Information Retrieval

Weike Pan

Exercise 2.1 [⋆]

Are the following statements true or false?

a. In a Boolean retrieval system, stemming never lowers precision. [FALSE]

b. In a Boolean retrieval system, stemming never lowers recall. [TRUE]

c. Stemming increases the size of the vocabulary. [FALSE]

d. Stemming should be invoked at indexing time but not while processing a query. [FALSE]

Exercise 2.3 [⋆]

The following pairs of words are stemmed to the same form by the Porter stemmer. Which pairs would you argue shouldn't be conflated (合并). Give your reasoning.

a. abandon/abandonment

b. absorbency/absorbent

c. marketing/markets

d. university/universe

e. volume/volumes

Exercise 2.6 [⋆]

We have a two-word query. For one term the postings list consists of the following 16 entries:

[4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180]

and for the other it is the one entry postings list:

[47].

Work out how many comparisons would be done to intersect the two postings lists with the following two strategies. Briefly justify your answers:

a. Using standard postings lists. [11 comparisons]

b. Using postings lists stored with skip pointers, with a skip length of P^0.5, as suggested in Section 2.3. [6 comparisons]

[4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180]

6 comparisons: (4,47), (14,47), (22,47), (120,47), (32,47), (47,47)

Exercise 2.8 [⋆]

Assume a biword index. Give an example of a document which will be returned for a query of New York University but is actually a false positive which should not be returned.

Some alumni had arrived from New York. University faculty said that ...

Exercise 2.9 [⋆]

Shown below is a portion of a positional index in the format: term: doc1: <position1, position2, …>; doc2: <position1, position2, …>; etc.

angels: 2: <36,174,252,651>; 4: <12,22,102,432>; 7: <17>;

fools: 2: <1,17,74,222>; 4: <8,78,108,458>; 7: <3,13,23,193>;

fear: 2: <87,704,722,901>; 4: <13,43,113,433>; 7: <18,328,528>;

in: 2: <3,37,76,444,851>; 4: <10,20,110,470,500>; 7: <5,15,25,195>;

rush: 2: <2,66,194,321,702>; 4: <9,69,149,429,569>; 7: <4,14,404>;

to: 2: <47,86,234,999>; 4: <14,24,774,944>; 7: <199,319,599,709>;

tread: 2: <57,94,333>; 4: <15,35,155>; 7: <20,320>;

where: 2: <67,124,393,1001>; 4: <11,41,101,421,431>; 7: <16,36,736>;

Which document(s) if any meet each of the following queries, where each expression within quotes is a phrase query?

a. "fools rush in". Answer: 2,4,7.

b. "fools rush in" AND "angels fear to tread". Answer: 4.

Exercise 2.10 [⋆]

Consider the following fragment of a positional index with the format:

word: document: <position, position, ...>; document: <position, ...>

...

Gates: 1: <3>; 2: <6>; 3: <2,17>; 4: <1>;

IBM: 4: <3>; 7: <14>;

Microsoft: 1: <1>; 2: <1,21>; 3: <3>; 5: <16,22,51>;

The */k* operator, word1 */k* word2 finds occurrences of word1 within *k* words of word2 (on either side), where *k* is a positive integer argument. Thus *k = 1* demands that word1 be adjacent to word2.

a. Describe the set of documents that satisfy the query Gates /2 Microsoft.

b. Describe each set of values for *k* for which the query Gates /k Microsoft returns a different set of documents as the answer.

k=1: doc 3; k=1,3: docs 1,3; k>=5: docs 1,2,3