



PennState

# CMPSC 311 - Introduction to Systems Programming

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Regular Expressions

Add WeChat edu\_assist\_pro

Professors:

Suman Saha

(Slides are mostly by *Professor Patrick*

*McDaniel* and *Professor Abutalib Aghayev*)

# Regular expressions



PennState

- Often shortened to “regex” or (more rarely) “regexp”
- Regular expressions are a [language for matching patterns](#)
  - Super-powerful find and replace tool
  - Can be used on the command line, as a text editor feature, or as part of a program language

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# What are they good for?



- Searching for specifically formatted text
  - Email address
  - Phone number
  - Anything that follows a
- Validating input
  - Same idea
- Powerful find-and-replace
  - E.g. change “X and Y” to “Y and X” for any X, Y

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# Regex “flavors”



PennState

(source: <https://gist.github.com/CMCDragonkai/>)

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# On the command line



PennState

- The grep command is a regex filter
  - That's what the “**re**” (regular expression) in the middle stands for
  - grep **-F** looks for literal strings
- Today we will use grep
  - E for “extended” regular expressions
  - Very close to other languages' flavors
  - Or egrep

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# grep command syntax



PennState

- To find matches in files:
  - `grep -E regex file(s)`
- To filter standard input:
  - `grep -E regex`
  - where `regex` is a regular expression and `file(s)` are the files to search
- Options (aka “flags”):
  - `-i`: ignore case
  - `-v`: find non-matching lines (inverted search)
  - `-r`: search entire directories
  - `-l`: list the files that match, not the lines

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# Playing along ...



PennState

- First, go to the course website and grab the `grep-text.txt` file and put it in your VM (it is under the files tab of canvas)
  - **E1\***: Now test it out:

Assignment Project Exam Help

<https://eduassistpro.github.io/>

```
% grep -E this grep-text.txt
this is a text file
this is a text file text file
%
```

Add WeChat edu\_assist\_pro

\*These numbered exercises are used throughout this lecture to help illustrate regular expression use. Answer key given later.

# First lesson



PennState

- Letters, numbers, and a few other things match literally

- **E2**: Find all the lines containing “fgh”

- **E3**: Find all the lines containing “lmn”

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

- Note: a regex can match **anywhere** in the string
  - Doesn't have to match the whole string



# Anchors



PennState

- Caret `^` matches at the beginning of a line
- Dollar sign `$` matches at the end of a line

- Use `"..."` to protect characters from the shell, e.g.,

```
grep -E "blah" https://eduassistpro.github.io/
```

- Try it

- **E4**: Find words ending in `"gry"`
  - **E5**: Find words starting with `"ah"`

- What happens if we use both anchors?

# Single-character wildcard



PennState

- Dot . matches any single character (exactly one)
  - E6: Find the lines line with 6-letter word where the second, fourth, and sixth letters are "c"  
Assignment Project Exam Help
  - E7: Find lines with word end with "at" with example middle  
<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# Multi-character wildcard



- Dot-star `.*` will match 0 or more characters
  - We'll see why on the next slide
  - **E8**: Find any words that start with `gr` and end with `at` with any number of characters in the middle

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# Quantifiers



PennState

- How many repetitions of the previous thing to match?

- Star `*`: 0 or more

- Plus `+`: at least 1

- Question mark `?`: 0 or

- Brackets `{#}`: exactly

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

- Try it out

- **E9**: Spell check: `necc?ess?ary`

- **E10**: Outside the US: `colou?r`

- **E11**: Find lines with words with `u`, `o`, `i`, `e`, `a` in that order and at least one letter in between each

# Character classes and spaces



PennState

- Spaces can be specified with “\s”
- Square brackets [abc] will match any one of the enclosed characters

- You can use quantifiers on char

- E12: Find words st the rest of the letters are c, h, or s

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# Ranges



PennState

- Part of character classes
- You can specify a range of characters with `[a-z]`
  - One hex digit: `[0-9a-f]`
  - Consonants: `[b-df-h]`
  - **E13**: Find lines that have a character that can be made exclusively with `a` through `e`.

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# Negative character classes



PennState

- If the first character is a caret, matches anything except these characters

- Consonants: `[^aeiou]`
- Can be combined with `r`
  - Any character that i
- **E14**: Find words that contain “`aq`”, followed something other than “`u`”

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# Groups



PennState

- Parentheses (...) create groups within a regex
  - Quantifiers operate on the entire group
  - E15: Find words with an "ill", followed by "ach" one or more times, followed by "e"
  - E16: Find words where the first, is an "e"

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro



# Branches



- The vertical bar | denotes that either the left or right side matches
  - It's the “or” operator
  - Useful inside parentheses
  - **E17**: Find lines with the

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# Special characters



PennState

- We've seen a lot already
  - `^$. *+?[ ]() | \`
- Backslash `\` will escape a special character to search for it literally
  - **E18**: Search for all line

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# Backreferences



PennState

- Groups in ( ) can be referred to later
  - Must match the exact same characters again
  - Numbered \1, \2, \3 from the start of the regex
  - E19: Find lines that have a sequence repeated immediately

Assignment Project Exam Help

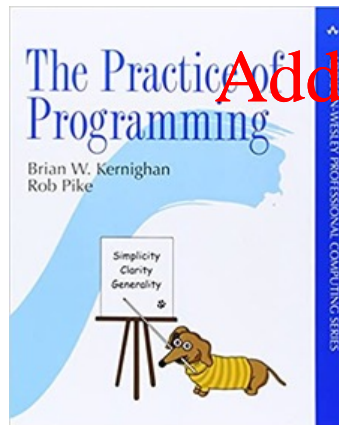
<https://eduassistpro.github.io/>

Add WeChat edu\_assist\_pro

# Want to learn more?



- [regexcrossword.com](https://regexcrossword.com)
  - Great way to practice your regex-fu
  - Starts with simpler tutorial puzzles and works up
- Write your own re <https://eduassistpro.github.io/>



Add WeChat edu\_assist\_pro