Questions

Consider a corpus of documents about American pop singers and Japanese cultural heritage sites. The corpus term frequencies are given in the following table:

Britney	Spears	Itsukushima	Floating	Shrine
78	76	2	114	32

This table represents the entire corpus, so $|C|$ = 302. Document 1 is the string "Britney Shrine" and document 2 is the string "Itsukushima Floating." A query is issued consisting of the string "Itsukushima Shrine." The user is clearly more interested in Japan's "floating" Itsukushima shrine than Britney Spears.

1) Using the LM model with Dirichlet corpus-dependent priors, compute the ranking score that should be given to each of the two documents in question (use $\mu = 10$). Which document would be returned to the user first?

2) Now consider a simpler model that performs smoothing by adding a universal constant to the $tf_j(d)$ in the numerator of each multinomial probability term. If the constant added is μ, what term should be added to the denominator to ensure that the terms form a valid probability distribution?

3) Under the uniform smoothing model, what scores are assigned to each document (in terms of μ)? Which document would be returned to the user first?

4) Are there situations in which uniform smoothing would yield the same results as the more sophisticated model? What is a simple example given the corpus under consideration?

Solutions:

1)
$$Score\ for\, d_1 = \frac{0+10(2/302)}{2+10} \cdot \frac{1+10(32/302)}{2+10} = .00552 \cdot .172 = .00095$$
$$Score\ for\, d_2 = \frac{1+10(2/302)}{2+10} \cdot \frac{0+10(32/302)}{2+10} = .0889 \cdot .0883 = .0078$$

So, d_2 is returned first, as would be expected of any good retrieval system.

2) $\mu \times M$, where M is the size of the vocabulary.

3) The scores are the same. For each document, a query term that appears once in the document contributes a $(1+\mu)/(2+5\mu)$ term to the document's score; a query term that does not appear in the document contributes a $(0+\mu)/(2+5\mu)$ term to the score. Each document contains exactly one query term and one non-query term, so they both get scored as:
$$\left(\frac{\mu}{2+5\mu}\right)\left(\frac{1+\mu}{2+5\mu}\right)$$

Either document could be returned first, which may not produce the desired results.

4) Yes -- if the distribution of words over documents is close to uniform, the results will be the same. This is not likely to happen in practice; however, the importance of the IDF term also decreases as queries become more explicit and comprehensive, and as documents become more “dogmatic.” In the corpus given in the example, the document “Itsukushima Shrine” would be ranked higher than “Britney Shrine” using both uniform smoothing and corpus-dependent prior smoothing.