Information Retrieval and Text Mining
Assignment 2

Fabian Leeske, 3478150, st169266@stud.uni-stuttgart.de
Sarah Breckner, 3425446, st163632@stud.uni-stuttgart.de
Kim Lingemann, 3380756, st160814@stud.uni-stuttgart.de

Task 1

What information does the task description contain that the master gives to a parser?

The parser assigns a split (= a set of documents) to the parser.

What information does the parser report back to the master upon completion of the task?

The parser reads each document in the split one at a time and creates (term, docID)-pairs for it. It then writes these pairs into term-partitions, which the parser returns to the master.

What information does the task description contain that the master gives to an inverter?

The master assigns a partition the inverter should work on.

What information does the inverter report back to the master upon completion of the task?

The inverter takes all postings for one partition and writes a sorted postings list. This list is then returned to the master.

How would you specify the number of parsers? Can you estimate this somehow (helpful corner case scenario to think about: is one parser for each term a good choice? Is just one parser overall a good choice?)

How would you specify the number of partitions from which the inverters put the index together? (corner cases: only one partition? Or one partition per term?)

Task 2

Task 3

Heaps' Law:

$$M = k \cdot T^b$$

M: size of the vocabulary (terms)
T: number of tokens in the collection

Subtask 3.1

$$10,000 = k \cdot 1,000,000^b$$

$$3,000 = k \cdot 100,000^b -> k = \frac{3,000}{100,000^b}$$

$$10,000 = \frac{3,000}{100,000^b} \cdot 1,000,000^b$$

$$10,000 = \frac{3,000}{10^b}$$
$$\frac{10}{3} = 10^b$$
$$b = 1\text{–}log_{10}(3)$$
$$b \approx 0,52$$

$$k = \frac{10,000}{1,000,000^{0,52}}$$
$$k \approx 7,56$$

$$-> b \approx 0,52 \, and \, k \approx 7,56$$

Subtask 3.2

$$M = 7,56 \cdot 100,000,000^{0,52} \approx 109,275$$

Task 4

Variable byte code

$216_{10} = 11011000_2$: 00000001 11011000

Gamma code

$216_{10} = 11011000_2$
Offset: 1011000
Length: $7_{10} = 11111110_1$
Gamma Code: 11111110 1011000

Task 5

11110 1100 010 0 110 00 0
$11110_1 = 4_{10}$, so the first number is: 11110 1100 which is $11100_2 = 28$
$010_1 = 1_{10}$, so the next number is: 010 0 which is $10_2 = 2$
$110_1 = 2_{10}$, so the next number is: 110 00 which is $100_2 = 4$
This means the encoded postings sequence is 28, 2, 4.

11110 1100 0 10 0 110 00 0
$11110_1 = 4_{10}$, so the first number is: 11110 1100 which is $11100_2 = 28$
$0_1 = 0_{10}$, so the next number is 0
$10_1 = 1_{10}$, so the next number is: 10 0 which is $10_2 = 2$

$110_1 = 2_{10}$, so the next number is: 110 00 which is $100_2 = 4$
$0_1 = 0_{10}$, so the next number is 0
This means the encoded postings sequence is 28, 0, 2, 4, 0.

Programming Task

Python Code