**CS671A: Introduction to Natural Language Processing**

**Assignment 1: Explanation sheet**

Name: Piyush Bagad

Roll no.: 150487

**Q1(a)(1):** I have handled conversational quotes under the following two cases:

1. Type:(, or newline or full stop) 'Capital letter -----(some ending)'
2. Type:, 'small letter -----(some ending)'

I have used ‘,’ before the quotes with a small letter as the first character so as to exclude examples like: He conveyed a big ‘thank you’ to all of us.

**Q1(a)(2):** I have tried to capture the end of each sentence as a ‘. ’ or “.’” or ‘? ’ or “?’” or ‘! ’ or “!’” followed by a capital letter or newline.

**Q1(b)(1):** For each of the punctuations, I have written a program that learns (using standard libraries) a vector of fixed length which basically captures the context (surrounding) around the punctuation. For a new input, given a punctuation and its context (surrounding), the program predicts whether the punctuation is at a termination of a sentence or not. Description of my approach is described below:

General structure of Feature vector for each of the punctuations:

I used a **7-dimensional feature vector** for each punctuation.

And denotes the value of character at position in the raw string. The values chosen for different characters and set of possible characters that can appear in surroundings of a punctuation are specified below. I have not considered separate features for characters at distance more than 3 from the punctuation because they will mostly be redundant and will not reveal too much information.

Character set = { capital letter, small letter, space, inverted comma (quote), newline }

|  |  |
| --- | --- |
| **Character** | **Value** |
| Capital letter | +2 |
| Small letter | -1 |
| Space | +3 |
| Inverted comma | +4 |
| Newline | +5 |

PERFORMANCE:

The following methods were used for the binary classification. Performance details are tabulated as follows:

1. Multi-layered Perceptrons

|  |  |
| --- | --- |
| **Punctuation** | **Accuracy** |
| . | 99.32 |
| ! | 98.50 |
| ? | 94.74 |

1. SVM

|  |  |
| --- | --- |
| **Punctuation** | **Accuracy** |
| . | 99.15 |
| ! | 98.51 |
| ? | 93.42 |