Projekt

Eksploracyjna analiza tekstu w R

Grupa projektowa: Projekt 20

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Imię | Nazwisko | Numer albumu | Grupa dziekańska | Wkład w prace nad projektem (zadania) | Wkład w prace nad projektem (procentowo) |
| Klaudia | Złamaniec | 206302 | ZIISN2-2411IS |  |  |
| Piotr | Kotarba | 210007 | ZIISN2-2411IS |  |  |

|  |
| --- |
| \_\_/100 pkt |

# 1. Opis utworzonego zbioru dokumentów

W ramach badania utworzony został zbiór dokumentów składający się z 5 zestawów popularnych cyklów literatury fantasty po 4 streszczenia wybranych książek należących do każdej z nich.

1. „Cykl demoniczny” autorstwa Peter V. Brett:

* „Ostatni most”
* „Pustynny król”
* „Wielki władca”
* „Wojna światła”

1. „Eragon” autorstwa Christopher Paolini:

* „Eragon”
* „Najstarszy”
* „Brisingr”
* „Dziedzictwo”

1. „Wiedźmin” autorstwa Andrzej Sapkowski:
   * „Krew elfów”
   * „Czas pogardy”
   * „Chrzest ognia”

* „Wieża jaskółki”

1. „Władca Pierścieni” oraz „Hobbit” autorstwa J. R. R. Tolkien:

* „Drużyna pierścienia”
* „Dwie wieże”
* „Powrót króla”
* „Hobbit czyli tam i z powrotem”

1. „Zwiadowcy” autorstwa John Flanagan:

* „Ruiny Gorlanu”
* „Płonący most”
* „Ziemia skuta lodem”
* „Bitwa o Skandię”

Każdy z wybranych cykli charakteryzuje się swoim własnym światem i fabułą, które bardzo często są unikalne dla ich uniwersów. Fakt posiadania braku prostych do zdefiniowania punktów wspólnych powinien być pomocny w uzyskaniu oczekiwanych wyników analizy.

# 2. Opis przeprowadzonych eksperymentów

Analizę zebranego zbioru danych rozpoczęto od utworzenia korpusu dokumentów i jego wstępnego przetworzeniu w celu wyeliminowania zbędnych elementów takich jak: Białe znaki, liczby, wyrazy bez znaczenia w tym łączniki itp. Itd. Kolejnym etapem była lematyzacja, czyli sprowadzenie słów wchodzących w skład korpusu do ich formy podstawowej. Kolejnym etapem było przygotowanie zestawu macierzy częstości, które w późniejszych etapach analizy były wykorzystywane jako dane wejściowe.

Jednym z pierwszych analiz była analiza tematów za pomocą metody ukrytej alokacji Diricleta. Celem tej analizy było wyodrębnienie 5 tematów(1 temat odpowiada 1 cyklowi, jako iż przyjmuje się, że poruszają ten sam problem). Uzyskanie odpowiednich zestawów słów w każdym temacie i wysokie współczynnika „topic coherance” pozwoliłoby w poprawnym sposób odczytać co jest tematem przewodnim każdego tekstu oraz ich przynależność do każdego cyklu. W ramach badania słów kluczowych wykorzystano tzw. Chmury słów, w celu odnalezienia słów najczęściej występujących i porównania ich z słowami uznanymi jako definicje tematów.

# 3. Wyniki i wnioski

Etap przygotowania korpusu dokumentów do analizy przeszedł pozytywnie, jednakże etap lematyzacji nie poradził sobie z cyklem „Eragon”, gdyż wyrazy: eragon, eragonowi, eragona zostały uznane jako osobne a w rzeczywistości są tylko formą odmiany imienia Eragon.

Analiza ukrytej alokacji Dirichlet’a przyniosła podobne skutki co lematyzacja. Utworzonych zostało 5 zestawów tematów zgodnie z oczekiwaniem jednakże, zamiast otrzymać 5 tematów, które charakteryzowały by każdy cykl algorytm utworzył na podstawie sreszczenia „Hobbit, czyli tam i z powrotem” osobny temat aniżeli z serii Cykl demoniczny. Mowa tutaj o temacie 2, w skład którego należą między innymi wyrazy: krasnolud, bilbo, góra, elf, Hobbit, co w jasny sposób pokazuje, że został on utworzony na podstawie wyżej wspominanego streszczenia. Biorąc pod uwagę utworzone tematy algorytm prawidłowo przypisał tematy do streszczeń poza serią „Cykl demoniczny” oraz „Hobbit” (pod warunkiem, że oczekiwanym wynikiem było przypisanie do wspólnego tematu z władcą pierścieni). Co ciekawe, wszystkie 4 streszczenia cyklu demonicznego wykazują najwyższe dopasowanie do tematu nr 5, czyli tematu utworzonego i odnoszącego się do serii „Zwiadowcy”. Wynik ten jest dość sporym zaskoczeniem, gdyż oba światy diametralnie się od siebie różnią jak i temat przewodni obu serii. Prawdopodobnie powodem otrzymanego wyniku jest fakt, że bohaterowie obu serii zostali byli nauczani swojego fachu przez mistrzów oraz żę oba światy wykazywały elementy średniowieczne takie jak zamki. Analizując słowa kluczowe z każdego z tematów, można zauważyć, że jako słowa najważniejsze zostały uznane imiona występujących bohaterów, co wynika z ich częstego występowania w tekstach. Częste występowania imion w tekstach widać szczególnie na utworzonych chmurach słów, gdzie zdecydowanie dominują imiona. Poza imionami, najczęściej pojawiającymi się słowami były między innymi „demon”, co w mocnym stopniu definiuje temat „Cylku demonicznego”, pierścień – będący tematem przewodnim serii Władcy pierścienia wraz z Hobbitem(choć tutaj jest elementem pobocznym), oraz zwiadowca, który odnosi się do serii Zwiadowcy.