Marek Czerniawski

Rafał Gajewski

Robert Kozakiewicz

Inteligentne Wyszukiwanie Informacji

Projekt

Wyszukiwanie podobieństwa pomiędzy kategoriami na Wikipedii

Naszym zadaniem było odtworzenie drzewa powiązań między kategoriami na Wikipedii oraz odnalezienie nowych powiązań pomiędzy kategoriami z pomocą wybranych metod. Do zadania wykorzystaliśmy dane z Wikipedii w języku Basic English, czyli uproszczonym języku angielskim, dzięki czemu były one łatwiejsze w analizie i miały stosunkowo niewielki rozmiar w porównaniu do standardowej wersji Wikipedii. Dane do analizy wygenerowaliśmy za pomocą programu MATRIXu.

W naszym projekcie użyliśmy dwóch metod:

**1. Wyszukiwanie nazwy kategorii**.

Metoda ta polegała na przetworzeniu wszystkich artykułów z danej kategorii oraz wyszukaniu w nich słów tworzących nazwę innej kategorii. Im więcej słów tworzących nazwę jednej kategorii występuje w artykułach z drugiej kategorii, tym większe podobieństwo między tymi kategoriami.

Na potrzeby tej metody wygenerowaliśmy w programie MATRIXu rozkład słów w każdym artykule, następnie napisaliśmy program analizujący te dane, czyli szukający słów składających się na nazwy kategorii.

Podczas generowania danych i wyszukiwania używaliśmy stemmingu, czyli redukowania słów do ich najbardziej podstawowej formy. Użyliśmy w tym celu funkcjonalności zaimplementowanej w programie MATRIXu.

Tabela poniżej przedstawia przykładowe wyniki. Siła powiązania to liczba wystąpień słowa tworzącego kategorię (w przypadku wielowyrazowych kategorii uwzględniane jest tylko słowo najrzadziej występujące) w artykułach z kategorii powiązanej.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kategoria | Kategoria podobna | Siła powiązania | Liczba artykułów tworzących powiązanie | Odległość w drzewie powiązań Wikipedii |
| Religion | Bad Religion | 54 | 5 | 5 |
| Twin people | Genetics | 15 | 2 | 5 |
| Weapons | Terrorism | 39 | 4 | 8 |
| Wind energy | Wind farms | 43 | 12 | 9 |
| Warsaw | 1944 in Europe | 80 | 1 | Nie znaleziono |

Jeśli chodzi o odtworzenie oryginalnych powiązań pomiędzy kategoriami, sprawdziliśmy jaki procent kategorii odległych o 1 od rozważanej kategorii (rodzica lub dzieci) udało nam się odtworzyć. Uzyskany wynik to ok. 40%, jednak prawdopodobnie jest on nieco zaniżony przez kategorie, w których znajduje się mało artykułów (im mniej artykułów i słów w danej kategorii, tym trudniej odtworzyć powiązanie). Udało się więc odtworzyć stosunkowo niewiele powiązań, natomiast zostało utworzonych dużo nowych. Część powiązań wydaje się dość przypadkowa i wynika prawdopodobnie z dużej popularności niektórych słów, natomiast duża część nowych powiązań jest sensowna, mimo dużej odległości w oryginalnym drzewie powiązań.

**2. Odwołania do innych kategorii.**

Ta metoda polegała na analizie linków prowadzących do innych artykułów. Im więcej linków prowadzących do artykułów z innej kategorii tym silniejsze powiązanie z tą kategorią. Na użytek tej metody wygenerowaliśmy odległości pomiędzy artykułami za pomocą programu MATRIXu (odwołanie do innego artykułu może być bezpośrednie lub prowadzić przez kilka artykułów pośrednich).

Ponieważ wygenerowanie danych z całej Simple Wikipedii zajmowało zbyt dużo czasu, wygenerowaliśmy dane dla wszystkich kategorii głównych po czym uruchamialiśmy program analizujący na każdej z nich osobno.

Tabela poniżej przedstawia przykładowe wyniki. Siła powiązania to odległość pomiędzy znalezionymi artykułami (bezpośrednie odwołanie: siła 1, odwołanie przez 1 artykuł pośredni: siła 0.5, itd.), uśredniona po liczbie powiązanych artykułów.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kategoria | Kategoria podobna | Siła powiązania | Liczba artykułów tworzących powiązanie | Odległość w drzewie powiązań Wikipedii |
| Philosophers | Acids | 0,56 | 2 | 9 |
| Physics | Nobel Prize in Physics winners | 0,19 | 24 | 9 |
| Diseases | Microbiology | 0,33 | 12 | 9 |
| Video games | Fantasy movies | 0,60 | 5 | 7 |
| American movies | American screenwriters | 0,46 | 6 | 10 |

Metoda ta słabiej poradziła sobie z odtworzeniem oryginalnych powiązań (ok. 20%), wyraźnie gorsza jest również jakość nowych powiązań.