Web Search and Mining (CS 610) **Homework 2**
Computer Science and Engineering **Out 2007-01-06**
Indian Institute of Technology Bombay **Due 2007-02-28**

1. Sometimes, posting lists are compressed using a Golomb code, which has a tunable positive integer m. The Golomb code for gap n is found by first expressing n as $qm + r$ with $0 \leq r < m$, then writing down q in unary (as in gamma codes), and finally writing down r using a fixed prefix code (see example for $m = 5$ below) that takes about $\log_2 m$ bits.

r (decimal)	0	1	2	3	4
Code for r (bits)	00	01	10	110	111

E.g., the Golomb codes for (decimal) numbers 6, 9, 10, 27, using $m = 5$ are 1001, 10111, 11000 and 11111010 respectively (in binary).

Write code to create posting lists from the standard Reuters dataset (the TA/s will make the corpus available to you). There is no need to do external sorts and inversion; in-memory structures are ok if your PC's RAM will hold it. Write code to compress the postings using Golomb codes, given a specific value of m. Report on how you tuned the value of m for best compression, and how it relates to any corpus statistics.

2. Given a small corpus such as Reuters (see above), design and implement a preprocessing and indexing scheme that will create a data structure to support queries about the (relative) frequency of word trigrams (three consecutive words). Here are some properties your system should satisfy:

- The preprocessing and indexing time should be comparable to (say, within 5×) the time taken by Lucene to index the corpus.
- The space needed to store your data structure on disk should be comparable to (say, within 20×) the space needed by a Lucene index. The space in RAM should be comparable to a hash map from single words to numbers.
- The query will be of the form "w_1" or "$w_1 w_2$" or "$w_1 w_2 w_3$". It must be processed using a constant number of disk seeks at worst.
- The output should be an estimated number of times the word or exact phrase occurs in the corpus. It is ok if all response numbers are scaled up or down with the same scale factor, i.e., relative frequencies are ok.

Present a study of the tradeoff between index space and response accuracy.

3. We designed a min-hash family $\mathcal{F}$ with size at most 4^n using the following steps: Let $n = 2^r$. Construct $\mathcal{F}$ in stages recursively. In the first stage, divide $[n]$ into two halves, top and bottom. There are $\binom{n}{n/2}$ ways to do this. A specific choice can be represented by an n-bit long string, $n/2$ 0s, $n/2$ 1s. In the second stage, divide the halves into quarters, etc. The number of permutations supported is

$$|\mathcal{F}| = \prod_{i=1}^{\log n} \binom{n/2^{i-1}}{n/2^i} \leq \prod_{i=1}^{\log n} 2^{n/2^{i-1}} \leq 2^{n(1+1/2+\cdots)} \leq 4^n$$

Show that the in the product above, no 2^{i-1} term is needed, i.e., we can reuse one $n/2$-bit string across both halves and so on, while still satisfying the *exactly min-wise independent* property: for any $X \subseteq [n]$ and any $x \in X$, when π is chosen uar from $\mathcal{F}$, we have

$$\Pr(\min\{\pi(X)\} = \pi(x)) = \frac{1}{|X|}$$

4. Starting from the data log likelihood expression in the aspect model,

$$\log \Pr(X|\Theta) \propto \sum_{d,w} n(d,w) \log \Pr(d,w|\Theta),$$

do you get the EM update equations proposed by Hofmann?

$$\text{E-update} \qquad \Pr(c|d,w) = \frac{\Pr(c,d,w)}{\Pr(d,w)} = \frac{\Pr(c)\Pr(d,w|c)}{\sum_\gamma \Pr(\gamma,d,w)} = \frac{\Pr(c)\Pr(d|c)\Pr(w|c)}{\sum_\gamma \Pr(\gamma)\Pr(d|\gamma)\Pr(t|\gamma)}$$

$$\text{M-updates} \qquad \Pr(c) = \frac{\sum_{d,w} n(d,w)\Pr(c|d,w)}{\sum_\gamma \sum_{d,w} n(d,w)\Pr(\gamma|d,w)}$$

$$\Pr(d|c) = \frac{\sum_t n(d,w)\Pr(c|d,w)}{\sum_\delta \sum_w n(\delta,w)\Pr(c|\delta,w)}$$

$$\Pr(w|c) = \frac{\sum_d n(d,w)\Pr(c|d,w)}{\sum_\tau \sum_d n(d,\tau)\Pr(c|d,\tau)}$$

If yes, show all the steps; if not, explain why and offer an alternative. Here Θ includes the M-parameter sets $\{\Pr(c)\}$, $\{\Pr(d|c)\}$ and $\{\Pr(w|c)\}$.

5. Give small, simple examples to show that the aspect model has local optima problems. Specifically, propose a reasonable parameter set Θ. Now show that a (large) number of different Θ' will also maximize the total data likelihood in EM. The simpler you make your example, the more reasonable Θ is, and the more unreasonable Θ' is/are, the more credit you will get.

6. Consider two vectors $x_1, x_2 \in \mathbb{R}^d$ in some high-dimensional space. In class we demonstrated for $d = 2$ that, if you throw down a random hyperplane through the origin with normal vector h, then

$$\Pr(\text{sign}(h \cdot x) = \text{sign}(h \cdot y)) = 1 - \frac{\angle(x,y)}{\pi}$$

Argue that this holds for any d.