

(see english version below)

Z uwagi na wiele pytań odnośnie znalezienia powiązania film - kraj pochodzenia dajemy odpowiedź.

Plik „przypadek.csv” to fragment danych z tabeli title.akas.tsv ograniczonych do jednego filmu tt0084549.

Kraj pochodzenia możemy wydedukować w następujący sposób:

- oryginalny tytuł to "Przypadek" (linia 2; *isOriginalTitle* = 1; *title* = Przypadek)
 - wyszukujemy wiersze, gdzie jest taka sama wartość w kolumnie *title*; mamy linię 18
 - w linii 18 wartość w kolumnie *region* jest określona i wynosi ona PL
- Korzystamy z zewnętrznej wiedzy, że PL, to Polska.

Uwaga, ze względu na specyfikę danych mogą być kolejne problemy.

Zasada jest taka, aby samodzielnie te problemy zidentyfikować, zanalizować, zaproponować rozwiązanie i opisać tą analizę podczas omawiania wyników.

Przykładowo mogą być filmy:

- gdzie nie uda się nam przyporządkować kraju/grupy krajów. Ile takich filmów jest?
- gdzie np. dla języka angielskiego może nie być możliwe określenie kraju pochodzenia:
 - można traktować, że ten film to koprodukcja wszystkich krajów
 - można skleić (niektóre) kraje w grupy, np Wielką Brytanię, USA, ... i traktować je razem jako jeden kraj (czyli wtedy porównujemy np anglojęzyczne westerny z włoskimi).

W podsumowaniu analizy oczekujemy informacji ile takich filmów niejednoznacznie przypisanych było.

I na koniec, jako ostateczność, możemy niektóre dane potraktować jako błędne i usunąć.

Proszę wtedy podać ile było takich danych. Ilość usuniętych danych nie powinna przekroczyć 10%.

Due to many questions regarding finding the connection between the film and its country of origin, we provide a hint.

The "przypadek.csv" file is a piece of data from the title.akas.tsv table limited to one movie tt0084549.

We can deduce the country of origin as follows:

- original title is "Przypadek" (row 2; *isOriginalTitle* = 1; *title* = Przypadek)
- find rows, where in column *title* we have the same value (here row 18)
- in row 18 the value in the column *region* is defined and equals PL

We use external knowledge that PL is Poland.

Please note that due to the nature of the data, there may be further problems.

The rule is to identify these problems yourself, analyze them, propose a solution and describe this analysis when discussing the results.

Example, there could be movies:

- for which we fail to assign a country/group of countries. How many such films are there?
- where, for example, for English it may not be possible to determine the country of origin:
 - such a film can be treated as a co-production of all countries
 - you can put (some) countries into groups, e.g. Great Britain, USA, ... and treat them together as one country (i.e. then we compare, for example, English-language westerns with Italian ones).

In the summary of the analysis, we expect information on how many such films were ambiguously assigned.

And finally, as a last resort, we may treat some data as incorrect and delete it.

Then please provide how much such data there was. The amount of data deleted should not exceed 10%.