58093 String Processing Algorithms (Autumn 2016)

Exercises 6 (Wednesday, December 7)

1. Describe a family of string pairs (A_i, B_i), $i \in \mathbb{N}$, such that $|A_i| = |B_i|$ and there is at least i different optimal edit sequences corresponding to $ed(A_i, B_i)$. Can you find a family, where the number of edit sequences grows much faster than the lengths of the strings?

2. Give a proof for Lemma 3.15 in the lecture notes.

3. Let $P = \texttt{evete}$ and $T = \texttt{neeteneeveteen}$. Simulate the operation of Myers' bitparallel algorithm when it computes column 5 for pattern P and text T.

4. A q-gram of a string is its factor of length q. Let $\gamma_q(A, B)$ denote the number q-grams shared by the strings A and B.

   For example, for $A = \texttt{varaurat}$ the 2-grams are va, ar, ra, au, ur, ra and at, and for $B = \texttt{ararat}$ they are ar, ra, ar, ra and at. The shared 2-grams are ra twice, ar and at, and thus $\gamma_q(A, B) = 4$.

   (a) Show that if $ed(A, B) \leq k$, then $\gamma_q(A, B) \geq |A| - q + 1 - kq$.

   (b) Design a filtering algorithm for approximate string matching based on the result of (a)-part.

5. Let T be a string and let R be a multiset of symbols. In *jumbled string matching*, a factor S of T is an occurrence of R if S consists of exactly the symbols of R. For example, if $T = \texttt{abahgcabah}$ and $R = \{\texttt{a}, \texttt{a}, \texttt{b}, \texttt{c}\}$, the only occurrence of R in T is the factor $S = \texttt{caba}$. Describe an algorithm for finding all occurrences of R in T. The time complexity should be $\mathcal{O}(|T| + |R|)$ on an alphabet of constant size.