Not LSA

July 2015

1 July 2015

NotLSA – a way to do LSA-like things without actually using LSA (Latent Semantic Analysis). Two very low-brow approaches, maybe well-known in the industry; I have no idea. Both of these approaches attempt to automatically extract keywords from documents. What cool about this is that its ... unsupervised; requires no training, and is based on very simple, proven ideas. Obvious, even: compute the mutual information between pairs of things ... between words and documents, between words and word-pairs, etc. Heh.

But how do we do this? How do we compute the MI between a page of text, and a word? No way to answer this without diving into the details.

Text-keyword correlation

Lets take a text, say – 1000 pages of .. something. Some corpus. We want to compute the mutual information between the page itself, and the words on the page. We do this by analogy to MI of word pairs.

Call the k'th page g_k. Count the number of times that word w_m appears on this page; let this count be N_{mk}. Define $N_m = \sum_k N_{mk}$ be the total number of times that the work w_m appear in the document, and let $N = \sum_m N_m$ be the total number of words in the document. Then, as usual, define probabilities, so that

$$p_m = P(w_m) = N_m/N$$

is the frequency of observing word w_m in the entire corpus, and

$$p_{mk} = P(w_m|g_k) = N_{mk}/\sum_m N_{mk}$$

be the (relative) frequency of the same word on page g_k. Notice that the definition of p_{mk} is independent of the page size. Pages do not all have to be of the same size. Define the mutual information as

$$\text{MI}(g_k, w_m) = -\log_2 \frac{p_{mk}}{p_m} = -\log_2 \frac{N_{mk}N}{\sum_m N_{mk} \sum_k N_{mk}} = -\log_2 \frac{p(m,k)}{p(m,*)p(*,k)}$$

This is essentially a measure of how much more often the word w_m appears on page g_k as compared to its usual frequency. The highest-MI words are essentially the topic words for the page. The right-most form introduces a new notation, to make it clear that it resembles the traditional pair-MI expression. The notation is

$$p(m,k) = \frac{N_{mk}}{N}$$

so that

$$p(m,*) = \sum_k p(m,k) \qquad \text{and} \qquad p(*,k) = \sum_m p(m,k)$$

are the traditional-looking pair-MI values.

TODO: – this does not have the feature-reduction/word-combing aspects of LSA...

Variants

Instead of working with words, we could work with word-pairs, which is a stand-in for working with (named) entities. Thus, we can identify if a named entity occurs in a document more often than average.