Exercise 5.1

$P(c) = \prod_{d=0}^{m} P(w_{1,l_d}^{d}) P(l_d)$

Exercise 5.2

Yes; as long as one of the distributions has some variation you will avoid the saddle point. If both the document/topic and word/topic distributions were uniform, then both p and q in the E step of the algorithm in section 5.3 would simplify to a constant, rendering the M-step useless. Agitating the document/topic or word/topic distributions a little bit prevents this.

Exercise 5.3

When theta equals zero, this equation is just the maximum likelihood estimate of t conditioned on w. However, as theta grows very large, the right side of the equation approaches 1/N, yielding a uniform distribution over topics given any word. This would not be a good model, but having a smaller (more reasonable) theta accounts for the fact that, for any given topic, no word is explicitly impossible to encounter.

In any case, the words with high probabilities according to to this equation will be the ones that are the most *informative* of the topic; that is, they have a high likelihood of occurring given one topic and a low likelihood of occurring in others. As theta becomes higher and higher, the high scoring words become more and more random (as the score for all words approaches 1/N).

Exercise 5.4

Rather than using PLSA to get topics, you could simply look for words that have the highest likelihoods of occurring in the same document (i.e. "Puerto" and "Rico", or "Hong" and "Kong"). This would increase the coherence score dramatically (for a topic comprising "hong" and "kong" for instance) but would not make a good model because it is too strongly affected by the lexical and idiomatic associations between some words, drowning out the purely semantic associations between words that we are trying to capture.