

Data Mining

Thomas Carriero
carriero@fas

Why?

- Tons of data on the Internet!
- Data is often in a bad format for us to use
- Data mining can be “tricky”, but not too “hard”
- Will hopefully be useful for the final project
- Will definitely be useful for HW2 Extra Credit!

How?

- Python
 - High level language
 - Like PHP, Ruby, Perl
 - Has special packages for webscraping!
- Read HTML page source
 - Python package called BeautifulSoup helps with this part
- Regular Expressions to extract only the information we want!

Introduction to Python

- Getting the python programming language
 - Windows: <http://www.python.org/download/>
 - Mac and Linux: You already have it! Type “python” at your command line
- Interpreted language (can also be compiled)
- No explicit types—all of them are figured out during runtime

Python Syntax

- Whitespace matters
- No {}'s, just :'s
- For loops are a little different
- Never write types (eg., “int x = 5”; should be “x = 5” in python)
- ;'s at the end of lines is optional

Factorial in Python

```
def fact(n):  
    if(n == 0):  
        return 1  
    else:  
        return n * fact(n-1)
```

```
> fact(5)
```

```
120
```

For loops in Python

```
a = [] #makes an array
for i in range(10):
    a.append(i*2) #fills the array with even numbers
```

```
> range(5)
[0, 1, 2, 3, 4]
> a
[0, 2, 4, 6, 8, 10, 12, 14, 16, 18]
```

Can also do:

```
for ele in a:
    print ele
```

Regular Expressions

- Pattern-matching tool
- Very powerful, we will only explore a small subset
- Intro: http://en.wikipedia.org/wiki/Regular_expression
- Intro in Python:
<http://www.amk.ca/python/howto/regex/>
- Regular expression checker:
<http://eclos.free.fr/utilLib/RegExpChecker.html>
 - Note that the matches are implemented slightly differently in this applet than in Python
 - Useful for exploring regular expressions before you write them in the code!

Time to write some code!

- (Code we wrote in class will be available on the website.)