Lecture 12: Shell Scripting, SSH, Super-Computing

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

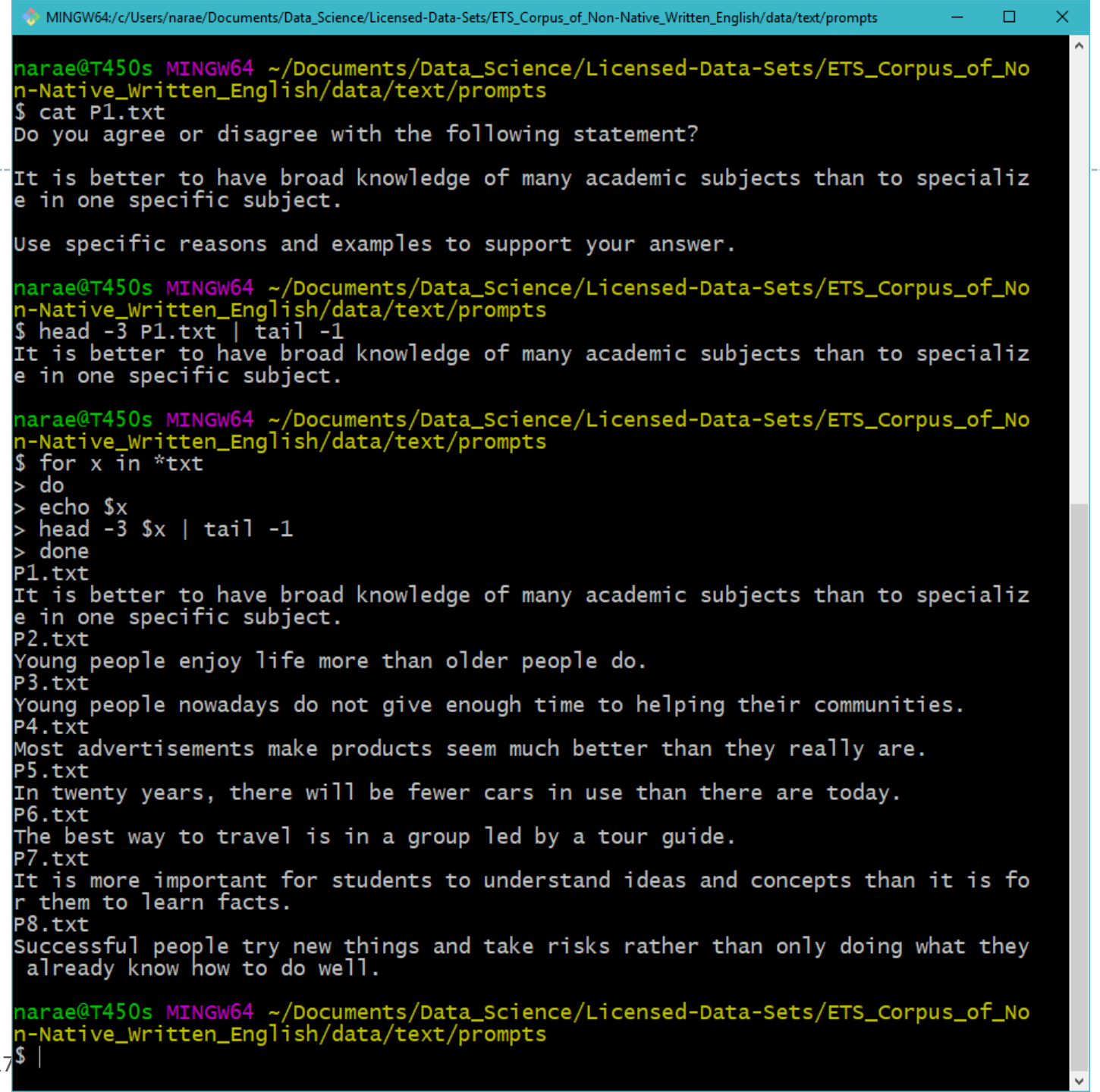
- Batch processing through for loop
- Shell scripting
- Server access through SSH
  - Pitt's timeshare account
  - nano: a simple command-line editor
- Supercomputing at CRC

Batch processing through for loop

- Your command line is actually running a programming environment: bash shell.
- You can *program* in command line, even for loops!

```
narae@T450s MINGW64 ~/Desktop/inaugural
$ for file in *.txt
> do
> iconv -f US-ASCII -t UTF-16 $file > try/$file
> echo $file complete
> done
1789-Washington.txt complete
1793-Washington.txt complete
1797-Adams.txt complete
1801-Jefferson.txt complete
1805-Jefferson.txt complete
1809-Madison.txt complete
1813-Madison.txt complete
1817-Monroe.txt complete
1821-Monroe.txt complete
1825-Adams.txt complete
```

Slide from October 5


MINGW64:/c/Users/narae/Documents/Data_Science/Licensed-Data-Sets/ETS_Corpus_of_Non-Native_Written_English/data/text/prompts
narae@T450s MINGW64 ~/Documents/Data_Science/Licensed-Data-Sets/ETS_Corpus_of_Non-Native_Written_English/data/text/prompts
$ cat P1.txt
Do you agree or disagree with the following statement?
It is better to have broad knowledge of many academic subjects than to specialize in one specific subject.
Use specific reasons and examples to support your answer.
narae@T450s MINGW64 ~/Documents/Data_Science/Licensed-Data-Sets/ETS_Corpus_of_Non-Native_Written_English/data/text/prompts
$ head -3 P1.txt | tail -1
It is better to have broad knowledge of many academic subjects than to specialize in one specific subject.
narae@T450s MINGW64 ~/Documents/Data_Science/Licensed-Data-Sets/ETS_Corpus_of_Non-Native_Written_English/data/text/prompts
$ for x in *txt
> do
> echo $x
> head -3 $x | tail -1
> done
P1.txt
It is better to have broad knowledge of many academic subjects than to specialize in one specific subject.
P2.txt
Young people enjoy life more than older people do.
P3.txt
Young people nowadays do not give enough time to helping their communities.
P4.txt
Most advertisements make products seem much better than they really are.
P5.txt
In twenty years, there will be fewer cars in use than there are today.
P6.txt
The best way to travel is in a group led by a tour guide.
P7.txt
It is more important for students to understand ideas and concepts than it is for them to learn facts.
P8.txt
Successful people try new things and take risks rather than only doing what they already know how to do well.
narae@T450s MINGW64 ~/Documents/Data_Science/Licensed-Data-Sets/ETS_Corpus_of_Non-Native_Written_English/data/text/prompts
$

Shell scripting

- Bash commands can be saved into a **shell script** file.
  - Can be run later, any time
  - Can be customized to take different file arguments, etc.
- Software Carpentry's tutorial:
  - http://swcarpentry.github.io/shell-novice/06-script/
- Running a script file
  - Option 1: `bash myscript.sh`

    ← This is how it's done in SC's tutorial
  - Option 2: `myscript.sh`
    - Put in a shebang line on top of your script file:

      `#!/bin/bash`
    - And then change permission of your file to make it executable:

      `chmod u+x myscript.sh`

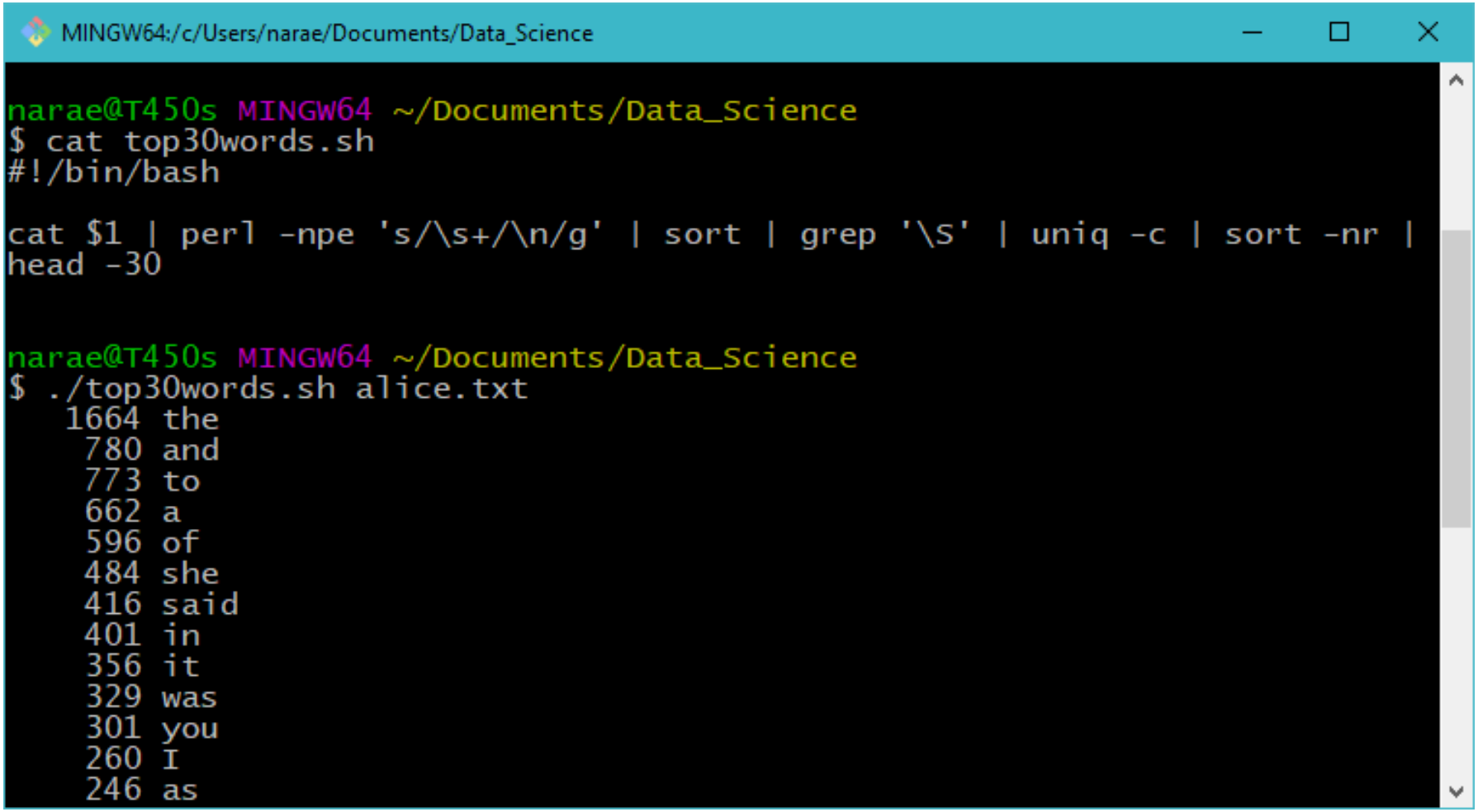
top30words.sh

Activity
5 minutes

- Create a shell script called `top30words.sh`
  - Takes a single text file as argument
  - prints out 30 most frequent words along with counts

```
narae@T450s MINGW64 ~/Documents/Data_Science
$ ./top30words.sh alice.txt
   1664 the
    780 and
    773 to
    662 a
    596 of
    484 she
    416 said
    401 in
    356 it
```

  - HINT: Page 23 of last class's slides "Piping gone mad"


MINGW64:/c/Users/narae/Documents/Data_Science
narae@T450s MINGW64 ~/Documents/Data_Science
$ cat top30words.sh
#!/bin/bash
cat $1 | perl -npe 's/\s+/\n/g' | sort | grep '\S' | uniq -c | sort -nr |
head -30
narae@T450s MINGW64 ~/Documents/Data_Science
$ ./top30words.sh alice.txt
1664 the
780 and
773 to
662 a
596 of
484 she
416 said
401 in
356 it
329 was
301 you
260 I
246 as

Accessing your Pitt server account

- Everyone at Pitt has a Unix timeshare account. (Bet you didn't know.)
- My own home page is hosted on it:
  - http://www.pitt.edu/~naraehan/
- You too can make your own home page!

Accessing Pitt server

Activity
5 minutes

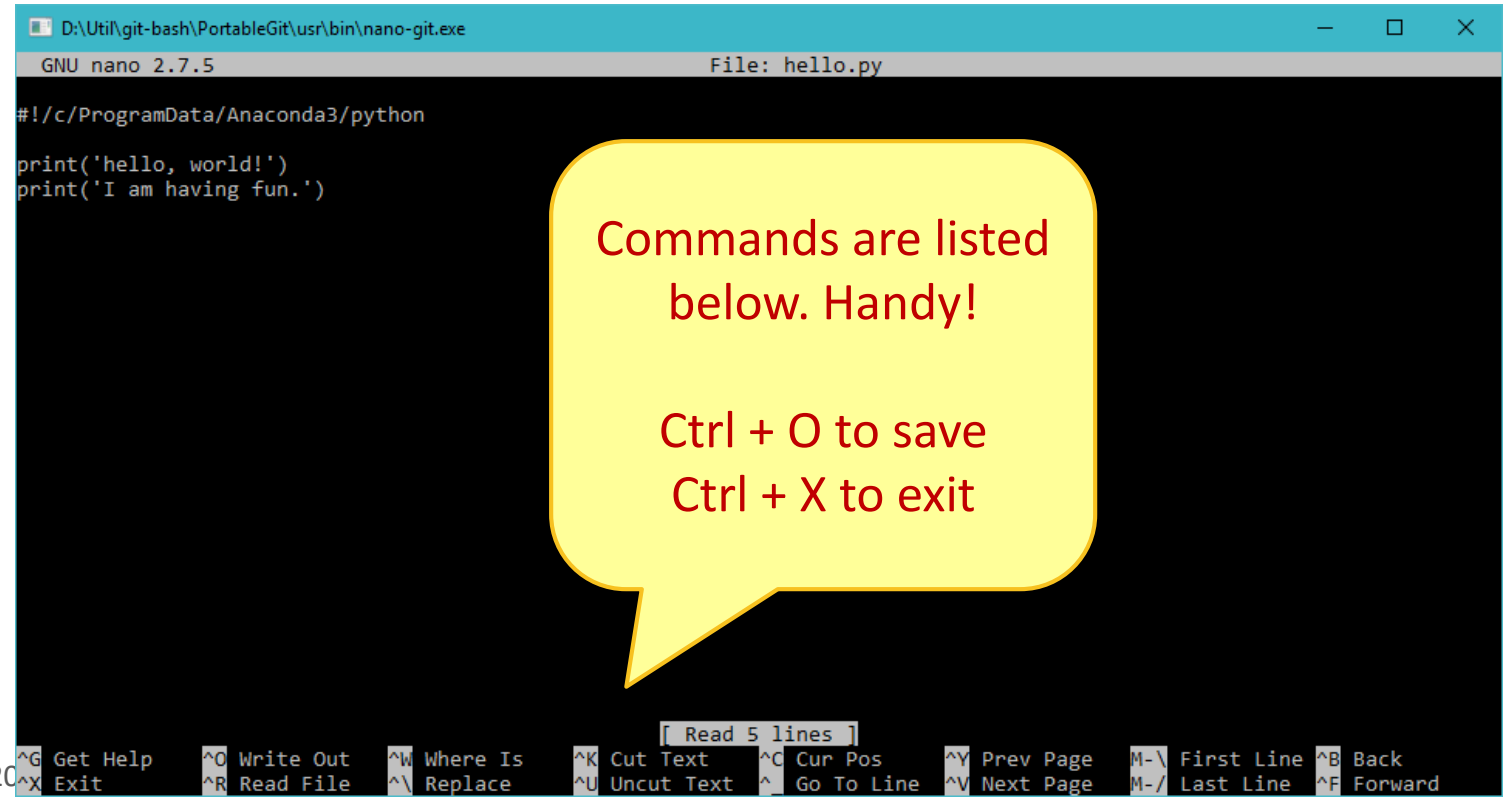
- Remote-access your account via SSH:
  - `ssh yourpittid@unixs.cis.pitt.edu`
- Move into `public/` directory. Use `cd`.
- Create a directory named `html/`. Use `mkdir`.
- Inside the `html/` directory, using the `nano` editor, create and edit a file named `index.html`. Put these lines:

```
<html>
<body>
Welcome to so and so's home page.
</body>
</html>
```

- Open up a browser and navigate to your home page address:
  - http://www.pitt.edu/~yourpittid
- What mischief can you do on this server? Find out.

nano

- nano is a simple command-line-based editor. It is found on all Linux distros.
  - Already present on Macs.
  - Windows users: you downloaded it and set it up.



Let us now supercompute.

- By Argonne National Laboratory's Flickr page - originally posted to Flickr as Blue Gene / PFrom Argonne National LaboratoryUploaded using F2ComButton, CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=6412306

You got a supercomputing account.

- You received this mysterious email:

From: <no-reply@core.sam.pitt.edu>
Subject: Center for Simulation and Modeling (SaM) Account
Date: November 2, 2017 at 10:49:31 AM EDT
To: <blh82@pitt.edu>

Dear user,

Welcome to SaM!

An account has been created for you on Center for Simulation and Modeling (SAM) resources. Your username is "blh82" (without the quotes). All authentication is through Pitt's Active Directory. Therefore, your password is the password associated with your Pitt account.

I got you all an account at Pitt's **Center for Research Computing** (formerly known as SAM).

CRC: Center for Research Computing

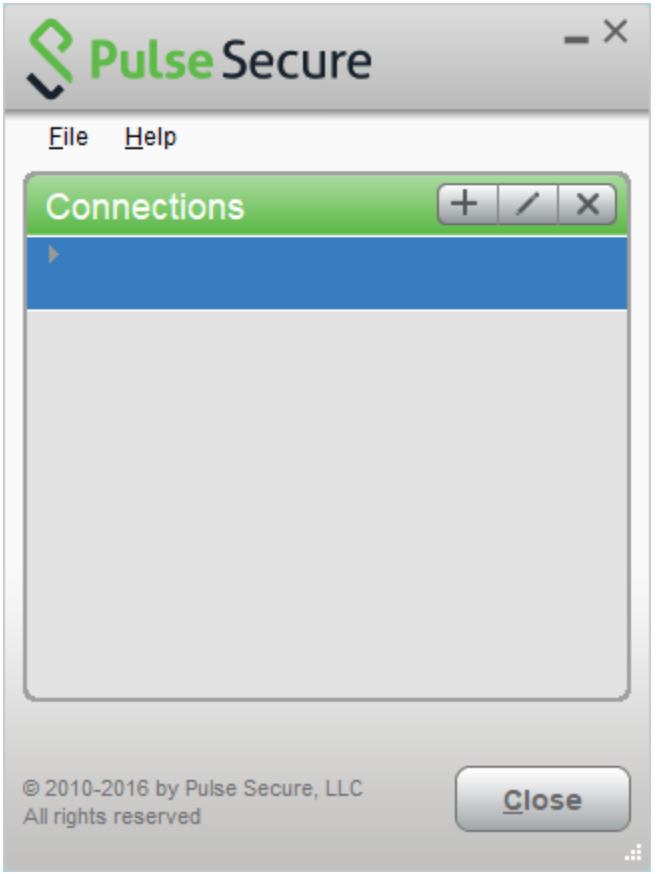
- https://crc.pitt.edu/

- New User Guide here:
  - https://crc.pitt.edu/documentation/
    - READ them!
- Information on h2p (Hail 2 Pitt)
  - https://crc.pitt.edu/documentation/h2p/
- Introduction by Barry Moore II on Thursday.

CRC machines require secure access

- Unless we are accessing from a wired connection on Pitt's campus, out laptop should be running a Secure Remote Access client.
  - Download and install **Pulse Secure Client** http://technology.pitt.edu/services/secure-remote-access
  - Add connection name "Pitt VPN", server "sremote.pitt.edu"
  - For VPN connection, choose:

    Firewall-SAM-USERS-NetworkConnect
  - If prompted for secondary password, type in "push" (this triggers Duo multi-factor authorization)



Accessing CRC server

- Remote-access your account via SSH:
  - `ssh yourpittid@h2p.crc.pitt.edu`
- Getting your bearings:
  - Where are you? `pwd`
  - What is your user 'group'? `groups`
  - Is python installed on this machine? `which python`
  - What are your configuration files:
    - .bash_profile
    - .bash_history

      ← Bash commands you typed in are logged here.

Grepping the inaugural

Activity
15 minutes

- Download inaugural.zip from NLTK's data page. How?
- Unzip the .zip archive. How?
- Grep for 'prosperity'. Hmm lines are too long...
- Use `fold` to fold long lines.
  - Line breaks in the middle of words! How to break along space? Use man page to find out.
  - Create another version inaugral2 with folded lines.
- Which presidents talked about 'Russia'? 'war'? 'unity'?
- How about 'God bless'?
- Which presidents used split infinitives?
  - How to print out more context: 2 lines before and after?

Grepping the inaugural

Activity
15 minutes

- Download inaugural.zip from NLTK's data page. How?
  - wget https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/inaugural.zip
- Unzip the .zip archive. How?
  - unzip inaugural.zip
- Grep for 'prosperity'. Hmm lines are too long...
- Use `fold` to fold long lines.
  - Line breaks in the middle of words! How to break along space? Use man page to find out.
  - Create another version inaugral2 with folded lines.
    - mkdir inaugural2
    - cd inaugural
    - for x in *.txt; do fold -s $x > ../inaugural2/$x; done
- Which presidents talked about 'Russia'? 'war'? 'unity'?
- How about 'God bless'?
- Which presidents used split infinitives?
  - How to print out more context: 2 lines before and after?
    - grep -P -C 2 '\bto \w+ly' *.txt

Before you get carried away

- Do NOT yet run any jobs that may be resource-intensive.
- This is a powerful super-computer, shared by many research groups at Pitt.
  - Our class as a group has a limited, shared allocation.
  - You do not want to accidentally initiate a run-away process and hog resources.
- There are PROPER ways to run jobs.
- We will learn all about it from Barry Moore II on Thursday!

Wrapping up

- To-Do 11
  - Fun with big(ish) data -- Yelp Dataset!
  - Downloading data alone takes about 25 minutes. Allocate enough time for this assignment, especially if you are new to command line.
- Next class
  - Supercomputing at CRC