Auxiliary notation: subset of k-minimal values

Given the set $X \subset \mathbb{N}$, and some $k \in \mathbb{N}$, for the sake of convenience let's introduce an auxiliary notation for the subset of k-minimal values: $min_k(X) \subseteq X$, such that $\forall a \in min_k(X), \forall b \in X \setminus min_k(X)$ it follows, that $a < b$, and $|min_k(X)| = k$ (in case if $|X| < k$, then $min_k(X) = X$).

K smallest edit distances

Let $lev_{a,b}(i, j)$ be the set of k smallest edit distances between the first i characters of the string a and the first j characters of the string b.

Let's introduce the following sets:

(1)

$I_{a,b}(i,j) := \{d + 1 \mid d \in lev_{a,b}(i-1, j)\}$	The set of k edit distances obtained through insertions
$D_{a,b}(i,j) := \{d + 1 \mid d \in lev_{a,b}(i, j-1)\}$	The set of k edit distances obtained through deletions
$S_{a,b}(i,j) := \{d + 1_{(a_i \neq b_j)} \mid d \in lev_{a,b}(i-1, j-1)\}$	The set of k edit distances obtained through substitutions
$E_{a,b}(i,j) := I_{a,b}(i,j) \cup D_{a,b}(i,j) \cup S_{a,b}(i,j)$	The set of all $3k$ edit distances

Where $1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise.

Then the set of k smallest edit distances between the first i characters of the string a and the first j characters of the string b is defined as follows:

$$(2) \quad \begin{cases} lev_{a,b}(i,j) := \{max(i,j)\} & \text{when } i = 0 \text{ or } j = 0 \\ lev_{a,b}(i,j) := min_k(E_{a,b}(i,j)) & \text{otherwise} \end{cases}$$

Proof of correctness

Lemma 1

Whenever the arbitrary natural number a doesn't belong to the set $min_k(X)$ and there exists some item of $min_k(X)$ which is greater than a, it means that $a \notin X$.

More formally, let's prove the following statement for any $a \in \mathbb{N}$ and any $X \subset \mathbb{N}$:

$$(3) \quad \Big((a \notin min_k(X)) \land (\exists b \in min_k(X), a < b)\Big) \Rightarrow \Big(a \notin X\Big)$$

Proof

According to the contrapositive proof scheme, let's show that:

$$(4) \quad \neg\Big(a \notin X\Big) \Rightarrow \neg\Big((a \notin min_k(X)) \land (\exists b \in min_k(X), a < b)\Big)$$

Which is equivalent to:

$$(5) \quad \Big(a \in X\Big) \Rightarrow \Big((a \in min_k(X)) \lor (\forall b \in min_k(X), a \geq b)\Big)$$

The latter statement is equivalent to:

$$(6) \quad \Big(a \in X\Big) \Rightarrow \Big((a \in min_k(X)) \lor (a \in X \setminus min_k(X))\Big)$$

Which is a tautology. ■

Proof by the smallest counterexample

Induction Basis:

In case if $i = 0$ or $j = 0$ there is possible only one edit distance, hence the set $lev_{a,b}(i,j)$ contains only one item, namely $max(i,j)$. As far as there is only one possible edit distance - it means, that $lev_{a,b}(i,j) = max_k(\{max(i,j)\}) = \{max(i,j)\}$, which complies to the definition of the subset of k-minimal values.

Induction Hypothesis:

- The set $lev_{a,b}(i-1,j)$ contains the k minimal edit distances between the first $i-1$ characters of the string a and the first j characters of the string b.
  Hence, for every edit distance y between the first $i-1$ characters of the string a and the first j characters of the string b, such that $y \notin lev_{a,b}(i-1,j)$ it follows, that $\forall x \in lev_{a,b}(i-1,j), y > x$.
- The set $lev_{a,b}(i,j-1)$ contains the k minimal edit distances between the first i characters of the string a and the first $j-1$ characters of the string b.
  Hence, for every edit distance y between the first i characters of the string a and the first $j-1$ characters of the string b, such that $y \notin lev_{a,b}(i,j-1)$ it follows, that $\forall x \in lev_{a,b}(i,j-1), y > x$.
- The set $lev_{a,b}(i-1,j-1)$ contains the k minimal edit distances between the first $i-1$ characters of the string a and the first $j-1$ characters of the string b.
  Hence, for every edit distance y between the first $i-1$ characters of the string a and the first $j-1$ characters of the string b, such that $y \notin lev_{a,b}(i-1,j-1)$ it follows, that $\forall x \in lev_{a,b}(i-1,j-1), y > x$.

Inductive Step:

We want to show, that the *Induction Hypothesis* implies, that the set $lev_{a,b}(i,j)$ contains the k smallest edit distances between the first i characters of the string a and the first j characters of the string.

For the sake of contradictions let's assume, that the set $lev_{a,b}(i,j)$ doesn't contain the k smallest edit distances. Thus, there there exists some edit distance $y \notin lev_{a,b}(i,j)$ between the first i characters of the string a and the first j characters of the string b, such that $\exists x \in lev_{a,b}(i,j)$ for which $y < x$:

$$\Big(y \notin lev_{a,b}(i,j)\Big) \wedge \Big(\exists x \in lev_{a,b}(i,j), y < x\Big) \tag{7}$$

Let's rewrite the expression in a following way:

(8)

$$
\begin{aligned}
& \Big(y \notin lev_{a,b}(i,j)\Big) \wedge \Big(\exists x \in lev_{a,b}(i,j), y < x\Big) && \\
\Leftrightarrow\ & \Big(y \notin min_k(E_{a,b}(i,j))\Big) \wedge \Big(\exists x \in min_k(E_{a,b}(i,j)), y < x\Big) && \text{By definition of } lev_{a,b}(i,j) \\
\Leftrightarrow\ & \Big(y \notin min_k(E_{a,b}(i,j))\Big) \wedge \Big(\exists x \in min_k(E_{a,b}(i,j)), y < x\Big) \wedge \Big(\exists x \in min_k(E_{a,b}(i,j)), y < x\Big) && \text{Idempotence} \\
\Rightarrow\ & \Big(y \notin E_{a,b}(i,j)\Big) \wedge \Big(\exists x \in min_k(E_{a,b}(i,j)), y < x\Big) && \text{According to the Lemma 1} \\
\Leftrightarrow\ & \Big(y \notin I_{a,b}(i,j) \cup D_{a,b}(i,j) \cup S_{a,b}(i,j)\Big) \wedge \Big(\exists x \in min_k(E_{a,b}(i,j)), y < x\Big) && \text{By definition of } E_{a,b}(i,j) \\
\Leftrightarrow\ & \Big(y \notin I_{a,b}(i,j)\Big) \wedge \Big(y \notin D_{a,b}(i,j)\Big) \wedge \Big(y \notin S_{a,b}(i,j)\Big) \wedge \Big(\exists x \in min_k(E_{a,b}(i,j)), y < x\Big) && \text{De-Morgan's law} \\
\Leftrightarrow\ & \Big((y-1) \notin lev_{a,b}(i-1,j)\Big) \wedge \Big((y-1) \notin lev_{a,b}(i,j-1)\Big) \wedge && \text{By definition of } I_{a,b}(i,j), \\
& \wedge \Big((y-1_{(a_i \neq b_j)}) \notin lev_{a,b}(i-1,j-1)\Big) \wedge \Big(\exists x \in min_k(E_{a,b}(i,j)), y < x\Big) && D_{a,b}(i,j) \text{ and } S_{a,b}(i,j) \\
\Rightarrow\ & \Big(\forall m \in lev_{a,b}(i-1,j), y-1 > m\Big) \wedge \Big(\forall n \in lev_{a,b}(i,j-1), y-1 > n\Big) \wedge && \text{By Induction Hypothesis} \\
& \wedge \Big(\forall u \in lev_{a,b}(i-1,j-1), y-1_{(a_i \neq b_j)} > u\Big) \wedge \Big(\exists x \in min_k(E_{a,b}(i,j)), y < x\Big) && \\
\Leftrightarrow\ & \Big(\forall m \in I_{a,b}(i,j), y > m\Big) \wedge \Big(\forall n \in D_{a,b}(i,j), y > n\Big) \wedge && \text{By definition of } I_{a,b}(i,j), \\
& \wedge \Big(\forall u \in S_{a,b}(i,j), y > u\Big) \wedge \Big(\exists x \in min_k(E_{a,b}(i,j)), y < x\Big) && D_{a,b}(i,j) \text{ and } S_{a,b}(i,j) \\
\Leftrightarrow\ & \Big(\forall m \in I_{a,b}(i,j) \cup D_{a,b}(i,j) \cup S_{a,b}(i,j), y > m\Big) \wedge \Big(\exists x \in min_k(E_{a,b}(i,j)), y < x\Big) && \text{Reordering} \\
\Leftrightarrow\ & \Big(\forall m \in E_{a,b}(i,j), y > m\Big) \wedge \Big(\exists x \in min_k(E_{a,b}(i,j)), y < x\Big) && \text{By definition of } E_{a,b}(i,j) \\
\Rightarrow\ & \Big(\forall m \in min_k(E_{a,b}(i,j)), y > m\Big) \wedge \Big(\exists x \in min_k(E_{a,b}(i,j)), y < x\Big) && \text{Because } min_k(X) \subseteq X \\
\Leftrightarrow\ & \Big(\forall m \in lev_{a,b}(i,j), y > m\Big) \wedge \Big(\exists x \in lev_{a,b}(i,j), y < x\Big) && \text{By definition of } lev_{a,b}(i,j) \\
\Rightarrow\ & \mathit{Contradiction.} &&
\end{aligned}
$$

Hence, our assumption was wrong. Consequently, the set $lev_{a,b}(i,j)$ contains the k-minimal edit distances.
■