Assignment 1

CS671

Jayant Agrawal

14282

Problem 1 (a)

1. Using Regular Expressions, find all non-conversational single quotes (*‘s*,*’t, I’m, I’ve* etc).
2. Replace all such instances with ‘$’, temporarily.
3. Replace all the other single quotes with double quotes.
4. Replace the ‘$’ symbols back with single quotes.
5. Print the output.

Running Instructions: python p1a.py <source_file> <output_file>
Source File: test.txt
Output File: out_p1_a.txt

Problem 1 (b)

1. Identify all the instances of sentence terminators with the following conditions using a regular expression.
    a. The previous alphabet must be in lowercase.
    b. The next character must be one of the following: *space, double quote, newline*
    c. The next alphabet after one of the above characters must be in uppercase.
2. Split the entire text using the indices retrieved above.
3. Add the sentence markers.

Running Instructions: python p1b.py <source_file> <output_file>
Source File: out_p1_a.txt
Output File: out_p1_b.txt

Problem 2

1. Extract all instances of potential sentence end markers (‘.’, ‘?’, ‘!’).
2. For each such instance, make a binary feature vector with the following features:
    a. Whether the previous character is a lowercase alphabet.
    b. Whether the next character is a *space.*
    c. Whether the character after the next character is an uppercase alphabet.
3. The labels for each marker can be extracted using the solution of *problem 1 (b).*
4. Split the *fullTest.txt* in a 3:2 ratio to make the training and the testing set respectively.
5. Fit an SVM using the above features and labels.
6. Final Accuracy on Test Set : 94.50%.

Running Instructions: python p2.py <train_file_path> <test_file_path>
Train File: fullTrain.txt
Test File: fullTest.txt