601.665 — Natural Language Processing
Assignment 3: Probabilities and Vectors

Aaron Mueller

26 September 2018

1. The per-word cross-entropy of a language model with add-0.01 smoothing built from `switchboard-small` on the `sample` test files are as follows:

$$H(p, q) = -\frac{1}{m} \sum_{i=0}^{m} p(w_i) \log(q(w_i))$$

Where p is the language model and q is the `sample` file. We already have the sum of the log probabilities per word, calculated by `fileprob.py`; thus, we can calculate H for all the sample files:

$$\begin{aligned}
H(p, \texttt{sample1}) &= -\frac{1}{m} \cdot -12121 \\
&= \frac{12121}{1686} \approx 7.189 \\
H(p, \texttt{sample2}) &= -\frac{1}{m} \cdot -7398.55 \\
&= \frac{7398.55}{978} \approx 7.565 \\
H(p, \texttt{sample3}) &= -\frac{1}{m} \cdot -7477.99 \\
&= -\frac{7477.99}{985} \approx 7.592
\end{aligned}$$

Then, for the per-word perplexity, we exponentiate 2 to the power of these cross-entropies:

$$\begin{aligned}
perplexity(p, \texttt{sample1}) &= 2^{7.189} = 145.9 \\
perplexity(p, \texttt{sample2}) &= 2^{7.565} = 189.4 \\
perplexity(p, \texttt{sample3}) &= 2^{7.592} = 192.9
\end{aligned}$$

When we build our language model instead on the larger `switchboard` corpus, we have slightly lower $\log_2$ probabilities (i.e., negative $\log_2$ probabilities with greater magnitude), which thus leads to higher cross-entropies and higher perplexities s well. This is because our language model built on the larger corpus has seen a greater number of word types, which thus slightly reduces the probability of seeing any given word type. Since we have these smaller probabilities, our $\log_2$

probabilities on the `sample` files will be somewhat more negative since our summation will see smaller probabilities that are put through a logarithmic function. Because of these more negative $\log_2$ probabilities, we also see greater cross-entropies since m, the number of words in the file, stays the same, while the numerator increases in magnitude. This higher cross-entropy then leads to higher perplexities upon exponentiation.

2.