TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

Exercise 4

Information Retrieval

Dresden Database
Systems Group

6. Implementing IR-Systems I
Index Compression

TECHNISCHE UNIVERSITÄT DRESDEN

Warm-Up

Dresden Database Systems Group

Exercise 6.1

- Are the following statements true or false? Give reasons for your answer.

a) Heaps' Law assumes that the vocabulary can grow infinitely.

b) Zipf's Law states that the i^{th} most frequent term has a collection frequency proportional to $1/i$.

c) Large document collections contain many frequent and few rare terms. f

d) We compress the index in order to save space.

e) We compress the index in order to speed up queries.

f) Elias Gamma Coding is a technique for dictionary compression. only f

g) Front Coding is a technique for dictionary compression.

Heaps' Law $M = k \cdot T^b$

Dresden Database Systems Group

做过，在pad上，还没对答案

Exercise 6.2

- While indexing a collection of web pages, you find out the following

In the first	10,000	tokens, there are	3,000	terms
In the first	1,000,000	tokens, there are	30,000	terms

a) For these two samples, compute the parameters k and b used in Heaps' law

b) Assume a search engine indexes a total of 20,000,000,000 pages, containing 200 tokens on average. What is the size of the vocabulary of the indexed collection as predicted by Heaps' law?

Elias Gamma Coding

Dresden Database Systems Group

Exercise 6.3

- Consider the following postings list:

1060	1078	1111	1115

a) What is the corresponding gap sequence?

b) Compress this gap sequence using Elias Gamma Coding

c) What is the resulting compression ratio?
Assume that a standard 32-bit integer is used for each entry of the uncompressed postings list.

d) Gamma coding cannot encode the number zero. Is this a problem for compressing postings lists?

ithk :NO

Variable Byte Coding

Dresden Database Systems Group

Exercise 6.4

- Consider the following sequence of variable-byte-coded gaps:

Byte 0	Byte 1	Byte 2	Byte 3	Byte 4	Byte 5
0001 0000	1000 0001	1000 0101	1010 0010	0000 0001	1000 0010

a) How many entries does this compressed postings list have?

b) Reconstruct the original postings list