Information Retrieval: Assignment 2

Problem 1. (10 points)

Assume (i) that machines in MapReduce have 100 GB of disk space each; (ii) that the postings list of the term THE has a size of 180 GB for a particular collection; (iii) that we do not use compression. Then the MapReduce algorithm as described in class cannot be run to construct the inverted index. Why? How would you modify the algorithm so that it can handle this case?

Problem 2. (10 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 3. (10 points)

We define a hapax legomenon as a term that occurs exactly once in a collection. We want to estimate the number of hapax legomena using Heaps' law and Zipf's law. (i) How many unique terms does a web collection of 600,000,000 web pages containing 600 tokens on average have? Use the Heaps parameters $k = 100$ and $b = 0.5$. (ii) Use Zipf's law to estimate the proportion of the term vocabulary of the collection that consists of hapax legomena. You may want to use the approximation $\sum_{i=1}^{n} 1/i \approx \ln n$. (iii) Do you think that the estimate you get is correct? (iv) Discuss what possible reasons there might be for the correctness / incorrectness of the estimate.

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

unary-code(length(offset(G))),offset(G)

We define the δ code of G as

($*$) gamma-code(length(offset(G+1))),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 $G = 7$.

Compute the δ-codes for 1, 2, 3, 4, 31, 63, 127 and 1023. Note that there are different definitions of the δ-code. You must use definition ($*$) given above.

Due date: Thursday, May 16, 2013, 12:15