Slithering over the web with Python

Troy Vasiga
Lecturer, UW
Director, Canadian Computing Competition

(non-dot version)

Outline

- Present future of programs
- Processing information
- Java vs. C++ vs. Python
- Structure of web/data
- Examples:
  - Simple Processing
  - Shakespeare
  - Weather

Present future of programs

- Thousands of “apps”
  - Blackberry
  - iPhone
  - Process information into a reasonable format
- Very few operating systems or extremely large programs: these are increasingly inaccessible
- Students can make “real” applications using very simple tools

Java vs C++ vs Python

- C/C++ has no usable web support (C# does)
- Java has web support, but limited text processing
- Python has easy URL connection capability and easy text processing/parsing

Structure of web/data

- HTML tags are a useful exercise:
  - Vast majority of data is “formatted”
  - Teaches repetition/recursion
    - <ul><li>...</li><li>...</li>...</ul>
  - Converting tags into other formats is a very useful process

Processing information

- Input
- Processing
- Output

- “Circle of life”

Tools

- Python 2.6.2 (quite a bit different than Python 3.x)
- Wing IDE 101
- Both free

Opening a URL connection

```
import urllib

testURL =
  "http://www.cs.uwaterloo.ca/~tmjvasig/talks/cascon09/index.htm
  l"

testCall = urllib.urlopen(testURL).read()

print testCall
```

Processing

1. Structure of Python
2. Lists in Python
3. Sets in Python
4. Strings in Python
5. split() and split(“string”)
6. Extracting just links from the page

Crawling

- Start somewhere
- Follow a link
- Repeat

Extracting Useful Information

- What is the volume of rain that fell on Canada in October 2009?
- Based on past performance, who will win tonight's World Series game?