Odległości

20 marca 2025

Słowo wstępne

Lista 1. pozwoliła na zapoznanie się z algorytmem *Odległości Edycyjnej* tj.: dla dwóch napisów s_1 i s_2, określ minimalny koszt, który pozwoli na przekształcenie s_1 w s_2. Dopuściliśmy trzy operacje:

- Dodanie znaku
- Usunięcie znaku
- Zamiana znaku

Następnie rozważyliśmy kolejną operację:

- Transpozycja dwóch sąsiednich znaków

Można dostrzec przynajmniej dwa mankamanety naiwenego działania algorytmu:

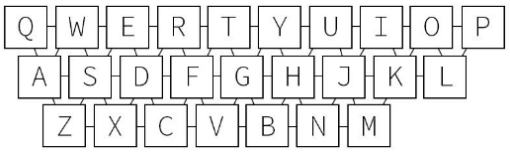
- Wszystkie możliwości były traktowane na równi
- Czas przeszukiwania przestrzeni był niezadowalający

W poniższych zadaniach, dla uproszczenia, rozważamy tylko język angielski (słownik jest podany w materiałach do zadania).

Zadanie 1. Ciężar Wagi Edycyjnej (2 pkt)

Wagi w obu powyższych wagi przypadkach były określane arbitralnie. Rozważmy następującą propozycję wag, opartą na układzie klawiatury *QWERTY*.

Odległość w tym układzie pomiędzy poszczególnymi klawiszami wynosi od 1 do 9, żeby uzyskać wartości od 0 do 2 każdą z tych wartości mnożymy przez $\frac{2}{9}$. Przy tak ujętej odległości możemy zdefiniować konszty następująco:



Rysunek 1: Układ klawiatury QWERTY wraz z zaznaczonym sąsiedztwem

- Dodanie znaku - bez zmian, 1.
- Usunięcie znaku - koszt usunięcia poszczególnego znaku jest równy średniej z odgległości (na klawiaturze) sąsiednich znaków.
- Zamiana znaku - koszt zamiany znaków jest równy odległości (na klawiaturze) pomiędzy wstawianym znakiem a zastępowanym znakiem.

Następnie, dla otrzymanej macierzy kosztów, policz średnią wagę i każdą wartość podziel przez tę średnią.

Zaimplementuj powyższe rozwiązanie i porównaj wyniki otrzymane z implementacją z **Listy 1.** (bez transpozycji).

Zadanie 2. O Wagach Raz Jeszcze (2 pkt)

Wariacją na temat obliczania wag jest następująca idea:

- Dla par klawiszy, które są sąsiadami, przypisz wagę x
- Dla pozostałych par (klawisze, które nie są sąsiadami), przypisz wagę y

Wartości x i y są podawane jako dane wejściowe. Niższa wartość oznacza mniejszy koszt (bardziej preferowany zamiennik). Następnie, dla otrzymanej macierzy kosztów, policz średnią wagę i każdą wartość podziel przez tę średnią.

Zaimplementuj powyższe rozwiązanie (z transpozycjami) i porównaj wyniki otrzymane z implementacją z **Zadania 1.** (z transpozycjami).

Zadanie 3. O Wagach Inaczej (2 pkt)

Poniższe zadanie jest rozwinięciem **Zadania 1.** (powyżej) oraz **Zadania 6.** z **Listy 1.**. W tej wersji problemu rozważymy, w jakis sposób możemy modelować koszty biorąc pod uwagę częstotliwość występowań konkretnych znaków.

Dane pochodzą z pliku tekstowego, możesz założyć, że nie zawierają cyfr, znaków specjalnych ani interpunkcji.

Wzbogać swoje rozwiązanie **Zadania 1.** o koszt oparty na częstotliwościach:

- Dla każdego znaku zlicz jego wystąpienia i podziel przez średnią liczbę wystąpień

W takim ujęciu koszty operacji przedstawiają się następująco:

- Wstawianie znaku *'A'* $= \omega(A)$
- Usunięcie znaku *'A'* $= \omega(A)$
- Zamiana znaku *'A'* na *'B'* $= d(A, B) + \omega(B)$

gdzie: $d(A, B)$ jest znormalizowaną odległością pomiędzy klawiszami A i B na klawiaturze, a $\omega(A)$ jest znormalizowanym kosztem opartym na wystąpieniach.

Zaimplementuj powyższe rozwiązanie i porównaj wyniki otrzymane z implementacją z **Listy 1.** (bez transpozycji).