Discussion 4

Tmux and Running the Web Crawler

Welcome! Information Retrieval - Discussion 4

Itinerary October 20, 2021

- Tmux overview and demo
- Web crawler GitHub project

Deadlines this week

- Nothing!

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

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 as a neural networks researcher
- You have to connect to the remote UCI server to train your networks
- It takes several days to run!
- You launch your program, and sit back.
- 12 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!

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

Easiest Way to Install

- Use a package manager
- Helpful guide

For Linux…

```
sudo apt install tmux
```

For Mac OS…

```
brew install tmux
```

Learning Resources

- Quick reference of commands: https://tmuxcheatsheet.com/
- Really good guide: https://github.com/ole3021/Resources/blob/master/Ebooks/Tools/tmux%20-%20Productive%20Mouse-Free%20Development.pdf

Essential Commands

Create a session

```
tmux new -s session_name
```

Detach from session

```
Ctrl+b d
```

List sessions

```
tmux ls
```

Re-attach to target session

```
tmux attach -t session_name
```

Web Crawler

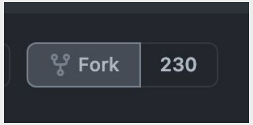
Assignment 2 Setup

Forking the project from GitHub

Navigate to the GitHub project:

https://github.com/Mondego/spacetime-crawler4py

Click “Fork” in the upper right corner



Now, the forked project will be included in your own projects.

Checkout the forked project for local development

Copy the link of the forked repository in your GitHub projects.

https://github.com/YOUR_USERNAME/spacetime-crawler4py

Open Terminal and paste the command:

```
git clone https://github.com/YOUR_USERNAME/spacetime-crawler4py
```

Now, the code is on your local machine ready for development!

Demo

- How to stop (Ctrl-c) and restart the crawler (run the program again)
- How restart crawling from scratch (delete the file frontier.shelve, and run it again)

Next Week's Discussion

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

- Homework 2

Recommended Homework

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

- **Get started** on your web crawler homework!
  - This homework is more **ambiguous** and there is **no** one right solution
  - Very important you begin development and testing
  - Do **not** leave it to the last minute!!