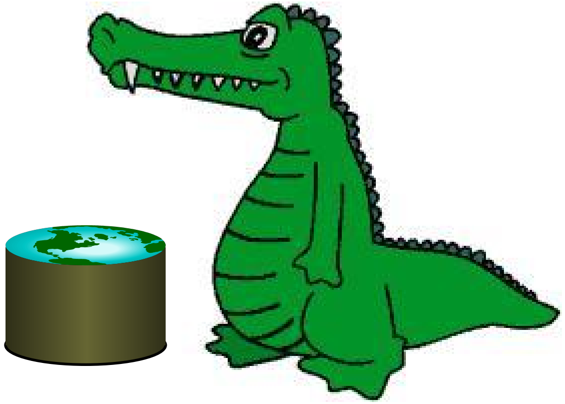
CAP4770/5771
Lab 0 part 1
UNIX command line utilities

University of Florida, CISE Department
TA: Xiaofeng Zhou



Goal

Use UNIX command line utilities to do:

- Some data cleaning
- Basic analysis

Outline

0. Set up your environment
1. UNIX command line utilities
    a. File system analytics
    b. Log processing with command line tools
    c. Data transformation with sed

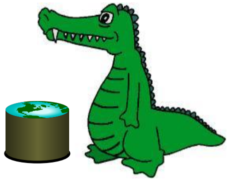
Set up your environment

1. Follow the instructions in ‘Lab Setup’ section on the canvas course homepage

   https://ufl.instructure.com/courses/320501
2. Please note: you are expected to use the VM provided for labs unless otherwise required

UNIX command line utilities

1. File system analytics
   a. Navigating the file system
   b. Sorting
2. Log processing with command line tools
   a. Downloading data files
   b. Exploring http logs dataset
3. Data transformation with sed
   a. Regular expression substitution in sed
   b. Backreferences in sed



File system analytics

Open a terminal(ctrl + alt + t)

1. Navigating the file system
   a. useful commands: `ls, cd, find, du (head, tail)`
   b. e.g. `$find /usr/bin -type f -name 'py*' -exec du {} \;`
2. Sorting
   a. useful commands: `sort`
   b. e.g. `$ls -l | sort -k5 -n -r` (like `$ls -lS`) (`|Pipe operator`)

Log processing w/ command line tools

1. Downloading data files
   a. useful commands: `curl, wget`
   b. e.g.(web server log)

```
$curl -L https://raw.github.com/biddata/datascience-fa14/master/lab1/data/wc_day6_1_log.tar.bz2  -o wc_day6_1_log.tar.bz2
```

Log processing w/ command line tools - contd

2. Exploring http logs dataset
   a. useful commands: `tar, less, head, tail, cut, uniq, wc, grep`
   b. e.g.
      ```
      $less wc_day6_1.log
      $grep "tickets.*html" wc_day6_1.log | wc -l
      $head -50 wc_day6_1.log  | cut -d ' ' -f 7 | sort | uniq -c | tail -10
      ```

Data transformation with sed

- Goal:

  Convert the log file into csv file with the fields:

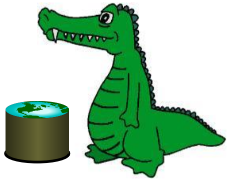
  ```
  ClientID, Date, Time, URI, Response code, Response size,
  Method
  ```

- e.g. from:

  ```
  0 - - [30/Apr/1998:22:00:02 +0000] "GET
  /images/home_intro.anim.gif HTTP/1.0" 200 60349
  ```

  to:

  ```
  0,1998-04-30,22:00:02,/images/home_intro.anim.gif,
  200,60349,GET
  ```



Data transformation with sed - contd

1. Regular expression substitution
   a. usage:

   ```
   $sed 's/regexPattern/replacementString/flags'
   ```

   b. e.g.

   ```
   $echo "The quick brown fox jumps over the lazy dog."
   | sed 's/[tT]he [a-z]*/The yellow/'

   $echo "The quick brown fox jumps over the lazy dog."
   | sed 's/[tT]he [a-z]*/The yellow/g'
   ```

Data transformation with sed - contd

2. Removing cruft in sed
   a.
   ```
   $head wc_day6_1.log | sed 's/ +0000]//; s/\[//'
   ```

3. Backreferences in sed
   a. usage example:
   ```
   $echo "The quick brown fox jumps over the lazy dog."
   | sed 's/\([Tt]he\) \([a-z]*\)/\1 "\2"/g'
   ```

Please remember go to course web site to finish the assignment for lab 0-1.

Deadline: Thu(9/3) 5:00PM