Web Information Retrieval

Exam
September 21st, 2017
Time available: 90 minutes
5 points for each problem

Problem 1

1. Are the following statements true or false? *Briefly motivate all your answers.*
   (a) In a Boolean retrieval system, stemming always increases precision.
   (b) In a Boolean retrieval system, stemming increases recall.
   (c) Stemming reduces the size of the dictionary.
2. Are skip pointers useful for queries of the form `x AND NOT y`?
3. Assume a biword index. Give an example of a document which will be returned for a query of `new york university` but is actually a false positive which should not be returned.

Problem 2

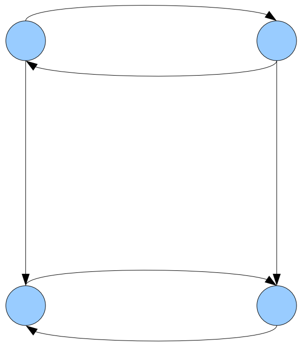
Answer the questions below:

1. Assume the *gaps* of a posting list containing n (strictly positive) DocIds obeys a Zipf's distribution. In particular, the probability that the generic i-th gap Δ_i has (integer) size x is $\mathbf{P}(\Delta_i = x) = \frac{1}{H_L \cdot x}$. Here, L is the maximum gap value and $H_L = \sum_{x=1}^{L} \frac{1}{x}$ is the x-th harmonic number. Denote by S_i the number of bits necessary to represent the i-th gap (so, for example, $S_i = \lceil \log_2 x \rceil$ if $\Delta_i = x$). Under these assumptions, give a good upper bound on $\mathbf{E}[S_i]$.
2. Under the same assumptions and using your answer to the former point, give an upper bound on the expected overall number of bits necessary to represent the whole postings list (consider only gap information).
3. Show how we can compress the list `[5, 7, 18, 19, 28, 40, 52, 80]` using variable byte encoding.

Note: for questions 1 and 2, use $\lceil \log_2 x \rceil \leq \log_2 x + 1$ and $\sum_{x=1}^{L} \frac{\log_2 x}{x} \approx \frac{\ln^2 L}{\ln 4}$.

Problem 3

1. What is the importance of the teleporting probability with respect to the convergence of pagerank?
2. We are given the following graph. Write down all the necessary equations needed to calculate the pagerank, for a general teleporting probability α.
3. Compute the pagerank of each node for teleporting probability $\alpha = 1/2$.



Problem 4

1. Explain briefly how the k-means algorithm works. Write the algorithm.
2. You are given the following example. Show that if the initial cluster assignment is unlucky the k-means solution might be bad.

3. Explain briefly why the k-means algorithm converges.

I consent to publication of the results of the exam on the Web

Firstname and Lastname in block letters..

Signature