Nr rejestracyjny: 2016/23/B/HS2/01323; Kierownik projektu: prof. dr hab. Adam Tomasz Pawłowski
Metody i narzędzia lingwistyki korpusowej w badaniach bibliografii polskich wydawnictw zwartych z lat 1997-2017.

Streszczenie popularnonaukowe

Historie państw, społeczeństw i narodów pisane są nie tylko przez profesjonalnych historiografów. Podobnie wiedza o otaczającym nas świecie mieści się nie tylko w źródłach naukowych tworzonych przez badaczy. Wiedzę taką można także wydobywać ze źródeł tekstowych, które tworzono w jakimś praktycznym celu – na przykład po to, by ułatwić korzystanie z zasobów bibliotek. Właśnie bibliografia będzie przedmiotem analiz w proponowanym projekcie. Jednak nie jej wąski wycinek, liczący tysiące lub dziesiątki tysięcy pozycji, lecz pełne dwadzieścia lat funkcjonowania polskiego rynku wydawniczego (1997–2017), pozwalające na stworzenie zbioru setek tysięcy krótkich opisów książek. Bibliografie jako wielkie zasoby informacji są niezwykle bogatymi źródłami wiedzy. Pozwalają na przykład na wskazanie trendów cywilizacyjnych (można zastanowić się, jak będzie przebiegać krzywa frekwencji wyrazów *terroryzm* lub *komputer* w tekstach kilkudziesięciu ostatnich lat, można sprawdzić, czy w dobie pośpiechu tytuły stają się coraz dłuższe itd.). W wielkich bibliografiach możliwe jest odkrywanie wzajemnych relacji między tytułami (będziemy testować algorytmy automatycznego grupowania ich według dziedzin). Można też wskazać na pewne cechy językowe, które w wielu wypadkach pozwalają, na podstawie samego tytułu, określić płeć autora.

Czy niewielki zespół badaczy może podołać takiemu zadaniu? Przetwarzanie zbioru ponad pięciuset tysięcy pozycji (a przeprowadzone testy wskazują, że tyle rekordów będzie można wygenerować z baz Biblioteki Narodowej) – praktycznie niewykonalne metodą ręczną – jest możliwe dopiero dzięki połączeniu narzędzi automatycznego przetwarzania tekstu i doświadczenia zdobytego przy pracy nad wielkimi korpusami tekstów z wiedzą informatologa i specjalisty z zakresu bibliografii. Taki jest też cel projektu. Mamy zamiar połączyć metody NLP (Natural Language Processing) i lingwistykę korpusową z przetwarzaniem szczególnie bogatych pod względem informacji „mikrotekstów" bibliograficznych. Dzięki temu chcemy sprawdzić, jak dobrze daje się przewidzieć na podstawie tytułu płeć autora, zamierzamy odtworzyć mapę (lub mapy) nazw własnych występujących w tak ogromnym zbiorze tekstów, spróbujemy poddać tytuły automatycznej klasyfikacji, oceniając następnie jej skuteczność, wygenerujemy także histogramy obrazujące trendy kulturowe. Nie zabraknie też wszechstronnych analiz statystycznych bibliografii.

Co stanowiło argument przemawiający za podjęciem tej problematyki? Przede wszystkim brak takich badań w Polsce, ale także wielkie możliwości humanistyki cyfrowej, stanowiącej zbiór praktyk badawczych, obejmujących metody komputerowe i statystyczne, pozwalające na automatyczne przetwarzanie wielkich mas tekstu.