Algorytmy Tekstowe

Laboratorium 4 – raport

Mateusz Kocot

1. Odległość edycyjna

Zaimplementowano algorytm obliczania odległości edycyjnej. Ponieważ możliwe ma być odtworzenie sekwencji, należy wypełnić całą tablicę *edit_table.* W związku z tym, złożoność pamięciowa wynosi tutaj $O(|x| \cdot |y|)$. Działanie algorytmu na danych przykładach przedstawiono poniżej. Zastosowano funkcję *delta*, która dla tych samych znaków zwraca 0, a dla znaków różnych – 1. Oczywiście, w zależności od potrzeb, funkcja *delta* może wyglądać inaczej. Można np. zmniejszyć zwracaną wartość dla podobnych znaków (np. l i ł albo ö i o).

```
kloc
los
edit distance: 2
1. *k*los <----- inserted k on idx 0
2. klo*c* <----- replaced s with c on idx 3

Lodz
Łódź
edit distance: 3
1. *L*ódź <----- replaced Ł with L on idx 0
2. L*o*dź <----- replaced ó with o on idx 1
3. Lod*z* <----- replaced ź with z on idx 3

quintessence
kwintesencja
edit distance: 5
1. *q*wintesencja <----- replaced k with q on idx 0
2. q*u*intesencja <----- replaced w with u on idx 1
3. quintes*s*encja <----- inserted s on idx 7
4. quintessenc*e*a <----- replaced j with e on idx 11
5. quintessence* <----- deleted a on idx 12

ATGAGGCTCTGGCCCCTG
ATGAATCTTACCGCCTCG
edit distance: 7
1. ATGA*G*TCTTACCGCCTCG <----- replaced A with G on idx 4
2. ATGAG*G*CTTACCGCCTCG <----- replaced T with G on idx 5
3. ATGAGGCT*C*TACCGCCTCG <----- inserted C on idx 8
4. ATGAGGCTCT*G*CCGCCTCG <----- replaced A with G on idx 10
5. ATGAGGCTCTG*G*CCGCCTCG <----- inserted G on idx 11
6. ATGAGGCTCTGGCC*CCTCG <----- deleted G on idx 14
7. ATGAGGCTCTGGCCCCT*G <----- deleted C on idx 17
```

2. Najdłuższy wspólny podciąg

Wykorzystano tokenizer *spaCy* w celu podzielenia tekstu *romeo-i-julia-700.txt* na tokeny – osobne słowa, a następnie stworzono dwa teksty, z których usunięto ok. 3% tokenów. Na tych tekstach przetestowano zaimplementowane algorytmy znajdujące najdłuższy wspólny podciąg. Algorytm

bazujący na tablicy okazał się być, zgodnie z przypuszczeniami, znacznie wolniejszy. Potrzebował on ok. $10s$. Dla porównania, algorytmowi „koralikowemu" wystarczyła ok. $1s$. Tego właśnie algorytmu użyto do znalezienia długości najdłuższego wspólnego podciągu rozpatrywanych tekstów. Wynik wyniósł 2157 słów w porównaniu do liczby wszystkich tokenów – 2272 słów. Zauważmy, że $\frac{2272-2157}{2272} = \frac{115}{2272} \cong 5{,}06\%$, co jest z oczekiwaniami.

3. Diff

Na podstawie wcześniejszych algorytmów, zaimplementowano funkcję działającą podobnie do narzędzia diff. Przetestowano ją na tekstach z poprzedniego punktu. Fragment wyniku zamieszczono poniżej. Cały wynik znajduje się w pliku *results.txt*.

```
>0 Shakespeare
<0 William Shakespeare
>5 978-83-288-2903-9
<5 ISBN 978-83-288-2903-9
>10   * ESKALUS — książę panujący Weronie
>11   * PARYS — młody Weroneńczyk szlachetnego rodu, krewny księcia
<10   * ESKALUS — książę panujący w Weronie
<11 * PARYS — młody Weroneńczyk szlachetnego rodu, krewny księcia
>17 * TYBALT — krewny Pani Kapulet
<17   * TYBALT — krewny Pani Kapulet
>24   TRZECH MUZYKANTÓW
<24   * TRZECH MUZYKANTÓW
>32   * Obywatele weroneńscy, różne osoby płci obojej, liczący się do
przyjaciół obu domów, maski, straż wojskowa i inne osoby.
<32   * Obywatele weroneńscy, różne osoby płci obojej, liczący się do
obu domów, maski, straż wojskowa i inne osoby.
>45 Dwa rody, zacne jednako i sławne —
>46 Tam, gdzie się rzecz ta rozgrywa, w Weronie,
<45 Dwa zacne jednako i sławne —
<46 Tam, się rzecz ta rozgrywa, w Weronie,
>55 Tej ich miłości zbyt bolesny
>56 I jak ojców nienawiść nie zmienia,
<55 Tej ich miłości przebieg zbyt
<56 I jak się ojców nienawiść nie zmienia,
>58 Dwugodzinnego treścią
<58 Dwugodzinnego treścią przedstawienia,
>60 Które otoczcie cierpliwymi względy,
>61 Jest w nim złego, my usuniem błędy…
<60 Które otoczcie cierpliwymi
<61 Jest w nim co złego, my usuniem błędy…
>77 Dalipan, Grzegorzu, będziem darli pierza.
<77 Dalipan, Grzegorzu, nie będziem darli pierza.
>102 Tak, ale nie zaraz się dać rozruchać.
<102 Tak, ale nie zaraz zwykłeś się dać rozruchać.
>117 Te psy z domu Montekich rozruchać mię mogą tylko do stania na
miejscu. Będę jak mur dla każdego mężczyzny i każdej kobiety z tego
domu.
<117 Te psy z domu Montekich rozruchać mię mogą tylko do stania na Będę
jak mur dla każdego mężczyzny i każdej kobiety z tego domu.
>122 To właśnie pokazuje twoją słabą stronę; mur dla nikogo niestraszny
i tylko słabi go się trzymają.
<122 To pokazuje twoją słabą stronę; mur dla nikogo niestraszny i tylko
go się trzymają.
```

>127 Prawda, dlatego to kobiety, najsłabsze, tulą się zawsze do muru. Ja też odtrącę od muru ludzi Montekich, a kobiety Montekich przyprę do muru.
<127 Prawda, dlatego to kobiety, jako tulą się zawsze do muru. Ja też odtrącę od muru ludzi Montekich, a kobiety Montekich przyprę do muru.
<140 GRZEGORZ
>141
>152 Tym lepiej, że się liczysz do zwierząt; bo gdybyś się liczył do ryb, to byłbyś pewnie sztokfiszem. Weź no się za instrument, bo oto nadchodzi dwóch domowników Montekiego.
<152 Tym lepiej, że się liczysz do zwierząt; bo gdybyś się liczył ryb, to byłbyś pewnie sztokfiszem. Weź no się za instrument, bo oto nadchodzi dwóch domowników Montekiego.
>159 Mój giwer już dobyty: zaczep ich, ja z tyłu.
<159 Mój giwer już dobyty: zaczep ich, ja stanę z tyłu.
<192 ABRAHAM
>193
>228 Zaczepki szukasz?
<228 Zaczepki waść szukasz?
>238 Jeżeli jej szukasz, to jestem na waścine usługi. Mój pan tak dobry jak i wasz.
<238 Jeżeli jej szukasz, to jestem waścine usługi. Mój pan tak dobry jak i wasz.
>257 Powiedz: lepszy. Oto nadchodzi jeden z krewnych mego pana.
<257 Powiedz: lepszy. Oto nadchodzi jeden z mego pana.
>279 Rozdziela ich swoim mieczem. /
<279 / Rozdziela swoim mieczem. /
>287 Do mnie, pilnuj swego życia.
<287 Do mnie, Benwolio! swego życia.
>292 Przywracam tylko pokój. miecz nazad
<292 Przywracam tylko pokój. Włóż miecz nazad
>300 Szatana, wszystkich Montekich ciebie.
<300 Szatana, wszystkich Montekich i ciebie.