Streszczenie

Ogromna ilość danych w postaci tekstowej przyczyniła się do rozwoju narzędzi automatycznie grupujących teksty, między innymi pod względem tematycznym. Dzięki temu znalezienie informacji w nich zawartych staje się łatwiejsze. Aby kategoryzować artykuły tematycznie, należy zwrócić uwagę na dwa elementy: sposób reprezentacji danych oraz algorytm używany do wykrywania skupień. Reprezentacja powinna zachować informacje o temacie. Stąd zazwyczaj przyjmuje się następujące założenia: po pierwsze kolejność słów nie gra roli, gdyż istotny jest jedynie ich rozkład, a po drugie fleksja słowa nie jest ważna, tylko jego znaczenie. Można zatem scharakteryzować każdy tekst jako wektor reprezentujący liczności każdego z wyrazów w nim występujących. Takie podejście jednakże jest wysoce nieskuteczne z powodu dużej wymiarowości danych (równej liczbie wszystkich słów występujących w tekstach) oraz ich rzadkości. Często stosowaną praktyką jest grupowanie wyrazów, by następnie reprezentować teksty przy użyciu wektorów liczności grup słów. Stąd niejednokrotnie stosowanym elementem przetwarzania danych tekstowych jest sprowadzenie słowa do jego rdzenia. Takie podejście nie uwzględnia jednak błędów w pisowni, literówek czy też braków znaków diakrytycznych, które często znajdują się w danych tekstowych. Można więc określić miary odległości na przestrzeni napisów (ciągów o dowolnej długości nad pewnym zbiorem skończonym, zwanym alfabetem), przyporządkowujące słowa do z góry określonych grup. Celem niniejszej pracy jest zbadanie wpływu doboru tychże odległości na jakość automatycznej kategoryzacji tematycznej tekstów na podstawie artykułów z polskiej Wikipedii. W pierwszym etapie grupowane są słowa przy użyciu wybranych odległości. Dalej reprezentujemy artykuły jako liczności występowania poszczególnych grup słów w danym tekście. W oparciu o tak uzyskane dane przeprowadzamy analizę skupień tekstów. Ocena wyników odbywa się na podstawie otrzymanych grup z prawdziwymi kategoriami przypisanymi do każdego tekstu Wikipedii. Zauważmy, że użycie takiego zbioru danych wiąże się z kilkoma istotnymi wyzwaniami. Po pierwsze język polski jest językiem gramatycznie bardzo złożonym i zawiera dużą liczbę słów. Dalej zbiór tekstów z polskiej Wikipedii jest względnie duży, co rodzi potrzebę odpowiedniego zarządzenia danymi oraz optymalizacji procesów z powodu ograniczonych zasobów obliczeniowych i pamięciowych. Co więcej niektóre metody analizy danych nie dają się efektywne stosować na dużych zbiorach danych.

Słowa kluczowe: odległości na przestrzeni ciągów znaków, napis, kategoryzacja tematyczna, analiza skupień