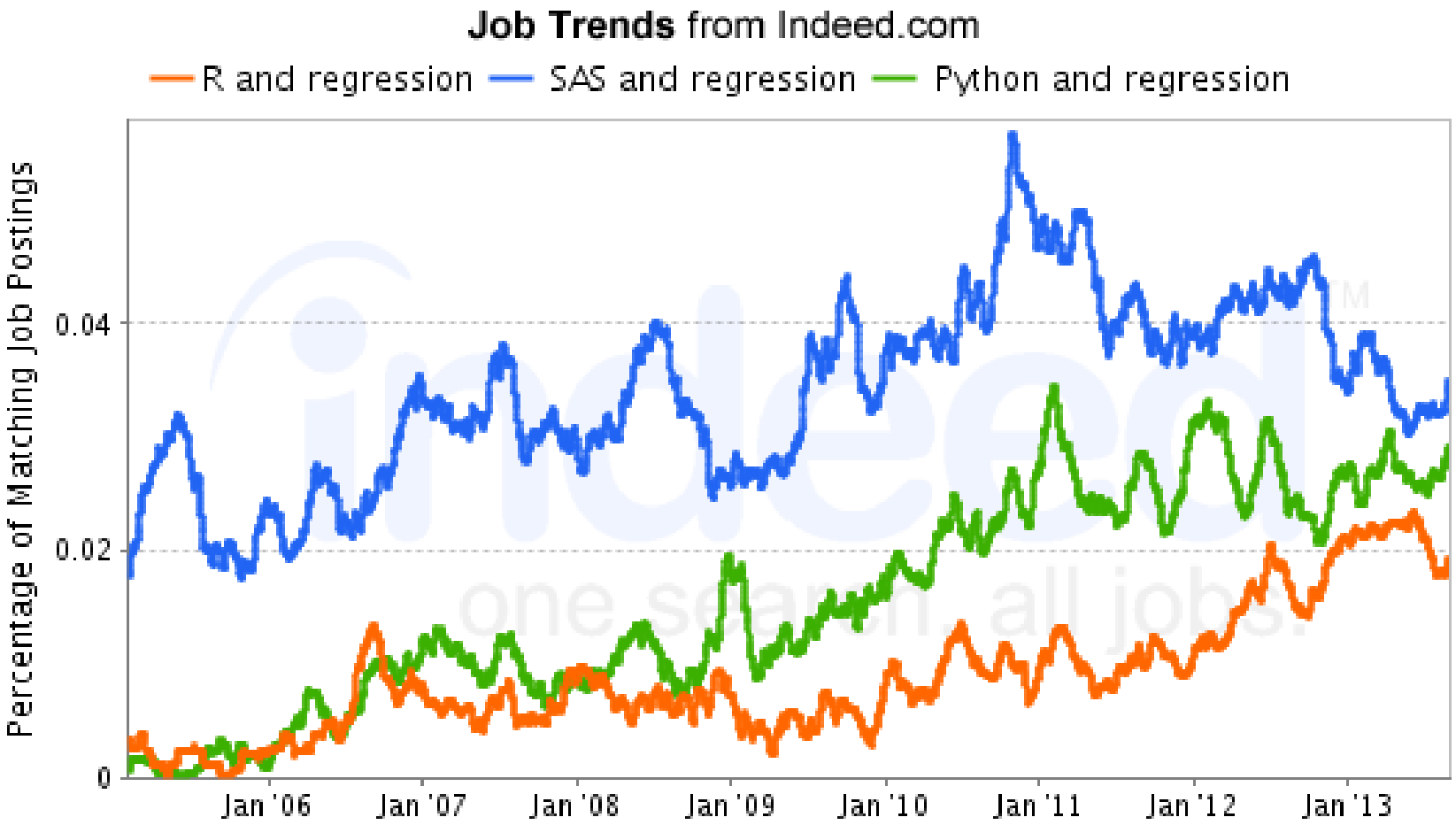
Plan

- Content
  - Why Python?
  - Working From the Command Line
  - Python
  - Webscraping
  - Discuss sites YOU want to scrape
  - Development environments
- Breaks
  - 10:30 (10 min)
  - 12:00 Lunch (30 min)
  - 1:30 (10 min)

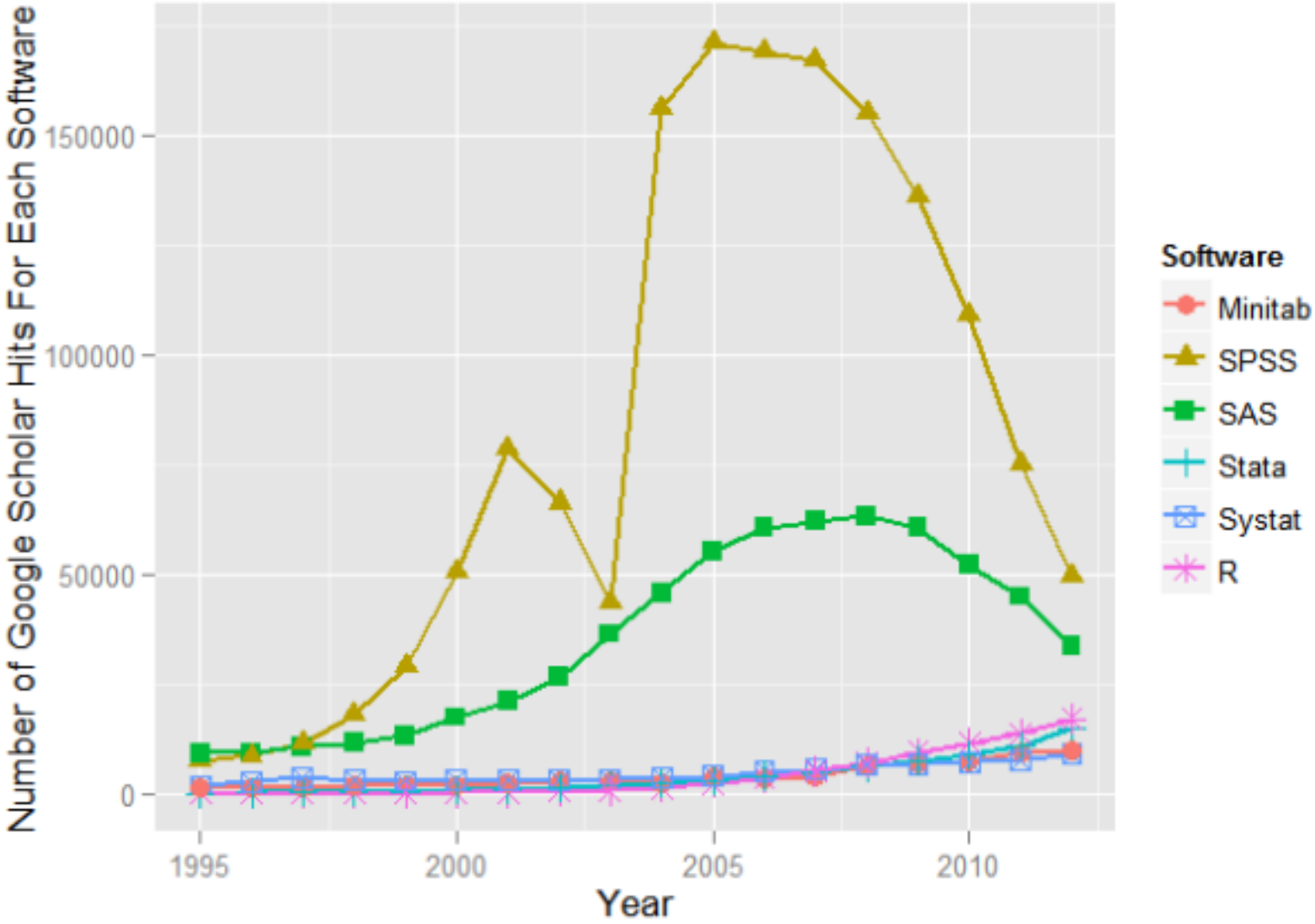
Opinionated History of Programming Languages

- Lisp
- C, C++
- Awk, Sed & shell scripts
- Practical Extraction and Reporting (perl)
- S (R precursor)
- Java
- Ruby
- R
- Haskell
- Clojure (Incanter)

Python and Stats



Python and Journals



Homogenization of Programming

http://www.talyarkoni.org/blog/2013/11/18/the-homogenization-of-scientific-computing-or-why-python-is-st

- TLDR: One tool for many problems

Python Considerations

Support For

- Readability & Consistency (pythonic)
- Fairly fast
- Not Java
- Used in biz ops & domains

Support Against

- Backward compatibility
- Fragile package dependencies
- Fragmentation
- Complementary Assets for Science

The many faces and versions of Python

- Cython (python to c to python)
- IronPython (.net)
- PyPy (JIT)
- Jython (Java)
- Ipython (scientific and interactive)

Version 2 vs 3

Python 3 is killing Python
`https://medium.com/@deliciousrobots/5d2ad703365d/`
Python 3 can revive Python `https://medium.com/p/2a7af4788b10`

Interactive Python (IPYTHON)

- Designed for interactive work & scientists
- Lots of useful features
  - Tab completion
  - object?, object??
  - %run scriptname
  - press up shows last command
  - %who shows all variables
  - !cmd lets you run terminal commands
- Terminal friendly

Why Terminals and Command Line Programs?

- Troubleshooting python programs
- Managing programs and files (very important for webscraping)
- Right tool for some jobs

CD - Change Directory

```
pwd #your current path or %pwd

mkdir test_dir #create directory

ls -laG #Show all files in directory

cd test_dir #folder = directory

cd ../../ #move up two directories

cd - #move back to last directory

cd #move to home directory

cd ~/test_dir #move to folder relative to home directory

touch test_dir/test_file.txt

rmdir test_dir #must be empty, so fails

rm -rf test_dir #-rf = recursive and force -- dangerous
```

Open files in text editor

- Mac

```
1 open -t filename.ext #default editor for extension
2 open -a TextEdit filename.ext #forces textedit
3 #alias textedit='open -a TextEdit' For .bashrc
```

- Windows

```
1 notepad filename.txt
```

- Terminal Viewer (useful for super large files)

```
1 less -SN filename.txt
```

Sudo, Elevated Rights, Admin

- Mac/Linux: sudo cmd file
- Windows: runas /user:admin
- Best to minimize programs running at elevated rights
- Modifying system files usually require this.

File Permissions

```
1 ls -laG #show all files and permissions
```

D = directory
4 = Read (r)
2 = Write (w)
1 = Execute(x)
777 = All rights for User, Group, Everyone <= BAD

- What is rwx-rw-r– in numerical permissions?
- Scripts will often need execution rights

```
1 chmod +x filename
```

Finding programs and scripts

- Depends on operating system

```
1 where programname
2 which programname #will give unix style path on windows
3 whereis programname #not on all programs
```

Simple Scripts

- Scripts should begin with #!PathToYourExecutable

```
#non-standard script
echo "print 'hello world'" > test.py
cat test.py # shows contents
ls -laG #look at the file

python test.py

echo -e "#\!PATHTOYOURPYTHON \n print 'hello world'" > test.py
less test.py #spits it out to terminal to viewer
./test.py
```

- How can we make the second way work?

```
which env
# !#/path/to/env python #absolute path not needed
```

Simple Scripts

- Scripts should begin with #!PathToYourExecutable

```
#non-standard script
echo "print 'hello world'" > test.py
cat test.py # shows contents
ls -laG #look at the file

python test.py

echo -e "#\!PATHTOYOURPYTHON \n print 'hello world'" > test.py
less test.py #spits it out to terminal to viewer
./test.py
```

- How can we make the second way work?

```
which env
# !#/path/to/env python #absolute path not needed
```

Notes:

Wget

- Flexible, fast tool for downloading & spidering

Find

http://www.tecmint.com/
35-practical-examples-of-linux-find-command/

Shells vs Terminals

- Shells are programs (like python) that help you interact computer.
  - csh (c shell, mostly seen on older servers)
  - bash (most common)
  - zsh (most convenient)
- Terminals are wrappers around shells (iterm2 for macs)
- .bashrc, .cshrc, .zshrc are configuration files for shells

Paths

- One of the biggest causes of angst
- Exists at system and user levels
- Order matters; read first > read second

```
#in bash, zsh

#in windows (dos)
path %path%;C:\Python #temp
# see control panel > environment variables for permanent
```

- Macs/linux

```
/etc/paths #admin levels for mac
/etc/environment #admin
~/.bashrc #user level for mac/linux
export PATH="$PATH:/usr/local/bin/python"
PATH=$PATH:/my/new/path #temporary
```

Anaconda and Spydyer

- Anaconda is a pre-packaged python distribution for scientists
- Spyder is an IDE (Integrated Development Environment)
- Open a terminal or click spyder

```
1 anaconda/bin/spyder
```

- Open terminal within spyder

Programming Concepts

- Types (int, strings)
- Data Structures
- Variables
- Flow structures
- Function, Objects and Modules
- Scripting and Programs

Hello World

Version 2 - Print Statement

```
1 print "hello world"
```

Version 3 - Print Function

```
1 print("hello world")
```

hello world

Hello World

Version 2 - Print Statement

```
1 print "hello world"
```

Version 3 - Print Function

```
1 print("hello world")
```

```
hello world
```

Note

- stuff and stuff

Comments in Python

```
1 # This is a single line comment
2 print "stuff" # This is also a comment
3
4 '''
5 Multiline comments
6 Are surround by triple-quoted strings
7 '''
```

Comments in Python

```
# This is a single line comment
print "stuff" # This is also a comment

'''
Multiline comments
Are surround by triple-quoted strings
'''
```

Notes:

- stuff and stuff2

Basic Types

- Numeric: int, float, long, complex
- Sequence: str, unicode, list, tuple, bytearray, buffer, xrange

```
1 var1 = "test strings"
2 var2 = 3
3 type(var1)
4 type(var2)
5 var3 = str(3) # conversion is possible, sometimes
6 type(var3)
```

```
1 <type 'str'>
2 <type 'int'>
3 <type 'str'>
```

Data Structures

- Often considered "types" or "compound types"
- Base python has
  - lists = ['apples',44, 'peaches']
  - tuples = read-only lists = ('apples',44,'peaches')
  - dictionaries = key:value pairs = {'firstname':'tom','lastname':'selleck'}

Lists: Slicing

- lists are flexible. They can be nested, shrunk, combined ...
- Indexed starting with 0
- Limitation: searching for elements when you don't know index #

```
1 ls = [1,"a",2,"b", 1]
2 ls[0]
3 ls[0:2]
4 ls[:]
5 ls[1:]
6 ls[1:4:2] #last element in step. Easy way to get odd
```

```
1 1
2 [1, 'a']
3 [1, 'a', 2, 'b', 1]
4 ['a', 2, 'b', 1]
5 ['a', 'b']
```

Lists: Adding and Removing Elements

```
1 ls # pre
2 ls.append("add to end")
3 ls.insert(1,"after second element")
4 ls.insert(-1, "after second to last")
5 ls.remove('a') # by value, not index
6 ls # post
7 ls.index('b')
8 ls.count(1)
```

```
1 [1, 'a', 2, 'b', 1]
2 >>> >>> >>> >>> [1, 'after second element', 2, 'b', 1, 'after
      second to last', 'add to end']
3 3
4 2
```

Lists: Whole List Operations

```
1 # Concatenate two lists
2 ls.extend(["newlist added to old"])
3 ls.sort()
4 ls
5 ls.reverse()
6 ls
```

```
1 [1, 1, 2, 'add to end', 'after second element', 'after second to
      last', 'b', 'newlist added to old']
2 ['newlist added to old', 'b', 'after second to last', 'after
      second element', 'add to end', 2, 1, 1]
```

Lists: List Comprehensions

- Functions on list elements, like loops
- Not recommended for complex scenarios

```
ls2 = [str(x) for x in ls]
ls2
## nested loop, += concat for strings
[[x+y for x in ls2] for y in ls2]
```

```
['1', 'a', '2', 'b', '1']
[['11', 'a1', '21', 'b1', '11'], ['1a', 'aa', '2a', 'ba', '1a'],
    ['12', 'a2', '22', 'b2', '12'], ['1b', 'ab', '2b', 'bb',
    '1b'], ['11', 'a1', '21', 'b1', '11']]
```

Sets

- Set are like lists, but must contain unique data and can't be nested
- Allows operations such a union and intersections

```
1 ls_dupes = [1,2,3,4,4,3]
2 st = set(ls_dupes)
3 print st
4 st2 = {1,2,3,5}
5 print st | st2 # union
6 print st & st2 # intersection
7 lss = list(st & st2) # convert back
```

```
1 >>> set([1, 2, 3, 4])
2 >>> set([1, 2, 3, 4, 5])
3 set([1, 2, 3])
4 >>> <type 'list'>
```

Tuples

- Tuples are like lists, but they are immutable
- Memory efficient because python knows how much memory to allocate

```
1 tp = () # empty tuple
2 tp1 = (1,) #tuple with one element (comma required)
3 tp2 = (1,2,3)
4 tp
5 tp1
6 tp2
7 tp2[2] #slicing uses [] not ()
```

```
1 ()
2 (1,)
3 (1, 2, 3)
4 3
```

Dictionaries

- Represented by key:value pairs. Know as hashes, maps, associative collections
- Key can be numbers or strings, but must be unique.
- Value can be mutable or not, can be combined with tuples
- Useful when you need a fast lookup based on custom key.

```
1 dct = {'first':1, 'second':2, 'third':3}
2 dct['second']
3 del(dct['third'])
4 dct.keys()
5 dct.values()
```

```
1 2
2 ['second', 'first']
3 [2, 1]
```

Operators

Control structures

Strings

Strings vs Numbers

```
1 string = "123456"
2 number = 123456
3 string is number
4 int(string) is number # different "objects"
5 int(string)==number # testing equality of value
```

```
1 False
2 False
3 True
```

Strings vs lists of strings

```
1 a = [string]
2 b = [string]
3 a == b # compares equality
4 a is b # compares whether objects
```

```
1 >>> T
```

Objects, Methods and Functions

- Methods are function that operate on objects
- Object: dog Method: eat
- Functions

http://stackoverflow.com/questions/8108688/in-python-when-should-i-use-a-function-instead-of-a-method

```
1 var1.capitalize() # method on object
2 len(var1) # also method, but functional looking
```

```
1 'Test strings'
2 12
```

Modules

Dates

Functions

- parameter order matters, unless name=paramater
- anonymous functions use lambda keyword
- return statements without value return nothing
- Variables within function have local scope

```
1 def printnum( x, y ):
2     """This passes a parameter to the print statement"""
3     print x, y
4     return
5
6 printnum(y=3, x="printing this:")
7 printnum("positional ordering matter if not named", 4)
```

```
1 printing this: 3
2 positional ordering matter is not named 4
```

Files I/O

CSV files - Basic

```
1 echo -e "header1, header2\n1,2\n3,4" > test.csv
```

```
1 import csv
2 fl = list(csv.reader(open("test.csv")))
3 header, values = fl[0], fl[1:]
4 header
5 values
6 fl
```

```
1 ['head1', 'head2']
2 [['1', '2'], ['3', '4']]
3 [['head1', 'head2'], ['1', '2'], ['3', '4']]
```

CSV files - Custom

```
1 class customcsv(csv.Dialect):
2     lineterminator = '\n'
3     delimiter = ','
4     quoting = csv.QUOTE_NONE
5
6 fl.csv = csv.reader("test.csv", dialect=customcsv)
7 fl.csv
```

CSV files - Pandas - read$_{csv}$

```
1 import pandas as pd
2 # header=none if not in file
3 # or read_table + sep(delimeter)
4 fldf = pd.read_csv("test.csv")
5 type(fldf) #type is different
6 fldf
```

```
1 <class 'pandas.core.frame.DataFrame'>
2       head1   head2
3 0       1       2
4 1       3       4
5
6 [2 rows x 2 columns]
```

CSV files - Pandas - More Options

- nrow=5 => read 5 rows
- na_rep='NULL' => set null to NULL else empty
- index=FALSE => no indices in output
- cols=['header1','header2'] => specify columns
- For all options:

`http://pandas.pydata.org/pandas-docs/version/0.13.1/generated/pandas.io.parsers.read_csv.html`

CSV files - Pandas - to_csv

- Many of the same options as read$_{csv}$

http://pandas.pydata.org/pandas-docs/version/0.13.1/generated/pandas.DataFrame.to_csv.html

```
1 import os #to see directory contents
2 fldf
3 fldf.to_csv("files/test_out.csv")
4 os.listdir('files')
```

```
1 head1   head2
2 0       1       2
3 1       3       4
4
5 [2 rows x 2 columns]
6 >>> ['test_out.csv']
```

Getting Help

- help(function) gets you the "docstring"

```
1 help(len)
```

```
1 Help on built-in function len in module __builtin__:
2
3 len(...)
4     len(object) -> integer
5
6     Return the number of items of a sequence or mapping.
```

Regular Expression

Expressions

Classes/Objects

Common Packages

Scientific

- Numpy: N-dimensional arrays, C integration, linear algebra
- SciPy: Numerical integration, optimization, depends on Numpy
- Matplotlib: 2d plotting
- Pandas: Approximates R/Stata, data cleaning, dataframes
- Statsmodels: For statistical models

Webscraping

- BeautifulSoup

HTML/XML/JSON

- HTML is an implementation of XML (a meta language)
- JavaScript Object Notation (JSON) is replacing xml for speed and readability (api)

Firebug

- Firebug is tool that allow you to inspect the elements of a webpage directly.

XPATH SQL for HTML/XML

- Xpath is a language that allows you to select "nodes" from xml
- Note: xpath 2.0 not implemented in all cases though many examples online
- Xpath 1.0 Tutorial

`http://www.zvon.org/comp/r/tut-XPath_1.html#Pages~List_of_XPath`

- Full reference

`http://www.w3.org/TR/xpath/`

XML - Loading

```
xml = """
    <root>
        <name type="superhero">Batman</name>
            <sidekick>Batty</sidekick>
        <contact type="email">riseup@batman.com</contact>
        <contact type="phone">555-1212</contact>
    </root>
            """

from lxml import objectify
root = objectify.fromstring(xml) #use parse from file

print root.tag
print root.text
print root.attrib

print root.name.tag
print root.name.text
print root.name.attrib

for con in root.contact:
    print con.text
    print con.attrib
```

view source

JSON - Loading

```
1  jsn = """
2      {"name":"batman",
3       "hobbies": ["fast cars", "fast planes", "spending money"],
4      "buddy":"robin",
5      "enemies": [{"name":"The Joker"},
6                  {"name":"The People of Gotham"}]
7                  }
8  """
9  import json
10 #NOTE: loads for strings, load for files
11 rslt = json.loads(jsn) #put this into a form for python
12 print rslt
13 jsn_again = json.dumps(rslt) #back to json
```

```
1  {u'buddy': u'robin', u'enemies': [{u'name': u'The Joker'},
       {u'name': u'The People of Gotham'}], u'name': u'batman',
       u'hobbies': [u'fast cars', u'fast planes', u'spending money']}
```

JSON - Converting to DataFrames

```
1 enemies = pd.DataFrame(rslt['enemies'], columns=['name'])
2 enemies
```

```
1         name
2 0          The Joker
3 1          The People of Gotham
4
5 [2 rows x 1 columns]
```

JSON - Converting to DataFrames

```
1 enemies = pd.DataFrame(rslt['enemies'], columns=['name'])
2 enemies
```

```
1          name
2 0              The Joker
3 1  The People of Gotham
4
5 [2 rows x 1 columns]
```

JSON - Example

```
1  import json
2  import urllib2
3  import pprint import pprint
4  import pandas as pd
5
6  prefix="http://maps.googleapis.com/maps/api/geocode/json?address="
7  suffix="&sensor=false"
8  address="165%20Whitney%20Avenue,%20New%20Haven,%20CT"
9  url = prefix+address+suffix
10 j = urllib2.urlopen(url)
11 js = json.load(j)
12 type(js) #if in doubt, check type
13
14 #pprint(js)
15
16 #notice nested list, so use index to get into it
17 rstadd = js['results'][0]['address_components']
18
19 for rs in rstadd:
20     print rs['short_name'], rs['types']
21
22 import pandas as pd
23 pd.DataFrame(rstadd)
```

view source

Regular Expressions (Regex)

- Regex came from perl, used to find text patterns
- To fragile for webscraping, but important complement

stuff

Stuff

Git

http://wildlyinaccurate.com/a-hackers-guide-to-git

Operators

Setting Up Your Development Environment

Top Aligned Blocks

Code

Cool Lots of Stuf
To talk
about

Result

pretty nice!

Beamer: Animated Bullets

- Trouble Shooting

Beamer: Animated Bullets

- Trouble Shooting
- A framework for thinking about programming

Beamer Columns

Stuff

- Truth is ephemeral
  - What is right?
  - What is Wrong?

setting python paths

```
:Setting environment variables (like PYTHONPATH)
:Create an emacs-lisp code block that looks like this:

:#+BEGIN_SRC emacs-lisp
:(setenv "PYTHONPATH" "/Users/neilsen/Development/obswatch-trur
:#+END_SRC
:Execute it, and it changes the environment accordingly.
:Note that you can also append to environment variables like th

:#+BEGIN_SRC emacs-lisp
:(setenv "PYTHONPATH" (concat (getenv "PYTHONPATH") ":" (getenv
:#+END_SRC
:#+END_SRC
```

How to use virtualenv & pip

```
## run this on the command line
## assuming you are in your projects folder, create a new folder
mkdir projects1

cd projects1

## now create your virtualenv environment
## this will create a folder called "env".
## this will house a local version of python.
virtualenv env

## IMPORTANT.
## Now you need to activate your environment.
source env/bin/activate

## now you will be using a local version of python instead of your
## system's python

## to deactivate, simply type
deactivate
```

How to Share Ipython Notebooks

How to share your vagrant box

Testing Python Output

```
1 a = ('b', 200)
2 b = ('x', 10)
3 c = ('q', -42)
4 return (a, b, c)
```

Python Output

```
a = ('b', 200)
b = ('x', 10)
c = ('q', -42)
return (a, b, c)
```

By removing the :exports both, you can export just the code and not the output. By replaceing it with :exports results, you can export the output without the source.

Using pip once virtualenv is activated

```
## again, these should be run on the command line.
## first, let's activate your virtual environment, if you haven't
## already
source env/bin/activate

## first, let's inspect what command are available in pip
pip help

## from this, we see that there are a number of commands we will
## find useful
pip list # this shows what programs are already installed
pip search numpy # this searches for packages named "numpy"
pip install numpy # this installs the numpy package.

## if you have many packages you want to install, you can
## create a requirements list
## this will create a file with a list of modules to install
## you can use your editor of choice to install this.
echo "numpy\nbeautifulsoup" > requirements.txt

## this will install all the packages in the text file.
## NOTE: you can specify the versions of module too. Sometimes
## this is important.
pip install -r requirements.txt

## now let's confirm that they installed correctly
pip list

## now if you are done with virtualenv remember to deactivate it
deactivate
```

Operators

Operator	Description	Example
+	Addition - Adds values on either side of the operator	a + b will give 30
-	Subtraction - Subtracts right hand operand from left hand operand	a - b will give -10
*	Multiplication - Multiplies values on either side of the operator	a * b will give 200
%	Modulus - Divides left hand operand by right hand operand and returns remainder	b % a will give 0
**	Exponent - Performs exponential (power) calculation on operators	a**b will give 10 to the power 20
//	Floor Division - The division of operands where the result is the quotient in which the digits after the decimal point are removed.	9//2 is equal to 4 and 9.0//2.0 is equal to 4.0