Natural Language Processing - Project 1

CSE 398/498-013

Name: Shihao Jing

Email: shj316@lehigh.edu

**2 Questions**

Answer the following two questions in your report

• Show that it makes sense to set C ∗ (w, v) = N1/ N for those unseen tokens ((w, v) with C(w, v) = 0).

*The probability mass for unseen tokens is .*

• Calculate the probability mass reserved for the unseen tokens when GT smoothing is used, and compare the mass to the mass reserved when Laplacian smoothing is used.

*Laplacian smoothing:*

*for ,*

*GT smoothing:*

For data given,

*Laplacian smoothing:*

*GT smoothing:*

Figure: Plot of frequencies of frequencies in the log scale, with a line fitted to the points.

