A few slides relevant to web crawler from late October or early November

CS 111: Program Design I

A few slides relevant to web crawler from late October or early November

Robert H. Sloan & Richard Warner

University of Illinois at Chicago

UIC

Break

- Causes immediate termination of innermost enclosing while or for loop
- Typical usage:

```
# We are inside a while or for
if <some particular case>:
    break
```

- Use carefully! Can make code very hard to read

Example

```
for n in range(2, 10):
    for x in range(2, n):
        if n % x == 0:
            print(n, 'equals', x, '*', n//x)
            break
        else:
            # loop fell through without finding  factor
            print(n, 'is a prime number')
```

Which of these will exit when x is initially 9?

A

```
while (x%2 == 1 and x%3 == 0):
    x = 9
```

B

```
while True:
    if (x%2 == 1 and x%3 == 0):
        break
    x = 9
```

C. Both D. Neither E. I don't know

Continue

- Continue continues with *next* iteration of loop *instead of finishing current iteration*

Continue example

```
for num in range(2, 6):
    if num % 2 == 0:
        print("Found an even number", num)
        continue
    print("Found a number", num)
```

- Found an even number 2
- Found a number 3
- Found an even number 4
- Found a number 5

Idiomatic Python: Empty list check

- To check whether list is empty vs. nonempty
  - Python in Boolean context after if or while treats empty list as False, all other lists as True
- So if want to keep processing ls as long as it's nonempty:

```
while ls:
    <process ls, remove, append, etc.>
```

I.e., Idiomatic test for truth

- Pythonistas do this

```
if item_ls:
    stuff
```

- *NOT* this

```
if len(item_ls)!=0:
    stuff
```

- And definitely *not*

```
if item_ls != []:
    stuff
```

Crawl all pages reachable from start

- List of pages to visit, initially start
- while that list is not empty:
  - Take a page from the list
  - Get its text # need to learn how to do this
  - remove that page from to-visit list, add it to already-visited list
  - Get all the links in that page
  - for each link
    - if not already in visited list
    - add it