Discussion 6

Assignment 2

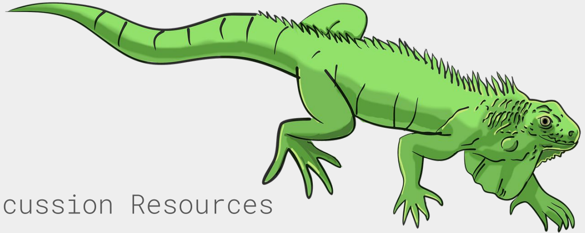
Welcome! Information Retrieval - Discussion 6

Itinerary November 03, 2021

- Assignment 2
  - Terminal skills
    - Tmux
    - OpenLab (identifying machine)

Deadlines this week

- Assignment 2
  - Report due 11:59pm Friday
  - Crawler must be **finished** by 11:59pm friday (**start** it before!)



Find these slides & Zoom recordings on Canvas → Pages → Discussion Resources

Assignment 2

Terminal Tools

IMPORTANT - Assignment 2 Deadlines

- It is **due** on Friday, November 5, 11:59PM
  - This means you must be done **crawling** by then
  - You probably need to plan to start your final crawl at the very latest on Thursday evening
    - You get 3 attempts (all three are recorded, cannot restart more than three times)
    - Remember, you need to run processing after you run your crawler, so allow time for that.
- If you want extra credit, AKA multi threading or exact/similar webpage detection, it is also due at this time.

Tools to run your crawler

- Your crawler should take anywhere between a few hours, to a full day to run
- **Strongly recommend** that you use tmux for running your crawler sessions
- Why?
    - Because if you lose SSH connection, **you lose an attempt!**
- Be aware also of what machine on openlab you use to run your crawler (you will need to use this!)

Tmux

Overview and Demo

What is tmux?

- Terminal multiplexer
- Allows for a single environment to launch multiple terminals or windows
  - Each runs its own process or program

Download from
Dreamstime.com
This watermarked comp image is for previewing purposes only.
ID 118552041
Milkovasa | Dreamstime.com

Scenario

- Let’s say you are working on your Assignment 2 crawler
- You have to connect to the remote UCI server to run your crawler
- It takes several hours to run!
- You launch your program, and sit back.
- 5 hours later, your connection to the remote server was lost!

If you used...

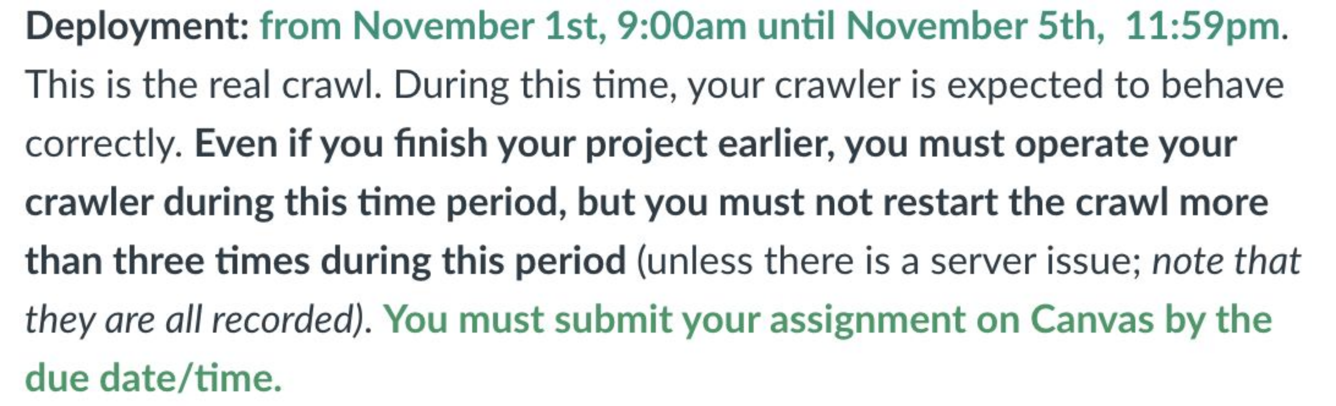
A regular terminal session

- Your work was completely **lost**!
  - The terminal you used to connect to SSH was the same terminal session running the program
  - So if you lost connection, you lose your session!
  - You also lose an **attempt** because of the server logs!

A **tmux** session

- Your task is **still running**!
  - Because tmux launches an independent terminal instance on the remote server
  - Allows you to keep things running persistently on servers, even if you disconnect

Losing SSH connection **does not count** as a server issue!



Deployment: from November 1st, 9:00am until November 5th, 11:59pm. This is the real crawl. During this time, your crawler is expected to behave correctly. **Even if you finish your project earlier, you must operate your crawler during this time period, but you must not restart the crawl more than three times during this period** (unless there is a server issue; *note that they are all recorded*). **You must submit your assignment on Canvas by the due date/time.**

- It is a “you” issue because you should have used tmux!

Example “crawler” program using tmux

- Going to ssh into openlab
  - We’re going to run one in tmux
  - One just directly on the terminal
- Observe results when we lose SSH connection
  - tmux will persist!!

infinite_loop.py

```
while(True):
    print("I'm crawling!!!!")
```

Tmux Demo

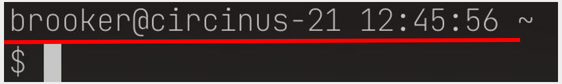
Tmux vs. no tmux when we lose connection

OpenLab - Keeping Track of the Machine

1. Log in to OpenLab as usual (if you don't remember how to SSH to openlab, please view my recording on this process)

   ```
   ssh USERNAME@openlab.ics.uci.edu
   ```

2. After you log in, take note of your username

   ```
   brooker@circinus-21 12:45:56 ~
   $ 
   ```

3. Start your tmux session `tmux new -s session_name`
   a. This is where you'd start your crawler
   b. To detach from your tmux session, press Control B quickly, then press D

OpenLab - Logging Back into Machine

1. Log into the **<u>specific</u>** machine from before

   (example): `ssh USERNAME@circinus-21.ics.uci.edu`

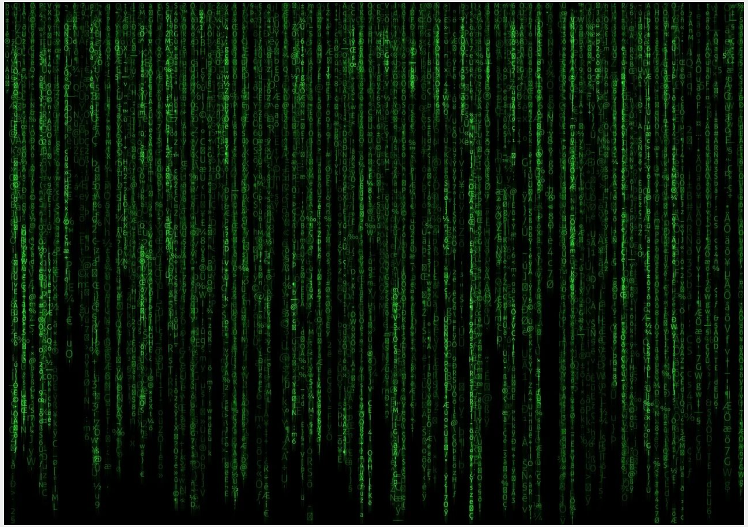
2. To find where your crawler is running:
   1. List the tmux sessions `tmux ls`
   2. Attach to session `tmux attach -t SESSION`

OpenLab Demo

Making sure you are on the right machine

Other Resources

- https://edstem.org/us/courses/13796/discussion/778849 Wonderful write-up from TA Rachel about terminal and tools we discussed
- Real server outages posted here: https://edstem.org/us/courses/13796/discussion/803024
- Tmux commands: https://tmuxcheatsheet.com/



Warning - Keep Your Crawler **Away** From

- UCI Machine Learning libraries
    - For those of you not familiar with Machine Learning, we use very **large** datasets
    - Don’t let your crawler download these!
    - (Here’s an example of a big one): https://archive.ics.uci.edu/ml/datasets/Letter+Recognition

Warning - Keep Your Crawler **Away** From

- Calendars
  - Things that end with a date
  - Things that end with events week
  - Oftentimes these can be a trap

Although time may be an illusion created by conscious agents, **calendars** are very real and can be traps for your crawler!!

Next Week's Discussion

Tentative plan for next week's discussion based on upcoming course deadlines.

- Half milestone 1, half quiz 3 review
  - Both are actually due next week

Recommended Homework

To best prepare for next week's session, I recommend you do the following.

- Rest of the course is going to be very fast-paced with the upcoming milestones + quizzes
- As always, **please**, I beg of you, **start early!!!**
  - I learned the hard way in undergrad you can't wait until the last minute with programming assignments
    - Don't let this be you!
    - Start this weekend on Assign. 3, Milestone 1