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.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. Dla pierwszych 5000 tokenów działał on ok. 45s. Dla porównania, algorytmowi „koralikowemu” wystarczyło ok. $5s$. Tego właśnie algorytmu użytu do znalezienia długości najdłuższego wspólnego podciągu rozpatrywanych tekstów. Wynik wyniósł 25024 w porównaniu do długości tekstów wynoszącej ok. 25850.

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.*

```
>11   * PARYS — młody Weroneńczyk szlachetnego rodu, krewny księcia
<11   * PARYS — Weroneńczyk szlachetnego rodu, krewny księcia
>14   * ROMEO — syn Montekiego
>15   * MERKUCJO — krewny księcia
<14 * ROMEO — syn Montekiego
<15   MERKUCJO — krewny księcia
>19   * JAN — brat z tegoż zgromadzenia
<19 * JAN — brat z tegoż zgromadzenia
>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, się do przyjaciół
obu domów, maski, straż wojskowa i inne osoby.
>37 Rzecz odbywa się przez większą część sztuki Weronie, przez część
piątego aktu w Mantui.
<37 Rzecz odbywa się przez większą część sztuki w Weronie, przez część
piątego aktu w Mantui.
>47 nowej zbrodni pchają dawne,
<47 Do nowej zbrodni pchają złości dawne,
>51 Pod najstraszliwszą z gwiazd, kochanków dwoje;
<51 Pod najstraszliwszą z gwiazd, kochanków
>55 Tej miłości przebieg zbyt bolesny
<55 Tej ich miłości przebieg zbyt bolesny
>61 Jest nim co złego, my usuniem błędy…
<61 w nim co złego, my usuniem błędy…
>87 Ale będziemy darli koty, jak z zadrą.
<87 Ale będziemy darli koty, jak z nami zadrą.
>112 Rozruchać się tyle znaczy co ruszyć się z miejsca; być walecznym
to stać nieporuszenie: pojmuję więc, że skutkiem rozruchania się twego
będzie - drapnięcie.
<112 Rozruchać się tyle znaczy co ruszyć się z miejsca; być to stać
nieporuszenie: pojmuję więc, że skutkiem rozruchania się twego będzie -
drapnięcie.
>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
miejscu. Będę jak mur dla każdego mężczyzny i każdej kobiety z tego
>122 To pokazuje twoją słabą stronę; mur dla nikogo niestraszny i tylko
słabi go się trzymają.
<122 To właśnie pokazuje twoją słabą stronę; mur dla nikogo niestraszny
tylko słabi go się trzymają.
>127 Prawda, dlatego to kobiety, jako najsłabsze, tulą się zawsze do
muru. Ja też odtrącę od ludzi Montekich, a kobiety Montekich przyprę do
muru.
```

<127 Prawda, dlatego to kobiety, jako najsłabsze, tulą się zawsze muru. Ja też odtrącę od muru ludzi Montekich, a kobiety Montekich przyprę do muru.
>132 Spór jest tylko między naszymi panami i między nami, ich ludźmi.
<132 Spór jest tylko między naszymi i między nami, ich ludźmi.
>147 Nie inaczej: wtłoczę miecz w każdą po kolei. że się do lwów liczę.
<147 Nie inaczej: wtłoczę miecz w każdą po kolei. Wiadomo, że się do lwów liczę.
>152 Tym lepiej, że się liczysz do zwierząt; bo 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 do ryb, to byłbyś pewnie sztokfiszem. Weź no się za instrument, bo oto nadchodzi dwóch domowników Montekiego.
>189 Nie jak chcą, ale jak śmią. Ja im gębę wykrzywię; hańba im, jeśli to ścierpią.
<189 Nie jak chcą, ale jak śmią. Ja gębę wykrzywię; hańba im, jeśli to ścierpią.
>211 Będziemy–ż mieli prawo za sobą, jak powiem: tak jest?
<211 Będziemy–ż mieli prawo za sobą, powiem: tak jest?
>221 Nie, mości panie; nie skrzywiłem się na was, tylko skrzywiłem się tak sobie.
<221 Nie, mości panie; nie się na was, tylko skrzywiłem się tak sobie.
>226 / do /
<226 / do Abrahama /
>248 Niech i będzie.
<248 Niech i tak
>257 Powiedz: lepszy. Oto nadchodzi jeden z krewnych mego pana.
<257 Powiedz: lepszy. Oto nadchodzi jeden z mego pana.
>272 Dobądźcie mieczów, jeśli macie serca. Grzegorzu, o swoim pchnięciu.
<272 Dobądźcie mieczów, jeśli macie serca. Grzegorzu, pamiętaj o swoim pchnięciu.
>286 Cóż to? krzyżujesz oręż parobkami?
<286 Cóż to? krzyżujesz oręż z
>293 Albo wraz ze mną rozdziel nim tych
<293 Albo wraz mną rozdziel nim tych ludzi.
>303 / Walczą. Nadchodzi kilku przyjaciół obu partii i mieszają się do zwady; wkrótce wchodzą mieszczanie z pałkami. /
<303 / Walczą. Nadchodzi kilku przyjaciół obu partii i mieszają się do zwady; wkrótce potem wchodzą mieszczanie z pałkami. /
>309 Precz z Montekimi, precz Kapuletami!
<309 Precz z Montekimi, precz z Kapuletami!
>330 / Wchodzą Monteki i Pani Monteki. /
<330 / Wchodzą Monteki i Pani /
>333 MONTEKI