TÜ Information Retrieval

Übung 1

Heike Adel, Sascha Rothe

Center for Information and Language Processing, University of Munich

April 24, 2014

Problem 1

(i) Find a query (two terms without quotes) which Google does not interpret as a conjunction.

Problem 1

(i) Find a query (two terms without quotes) which Google does not interpret as a conjunction.

e.g.: Maracujasaft retrieval

Problem 1

(i) Find a query (two terms without quotes) which Google does not interpret as a conjunction.

e.g.: Maracujasaft retrieval

Information

Differentiate boolean conjunctions and phrase queries!

Problem 1

(ii) Does Google interpret all queries as a Boolean conjunction?

Problem 1

(ii) Does Google interpret all queries as a Boolean conjunction?

In most cases: yes

Exceptions are for example:

- anchor text
- the page may have changed
- a morphological or orthographic variant of a search word may appear on the page
- a semantic equivalent may appear

Problem 2

Given the following positional index

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: <15,35,735>;

which documents match the phrase query "fools rush in"?

Problem 2

Given the following positional index

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: <15,35,735>;

which documents match the phrase query "fools rush in"?

$\Rightarrow$ document 2 matches "fools rush in" at position 1.

Problem 2

Given the following positional index

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: <15,35,735>;

which documents match the phrase query "fools rush in"?

$\Rightarrow$ document 2 matches "fools rush in" at position 1.
$\Rightarrow$ document 4 matches "fools rush in" at position 8.

Problem 2

Given the following positional index

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: <15,35,735>;

which documents match the phrase query "fools rush in"?

$\Rightarrow$ document 2 matches "fools rush in" at position 1.
$\Rightarrow$ document 4 matches "fools rush in" at position 8.
$\Rightarrow$ document 7 matches "fools rush in" at position 3 and 13.

Problem 2

Given the following positional index

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: <15,35,735>;

which documents match the phrase query "fools rush in" AND "angels fear to tread"?

Problem 2

Given the following positional index

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: <15,35,735>;

which documents match the phrase query "fools rush in" AND "angels fear to tread"?

$\Rightarrow$ document 4:8&12

Problem 2

Given the following positional index

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: <15,35,735>;

There is something wrong with this positional index. What is the problem?

Problem 2

Given the following positional index

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: <15,35,735>;

There is something wrong with this positional index. What is the problem?

$\Rightarrow$ Only one word can occur at position 15 of document 7.
But according to the index, two words occupy this position ("in" and "where")

Problem 3

Compute the Levenshtein matrix for the distance between the strings “apfel” (input) and “poems” (output).

Problem 3

Compute the Levenshtein matrix for the distance between the strings "apfel" (input) and "poems" (output).

$\Rightarrow$ Solution:

			p		o		e		m		s	
		0	1	1	2	2	3	3	4	4	5	5
a		1	1	2	2	3	3	4	4	5	5	6
		1	2	1	2	2	3	3	4	4	5	5
p		2	1	2	2	3	3	4	4	5	5	6
		2	3	1	2	2	3	3	4	4	5	5
f		3	3	2	2	3	3	4	4	5	5	6
		3	4	2	3	2	3	3	4	4	5	5
e		4	4	3	3	3	2	4	4	5	5	6
		4	5	3	4	3	4	2	3	3	4	4
l		5	5	4	4	4	4	3	3	4	4	5
		5	6	4	5	4	5	3	4	3	4	4

Problem 3

Find the shortest path in the matrix:

$\Rightarrow$ Solution:

			p		o		e		m		s	
		0	1	1	2	2	3	3	4	4	5	5
a		1	1	2	2	3	3	4	4	5	5	6
		1	2	1	2	2	3	3	4	4	5	5
p		2	1	2	2	3	3	4	4	5	5	6
		2	3	1	2	2	3	3	4	4	5	5
f		3	3	2	2	3	3	4	4	5	5	6
		3	4	2	3	2	3	3	4	4	5	5
e		4	4	3	3	3	2	4	4	5	5	6
		4	5	3	4	3	4	2	3	3	4	4
l		5	5	4	4	4	4	3	3	4	4	5
		5	6	4	5	4	5	3	4	3	4	4

Problem 3

Find the shortest path in the matrix:

$\Rightarrow$ Solution:

			p		o		e		m		s	
		0	1	1	2	2	3	3	4	4	5	5
a		1	1	2	2	3	3	4	4	5	5	6
		1	2	1	2	2	3	3	4	4	5	5
p		2	1	2	2	3	3	4	4	5	5	6
		2	3	1	2	2	3	3	4	4	5	5
f		3	3	2	2	3	3	4	4	5	5	6
		3	4	2	3	2	3	3	4	4	5	5
e		4	4	3	3	3	2	4	4	5	5	6
		4	5	3	4	3	4	2	3	3	4	4
l		5	5	4	4	4	4	3	3	4	4	5
		5	6	4	5	4	5	3	4	3	4	4

Problem 4

While the Levenshtein sequence of edit operations is not unique, the minimum number of operations is fixed. Let n_i, n_d, n_r be the number of inserts, deletes and replaces in a sequence of operations. Can you find a pair of strings and two different sequences of edit operations σ_1 and σ_2 such that $n_i(\sigma_1) \neq n_i(\sigma_2)$ or $n_d(\sigma_1) \neq n_d(\sigma_2)$ or $n_r(\sigma_1) \neq n_r(\sigma_2)$?

Problem 4

While the Levenshtein sequence of edit operations is not unique, the minimum number of operations is fixed. Let n_i, n_d, n_r be the number of inserts, deletes and replaces in a sequence of operations. Can you find a pair of strings and two different sequences of edit operations σ_1 and σ_2 such that $n_i(\sigma_1) \neq n_i(\sigma_2)$ or $n_d(\sigma_1) \neq n_d(\sigma_2)$ or $n_r(\sigma_1) \neq n_r(\sigma_2)$?

$\Rightarrow$ Consider the strings "ab" and "ba":
- Levenshtein distance: 2
- operation sequences: σ_1: replace a with b, replace b with a
  σ_2: delete a, copy b, insert a
- Hence: $0 = n_i(\sigma_1) \neq n_i(\sigma_2) = 1$ and $0 = n_d(\sigma_1) \neq n_d(\sigma_2) = 1$ and $2 = n_r(\sigma_1) \neq n_r(\sigma_2) = 0$

Problem 4

Information

We are looking at the **minimum** number of operations:
Each Levenshtein sequence of edit operations has the same total number of operations!

Problem 5

Permutation wildcard index: If you wanted to search for s*ng, what key(s) would you do the lookup on?

Problem 5

Permutation wildcard index: If you wanted to search for s*ng, what key(s) would you do the lookup on?

$\Rightarrow$ We would perform the lookup on the key: ng$s*

The end

Thank you for your attention!

Do you have any questions?