SI 330 Fall 2019 Bonus Assignment

Goal

The purpose of this project is to allow you to go above and beyond the course material and your demonstration of a single particularly challenging topic - Regular Expressions. This assignment is worth up to 5% of a bonus grade, and thus can allow you to achieve an A+. In addition, for students who do better on this assignment than the did on the midterm, their midterm grade can be discarded and replaced with the score from this assignment. I will only take the higher of the grade average when calculating your final grade in the course.

Overview

Regular expressions are hard to learn, and even once learned they are hard to remember if you don't practice them regularly. I want you to help me change that, and in doing so, demonstrate your deep knowledge of regular expressions. This will be done through two different activities:

Activity One: Worked Tutorials

A common pattern in data science education is the notion of the worked tutorial. This is commonly seen in medium-style posts, where an author picks a topic, a dataset, and a library, and demonstrates how the topic is explored in the dataset using the library. I want you to write four different computational narratives (Jupyter notebooks) teaching regex fundamentals in python. Each notebook should address one of the following topics (topics marked with a ★ are topics we did not cover in class, and I would be especially impressed if you did tackle them):

a) Matching characters within sets []. This should cover both set matching, negation in sets, and ranges in sets.
b) Metacharacters. This should include at a minimum \d \D \w \W \s \W.
c) Capture groups. This should include both regular capture groups (), named capture groups (?P<name>), and non-returning capture groups (?:).
d) Multiplicity and cardinality. This should include all of *?+ and {m.n}.
e) ★ Positive and negative look ahead assertions. This should include (?=) and (?!).
f) ★ Positive and negative look behind assertions. This should include (?<=) and (?<!).
g) ★ Backreferences. This should include (?P=name).
h) ★ Backtracking and greediness.

Each computational narrative should include one or more paragraphs of discussion in markdown text, followed by some interesting data where this might be demonstrated (with a preference for real data from wikipedia, please cite source regardless!) followed by two examples of using the concepts you described in python code. This should be followed by two questions which could be asked to students learning this concept to test their knowledge, along with your worked solutions. Each notebook must end with one of the reuse statements indicating whether I can reuse your notebook to help teach others or not. Approval or not has no bearing on your grade.

I have included an example notebook demonstrating what I expect (using the simplest aspect of regex, matching specific strings, this is not a choice in the above list!).

This activity must be handed in by December 12th, midnight.

Activity Two: Learning Together

In this activity, which opens on December 13th, you are going to be given eight notebooks to learn from which your peers have put together. You must complete those notebooks, specifically you must:

1. Attempt both problems in the solution space which is given. The notebooks you will be given will have the solutions removed from them. Your solutions must be reasonable and demonstration knowledge of the

content, however, if this was a notebook testing a technique from outside of the course (marked with a ★ above) then your solution doesn't have to be correct but a solid demonstration of effort.

2. Write a short paragraph at the end of how the notebook could be improved, be it the examples, the problem sets, or the description itself.
3. Provide a score out of 100% as to how reasonable this notebook was. If you handed this notebook in to me, what evaluation do you think would be appropriate? Use the following rubric in your own grading
   a. 20% Introductory discussion (the "lesson") and overall quality/professionalism
   b. 40% The examples (Are they easy to follow? Are they interesting examples? Do they demonstrate the lesson appropriately?)
   c. 40% The problem sets (Are the questions clear? Do they test the material? Are they sufficiently difficult?

This activity must be handed in by December 18th, midnight.

Grading

The grading rubric for this assignment is as follows:

1. 25%: Your solutions to the four notebooks you were assigned, as evaluated by the course grader(s)
2. 25%: Your feedback on the four notebooks you were assigned, as evaluated by the course grader(s)
3. 50%: The average of the scores assigned to you by your peers for your notebooks