58093 String Processing Algorithms (Autumn 2013)
Exercises 5 (26 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. Let $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$. Define the function $\gamma : \Sigma \times \Sigma \to \mathbb{R}_{\geq 0}$ as follows

$$\gamma(\mathtt{a}, \mathtt{a}) = \gamma(\mathtt{b}, \mathtt{b}) = \gamma(\mathtt{c}, \mathtt{c}) = 0$$
$$\gamma(\mathtt{a}, \mathtt{b}) = \gamma(\mathtt{b}, \mathtt{c}) = \gamma(\mathtt{c}, \mathtt{a}) = 0.5$$
$$\gamma(\mathtt{b}, \mathtt{a}) = \gamma(\mathtt{c}, \mathtt{b}) = \gamma(\mathtt{a}, \mathtt{c}) = 1.5$$

   Let ed_γ be a *weighted edit distance*, where the cost of substituting a character x with a character y is $\gamma(x, y)$. The cost of insertions and deletions is 1.

   (a) It might seem that we can compute $ed_\gamma(A, B)$ using the recurrence for the standard edit distance (page 112 on the lecture notes) except δ is replaced by γ. Show that this is not the case by providing an example for which the recurrence produces an incorrect distance.

   (b) Is ed_γ a metric?

3. 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?

4. 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, $lcss(\mathtt{berlin}, \mathtt{helsinki}) = 4$ since $\mathtt{elin}$ 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 $\mathcal{O}(|A||B|)$.

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

6. Let $P = \mathtt{evete}$ and $T = \mathtt{neeteneeveteen}$.

   (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.