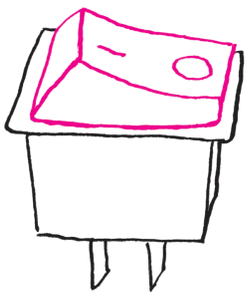
Informatyczny kącik olimpijski (71): Różne słowa

W tym kąciku omówimy zadanie *Różne słowa* z Obozu Naukowo-Treningowego im. A. Kreczmara w 2013 roku. Dane jest n słów o długości $k = 5$. Słowa składają się z małych i wielkich liter alfabetu łacińskiego. Naszym zadaniem jest stwierdzić, czy istnieje wśród nich para *kompletnie różnych* słów, czyli słów, które na odpowiadających sobie pozycjach mają różne litery. Słowa $u = u_1 \dots u_k$ i $v = v_1 \dots v_k$ są więc kompletnie różne, jeśli $u_i \neq v_i$ dla każdego $i = 1, \dots, k$. Mimo prostego sformułowania rozwiązanie zadania wymaga pewnej pomysłowości. Przedstawimy dwa różne podejścia do rozwiązania.



Pierwszy pomysł będzie opierał się na zasadzie włączeń-wyłączeń. W pierwszym kroku dla każdego słowa wyznaczymy wszystkie *wzorce*, do których ono pasuje. Wzorcem nazywamy tutaj słowo długości k, które oprócz liter może zawierać znaki zapytania – znak zapytania zastępuje we wzorcu dowolną literę. Przykładowo, ze słowa `abcab` można otrzymać m.in. wzorce `a?c??` i `????b`. Zastępując każdą możliwą kombinację liter w słowie znakami zapytania, dla jednego słowa otrzymamy 2^k różnych wzorców. Wszystkie te informacje możemy następnie połączyć w jedną tabelę, która dla każdego wzorca zapamięta liczbę wejściowych słów pasujących do niego. Ze względu na konieczność posortowania par: wzorzec-słowo złożoność czasowa konstrukcji takiej tabeli wyniesie $O(n2^k \cdot \log(n2^k) \cdot k) = O(n \log n \cdot 2^k k^2)$.

Zaopatrzeni w tabelę wzorców możemy już zastosować zasadę włączeń-wyłączeń. Dla każdego z wejściowych słów chcemy wyznaczyć liczbę par kompletnie różnych słów, w skład których to słowo wchodzi. Ustalmy jedno słowo wejściowe w i oznaczmy przez A_i liczbę wejściowych słów, które zgadzają się ze słowem w na i-tej pozycji ($i \in \{1, \dots, k\}$). Wówczas wynik dla słowa w możemy obliczyć ze wzoru:

$$n - |A_1 \cup A_2 \cup \ldots \cup A_k| = n - \sum_i |A_i| + \sum_{i_1 < i_2} |A_{i_1} \cap A_{i_2}| - \sum_{i_1 < i_2 < i_3} |A_{i_1} \cap A_{i_2} \cap A_{i_3}| + \ldots$$

Zauważmy, że składnik $|A_{i_1} \cap \ldots \cap A_{i_r}|$ w powyższej sumie odpowiada liczbie słów wejściowych pasujących do wzorca powstałego z w poprzez zastąpienie liter na wszystkich indeksach poza $i_1, \dots, i_r$ znakami zapytania. Liczbę takich słów możemy odczytać bezpośrednio z tabeli wzorców. Łączny koszt drugiej fazy rozwiązania to $O(n \cdot 2^k k)$, jest on zdominowany przez koszt pierwszej fazy. Zauważmy, że opisane rozwiązanie pozwala nie tylko stwierdzić, czy wśród podanych słów jest jakaś para kompletnie różnych słów, lecz także wyznaczyć liczbę takich par.

Drugie rozwiązanie koncentruje się na alfabecie, czyli na zbiorze liter występujących w słowach. W naszym zadaniu alfabet ma $A = 52$ litery. Zastanówmy się, co by było, gdybyśmy mieli do czynienia z dużo mniejszym alfabetem: alfabetem dwuliterowym. Wówczas mielibyśmy tylko 2^k różnych słów i dla każdego słowa umielibyśmy wskazać jedyne słowo, które tworzyłoby z nim parę kompletnie różnych słów; byłoby to słowo stanowiące „negację" pierwszego. Rozwiązanie zadania w tym przypadku nie przedstawiałoby żadnych trudności. W naszym zadaniu nie mamy do czynienia z tak prostym przypadkiem. Spróbujemy jednak sprowadzić je do tego przypadku, wprowadzając do rozwiązania element losowości.

Każdej literze alfabetu przyporządkujemy losowo bit 0 lub 1. Co więcej, uczynimy to osobno dla każdej pozycji w słowach, przy czym losowania na poszczególnych pozycjach będą niezależne. W ten sposób sprowadzimy zadanie do przypadku binarnego, lecz, niestety, utracimy pewien zasób informacji. Dokładniej, wiemy, że jeśli po tym przyporządkowaniu jakieś dwa słowa są wzajemnie negacjami, to przed zamianą liter były one kompletnie różne, jednak implikacja odwrotna nie musi zachodzić. Zastanówmy się, jakie jest prawdopodobieństwo tego, że dana para kompletnie różnych słów przeszła na parę słów będących wzajemnie negacjami. W przypadku jednej pozycji jest ono równe $\frac{1}{2}$, gdyż taka jest szansa na to, że dwie ustalone różne litery oryginalnego alfabetu otrzymały w losowaniu różne bity. Ponieważ losowania na poszczególnych pozycjach były niezależne, więc szukane prawdopodobieństwo z uwzględnieniem wszystkich pozycji jest równe $\frac{1}{2^k}$.

Aby zwiększyć nasze szanse, możemy całe losowanie wielokrotnie powtórzyć. Szansa na to, że po wykonaniu p prób dane dwa kompletnie różne słowa ani razu nie okażą się wzajemnie negacjami, wynosi $(1 - \frac{1}{2^k})^p$. Jeśli zatem wykonamy $p = C \cdot 2^k$ losowań, gdzie $C = 20$, prawdopodobieństwo porażki będzie znikome:

$$\left(1 - \frac{1}{2^k}\right)^{C \cdot 2^k} = \left(\left(1 - \frac{1}{2^k}\right)^{2^k}\right)^C < \frac{1}{e^C} < 10^{-8}.$$

Złożoność całego rozwiązania wynosi $O((n + A) \cdot 2^k k \cdot C)$. Za pomocą tego samego podejścia można bez problemu wyznaczyć jakieś m par kompletnie różnych słów z wejścia (lub wszystkie takie pary, jeśli jest ich mniej niż m), czego wymagało oryginalne polecenie omawianego zadania.

Jakub RADOSZEWSKI