CS221 Fall 2018 Homework 3

SUNet ID: 05794739
Name: Luis Perez
Collaborators:

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Problem 1

1. To show that this greedy algorithm is suboptimal, simply consider the following set-up. First, we work with a 1-gram model, and consider the input:

   “thesecount’

   The greedy algorithm will compare the following on the first iteration:

   $$\begin{gathered} u(\text{“t”}) \\ u(\text{“th”}) \\ u(\text{“the”}) \\ u(\text{“thes”}) \\ u(\text{“these”}) \\ u(\text{“thesec”}) \\ u(\text{“theseco”}) \\ u(\text{“thesecou”}) \\ u(\text{“thesecoun”}) \\ u(\text{“thesecount”}) \end{gathered}$$

   From the above, it’s reasonable to have our 1-gram model such that $u(\text{“the”}) < u(\text{“these”})$, and both of these will obviously have lower cost than the other non-English words. Therefore, on the first iteration, our greedy algorithm will select the split:

   “the secount″

   On the second iteration, the algorithm will consider:

   $$\begin{gathered} u(\text{“s”}) \\ u(\text{“se”}) \\ u(\text{“sec”}) \\ u(\text{“seco”}) \\ u(\text{“secou”}) \\ u(\text{“secoun”}) \\ u(\text{“secount”}) \end{gathered}$$

Note that non of these are English words, and therefore we define them to have extremely high cost. For the sake of simplicity, we'll have $u(\text{“secount”})$ have the lowest cost amongts the above, but still have an extremely high cost since it's not an English word. As such, the final output of our algorithm will be:

$$\text{“the secount”}$$

With cost $u(\text{“the”}) + u(\text{“secount”})$. However, note that the optimal split point would actually be:

$$\text{“these count”}$$

with cost $u(\text{“these”}) + u(\text{“count”})$. We note that:

$$u(\text{“these”}) + u(\text{“count”}) < u(\text{“the”}) + u(\text{“secount”})$$

mainly because of an extremely high cost associated with $u(\text{“secount”})$.