**Frequency Weighing in SAS Text Mining**

The Weiss, Indurkhya, and Zhang text does not provide details on frequency weighing as is done in the SAS Filter node. The ones included in SAS EM include entropy (default), inverse document frequency, and mutual information.

## Entropy

This is based on information theory as developed from work by Claude Shannon at Bell Labs to provide information on channel capacities and coding schemes. The basic formula is often used in thermodynamics. The weights, wt, where t is the term index, assigned by entropy are:

wt = 1 + sum[j = 1…n | (ftj/gt)\*log2(ftj/gt)/log2n]

Here gt is the number of times term t appears in the document collection and n is the number of documents in the collection. The product in the numerator is set to 0 if ftj = 0.

Entropy is 0 if the term appears only once in the entire collection and is 1 if the term only appears once in one document. This provides the end points of the entropy weight scale. Frequent terms get lower weights since they are likely to appear in many document which does not help in distinguishing documents.

## Inverse Document Frequency

This is based on the tf-idf as discussed in the text on page 25. The weight assigned is:

wt = log(n/dft) + 1

and is just idf adjusted for zero dft.

## Mutual Information

This is based on an old formula called lift which is usually the joint probability of two events divided by the product of the individual probabilities, the latter representing the probability of independent events. It is used with a category variable allowing some measure of how likely the frequency of a term helps in specifying to which category (ck) the document containing the term should belong. The formula used in SAS is:

wt = maxck[ log (P(t, ck)/(P(t)\*P(ck)) ]

Here t is the term index and ck is the category index. The argument of the log() is the lift. If the numerator of the log() is greater than the denominator then the term suggests the category to a greater extent than randomly and *vice versa*. The maximum of the log(lift) over all categories is the weight assigned to the term.