Conceptual Overview

Regular Expressions (Regex) is syntax and format for describing patterns and formats for text data. Once you can describe a pattern or format using regex, you can use it for several useful text processing tasks, including:

- Validating that some text matches a given format
- Finding text that matches a given format within a larger body of text
- “Parsing” text into another structure that is easier to work with in code.

You can use regular expressions in almost any programming language, as most will come with the ability to work with regular expressions as part of a built-in library.

In this project, we’ll use regex to clean, read, and perform computations against some data that we read from a text file. The text file will include information about a ship’s navigation across a body of water, in a pre-defined format.

Prerequisites

Students should:
- Have familiarity with File I/O, reading text files in Java
- Have familiarity with enums
- Have familiarity with importing classes from the Java standard libraries

Students should know:
- How to create Classes and instantiate Objects.
- How to work with Strings and String methods
- How to create, iterate over, and manipulate Java Lists (and ArrayLists)
- How to call and pass data to methods.
- How to run code from a main method

Learning Objectives

Students will learn:

- Reading data from files into strings
- Parsing strings into classes using regex
- Regex for digits
- Regex for words
- Regex capture groups
- Regex named capture groups
- Regex for whitespace
- Regex quantifiers

Guided Project Activity

Introduction:

In this lesson, we will read information from a file containing information about a ship's navigation from Palermo in Sicily, Italy, to the island of Malta. Using regex, we'll extract the important information from this file to visualize what route it took, how far it traveled, and how long it took. Students will learn to use different parts of the Java regex implementation to search and extract different kinds of data out of string input. The project will build iteratively, starting from basic regex matching capabilities like digit, word, and whitespace matching, and extending to named capture groups to make parsing the data into classes more easy.

Running the Lesson:

1. Introduce regular expressions
  - Describe how regular expressions are used with real-world examples
  - Describe different kinds of input data files that regex can be used to work with
  - Explain how regex is implemented in different programming languages, and introduce the java.util.regex library
2. Show some starter code that contains a class with some utility methods to read data from a file into a string.

- Run the application and print the content from the file
- Explain the code required to strip whitespace from a file, and have the learner clean the input data accordingly.
- Explain to the learner how a List will be used to store the individual “headings” in the input data.

3. Show the learner an example that helps them to parse digits and strings out of a file
- Have the learner print the digits and strings separately as they iterate over their list of headings (cleaned data)

4. Converting the individual lines into instances of a class.
- Learner uses capture groups and named capture groups in this stage.

5. Perform computations over the list of headings, and explain how this is easier with classes than it would have been with the raw input data.

The rest of the lesson could discuss:

- other good use cases (maybe even some publicly available datasets) for using regex to parse information (data visualization, compression, computation)
- extensions for advanced students e.g. plot the ship’s path, update the regex to be able to read extra information from the file (fuel levels)