

Antyplagiat 1

<https://szkopul.edu.pl/problemset/problem/ap1/site>

Znajdowanie niesamodzielnych prac to problem nie tylko szkolnych wypracowań. Przemówienia, powieści, programy – w każdej niemal dziedzinie jest pokusa by podpisać się pod nieswoim dziełem.

Jednym z najślawniejszych przypadków sporu o plagiat jest piosenka “Blurred Lines” napisana i wykonana przez Robina Thicke’a oraz Pharrella Williamsa:

<https://youtu.be/yyDUC1LUXSU>

Ta piosenka była 12 tygodni numerem 1 na liście przebojów Billboard Hot 100 – amerykańskiego, muzycznego tygodnika Billboard. To najdłużej emitowany singel w Stanach Zjednoczonych.

W 2013 roku dzieci Marvinina Gaye'a oskarżyły twórców o plagiat. Piosenka miała być kopią przeboju ich ojca „Got to Give It Up” z 1977 roku:

<https://youtu.be/Ayyy-03ITDg>.

Sąd przyznał rację spadkobiercom Gaye’a zasądzając 5 milionów USD odszkodowania. A Ty co sądzisz?

Bez względu na wszystko systemy które sprawdzają na ile podobne są dwa utwory stają się coraz ważniejsze. Napisz swoją wersję!

Wejście

Wejście składa się z dwóch tekstów, których podobieństwo mamy analizować. W pierwszej linii wejścia znajduje się jedno słowo które może mieć obok siebie spacje i tabulatory:

TEXT1

oznaczające początek tekstu TEXT1. Nie jest znana ilość linii tekstu TEXT1.

Koniec tekstu TEXT1 a zarazem początek tekstu TEXT oznacza linia z jednym słowem:

TEXT2

Podobnie słowo TEXT2 może mieć dowolną ilość spacja po lewej i prawej stronie.

Nie jest znana ilość linii tekstu TEXT1.

Linie tekstów TEXT1 oraz TEXT2 mogą być puste, zawierać tylko białe znaki – spacje czy tabulatory. Takie linie pomijamy w analizie.

Wiemy, że łącznie w TEXT1 będzie maksimum milion znaków – nie wliczając linii z TEXT1. Podobnie, niezależnie w TEXT2 będzie maksimum milion znaków

Wyjście

Twój program powinien wypisać te linie tekstu TEXT2 które są identyczne z TEXT1. Dwie linie są identyczne, jeśli po pominięciu pomijając białych znaków – spacji, tabulatorów – każdy kolejny znak pierwszej linii jest identyczny z każdym kolejnym znakiem drugiej linii.

Jeśli TEXT2 ma kilka identycznych linii z TEXT1 to wypisujemy taką linię jednokrotnie z pominięciem białych znaków.

Linie tekstu TEXT2 które są identyczne z TEXT1 powinny być wypisane w następującej kolejności:

- Najpierw najdłuższe linie
Im dłuższa identyczna linia w obydwu tekstach tym większe prawdopodobieństwo plagiatu
- Jeśli dwie linie mają identyczną długość najpierw wypisujemy tą która jest pierwsza leksykograficznie.

Przykład 1

Wejście

TEXT1

Kochaj, zeby zyc, i zyj, zeby kochac.

Mysle wiec jestem.

Wyznacz sobie cel, a na pewno sie uda.

M y s l e w i e c j e s t e m .

Jestem wiec mysle.

Mysle wiec CHYBA jestem.

Wyzna cz sobie cel, a n a pewno sie ud a.

Mysle wiec jestem..

TEXT2

Jajko czy kura - co pierwsze?

Jestem wiec my sle.

W y z n a c z sobie cel, a na pewno sie uda.

Mysle wiec jes tem..

Mysle w i e c j e s t e m.

Mysle w i e c j e s t e m.

Mysle wiec chyba jestem.

M y s l e w i e c j e s t e m.

Wyznacz s obie cel, a na pewno sie uda.

Jestem wiec my sle.

Wyjście

Wyznaczsobiecel,anapewnosieuda.

Myslewiecjestem..

Jestemwiecmysle.

Myslewiecjestem.

Wyjaśnienie

Po usunięciu białych znaków oraz pustych linii TEXT2 zawiera następujące identyczne linie z TEXT1:

Wyznaczsobiecel,anapewnosieuda.

Myslewiecjestem..

Jestemwiecmysle.

Myslewiecjestem.

Pomijając puste znaki, linia

Myslewiecjestem.

powtarza się trzykrotnie w TEXT2 ale wypisujemy ją raz.

Linie wypisujemy od najdłuższych – czyli najpierw będą linie

Wyznaczsobiecel,anapewnosieuda.

Myslewiecjestem..

jako najdłuższe.

Linie

Jestemwiecmysle.

Myslewiecjestem.

Są tej samej długości więc wypisujemy w kolejności leksykograficznej – zgodnej ze standardowym porównywaniem kodów ASCII. To znaczy

- Najpierw jest wypisywane to słowo które ma pierwszy mniejszy znak ASCII
- Jeśli pierwsze znaki mają identyczny kod ASCII, porównujemy kod ASCII drugiego znaku
- itd