Saransh Bhatnagar (150637)

Q1.(a)
1. substituted ‘ with “ which occur at the word boundary, i.e either at the starting or at the end of a word. Therefore the regex itself takes cares of apostrophe.

Q1.(a)
2.
-- allocating </s>
first regex helps in allocating </s> at the end of a sentence which ends with ‘.’ or ‘?’. Various checks used are for:
(?<![A-Z]\.) for A. , B. (kind of words); (?<![A-Z][a-z][a-z]\.) for Mrs. kinds ; (?<![A-Z][a-z]\.) for Mr. , Dr. kind of words and a negative lookahead for ‘

Second regex deals with the sequence that ends with someone saying a sentence therefore will have a \n succeeded by the single inverted commas enclosed sentence itself.

-- allocating <s>
Now when the sentence starts with the end of another sentence therefore <s> will be just after </s> except for the boundary cases which are handled explicitly in the code successively.
The \n between <s> and <\s> are dealt in the regex itself and are substituted accordingly

Additionally removing \n which occur at random places in between a sentence. Therefore a regex that matches with <s> words + spaces <\s> is used to make and array and then \n is replaced with ‘ ’ and \n\n is replaced with \n one by one and hence the output created is written in output file.

Q1.(b)
Code from the Q1.a 2 is used to make array of sentences and inbuilt functions from scikit learn library are used to make a vocabulary of words and also to create one hot vector.
There are three kind of feature vectors used to train the classifier but only two kind are used for testing.
Logic used is that certain words occur more frequently at the end of a sentence than that at the start or at the middle.
1st: taking three words from the end of a sentence i.e before sentence terminator and creating a one hot vector out of it.
2nd: taking three words before a punctuation character that is not a sentence terminator.
3rd: to further train the classifier, a vector of three words from the starting of each sentence is used(using the fact that words used in the start would have less probability to be at the end of sentence)

The efficiency calculated for svm and for logistic classifiers were around 85% and is larger if larger training set is provided.

Set corresponding to punctuation that are not sentence terminator are small therefore there is certain bias which can be rectified with larger training set and is quite visible when fulltext.txt is used.
1’s represent that the punctuation is a sentence terminator and 0’s represent that they are not. The input output set generated is divided into ⅔ training and ⅓ testing for 1’s and 0.9 training and 0.1 testing in the script the corresponding parameters of the classifiers are adjusted for better outputs.