

Wstęp do technologii baz danych (2016Z):

zadanie zaliczeniowe

Pliki `hocl-ge2017-results-full.csv` i `hocl-ge2017-results-summary.csv` zawierają informacje o wynikach wyborów powszechnych, jakie odbyły się w Zjednoczonym Królestwie Wielkiej Brytanii i Północnej Irlandii 8 czerwca 2017 r. (źródło: <http://researchbriefings.parliament.uk/ResearchBriefing/Summary/CBP-7979>).

Zadanie polega na przetworzeniu tych plików, za pomocą programu w języku Python, w bazę danych SQLite o znormalizowanej strukturze. Stworzony program powinien zawierać definicję danych i procedury analizujące treść plików i wypełniające tabele bazy na jej podstawie.

Oddzielny program (lub ten sam, wywołany z odpowiednim argumentem) powinien z wykorzystaniem zapytania SQL wypisać na podstawie zawartości stworzonej bazy wynik wyborów, tzn. listę partii ubiegających się o mandaty wraz z liczbą uzyskanych przez każdą z nich mandatów. W wyborach parlamentarnych w UK okręgi są jednomandatowe i obowiązuje zasada prostej większości, tzn. kandydat który uzyskał najwięcej głosów w swoim okręgu uzyskuje mandat. Kolejną funkcją programu niech będzie wyświetlenie histogramu rozkładu procentu głosów ważnych jaki w swoich okręgach uzyskali zwycięscy kandydaci.

Wskazówki:

- UK składa się z krajów (*country*): England, Wales, Scotland, ...
- każdy kraj składa się z regionów (niektóre być może tylko z jednego)
- każdy region ma, oprócz nazwy, kod identyfikacyjny *ons_region_id*
- regiony podzielone są na okręgi wyborcze (*constituency*)
- jednoznacznym identyfikatorem okręgu jest *ons_id*, a pozostałe atrybuty to nazwa (*constituency_name*) i typ (*constituency_type*) - wiejski (*county*) lub miejski (*borough*)
- oprócz już wymienionych kolumn, z pliku `hocl-ge2017-results-summary.csv` należy wykorzystać jeszcze wartości *declaration_time* (data i godzina ustalenia wyniku wyborczego) oraz *electorate*, *valid_votes* i *invalid_votes* (odpowiednio: liczba zarejestrowanych wyborców, oddanych głosów ważnych, i głosów nieważnych). Dane z pozostałych kolumn można pominąć
- w drugim pliku - `hocl-ge2017-results-full.csv` znajdziemy nazwy i kody poszczególnych partii (*party_name* i *party_abbreviation*), kandydatów poszczególnych partii w poszczególnych okręgach - imię, nazwisko, płeć, oraz dwie flagi: *sitting_mp* (czy jest to poseł/posłanka ubiegający się o reelekcję) i *former_mp* (czy był/a już posłem w jakiegokolwiek poprzedniej kadencji), oraz liczbę głosów (*votes*). Dalszych kolumn nie wykorzystujemy.
- zbudowana baza powinna być znormalizowana, w szczególności unikać redundancji danych obecnej w plikach csv.