DB Management Systems
Regular Expressions

Joel Klein – jdk514@gwmail.gwu.edu

Regular Expressions - Regex

- Regular expressions are character patterns that define a set of valid character sequences.
  - Essentially a means to define pattern matching for text/strings

- Regex logic is usually implemented in programming languages and text editors
  - Python has the **re** package
  - Visual Studio Code we can enable **find/replace** with regex capabilities

Implementation Specific Syntax

- Regexes can take different forms depending on the system that is implementing them:
  - Mongo - /pattern/
    - Ex – to find 'hello' in a sequence of text we would use '/hello/'
  - Python – pattern
    - re.complie(*{pattern}*) – compiles a pattern for use
    - re.match*({pattern}, {str}*) – finds the first occurrence of a pattern

Power of Regex

- Regular expressions aren't really used for finding simple patterns like 'hello', rather they are used for describing complex patterns
  - A lot of email validation on forms leverage regex
- Regex solve the problem of describing a pattern of characters used to represent more abstract concepts
  - What does a phone number look like, what does a datetime string look like, etc.
- To enable these use cases, regex leverage special characters/sequences to represent these patterns

Special Characters

Brackets

- Brackets define sets of valid values:
  - **[]** define what characters are acceptable at a given location in the pattern:
    - [abc] – accept a, b, or c
    - [0-9] – accept any numeric value
  - **()** define alternative patterns to match
    - (abc|123) find abc or 123

Metacharacters

- Metacharaters provide ways to describe more generalized concepts

- Some examples:
  - . – match any character
  - \w – match a word character ([a-zA-Z0-9_]
  - \t – a tab
  - \n – a newline

Quantifiers

- Quantifiers are used to describe the number of repetitions that are allowed for a given pattern. Some examples:
  - n+ - match one or more instances of n
  - n* - match zero or more occurrences (typically used as .* to mean match anything)
    - Ex - '.*hello.*' would find any text that contains 'hello'
  - ^n - match any occurrence that starts with n

End Slide

EMSE 6992 – DBMS for Data Analytics