Assignment 1 - Solutions

Exercise 1 (IIR 1) [1 P.]

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) match each of the following queries at which positions, where each expression within quotes is a phrase query? (i) "fools rush in" (ii) "fools rush in" AND "angels fear to tread".

Solution

(i) doc2:1, doc4:8, doc7:3,13 (ii) doc4:8 & 12

Exercise 2 (IIR 1) [3 P.]

Write a (Python) program [...]

Solution

See `boolean.py` in the `assignment_1_ex_2_solution.zip` file on the course homepage.

Exercise 3 (IIR 2) [1 P.]

The following pairs of words are stemmed to the same form by the German version of the Porter stemmer[1]. Which pairs, would you argue, should not be conflated? Give your reasoning.

Solution

(a) geräumige/geräumigen (→ geraum)
OK

(b) musik/musiker (→ musik)
Sometimes good, sometimes bad. E. g. when somebody wants to find out more about a certain musical genre.

(c) neuer/neues (→neu)
Not OK, e. g. when somebody wants to get some information about the goalkeeper of the German National Soccer Team.

(d) persönlich/persönlichkeit (→ person)
The two words are thematically far away from each other.

[1] http://snowball.tartarus.org/algorithms/german/stemmer.html

(e) schlaf/schlafen (→ schlaf)
OK.

(f) unternehmer/unternehmung (→ unternehm)
Semantically too far away.

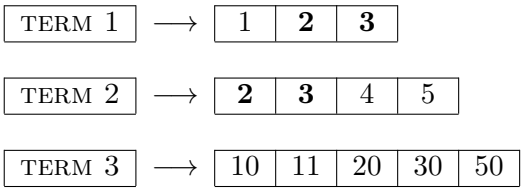
(g) wetten/wetter (→ wett)
Completely different words.

Exercise 4 (IIR 1) [1 P.]

For a conjunctive query, is processing postings list in order of size guaranteed to be optimal? Explain why it is, or give an example where it is not.

Solution

Processing postings list in order of size (i.e. the shortest postings list first) is usually a good approach. But it is not optimal e. g. in a conjunctive query with three terms:



As we can see there is no document containing all three query terms. If we would have checked the first posting of the third list right at the beginning, we would have noticed that there is no intersection between the first and the third postings list. That would make any further search superfluous.