Lecture 3:

Last time

• A quick introduction to the Unix system; basic commands for navigating the filesystem and simple, yet powerful, tools for manipulating text files

We also described pipes, the Unix metaphor for chaining a series of commands, redirecting the output from one into the input of another

Clarification

cut -d " " -f1,5,9  select the first, fifth and ninth
cut -d " " -f1-5     select the first through the fifth
cut -d " " -f1        select just the first

Today

• Discuss your experiments from last time

  How would we get a time series of hits per day?
  Who had the largest number of hits?
  What can you say about the files being accessed?

The command grep let us extract lines from a file that contained a string of characters; as we started digging into the data, we wanted a more expressive tool for defining patterns

Today we will start on so-called regular expressions, a language for describing patterns in text data

But first, we make a few more broad comments on operating systems
Operating systems

An operating is a piece of software (code) that organizes and controls hardware and other software so your computer behaves in a flexible but predictable way.

In general, an operating system performs six basic tasks (this list is good enough for us; some may debate whether it should do more):

- Processor management
- Memory management
- Storage management
- Device management
- Application interface
- User interface

Running top on compute.stat.ucla.edu
How many processes are running?
How much RAM is available?

Running top on taia.stat.ucla.edu
How many processes are running?
How much RAM is available?

What do you reckon this computer does?
Another way to get at processes

While `top` gives you a dynamic, constantly updating, view of what the processor is doing, you can use the command `ps` to give you a snapshot.

The command `ps` has lots and lots of options; it lets you look at all users, just a specific user and control the format of the output.

Just typing `ps` will give you the processes that you started (or were started on your behalf); we can also see what others are up to (ah, the joys of a multi-user system).

Running `ps -aux` on `lab-compute.stat.ucla.edu`

What can we see?
Who are the users?
* the option `-a` gives you information on all users, `-u` gives you a popular view (fields) of the processes, and `-x` gives you processes that aren't necessarily associated with a terminal.

Job control

Unix allows you to run several processes at once; each process is given a number which you can use to change the status of the process.

Because many jobs are running on the computer, the amount of “attention” they get from the central processing unit is controlled by their priorities (-20 to 20, with the higher the number meaning the lower the priority).

nice and renice lets you lower the priority on a job that you know will run for a long time, freeing system resources for others; `kill` can be used to end processes (politely or with a greater sense of urgency).
Job control

C-z stops jobs, C-c kills them, and C-d kills your shell.

You can also set jobs to run in the **background** (which means your prompt returns).

The command `jobs` lets you see what jobs you have running.

If you stop a job, you can restart it or restart it in the background using the commands `bg` and `fg`.

---

In response, Unix stops the job and gives you a list of other jobs you have stopped or are running in the background.

---

Operating systems

**Device management**

A program called a **driver** translates data (files from the filesystem) into signals that devices like printers can understand; an operating system manages the communication between devices and the CPU, for example.

**Application interface**

An **API** (application program interface) lets programmers use functions of the computer and the operating system without having to know how something is done.

**User interface**

Finally, the operating system turns and looks at you; the UI is a program that defines how users interact with the computer -- some are graphical (Windows is a GUI) and some are text-based (your Unix shell).

---

Job control

You can place jobs in the background immediately by completing your command with a `&`.

```
% R --slave < commands.txt &
```

This bit tells R to perform the commands in the file, but to do so quietly; OK, maybe that’s not the best choice of argument name.
Operating systems

Most devices that contain a computer of some kind will have an OS; they tend to emerge when the appliance will have to deal with new applications, complex user-input and possibly changing requirements on its function.

Your Tivo, Treo and (soon) Peugeot will all have operating systems.

For personal computers, Unix, Windows, and Mac OS are the most popular (um, not in that order).

Back to your “assignments” from last time

Who had the most hits and how did you compute it?

How would you compute the number of hits per day?

Also...

I might not have emphasized enough that each of the commands we worked with can either take input from a file as in

```
sort one_week.txt
```

or from a pipe (so-called standard input), as in

```
grep dinov one_week.txt | sort
```
Also...

Notice that each line of our file is pretty ragged; we would not try to read this into R; there aren’t sufficiently regular fields

% head one_week.txt

203.145.158.2 - - [28/Sep/2005:00:00:05 -0700] "GET /theses/james.pdf HTTP/1.1" 206 13836 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0; .NET CLR 1.0.3705)"

66.249.71.10 - - [28/Sep/2005:00:00:12 -0700] "GET /library/index_body.php HTTP/1.0" 200 3248 "-" "Googlebot/2.1 (+http://www.google.com/bot.html)"

New data

We are now going to look at a series of log files containing data from chat sessions; recording began just before 9 and ended just before 11 this morning and captured 50,000 lines from 4,000 people

The data are stored on lab-compute.stat.ucla.edu and you can bring them to your desktop with the command

```
scp -r lab-compute.stat.ucla.edu:/Data/chat .
```

Rudimentary pattern matching

We have already seen some basic pattern matching notions; recall the command "wc *.txt"

In this expression "*" acts as a wildcard and matches anything

The files chat.1128956004.txt, chat.1128962490.txt, chat.1128967641.txt, chat.1128959603.txt, and chat.1128965050.txt will all be returned by this command
Rudimentary pattern matching

In the expression "*.txt" we can name two kinds of characters

The "*.txt" is made up of literal or normal text characters

The "*" is a metacharacter

Regular expressions

When working with text files (and indeed statistical data of many kinds) we are often presented with rather complex patterns; for example, to understand the kinds of files downloaded from each site (dinov, bk, etc.) we really want to isolate a file's suffix

What we would like is a rich and expressive language for recognizing these patterns

Regular expressions provide us with this capability

From last time:

grep can be used to skim lines from a file that have (or don't have) a particular pattern

Patterns are specified via regular expressions

The name comes from an editing operation: g/re/p
egrep

As with Unix itself, regular expressions have different flavors depending on their implementation.

The most direct way to use the tools we have discussed so far is with the command `egrep` rather than `grep`.

This specifies so-called extended regular expressions.

Regular expressions

Regular expressions have a rich set of metacharacters.

To continue with the language idea, think of literal text forming the words of this language, and the metacharacters defining its grammar.

egrep

To try out any of the expressions we will describe you can enter the command

```
egrep 'pattern' chat.1128962490.txt
egrep 'pattern' chat*
```

An, and again, the funny digits are a date

```
% date -r 1128962490
Mon Oct 10 09:41:30 PDT 2005
```

Regular expressions

Simplest pattern consists only of literals.

The literal "earth" would match to the following chat lines:

- "This earthquake is all americas fault"
- "I would be so scared in an earthquake"
- "2 100,000+ killer earthquakes in less than a year"
- "how much money is the most powerful nation on earth sending to the grief stricken area of Pakistan?"
Regular expressions

The literal "war" would match to the following chat lines

“Bush says Pakistan is our faithful ally in the war on terrorism.”
“we’re talking about a time of war here”
“do u know what happens to a slice of bread with butter and cheese on if u leave it somewhere warm for like 4 months?”
“David Lee Roth is taking Howard Stern’s place...”

Regular expressions

We need a way to express
white space
word boundaries
sets of literals
the beginning and end of a line
alternatives ("war" or “peace”)

Metacharacters to the rescue!

Regular expressions

Simplest pattern consists only of literals; a match occurs if the sequence of literals occurs anywhere in the text being tested

What if we only want the word “earth”? or sentences that end in the word "war”? or “War”, “WAR” and “war”?

Some metacharacters

^ represents the start of a line

^i think

will match the lines

i think i’m falling down, down, down
i think i’m falling for you, you, you
i think i’m falling in love
i think i better shut up.
i think it might be better
Some metacharacters

$ represents the end of a line

morning$

will match the lines

Richard good morning
I slept in til 8am this morning
i have a client in the morning
i just got back from edinburgh this morning
the visuals that go with that are just too much this morning
my niece, I somehow got suckered into babysitting this morning

Character classes with [ ]

As we saw in the last lecture, we can list a set of characters we will accept at a given point in the match

[Bb][Uu][Ss][Hh]

will match the lines

BUSH has done 0 to fight the drug war
To Mr. Bush: Efforts and courage are not enough without purpose and direction.
reason being some of the repubs are upset with bush's selection
bush is basically a lame duck as it were
BUSH MAKES NIXON LOOK GOOD

Character classes with [ ]

Similarly, you can specify a range of letters [a-z] or [a-zA-Z]; notice that the order doesn’t matter

^[0-9][a-zA-Z]

will match the lines

6foot4 250 very fit blackm
6pm here
3rd time lucky
3x your age
5k of what?

Character classes with [ ]

When used at the beginning of a character class, the “^” is also a metacharacter and indicates matching characters NOT in the indicated class

[^?.]$

will match the lines

how are you
raisin bran
really?? - I am in Houston
like drive by hunting... think that i isnt done in the south? lmao
Tom’s first movie, he played a juvenile delinquit
More metacharacters: .

"." is used to refer to any character. So

9.11

will match the lines

its stupid the post 9-11 rules
if any 1 of us did 9/11 we would have been caught in days.
NetBios: scanning ip 203.169.114.66

More metacharacters: |

This does not mean “pipe” in the context of regular expressions; instead it translates to “or”; we can use it to combine two expressions, the subexpressions being called
alternatives

flood|fire

will match the lines

Horses are forbidden to eat fire hydrants in Marshalltown, Iowa. was chilly here too but with fire hot hot hot waiting for goverment handouts when the basement floods puts flood pants on i case it happens again

More metacharacters: .

In my original grep example, I had us look for

203.145.158.2

What could it have matched? Why did this work in our context?

More metacharacters: |

We can include any number of alternatives...

flood|earthquake|hurricane|Starfire

will match the lines

Bush’s earthquake causing lever is right beside his hurricane causing lever earthquakes are like our tornados I guess
Global Warming is causing Tsunamis and hurricanes and Earthquakes...
FL usually gets some hurricanes.. we were lucky this time...
puts flood pants on i case it happens again
Not everyone gets it off work Starfire
More metacharacters: |

The alternatives can be real expressions and not just literals

\[^{Gg}ood|{Bb}ad\]

will match the lines

good to hear some good news from someone here
Good afternoon fellow American infidels!
good on you—what do you drive?
Katie... guess they had bad experiences...
my middle name is trouble, Miss Bad News

More metacharacters: ?

The question mark indicates that the indicated expression is optional

\[^{Gg}eorge\(\[Ww]\.)?{Bb}ush\]

will match the lines

i bet i can spell better than you and george bush combined
BBC reported that President George W. Bush claimed God told him to invade Iraq
a bird in the hand is worth two george bushes

More metacharacters: ( and )

Subexpressions are often contained in parentheses to constrain the alternatives

\[^{Gg}ood\(\[Bb]\)ad\]

will match the lines

bad habit
bad coordination today
good, because there is nothing worse than a man in kinky underwear
Badcop, it's because people want to use drugs
Good Monday Holiday
Good riddance to Limey

One thing to note...

In the following

\[^{Gg}eorge\(\[Ww]\.)?{Bb}ush\]

we wanted to match a “.” as a period; to do that, we had to “escape” the metacharacter, preceding it with a backslash

In general, we have to do this for any metacharacter we want to include in our match
More metacharacters: * and +

The * and + signs are metacharacters used to indicate repetition; * means “any number, including none, of the item” and + means “at least one of the item”

\( (.*\) \)

will match the lines

anyone wanna chat? (24, m, germany)
hello, 20.m here... (east area + drives + webcam)
(he means older men)
()

working as MP here 720 MP battallion, 42nd birgade
so say 2 or 3 years at colleage and 4 at uni makes us 23 when and if we finish
fixing to spend over $15,00 to fill two 2.5 gallon gas jugs
it went down on several occasions for like, 3 or 4 *days*
Mmmm its time 4 me 2 go 2 bed

More metacharacters: { and }

{ and } are referred to as interval quantifiers; the let us specify the minimum and maximum number of matches of an expression

\([Bb]\text{ush}( +[^ ]+ +\}{1,5} \text{ debate} \)

will match the lines

Bush has historically won all major debates he's done.
in my view, Bush doesn't need these debates..
bush doesn't need the debates? maybe you are right
That's what Bush supporters are doing about the debate.
Felix, I don't disagree that Bush was poorly prepared for the debate.
indeed, but still, Bush should have taken the debate more seriously.
Keep repeating that Bush smirked and scowled during the debate

More metacharacters: * and +

The * and + signs are metacharacters used to indicate repetition; * means “any number, including none, of the item” and + means “at least one of the item”

\([0-9]+ (.*[0-9]+\)

will match the lines

working as MP here 720 MP battallion, 42nd birgade
so say 2 or 3 years at colleage and 4 at uni makes us 23 when and if we finish
fixing to spend over $15,00 to fill two 2.5 gallon gas jugs
it went down on several occasions for like, 3 or 4 *days*
Mmmm its time 4 me 2 go 2 bed

More metacharacters: { and }

{m,n} means at least m but not more than n matches
{m} means exactly m matches
{m,} means at least m matches
More metacharacters: ( and ) revisited

In most implementations of regular expressions, the parentheses not only limit the scope of alternatives divided by a "|", but also can be used to “remember” text matched by the subexpression enclosed.

We refer to the matched text with \1, \2, etc.

Next time

This gives us a pretty rich language to start thinking about patterns and how to extract them from files.

On Wednesday, we will start at 4:30 and we will have a guest lecture back in the regular room (Haines A25); our speaker will be Neal Richman who will describe a community mapping too he created.

More metacharacters: ( and ) revisited

So the expression

\+([^a-zA-Z]+) +\1 +

will match the lines

- blah blah blah blah
- i was standing all alone against the world outside...
- hi anybody anybody at home
- what was that movie with with brad pitt & tom cruise
- ha ha ha bionicwoman, how observant