91.542 - Natural Language Processing
Problem Set 1

Chang Liu
chang_liu@student.uml.edu

September 24, 2015

1 Part I. Problem Setup, Language Modeling

1.1 Problem 1.[40 pts]

I).a) the instances are the 300 articles / 100,000 text lines / 900,000 words, each article is an instance.

b) the labels are the '<article>' and '<\article>'

c) 300 * $\frac{2}{3}$ = 200

d)

II).

a) the instances are the 300 articles, each article is an instance.

b) There're two labels I want to assign, the first one is the 'article', the second one is the 'title'. By using these two labels, I can select out each article and its title with the marker '<article>' and '<title>'. I think maybe the problem is not very clear or my understanding is correct, but if we have already known each separated article without worrying about dividing them, then the classifier is just to predict the '<title>' labels.

c) 200

d) The five boolean features for the label '<title>' can be:

1) Whether they're all upper characters, form 'A' to 'Z'.

2) Whether they're the first word in the paragraph.

3) Whether they're followed by comma.

4) Whether their previous line is the blank line.

5)