Podstawowy warsztat informatyka — lista 12

Zadanie 1 (1 punkt). Pod adresem `http://www.ii.uni.wroc.pl/~jmi/Dydaktyka/PWI/dane/` znajdują się pliki, które będą potrzebne w kolejnych zadaniach. Ściągnij je wszystkie na dysk jednym wywołaniem programu `wget`. Sprawdź programem `file`, jakiego kodowania używają pliki. Następnie wykorzystaj program `iconv` do przekonwertowania wszystkich plików na UTF-8 (oczywiście oprócz tych, które już są w tym formacie).

Jesteśmy gotowi do scalenia wszystkich plików w jeden. Przemyśl, w jakiej kolejności należy poskładać te pliki, aby wpisy były chronologicznie uporządkowane. Scal te pliki używając podstawowych poleceń linuksowych (`cat, >>, ...`). Plik wynikowy nazwij `dane`.

Zadanie 2 (1 punkt). Sprawdź, czy studenci częściej rozmawiają o teatrze, czy o operze. Uwzględnij przy tym różne odmiany tych słów. Jakie czynniki mogą zaburzyć wyniki Twojego testu?

Zadanie 3 (1 punkt). Dane z pliku `dane` były zbierane w kolejnych latach różnymi sposobami, z których jedne były lepsze, inne gorsze. Chcemy mieć pewność, że w tym pliku każda linia oznacza dokładnie jedną wypowiedź. Używając programu `grep` napisz polecenie, które usunie linie, które nie są w formacie wypowiedzi, tzn. nie są postaci „[(jakiś ciąg znaków)] nazwa użytkownika :”. *Rada: Nie zapisuj od razu do pliku, z którego czytasz, bo inaczej błędnym zapytaniem możesz zniszczyć swoje dane. Lepiej utwórz plik pomocniczy, a gdy uznasz, że jest wystarczająco dobry, użyj poleceń rm i mv.*

Zadanie 4 (3 punkty*). W dyskusjach internetowych znaki zapytania i wykrzyknienia są często wielokrotnie powtarzane. Powiemy (sugerując się informacjami ze strony internetowej PWN), że wypowiedź jest *przestankowo estetyczna*, jeśli każdy jej fragment składający się wyłącznie ze znaków zapytania i wykrzyknienia należy do zbioru $\{!, !!, !!!, ?, !?, ?!\}$. W tym zadaniu będziemy chcieli poprawić w pliku `dane` każdą wypowiedź tak, by była przestankowo estetyczna, ale tak, aby stracić przy tym jak najmniej informacji. W tym celu napisz skrypt (albo w bashu, korzystając z poleceń `sed, tr` i innych, albo w `awk`u), który:

- Usunie zbędne odstępy między znakami wykrzyknienia i zapytania, np. zamieni ciąg „? !? !” na „?!?!”.
- Usunie znaki powstające w wyniku puszczenia klawisza Shift, tzn. usunie jedynki z ciągu „!!!!!!!!!!!!1111” oraz ukośniki z ciągu „?????//”.
- Zamieni otrzymane wypowiedzi na wypowiedzi przestankowo estetyczne, zgodnie z następującymi zasadami: każdy ciąg znaków zapytania zamieniamy na jeden znak zapytania, ciąg przynajmniej czterech wykrzykników zamieniamy na trzy wykrzykniki, natomiast ciąg składający się z wykrzykników i znaków zapytania zamieniamy na „!?” lub „?!”, jako pierwszy ustawiając ten znak, który był pierwszy w zastępowanym ciągu (tzn. zamieniamy „!!!?!?!!!” na „!?” i „?!!!!!!!!!!!!!!!” na „?!”.

Zadanie 5 (1 punkt). Znajdź w pliku `dane` te wypowiedzi, które zawierają jakiś trzyliterowy ciąg powtórzony co najmniej trzy razy.