

Information Retrieval and Text Mining: Assignment 2

Problem 1. (8 points)

Given is a collection that contains 4 different words a, b, c, d and no other words. Frequency order is $a > b > c > d$. The total number of tokens in the collection is 5000. Assume that Zipf's law holds exactly for this collection. What are the frequencies of the four words?

Problem 2. (10 points)

Compute variable byte and γ codes for the postings list 777, 17743, 294068, 31251336. Use gaps instead of docIDs where appropriate. Give the solution for variable bytes as a sequence of 8-bit blocks. Give the solution for γ codes as a sequence of 4 pairs of bit strings, where the first bit string of each pair corresponds to a length and the second to an offset.

Problem 3. (5 points)

Consider this sequence of γ coded gaps: 11100011101010111111011011110110. (i) What is the sequence of gaps? (ii) What is the sequence of postings?

Problem 4. (15 points)

γ -codes are inefficient for large numbers (e.g., 1000 or 10,000) as they encode the length of the offset in unary code. δ -codes use γ code for encoding this length instead.

We defined the γ code of G as

$$\text{unary-code}(\text{length}(\text{offset}(G))), \text{offset}(G)$$

We define the δ code of G as

$$(*) \text{ gamma-code}(\text{length}(\text{offset}(G+1))), \text{offset}(G+1)$$

For example, the δ -code of $G = 6$ is 10,0,11 (as before, we add commas for readability only). 10,0 is the γ -code for *length* (2 in this case). The encoding of *offset* (11) is the same as it would be in the γ code for $G = 7$.

(i) Compute the δ -codes for 1, 2, 3, 4, 31, 63, 127 and 1023. (ii) For what range of numbers is the δ -code shorter than the γ -code? (iii) An alternative definition of the δ code is “gamma-code(length(offset(G))), offset(G)”. What is the disadvantage of this definition?

Note that there are several different definitions of the δ code. You must use definition (*) given above.

Problem 5. (12 points)

Compute the ltn.lnc similarity between the query “digital phones” and the document “digital phones and video phones and other phones” by filling out the empty columns in the table below. Assume $N = 10,000,000$. Treat *and* and *other* as stop words. What is the final similarity score?

word	query					document				product
	tf-raw	tf-wght	df	idf	weight	tf-raw	tf-wght	weight	n'lized	
digital			10,000							
video			100,000							
phones			50,000							

Due date: Tuesday, May 26, 2009, 15:45 (turn assignments in before class in V38.03 or by email)