Information Retrieval

Exercises week 3

Exercise 1

Are the following statements true or false?

a) In a Boolean retrieval system, stemming never lowers precision.

b) In a Boolean retrieval system, stemming never lowers recall.

c) Stemming increases the size of the vocabulary.

d) Stemming should be invoked at indexing time but not while processing a query.

Exercise 2

Given a biased coin with p(H)=0.25 and p(T)=0.75.

- Suppose we generate a series of symbols $s \in \{H, T\}$
- What is the theoretical minimum # bits per symbol required for a lossless compression?

Exercise 3

- Compress the fragment below of Jan Hanlo's poem "Oote" (1952)
- Use a word based version of Huffman coding (words are symbols)
- Assume a lowercase version of the poem, ignoring whitespace
- Provide code table and compute compression ratio (one ASCII character is 7 bits)

OOTE

Oote oote oote
Boe
Oote oote
Oote oote oote boe
Oe oe
Oe oe oote oote oote
A
A a a
Oote a a a
Oote oe oe
Oe oe oe
<…>

Exercise 4

Consider the postings list (4, 10, 11, 12, 15, 62, 63, 265, 268, 270, 400, 444) with a corresponding list of gaps (4, 6, 1, 1, 3, 47, 1, 202, 3, 2, 130, 44). Assume that the length of the postings list is stored separately, so the system knows when a postings list is complete. Using variable byte encoding:

(i) What is the largest gap you can encode in 1 byte?

(ii) What is the largest gap you can encode in 4 bytes?

(iii) How many bytes will the above postings list require under this encoding? (Count only space for encoding the sequence of numbers.)