An Introduction to Regular Expressions

BIOS 669

What is a regular expression?

- A character pattern used for searching and matching
- Can be used interactively or in an automated fashion

Uses of regular expressions

Text editing

Text searching

Textual analysis

Uses of regular expressions

Text editing – for more complicated replacements than simply one word for another

- It's simple to replace all instances of variable RACE in a program with variable ETHNICITY
- It's harder to replace all references to variable names with the pattern A followed by a number (A<n> or A<nn>) with B followed by that same number (B<n> or B<nn>)
- TextPad, UltraEdit, Notepad++, TextWrangler, SublimeText, etc.

Uses of regular expressions

Text editing – for more complicated replacements than simply one word for another

Text searching – could help with web scraping

- Web scraping means pulling information from the complicated text file that defines a web page

Uses of regular expressions

Text editing – for more complicated replacements than simply one word for another

Text searching – could help with web scraping

Textual analysis – could help with categorization

Uses of regular expressions

Text editing – for more complicated replacements than simply one word for another

Text searching – could help with web scraping

Textual analysis – could help with categorization

- ICD9 code vs. ICD10 code (sample ICD code: 410.x)

Uses of regular expressions

Text editing – for more complicated replacements than simply one word for another

Text searching – could help with web scraping

Textual analysis – could help with categorization

- ICD9 code vs. ICD10 code
- Medical record categorization based on notes

Uses of regular expressions

Text editing – for more complicated replacements than simply one word for another

Text searching – could help with web scraping

Textual analysis – could help with categorization

- ICD9 code vs. ICD10 code
- Medical record categorization based on notes
- Literary analysis (who is the author of this text?)

Why cover regular expressions in this course?

- So that you know regular expressions exist as a possible tool – as with SQL, they are available not just in SAS but in many other programming languages (R, Python, etc.)

- To provide you with practice with regular expressions so that if you have a need for them in the future, you won’t be starting from nothing

A few simple regular expressions

/HUNT/	finds the sequence of characters HUNT
/[HUNT]/	finds any of the characters H, U, N, or T
/\d/	finds any digit 0–9
/^\d/	finds any digit 0-9 at the very beginning of the text
/[0-9]/	finds any digit 0-9

An important principle to remember:

Your goal is to write regular expressions that both find all cases that you are looking for AND omit all other cases.

Example: In searching for all strings matching the pattern of a Social Security Number (nnn-nn-nnnn), you shouldn’t just look for three digits followed by a dash because this would also pick up phone numbers where the area code is written as three numbers followed by a dash.