Week 4

Announcements

- Advanced 3 and Basic 3 are out
- Lecture 3 survey is closing today

Review + Regular Expressions

Lecture 4

Overview

1. Bash review
2. Regular expressions

Bash review

- Stringing commands together
- Pipelines
- Redirection
- Expansion
- Quoting
- Control flow
- Functions
- Scripts

Regular expressions (regexes)

- A pattern that matches a set of strings
- Provide a (relatively) standardized way to perform matches on text
- Important to know as ***many*** tools and utilities make use of them
  - `grep`, `sed`, `find` to name a scant few
- Lots of different flavors, but they all encapsulate similar ideas
- You provide a **pattern** that is matched on the text
- The **pattern** can be a simple unassuming string or contain special characters that perform more powerful matching
- For this lecture, we'll be looking at POSIX BRE (basic regex) and ERE (extended regex)
  - `grep` is a utility that searches for patterns in a file via regexes
  - Defaults to BRE; `-E` flag (or `egrep`) for ERE
  - `ls /dev | grep tty`: list `/dev` directory, filtering by things that contain "tty"

Resources

- Online regex tester: https://regex101.com/ (one among many)
  - Can provide a breakdown of the regex
  - (`grep` can serve as an offline tester as well)
- GNU `grep`'s manual on regular expressions
- Highly detailed website: https://www.regular-expressions.info/

Regex basics

- Patterns are composed of smaller regexes that are concatenated
- The atomic regexes are those that match single characters
- The alphanumeric characters (A-Z, a-z, 0-9) act like normal characters
  - `hello` is a simple pattern that matches strings that contain "hello"
- There are also special functions denoted by special characters
  - `.` for any single character
  - `|` for an OR
  - `\` for special expressions/escapes
  - Quantifiers: how many to match
  - Brackets: a set of characters to match
  - Anchors: for *positional* matching
  - Backreferences: for matching a previous match
  - `^tty[0-9]+$` is a less simple pattern that matches lines that exactly compose of only "tty" and some numeric digits after it

Misc special characters

- `.` matches *any* single character
  - `...` matches strings containing three characters
- `|` for an OR between regexes
  - `hello|world` matches a string containing "hello" or "world"
- `\` for special expressions/escapes
  - `\b` matches empty string at the edge of a word
  - There's more: check the GNU `grep` manual for the rest
- `(`, `)` enclose a whole expression as a *subexpression*
  - `(Hello|Goodbye) (Arav|Sowgandhi)` matches:
  - "Hello Arav"
  - "Hello Sowgandhi"
  - "Goodbye Arav"
  - "Goodbye Sowgandhi"

Quantifiers

- Specify how many of a preceding regex to match
- `?`: ≤1 time
- `*`: ≥0 times
- `+`: ≥1 times
- `{n}`: *n* times
- `{n,}`: ≥*n* times
- `{,m}`: ≤*m* times
- `{n,m}`: ≥*n* and ≤*m* times

Examples

- `a{4}`: matches "aaaa"
- `ba+`: matches "ba", "baa", "baaa"...
- `(hello){3}`: matches "hellohellohello"

Brackets

- `[`, `]` enclose a set to match for one character
  - `[abc]` matches 'a', 'b', or 'c'

Special things you can put inside them:

- `-`: range
  - `[A-Za-z0-9]`: capital and lowercase numbers and digits
- `^`: not in set
  - `[^ab]`: everything not 'a' or 'b'
- Named classes
  - `[:alnum:]`: alphanumeric characters
  - `[:alpha:]`: alphabetic characters
  - `[:digit:]`: digit characters
  - `[:blank:]`: space and tab characters
  - ...and others (see the GNU `grep` manual)
  - Brackets are part of the class name: e.g. `[[:alnum:]]` to match alphanumerics

Anchors

- Perform *positional* matching
- `^`: match empty string at the beginning of a line
  - i.e. following regex must be at the beginning
  - `^hello`: "hello" must be at the beginning
- `$`: match empty string at the end of a line
  - i.e. preceding regex must be at the end
  - `world$`: "world" must be at the end
- `^hello world$`: entire string must be "hello world"

Backreferences

- Match previous parenthesized `()` subexpression
- `\n`: match *n* th parenthesized subexpression
  - `(123)testing\1` matches "123testing123"

Q: `<([[:alpha:]][[:alnum:]]*[^>])>.*</\1>`

- Match (simple) HTML/XML tags

Caveats

- GNU `grep` defaults to BRE flavor
  - Use `-E` flag or use `egrep` for ERE flavor
  - In ERE mode, use `[{]` to capture literal '{' for portability
- Other flavors may require escaping certain characters

BRE vs ERE

- In BRE `?`, `+`, `{`, `|`, `(`, and `)` must be escaped with `\`

Any other questions?