The Unigram Term Frequency Distribution

Jason D. M. Rennie
jrennie@gmail.com

June 18, 2005

The unigram posits that each word occurrence in a document is independent of all other word occurrences. I.e. we can think of the document generation process as a sequence of dice rolls, where there is a fixed probability of occurrence associated with each word. The chance observing a given document is simply the product of the word probabilities. To calculate the chance of observing a given set of word frequencies, we must count all the possible orderings that achieve that set of frequencies. Let $\{x_i\}$ be the observed frequencies for a set of words. There are $\frac{(\sum_i x_i)!}{\prod_i x_i!}$ word arrangements that achieve that set of word frequencies. Hence, the likelihood of generating a document with that set of frequencies is

$$P\left(\{x_i\}\middle|\sum_i x_i = l\right) = \frac{l!}{\prod_i x_i!}\prod_i \left(\frac{w_i}{\sum_i w_i}\right)^{x_i}. \tag{1}$$

Note that the unigram is conditional on document length; the above gives the conditional likelihood of generating a particular set of frequencies given that their sum is l. The $\{w_i\}$ are the unnormalized word occurrence probabilities.

To find maximum likelihood weights for a document set, it is easiest to consider minimization of negative log-likelihood. Let x_{ij} represent the number of times word j occurs in the i^{th} document; let $l_i = \sum_j x_{ij}$. Then, the negative log-likelihood is

$$J = \sum_{i,j} \log(x_{ij}!) - \sum_i \log(l_i!) + \sum_i l_i \log\left(\sum_j w_j\right) - \sum_{ij} x_{ij}\log w_j. \tag{2}$$

To minimize this quantity, we find settings which give us a zero gradient with respect to the weights. The partial derivative with respect to a weight is

$$\frac{\partial J}{\partial w_j} = \sum_i \frac{l_i}{\sum_j w_j} - \sum_i \frac{x_{ij}}{w_j}. \tag{3}$$

Note that the setting $w_j = \sum_i x_{ij}$ gives us a zero gradient.