CS671A: Introduction to NLP
Assignment #1: Text normailization, regular expressions

Due on: 30-1-2018, 23.59 20-1-2018
MM: 200

1. (a) Design and implement Python programs using the re module for the following text normalization tasks:
   1. The file `test.txt` contains an extract from Chesterton's 'The Man who was Thursday'. At several places in the text there is conversational text enclosed in single quotes. Single quotes are also used at other places in the text. Your program should identify converstational text and replace enclosing single quotes by double quotes.
      For example: 'An artist is identical with an anarchist,' he cried. should become "An artist is identical with an anarchist," he cried.
   2. Build a sentence terminator recognizer and tag each sentence in the file test.txt with a beginning of sentence and end of sentence tag as follows:
      `<s>'An artist is identical with an anarchist,' he cried.</s>`

   (b) In this question you must build a learning based sentence terminator classifier using the context around the terminator character (a punctuation character). You can use any binary classification algorithm available in the Python scikit-learn library (scikit-learn.org). You should be able to do this even if you have not done an ML course. Consult with your TA if you have difficulty. Invent a way to create a feature vector from the context of the punctuation character. Use the output of the sentence tagger in question 1(a)2 as the labelled set. You can use a small part of the tagged data for testing your model. If you want a larger tagged data set run your sentence tagger on file `fullTest.txt` which is the full text of Chesterton's book. Report the accuracy of your model.

[(50,50),100=200]