Lab 1: Regex and Evil Doings

September 14th 2017

In this lab you'll be parsing a book into xml. Turn this assignment in by midnight on Wednesday, 9/20.

The Situation

You've been forced to take an English Literature class in order to fulfill some arcane requirement. This displeases you. You didn't become a computer scientist in order to read English literature, did you? DID YOU?! No. You did not. So when your professor informs you that you'll be reading The *Blithedale Romance* by Nathaniel Hawthorne, you decide that drastic measures are in order.

The Assignment

You're not yet sure what your evil plan is (because you're only two weeks into your NLP course), but you know you need to be ready. And that means parsing the book into a more analytics friendly format, namely XML. You'll be using a combination of regular expressions, rule-based reasoning, and plain old manual effort. It is up to you to decide which of these tactics makes the most sense for each of you goals.

Part 1: 50 points

The book, and thus the corresponding quizzes and discussions, will likely be assigned in chapter increments. Your first goal is to capture this, and other major structural components. This includes the following XML tags:

```
<book>
<booktitle>
<author>
<chapter>
<chaptertitle>
<paragraph>
```

Part 2: 30 points

Surely you'll be tasked with writing some sort of deeply probing paper on the inner workings of Nathaniel Hawthorne's characters. And such a paper needs some juicy quotes straight from the horse's mouth. You'll need to augment your XML to include `<quote>` tags that capture the book's dialog. Thus:

"Mr. Coverdale," said he softly, "can I speak with you a moment?"

becomes...

`<quote>`Mr. Coverdale,`<\quote>` said he softly, `<quote>` can I speak with you a moment?`<\quote>`

Part 3: 20 points

Be clever. Incorporate additional structure of your own devising that will help you automate your way through English Lit. While this is, by design, an open ended task, please consider the following three points as guidance:

1. Think about a typical English class comprised of reading quizzes, discussions, and papers. What sort of eventual analytics might be useful? What are the preprocessing steps that would need to happen to enable those analytics? Focus on the latter.
2. We talked about some very common preprocessing steps in class this week. It would be reasonable to assume that this was not a coincidence.
3. The points to be earned in this lab are intended to be proportional to the effort required to earn them. There is no need to invent a new sub-field of machine learning here.

Step 7

In a single zip file, include your code and a text file describing (in English) the steps you took for Part 3. Submit this using the command:

provide comp150nlp regex regex.zip