Zadanie 13

Programowanie Java

Dzisiaj skonstruujesz prosty program antyplagiatowy. Celem jest porównanie skryptów tekstowych (lub tekstów) w celu wykrycia, czy nie są do siebie zbyt podobne. Oczywiście, nie muszą być identyczne, być może skopiowana została tylko część pliku lub zostały wstawione puste linie.

Przede wszystkim musisz być w stanie określić ilościowo podobieństwo dwóch ciągów znaków. Najprostszą miarą jest podobieństwa jest odległość Hamminga między dwoma ciągami znaków. Jest to liczba pozycji, na których ciągi znaków się różnią. Mówiąc bardziej technicznie, jest to miara minimalnej liczby zmian wymaganych do przekształcenia jednego ciągu w drugi. Na przykład jeden ciąg to **medication**, drugi to **meditation**, w tym przypadku odległość Hamminga wynosi 1. Wynosi ona 0, tylko wtedy, gdy ciągi są identyczne. Ciągi mogą mieć różną długość, na przykład **speed** i **speeding**, w tym przypadku odległość Hamminga wynosi 3.

Aby porównać ciągi, należy utworzyć klasę Hamming z publiczną metodą compare, która porównuje dwa ciągi znaków i zwraca odległość Hamminga. Metoda powinna zgłosić wyjątek, gdy jeden z ciągów jest pusty. Przetestuj metodę na kilku różnych przykładach.

Ciągi znaków mogły różnić się tylko spacjami, tabulatorami, itp. Aby poradzić sobie z tą sytuacją, należy usnąć z ciągu znaków, spacje, tabulatory, podkreślenia. Należy zatem wyczyścić oba ciągi, a następnie porównać je przy użyciu utworzonej metody.

Gdy będziesz w stanie określić ilościowo podobieństwo ciągu, skonstruuj klasę CheckPlagiarism. Klasa powinna mieć publiczną metodę CompareFiles , która ładuje dwa pliki i porównuje ich wiersze przy użyciu odległości Hamminga. Oczywiście podobne linie mogą znajdować się w różnych miejscach w plikach. Aby je wykryć, zacznij od pliku zawierającego więcej linii i porównaj dowolny wiersz z pierwszego pliku ze wszystkimi wierszami drugiego pliku, rejestrując minimalną odległość Hamminga. Następnie oblicz średnią z minimalnych obliczonych odległości Hamminga. Jeśli ta średnia wartość jest poniżej pewnego progu (reprezentowanego przez jakąś zmienną statyczną), możesz sklasyfikować oba pliki jako "Wykryto plagiat".

Utwórz raport dla porównywanych dwóch plików tekstowych, zawierający informacje jaka jest średnia wartość minimalnej odległości Hamminga i ile wierszy jest identycznych w obu plikach. Jeśli ta średnia wartość wynosi zero, oznacza to, że oba pliki są identyczne. Musisz również poprawnie przechwytywać wyjątki na wypadek, gdybyś miał puste wiersze w którymś z plików wejściowych.

W zadaniu skorzystaj z klas strumieni w java.io takich jak FileReader, FileWriter, StringReader, LineNumberReader. Sprawdź, czy program radzi sobie z obsługą strony kodowej (Unicode) w sprawdzanych tekstach.