Algorithm Engineering

1 June 2010

Exercise 1 [2+2+5+3 points]. Let us given the 4 texts:

- T1=″one rose two girls″
- T2=″one girl two girls″
- T3=″one one two two″
- T4=″one girl one rose two″

1. Describe the content of the Inverted List (no gap-coding) built on this collection, under the hypothesis that **only boolean** queries must be supported.
2. Describe the content of the Inverted List (no gap-coding) built on this collection, under the hypothesis that **phrase** queries must be supported.
3. Describe the TF-IDF weight of a <term,doc> pair, and compute the TF-IDF vector of each of those four texts (*pls attention to the vector size*).
4. Assume that you are given the text T=″rose rose girls girls″, how would you find the **most similar** text to this one within the above collection?

Exercise 2 [4+6+2 points]. Given a string T[1,t], and two patterns P and Q, we wish to establish whether it does exist an occurrence of P and Q in T at distance smaller than L from each other.

- Design an algorithm that solves the previous problem efficiently, when T is given in input together with <P,Q,L>.
- Design an algorithm that solves the previous problem efficiently, when T is given to be preprocessed, and <P,Q,L> are given at query time.
- Indicate the time and space complexities (in the worst case) of all your solutions.

Exercise 3 [6 points]. Given a text consisting of the symbols {a,b,c,d,e,f,g} occurring with frequency f(a) = f(b) = f(d) = f(e) = 0.1, f(c) = 0.11, f(f)= 0.21, f(g) = 0.28.

- Compute the Canonical Huffman code for the symbols of T
- Decode the first 3 symbols of compressed sequence: 00110101011011100….