

58093 String Processing Algorithms (Autumn 2012)

Exercises 5 (29 November)

1. Show that edit distance is a *metric*, i.e., that it satisfies the metric axioms:

- $ed(A, B) \geq 0$
- $ed(A, B) = 0$ if and only if $A = B$
- $ed(A, B) = ed(B, A)$ (symmetry)
- $ed(A, C) \leq ed(A, B) + ed(B, C)$ (triangle inequality)

2. Describe a family of string pairs (A_i, B_i) , $i \in \mathbb{N}$, such that $|A_i| = |B_i| \geq i$ 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?

3. A string S is a *subsequence* of a string T if we can construct S by deleting characters from T . Let $lcss(A, B)$ denote the length of the longest common subsequence of the strings A and B . For example, $lcss(\text{berlin}, \text{helsinki}) = 4$ since `elin` is a subsequence of both strings.

(a) Let $ed_{\text{indel}}(A, B)$ be a variant of the edit distance, where insertions and deletions (indels) are the only edit operations allowed (i.e., no substitutions). Show that

$$ed_{\text{indel}}(A, B) = |A| + |B| - 2 \cdot lcss(A, B)$$

(b) Give an algorithm for computing $lcss(A, B)$ in time $\mathcal{O}(|A||B|)$.

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

5. Let $P = \text{evete}$ and $T = \text{neeteneeveteen}$.

(a) Use Ukkonen's cut-off algorithm to find the occurrences of P in T for $k = 1$.

(b) Simulate the operation of Myers' bitparallel algorithm when it computes column 5 for P and T .

6. 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 $A = \text{varaurat}$ the 2-grams are `va`, `ar`, `ra`, `au`, `ur`, `ra` and `at`, and for $B = \text{ararat}$ 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.