0007 sheet 1 učo 422640 points

Recommend a query processing strategy [2 points] for the following boolean query:

(CLEOPATRA OR CALPURNIA) AND (BRUTUS OR CAESAR)

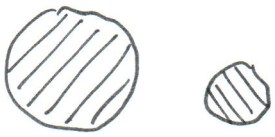
given the following document frequencies:

- CLEOPATRA: 21,502 documents
- CALPURNIA: 257,219 documents
- BRUTUS: 163,587 documents
- CAESAR: 175,843 documents

Several strategies for estimating the cardinality of the result for an "OR" query exist. Explain which one you used and what assumption your strategy makes. [2 points]

|CLEOPATRA OR CALPURNIA| = |CLEOPATRA| + |CALPURNIA|

Assumption: Documents containing CLEOPATRA and CALPURNIA are disjoint.



|CLEOPATRA| + |CALPURNIA| = 278 721 docs

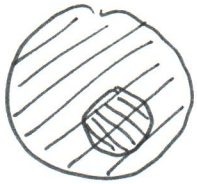
|BRUTUS| + |CAESAR| = 339 430 docs

Recommended query processing strategy:

(CLEOPATRA OR CALPURNIA) AND (BRUTUS OR CAESAR)

|CLEOPATRA OR CALPURNIA| = max(|CLEOPATRA|, |CALPURNIA|)

Assumption: Documents containing CLEOPATRA and CALPURNIA fully overlap.



max(|CLEOPATRA|, |CALPURNIA|) = 257 219 docs

max(|BRUTUS|, |CAESAR|) = 175 843 docs

Recommended query processing strategy:

(BRUTUS OR CAESAR) AND (CLEOPATRA OR CALPURNIA)

0007 sheet 2 učo 422640 points

Perform the following boolean query:

FRIENDS AND ROMANS

given the following inverted index:

- FRIENDS: 3, 5, 9, 15, 23, 39, 40, 41, 47, 49, 51
- ROMANS: 2, 3, 5, 7, 11, 23, 29, 31, 37, 39, 41, 43, 47

$|P| = 11 \rightsquigarrow \lfloor\sqrt{11}\rfloor = 3$
$|P| = 13 \rightsquigarrow \lfloor\sqrt{13}\rfloor = 3$

Present the result of the query and the number of comparisons with [2 points], and without [2 points] a skip list. Assume the skip list has frequency $\lfloor\sqrt{|P|}\rfloor$, where $|P|$ is the cardinality of a posting P and $\lfloor . \rfloor$ is the *floor function*. (*Example:* $P = \{1, 2, 3\}, |P| = 3, \sqrt{|P|} \approx 1.73, \lfloor\sqrt{|P|}\rfloor = 1$)

Number of comparisons:

- without a skip list: 16 comparisons

(3,2), (3,3), (5,5), (9,7), (9,11), (15,11), (15,23), (23,23), (39,29), (39,31), (39,37), (39,39), (40,41), (41,41), (47,43), (47,47)

- with a skip list: 18 comparisons

(3,2), (3,7), (3,3), (5,5), (9,7), (9,29), (9,11), (15,11), (15,23), (40,23), (23,23), (39,29), (39,39), (40,41), (49,41), (41,41), (47,43), (47,47)

Result of the query: 3, 5, 23, 39, 41, 47 (6 matches)

0007 sheet 3 učo 422640 points

Encode the following posting:

- COUNTRYMEN: 687, 1599, 1978

using the *variable byte code* [2 points], and the *gamma code* (γ), where $\gamma(n) = \langle\alpha(\text{length of offset}(n)), \text{offset}(n)\rangle$ [2 points].

$(687)_2 = (512 + 128 + 32 + 8 + 4 + 2 + 1)_2 = 10\ 1010\ 1111$

$(1599 - 687 = 912)_2 = (512 + 256 + 128 + 16)_2 = 11\ 1001\ 0000$

$(1978 - 1599 = 379)_2 = (256 + 64 + 32 + 16 + 8 + 2 + 1)_2 = 1\ 0111\ 1011$

Variable byte code:

0000 0101 1010 1111 0000 0111 1001 0000 0000 0010 1111 1011

Gamma code:

111 111 1110, 0 1010 1111 111 111 1110, 1 1001 0000

11 111 1110, 0111 1011

Write your solution only on this side of the sheet!

0007 *sheet* 4 *učo* 422640 *points*

Compute the Levenshtein distance between CRYPTO and CORRUPT using the matrix [2 points] and present at least one sequence of operations (insert, delete, replace, copy) that transform one term to the other [2 points].

	ε	C	O	R	R	U	P	T
ε	0	1	2	3	4	5	6	7
C	1	0	1	2	3	4	5	6
R	2	1	1	1	2	3	4	5
Y	3	2	2	2	2	3	4	5
P	4	3	3	3	3	3	3	4
T	5	4	4	4	4	4	4	3
O	6	5	4	5	5	5	5	**4**

	ε	C	R	Y	P	T	O
ε	0	1	2	3	4	5	6
C	1	0	1	2	3	4	5
O	2	1	1	2	3	4	4
R	3	2	**1**	2	3	4	5
R	4	3	2	2	3	4	5
U	5	4	3	3	3	4	5
P	6	5	4	4	3	4	5
T	7	6	5	5	4	3	**4**

[redacted]	OPER.	IN	OUT
	COPY	C	C
	DEL	R	*
	REPL	Y	O
	DEL	P	*
	INS	*	R
	INS	*	R
	INS	*	U
	INS	*	P
	COPY	T	T
	DEL	O	*

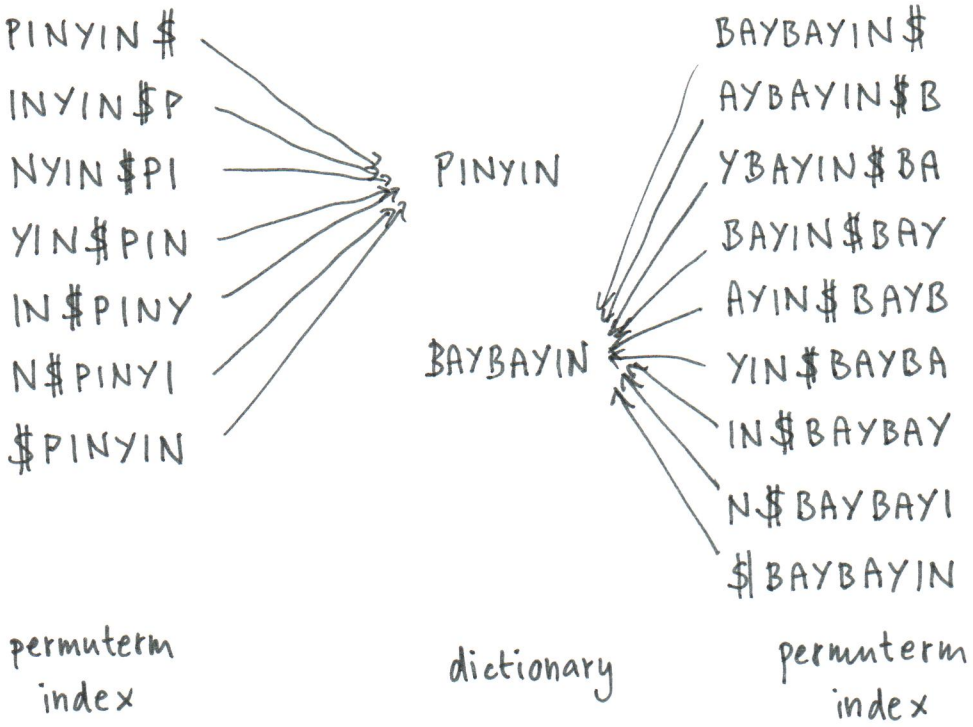
COPY	C	C
REPL	R	O
REPL	Y	R
REPL	P	R
REPL	T	U
INS	*	P
REPL	O	T

⋮

Write your solution only on this side of the sheet!

0007 *sheet* 5 *učo* 422640 *points*

Construct a permuterm index for the terms PINYIN, and BAYBAYIN. [2 points]



Consider a search engine that indexes a total of 1,000,000,000 pages, each page containing 250 tokens on average. Assume the parameters $k = 30$ and $b = 0.5$ for your computation.
What is the size of the vocabulary of the indexed collection as predicted by *Heaps' law*? [2 points]

$M = k \cdot T^b$

$M = 30 \cdot (250 \cdot 1\,000\,000\,000)^{1/2} = 30 \cdot (5^2 \cdot 10^{10})^{1/2}$

$= 30 \cdot 5 \cdot 10^5 = 15\,000\,000$

Vocabulary size: 15 000 000 terms

Write your solution only on this side of the sheet!