Information Retrieval 23-24

Exercises week 2

Deadline: 18 February 23:59

Exercise 1

Consider these documents:

Doc 1 new recipe for cookies

Doc 2 cookie dough recipe

Doc 3 new chocolate dough

Doc 4 recipe for chocolate cookies

a) Draw the term-document incidence matrix for this document collection

b) Draw the inverted index representation for this collection.

c) What is the result set for the query: recipe AND NOT (chocolate OR dough)

Exercise 2

Recommend a query processing order for the query below, given the postings list sizes in the table:

(rocket OR launch) AND (sky OR universe) AND (sun OR orbit)

Term	Postings size
rocket	81300
launch	109999
sky	43123
universe	48491
sun	62687
orbit	577513

Exercise 3

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 4

- Compress the fragment below of Jan Hanlo's poem "Oote" (1952), chosen because of its particular redundancy.
- 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 5

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 just the required #bytes for encoding the sequence of numbers.)