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	Wybór i dodanie dokumentów z 3 tematów, opis dokumentów, kod i eksperymenty do analizy skupień, pca i lsa, opis eksperymentów i analiza wyników analizy skupień, pca i lsa, podsumowania wniosków	50%
Piotr	Kotarba	210007	ZIISN2-2411IS	Wybór i dodanie dokumentów z 2 tematów, stworzenie korpusu dokumentów i macierzy częstości kod i eksperymenty do analizy Dirichleta i słów kluczowych, opis eksperymentów i analiza wyników lda i słów kluczowych.	50%

__/100 pkt

Spis treści

1. Opis utworzonego zbioru dokumentów

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

1. „Pieśń lodu i ognia" autorstwa George R.R. Martin:
   - ❖ Podgatunek: Epic Fantasy
   - ❖ Wybrane książki:
     - ➢ „Gra o tron"
     - ➢ "Starcie królów"
     - ➢ „Nawałnica mieczy"
     - ➢ „Uczta dla wron"
   - ❖ Kluczowe elementy:
     Skomplikowana intryga polityczna, złożone postacie, konflikty na dużą skalę, wiele perspektyw, surowe i realistyczne ujęcie.
   - ❖ Godne uwagi aspekty:
     Osadzona w średniowiecznym świecie o niskim poziomie magii, z naciskiem na polityczne manewry i walki o władzę.

2. „Eragon" autorstwa Christopher Paolini:
   - ❖ Podgatunek: High Fantasy, Dla młodzieży
   - ❖ Wybrane książki:
     - ➢ „Eragon"
     - ➢ „Najstarszy"
     - ➢ „Brisingr"
     - ➢ „Dziedzictwo"
   - ❖ Kluczowe elementy:
     Opowieść o dojrzewaniu, smoczy towarzysze, wybraniec, przeznaczenie, epickie bitwy, systemy magii,
   - ❖ Godne uwagi aspekty:
     Eragon, młody chłopiec z farmy, który został smoczym jeźdźcem, walczy z opresyjnym imperium i poznaje swoje przeznaczenie.

3. „Wiedźmin" autorstwa Andrzej Sapkowski:
   - ❖ Podgatunek: Dark Fantasy
   - ❖ Wybrane książki:
     - ➢ „Krew Elfów"
     - ➢ „Czas Pogardy"
     - ➢ "Chrzest Ognia"
     - ➢ „Wieża Jaskółki"
   - ❖ Kluczowe elementy:
     Dwuznaczność moralna, istoty nadprzyrodzone, mroczna atmosfera, antybohaterowie, złożona fabuła.
   - ❖ Godne uwagi aspekty:
     Skupia się na Geralcie z Rivii, łowcy potworów, poruszającym się po świecie pełnym potworów, konfliktów politycznych i złożonych relacji.

4. „Władca Pierścieni" oraz „Hobbit" autorstwa J. R. R. Tolkien:

- ❖ Podgatunek: High Fantasy
- ❖ Wybrane książki:
  - ➢ „Drużyna pierścienia”
  - ➢ „Dwie wieże”
  - ➢ „Powrót króla”
  - ➢ „Hobbit czyli tam i z powrotem”
- ❖ Kluczowe elementy:
  Narracja misji (ang. quest), mityczne stworzenia, bohaterskie postacie, epickie bitwy, szczegółowe budowanie świata, bogata mitologia.
- ❖ Godne uwagi aspekty:
  Przedstawia wielką bitwę między dobrem a złem, bada tematy przyjaźni, poświęcenia i korumpującego wpływu władzy.

5. „Zwiadowcy” autorstwa John Flanagan:
   - ❖ Podgatunek: Low Fantasy, Przygodowe, Dla młodzieży
   - ❖ Wybrane książki:
     - ➢ „Ruiny Gorlanu”
     - ➢ „Płonący most”
     - ➢ „Ziemia skuta lodem”
     - ➢ „Bitwa o Skandię”
   - ❖ Kluczowe elementy:
     Opowieść o dojrzewaniu, umiejętni bohaterowie, pełne akcji fabuły, szkolenie i mentoring, eksploracja, przetrwanie, niski poziom elementów magicznych.
   - ❖ Godne uwagi aspekty:
     Opowieści podróży Willa, ucznia Zwiadowcy, który wraz z przyjaciółmi uczy się chronić królestwo przed zagrożeniami i odkrywa po drodze tajemnice.

Przez wzgląd na to, iż wszystkie wybrane książki należą do gatunku fantasy są one pod pewnymi względami do siebie podobne, jednak należą one do innych podgatunków fantasy. Nie mniej jednak każdy z wybranych cykli charakteryzuje się swoim własnym światem i fabułą, które bardzo często są unikalne dla ich uniwersów.

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. Następnym etapem było przygotowanie zestawu macierzy częstości. Utworzone zostało 6 macierzy:

- tdm_tf_all
- tdm_tf_612
- tdm_tfidf_410
- dtm_tf_all
- dtm_tf_612
- dtm_tfidf_410

Macierze te zostaną wykorzystywane jako dane wejściowe w późniejszych etapach analizy.

Redukcja wymiarów macierzy częstości

Pierwszym etapem tego przedsięwzięcia jest zmniejszenie wymiarów macierzy częstości. Do tego celu zostaną zastosowane dwie metody: Analizy głównych składowych (ang. principal component analysis, PCA) oraz Dekompozycja według wartości osobliwych z wykorzystaniem Analizy ukrytych wymiarów semantycznych (ang. Latent Semantic Analysis, LSA).Aby sprawdzić efektywność tych metod przeprowadzono po 3 eksperymenty dla każdej metody.

Dla Analizy głównych składowych (PCA) sprawdzono następujące macierze:

1. Exp_pca_tf_all
   exp_matrix <- dtm_tf_all
2. Exp_pca_tf_612
   exp_matrix <- dtm_tf_612
3. Exp_pca_tfidf_410
   exp_matrix <- dtm_tfidf_410

Dla Analizy ukrytych wyrazów semantycznych sprawdzono następujące macierze, wraz podanymi terminami:

4. Exp_lsa_tf_all
   exp_matrix <- tdm_tf_all
   own_terms <- c("daenerys", "jon", "cersei", "eragon", "bilbo", "yennefer", "ciri", "jaskier", "geralt", "frodo", "gandalf", "halt", "ork", "horace")
5. Exp_lsa_tf_612
   exp_matrix <- tdm_tf_612
   own_terms <- c("elf", "czarodziej", "smok", "brat", "arya", "cel", "dawny")
6. Exp_lsa_tfidf_410
   exp_matrix <- tdm_tfidf_410
   own_terms <- c("catelyn", "brama", "bohater", "bagginsa", "bronić","bezpieczny", "atakować")

Powyższe eksperymenty zostały zaprojektowane tak, aby sprawdzić jaki wpływ na wyniki obu metod ma zmiana macierzy częstości.

Analiza skupień

Następnym etapem analizy eksploracyjnej jest przetestowanie hierarchicznych i niehierarchicznych metod analizy skupień. Na potrzeby tego przedsięwzięcia zaprojektowano następujące eksperymenty:

1. Exp_ward_all - Porównanie metod hierarchicznych:
   Parametry:
   a. macierz częstości: dtm_tfidf_all_m
   b. jednostka miary: euclidean
   c. metoda: ward
2. Exp_complete_all - Porównanie metod hierarchicznych:
   Parametry:
   a. macierz częstości: dtm_tfidf_all_m
   b. jednostka miary: euclidean
   c. wykorzystane metody: complete
3. Exp_ward_bounds
   Parametry:
   a. macierz częstości: dtm_tfidf_410_m
   b. jednostka miary: euclidean
   c. wykorzystane metody: ward
4. Exp_complete_bounds
   Parametry:
   a. macierz częstości: dtm_tfidf_410_m
   b. jednostka miary: euclidean
   c. wykorzystane metody: complete

Powyższe eksperymenty zostały skonstruowane tak aby porównać ze sobą dwie metody hierarchicznej analizy skupień (ward i complete), oraz sprawdzić jaki wpływ na wyniki będzie miała zmiana macierzy.

Metoda ukrytej alokacji Dirichleta

Kolejnym etapem eksploracyjnej analizy tekstów jest analiza tematów za pomocą metody ukrytej alokacji Dirichlet'a *(ang. Latent Dirichlet Allocation method)*. 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 wysokiego współczynnika „topic coherence" pozwoliłoby w poprawny sposób odczytać co jest tematem przewodnim każdego tekstu oraz ich przynależność do każdego cyklu.

Aby sprawdzić efektywność tej metody dla podanego zbioru dokumentów przeprowadzono poniższe eksperymenty:

1. Exp_lda_3_all
   Parametry:
   - liczba tematów: 3
   - macierz częstości: dtm_tf_all
2. Exp_lda_3_bounds
   Parametry:
   - liczba tematów: 3
   - macierz częstości: dtm_tf_612

3. Exp_lda_5_all
   Parametry:
   - liczba tematów: 5
   - macierz częstości: dtm_tf_all
4. Exp_lda_5_bouds
   Parametry:
   - liczba tematów:5
   - macierz częstości: dtm_tf_612
5. Exp_lda_7_all
   Parametry:
   - liczba tematów:7
   - macierz częstości: dtm_tf_all
6. Exp_lda_7_bounds
   Parametry:
   - liczba tematów:7
   - macierz częstości: dtm_tf_612

Powyższe eksperymenty zostały skonstruowane tak, aby porównać przypadki mniejszej (3), większej (7) oraz dokładnej (5) liczby tematów, a także zobaczyć jaki wpływ na wyniki ma wykorzystanie dwóch różnych macierzy częstości - całościowej (dtm_tf_all) oraz ograniczonej (dtm_tf_612).

Słowa i frazy kluczowe

W ramach badania słów kluczowych posłużono się dwoma metodami: Chmurze tagów stworzonej na podstawie korpusu oraz sortowania na podstawie macierzy. W tym celu zaprojektowano następujące eksperymenty:

1. Exp_cloud_tags
   Chmura tagów na podstawie korpusu
2. Exp_keywords_tf_all
   Najczęściej występujące słowa na podstawie macierzy dtm_tf_all_m
3. Exp_keywords_tf_612
   Najczęściej występujące słowa na podstawie macierzy dtm_tf_612_m

Powyższe eksperymenty zostały zaprojektowane tak, aby przetestować dwie metody wyznaczania słów kluczowych oraz sprawdzić jaki wpływ na wyniki sortowania ma zmiana rozmiaru macierzy częstości.

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.

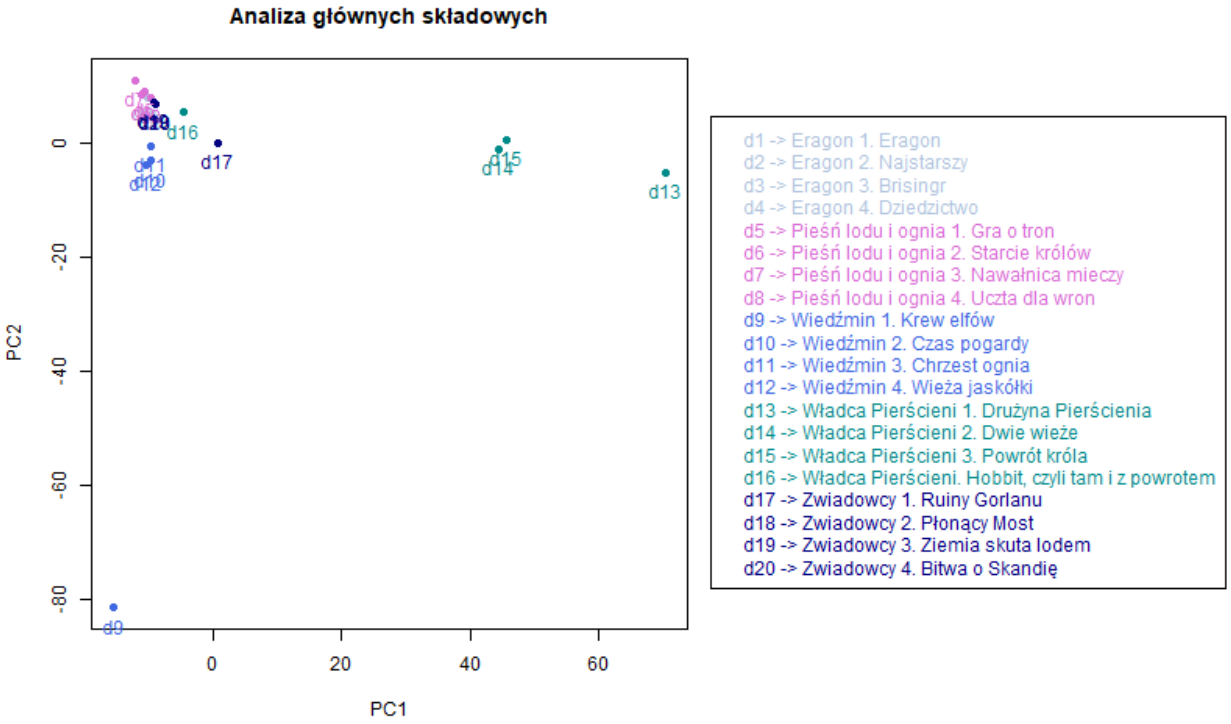
Redukcja wymiarów macierzy częstości

PCA

1. Exp_pca_tf_all

W pierwszym eksperymencie PCA została wykorzystana całościowa macierz częstości dtm_tf_all.

Na poniższym wykresie można zaobserwować, iż większość dokumentów jest zbita w jedną grupę w lewym górnym rogu wykresu, co oznacza, że na podstawie analizy tej macierzy, są one do siebie bardzo podobne. Wyjątkami są dokumenty d13, d14 i d15 należące do trylogii "Władca pierścieni" które są odsunięte trochę dalej ale można powiedzieć, że tworzą w 3 jedną grupę. Ciekawym przypadkiem jest, odsunięty maksymalnie jak się tylko da, dokument d9 czyli "Wiedźmin Krew elfów", co oznacza, że różni się on zdecydowanie od pozostałych. Może być to spowodowane tym, iż to streszczenie zostało napisane trochę inaczej, z podziałem na rozdziały.



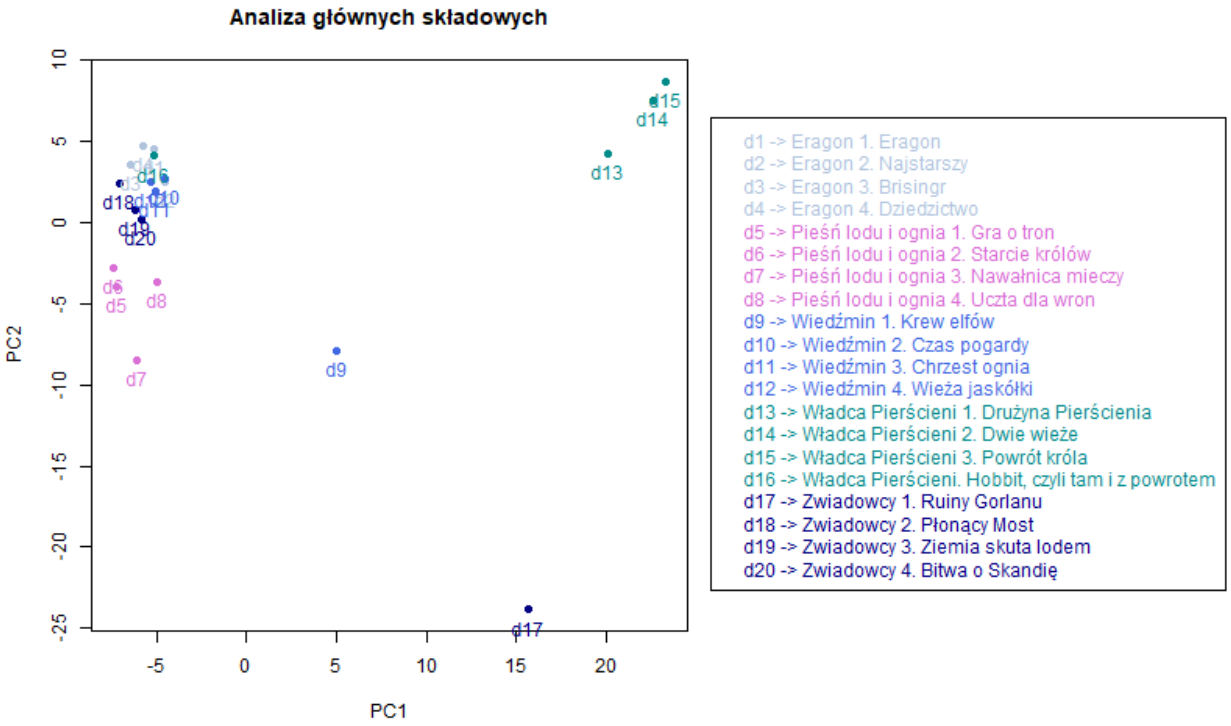
Rys. 1. Analiza głownych składowych dla Exp_pca_tf_all

2. Exp_pca_tf_612

W drugim eksperymencie PCA została wykorzystana ograniczona macierz częstości dtm_tf_612.

Poniższy wykres różni się od wykresu z pierwszego eksperymentu. Można zobaczyć, iż zbita grupa dokumentów zaczęła się trochę rozchodzić, pojawiła się wyraźna różowa grupa dokumentów z cyklu

"Pieśń lodu i ognia". Pozostałe dokumenty z tej grupy pozostały skupione w jednym miejscu, jednak na tyle luźno, że widać, iż kolory poszczególnych cykli tworzą własne grupy. Dokumenty trylogii "Władca Pierścieni" zostały dalej w dużym oddaleniu, jednak zbliżyły się do siebie nawzajem. Dokument d9 zbliżył się do pozostałych dokumentów, natomiast dokumentem który różni się w tej macierzy najbardziej jest d17, czyli "Zwiadowcy: Ruiny Gorlanu".

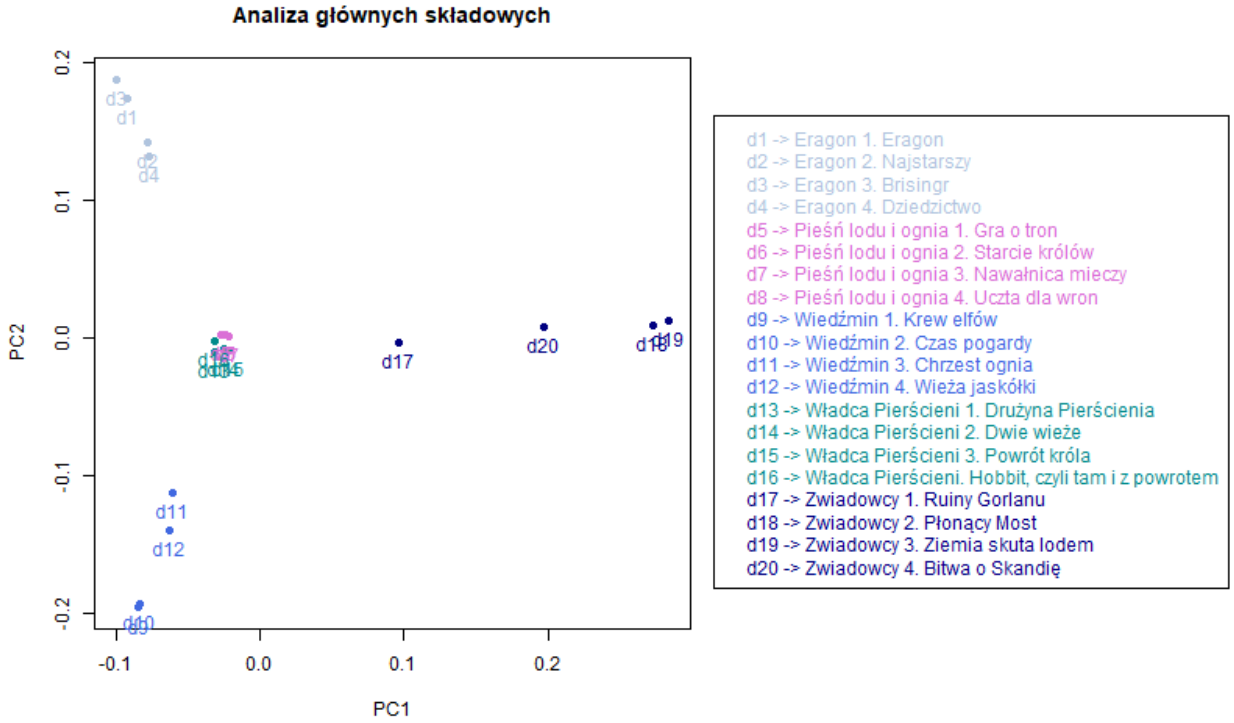


Rys. 2. Analiza głownych składowych dla Exp_pca_tf_612

3. Exp_pca_tfidf_410

W ostatnim eksperymencie PCS została wykorzystana ograniczona macierz częstości dtm_tfidf_410.

Wyniki tego eksperymentu zostały przedstawione na wykresie poniżej i są one najbardziej zadowalające. Wyraźnie widać podział dokumentów na grupy zgodne z odpowiadającymi im seriami książek. Na samym środku pozostała grupa w której skupione zostały seria "Władcy pierścieni" i "Pieśń lodu i ognia", co jest zastanawiające. Jednak widać, że również w tej grupie są one nieco między sobą podzielone. Za tak dobry wynik odpowiada najprawdopodobniej zmiana wagi macierzy z tf na tfidf, oraz to, że jest to macierz ograniczona.



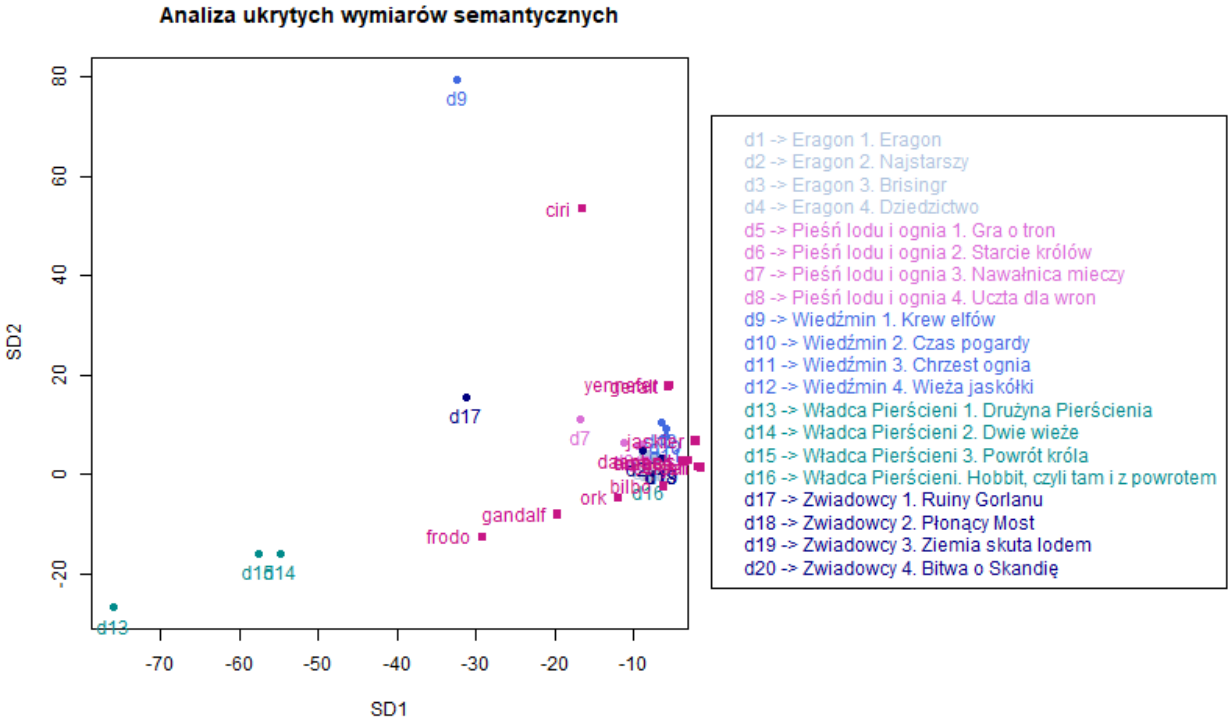
Rys. 3. Analiza głownych składowych dla Exp_pca_tf_410

LSA

4. Exp_lsa_tf_all
   own_terms <- c("daenerys", "jon", "cersei", "eragon", "bilbo", "yennefer", "ciri", "jaskier", "geralt", "frodo", "gandalf", "halt", "ork", "horace")

W pierwszym eksperymencie LSA została wykorzystana całościowa macierz częstości dtm_tf_all. Na wykres naniesiono również słowa kluczowe przekazane jako tablica own_terms, zapisane powyżej.

Na poniższym wykresie zostały przedstawione wyniki tego eksperymentu. Można na nim zaobserwować, iż większość dokumentów zbiła się w jedną grupę, podobnie jak w Exp_pca_tf_all. Kolejnym podobieństwem pomiędzy tymi dwoma eksperymentami jest odłączenie się od głównej grupy dokumentów d13, d14 i d15 w jedną stronę, a także dokumentu d9 w drugą stronę wykresu. Można zaobserwować, że wyrazy będące słowami kluczowymi dla trylogii "Władca Pierścieni", takie jak "frodo", "gandalf" i "ork" migrują wyraźnie w stronę dokumentów d13, d14 i d15. Nie występują one jednak jedynie w tych dokumentach, a również w dokumencie d16 "Hobbit". Z tego powodu słowa te znajdują się pomiędzy dwoma grupami - grupą zieloną "Władca Pierścieni", a grupą złożoną z większości dokumentów w której znajduje się również "Hobbit". Ta sama sytuacja pojawia się w przypadku punktu d9, który jest oddalony od większej grupy w drugą stronę. Wyrazy związane z sagą "Wiedźmin" t.j. "ciri", "yennefer" i "geralt", również migrują w stronę punktu d9 "Wiedźmin. Krew elfów". Pozostałe słowa znajdują się zbite w jednym miejscu razem z większością dokumentów przez co trudno je odczytać.

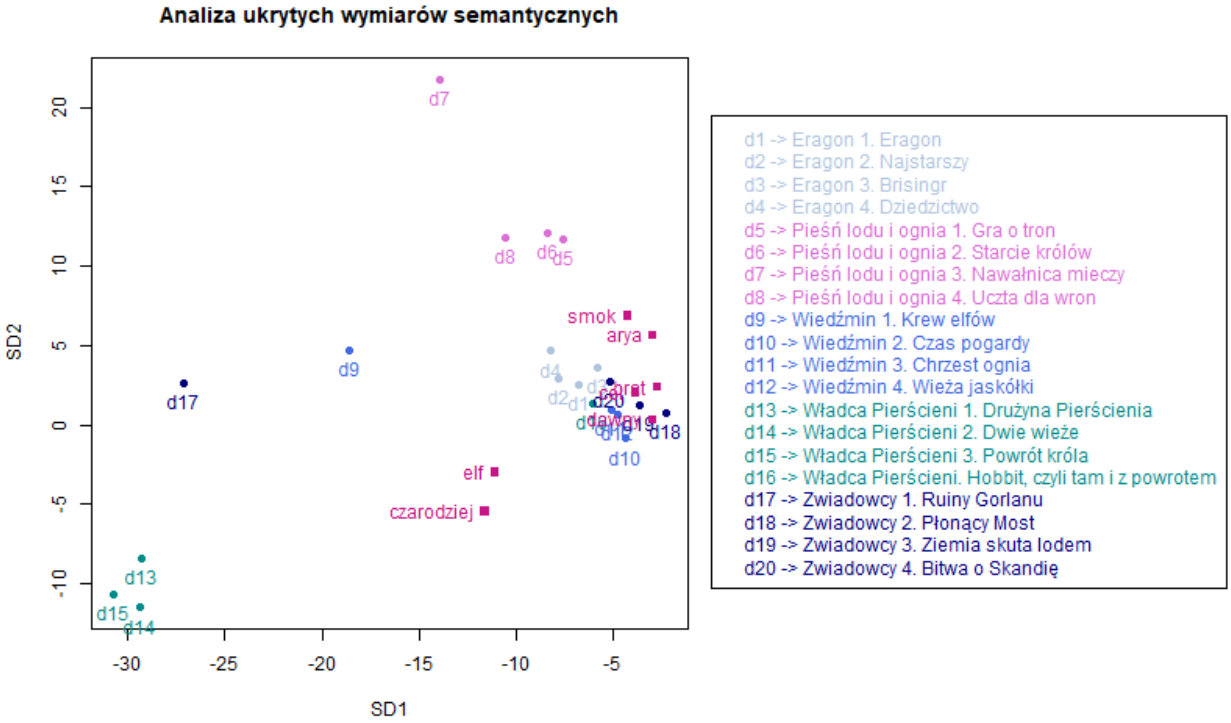


Rys. 4. Analiza ukrytych wymiarów sematycznych dla Exp_lsa_tf_all

5. Exp_lsa_tf_612
   own_terms <- c("elf", "czarodziej", "smok", "brat", "arya", "cel", "dawny")

W drugim eksperymencie LSA została wykorzystana ograniczona macierz częstości dtm_tf_612. Na wykres naniesiono również słowa kluczowe przekazane jako tablica own_terms, zapisane powyżej.

Na wykresie wynikowym przedstawionym poniżej, można zauważyć, że w porównaniu z pierwszym eksperymentem, punkty przedstawiające dokumenty są już bardziej rozrzucone i zaczynają się łączyć w bardziej widoczne grupy tematyczne. Tak samo jak wcześniej, widoczna jest zielona grupa trylogii J.R.R. Tolkien'a, jednak jest ona bardziej zbita w porównaniu z Exp_lsa_tf_all. Widocznie wyodrębniła się również grupa różowa - "Pieśń lodu i ognia" oraz szara - "Eragon". Można również zauważyć, że słowa kluczowe takie jak "smok" i "arya" znajdują się pomiędzy tymi dwoma grupami. Wynika to z faktu, iż oba te dzieła mają bohaterkę o imieniu "Arya" a także poruszają tematykę smoków. Nie dziwi również fakt, że w pobliżu szarej grupy znajduje się punkt d16 - "Hobbit", którego fabuła również skupiona jest wokół smoka. Wyniki tego eksperymentu są bardziej obiecujące, jednak w dalszym ciągu algorytm nie do końca radzi sobie z różnicami pomiędzy saga "Wiedźmin", a serią "Zwiadowcy" oraz książką "Hobbit". Obrazują to rozrzucone punkty d17, d9 i d16.

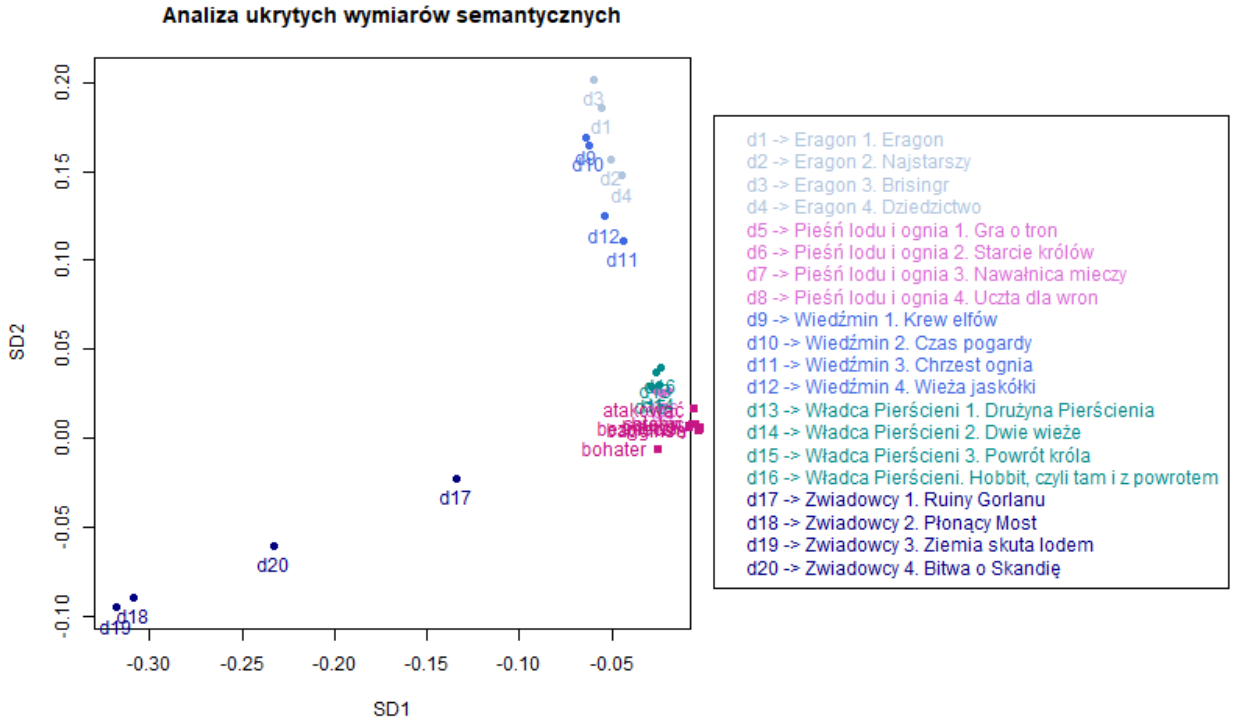


Rys. 5. Analiza ukrytych wymiarów sematycznych dla Exp_lsa_tf_612

6. Exp_lsa_tfidf_410
   own_terms <- c("catelyn", "brama", "bohater", "bagginsa", "bronić","bezpieczny", "atakować")

W ostatnim eksperymencie LSA została wykorzystana ograniczona macierz częstości dtm_tfidf_410. Na wykres naniesiono również słowa kluczowe przekazane jako tablica own_terms, zapisane powyżej.

Wykres przedstawiony poniżej jest częściowo podobny do wykresu Exp_pca_tfidf_410 (analizy PCA opratej na tej samej macierzy). Jednak analiza lsa, w tym wypadku, daje trochę gorsze wyniki. Można zaobserwować rozejście na osobne grupy punktów ciemno niebieskich, szarych oraz jasno niebieskich, czyli odpowiednio serii “Zwiadowcy”, “Eragon” oraz “Wiedźmin”. Natomiast punkty zielone i różowe zbiły się w jedną grupę razem z wszystkimi słowami kluczowymi, wybranymi do tego eksperymentu. W tym wypadku można stwierdzić, że wyniki tego wykresu są w pewien sposób odwrotnością wyników z poprzedniego eksperymentu (Exp_lsa_ts_612). Tam algorytm nie mógł poradzić sobie z rozdzieleniem tematycznym sagi “Wiedźmin” oraz serii “Zwiadowcy”. Tutaj są one wyodrębnione, jednak kosztem cykli oznaczonych zielonym i różowym kolorem, które były bardziej wyodrębnione w poprzednim eksperymencie lsa.



Rys. 6. Analiza ukrytych wymiarów sematycznych dla Exp_lsa_tf_410

Podsumowanie wyników PCA i LSA

Podsumowując, najbardziej zadowalające wyniki rozdzielenia tematycznego przy pomocy analizy głównych składowych (PCA), przyniósł eksperyment nr 3, czyli Exp_pca_tfidf_410. Natomiast w przypadku analizy ukrytych wymiarów semantycznych (LSA) dobre wyniki pokazały się przy dwóch macierzach ograniczonych, w Exp_lsa_tf_612 i Exp_tfidf_410. Ciężko jednak zdecydować, który z tych eksperymentów przyniósł lepsze wyniki, ponieważ były one dobre pod różnymi względami.

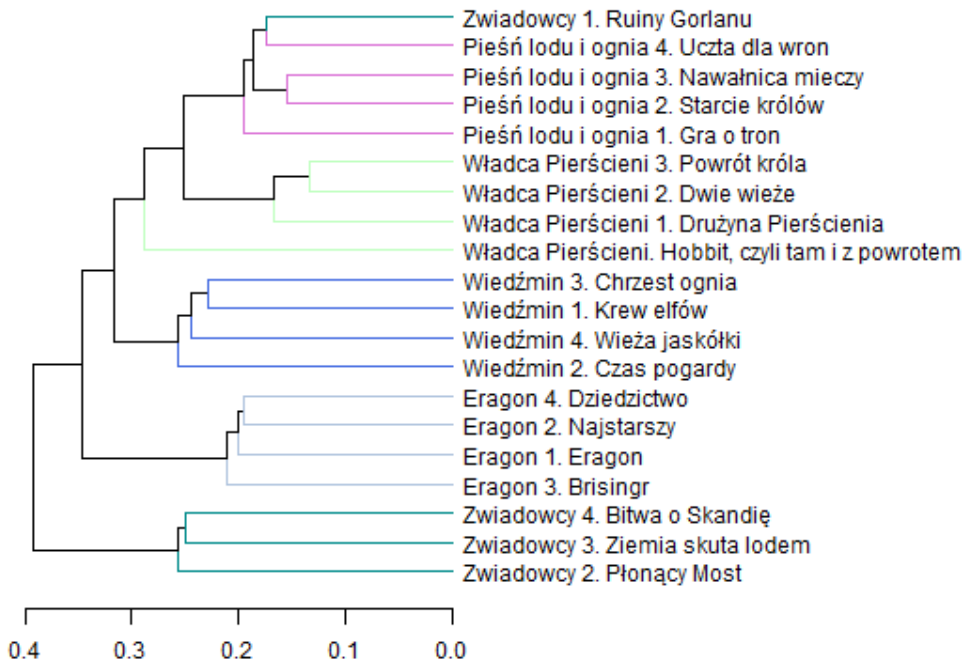
Analiza skupień

1. Exp_ward_all

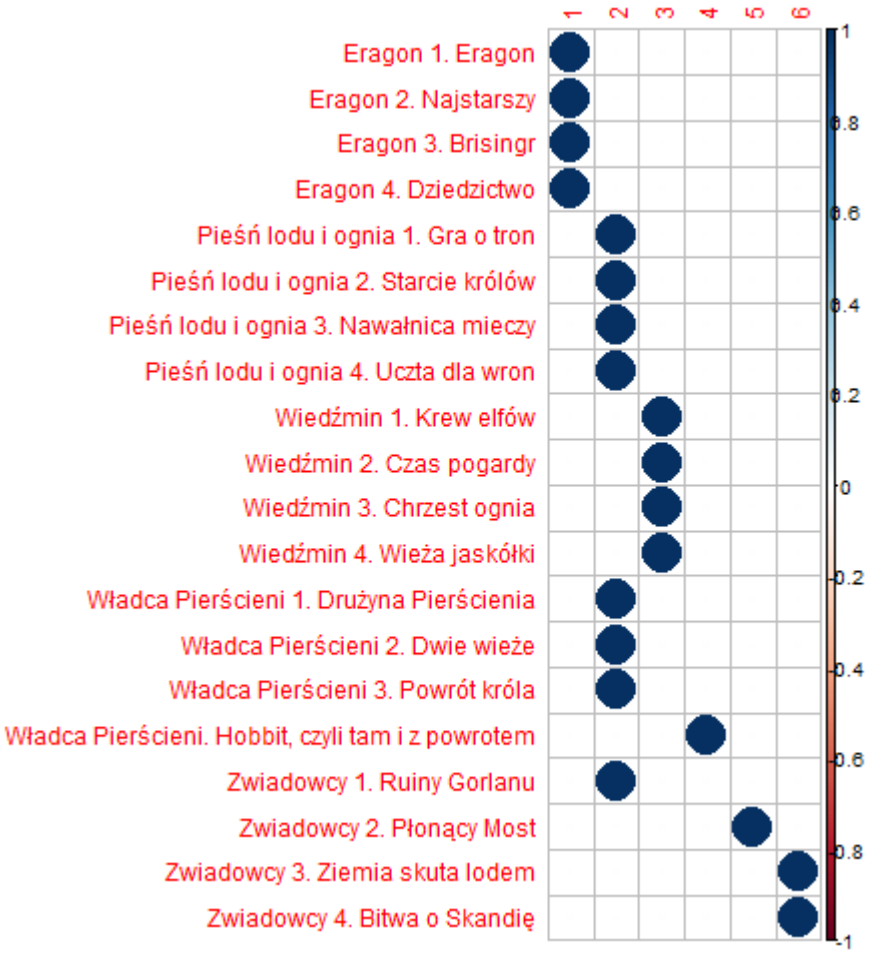
W pierwszym eksperymencie do przeprowadzenia analizy skupień została wykorzystana metoda “Ward” oraz całościowa macierz częstości dtm_tfidf_all.

W tym eksperymencie dość dobrze zostały wyróżnione grupy tematyczne związane z sagą “Wiedźmin” oraz cyklem “Eragon”.Co interesujące “Hobbit” został zidentyfikowany jako oddzielne skupienie pomiędzy “Władcą pierścieni” a “Wiedźminem”, co może być spowodowane tym, że mimo iż jest to historia osadzona w świecie stworzonym przez J.R.R. Tolkien’a, ma ona jednak inna formę niż sama trylogia tego autora, co może prowadzić do tego, że znajduje się w niej również dużo podobieństw, również do innych wybranych serii. Ciekawym wynikiem jest też przypisanie pierwszej części serii “Zwiadowcy” do grupy tematycznej cyklu “Pieśń lodu i ognia”, podczas gdy pozostałe książki z serii zwiadowcy zostały zgrupowane w jedno skupienie. Może to wynikać z faktu iż książka “Ruiny Gorlanu” dzieje się niemal w całości w jednym miejscu - Zamku Redmont, i poruszane są tam tematy rycerstwa, nauki, przygotowań, spisku itp. co może być tematycznie zbliżone właśnie do cyklu napisanego przez George'a Martina. Natomiast pozostałe 3 książki z serii “Zwiadowcy” są raczej opowieściami drogi.

Wyniki zostały przedstawione na poniższych wykresach.



Rys. 7. Dendrogram dla Exp_ward_all



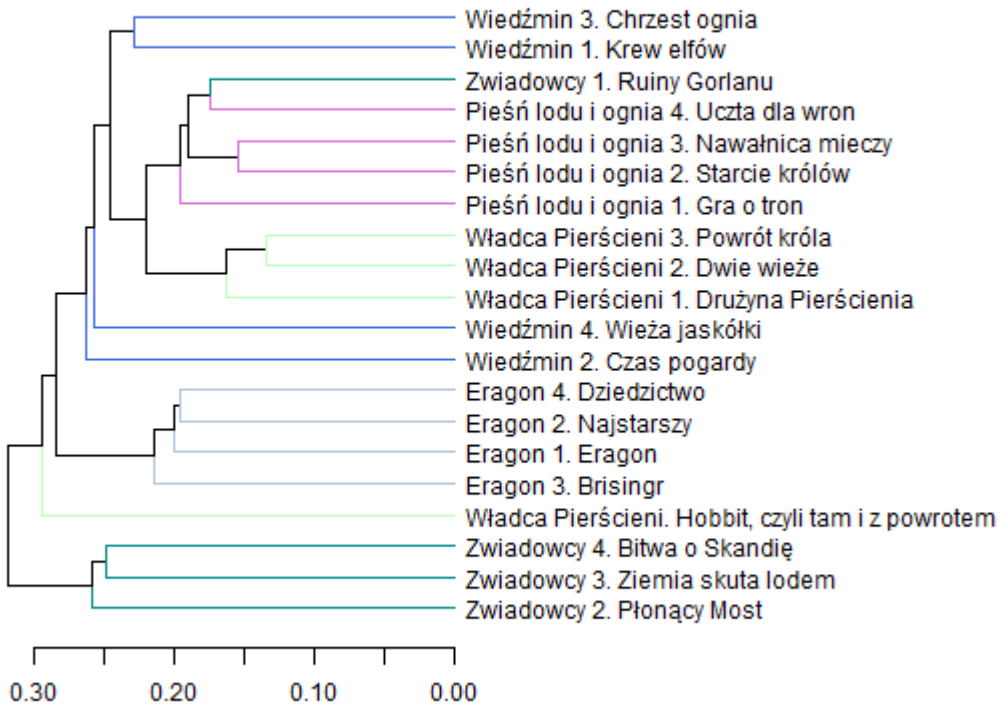
Rys. 8. Wykres dla Exp_ward_all

2. Exp_complete_all

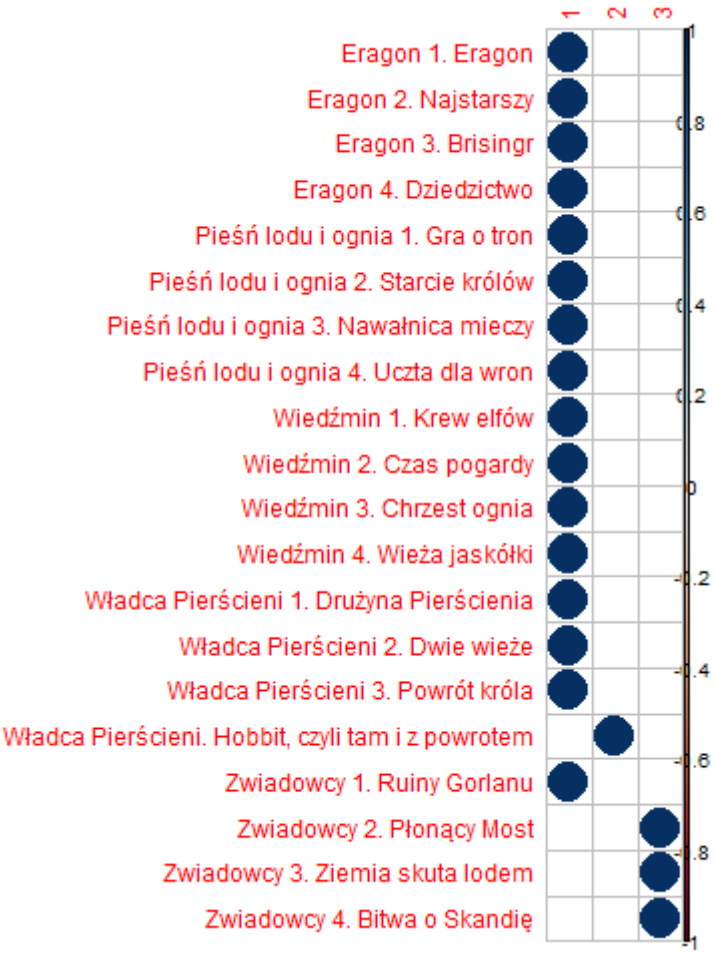
Drugi eksperyment analizy skupień wykorzystywał metodę "Complete" oraz całościowa macierz częstości dtm_tfidf_all. Wyniki zostały przedstawione na poniższych wykresach.

Wyniki tego eksperymentu są znacznie mniej zadowalające. Dokumenty zostały podzielone na trzy główne skupienia - "Hobbit", 2,3 i 4 część "Zwiadowców" oraz cała reszta. Tam, w pod skupieniach, poprawnie zostały wyodrębnione książki z serii "Eragon" oraz "Władca pierścieni". Jednak w przypadku wiedźmina coś poszło nie tak i został on bardzo podzielony na różne mniejsze grupy. Został także powtórzony przypadek z pierwszego eksperymentu, t.j. połączenie serii "Pieśń lodu i ognia" z książką "Ruiny Gorlanu".

Wyniki zostały przedstawione na poniższych wykresach.



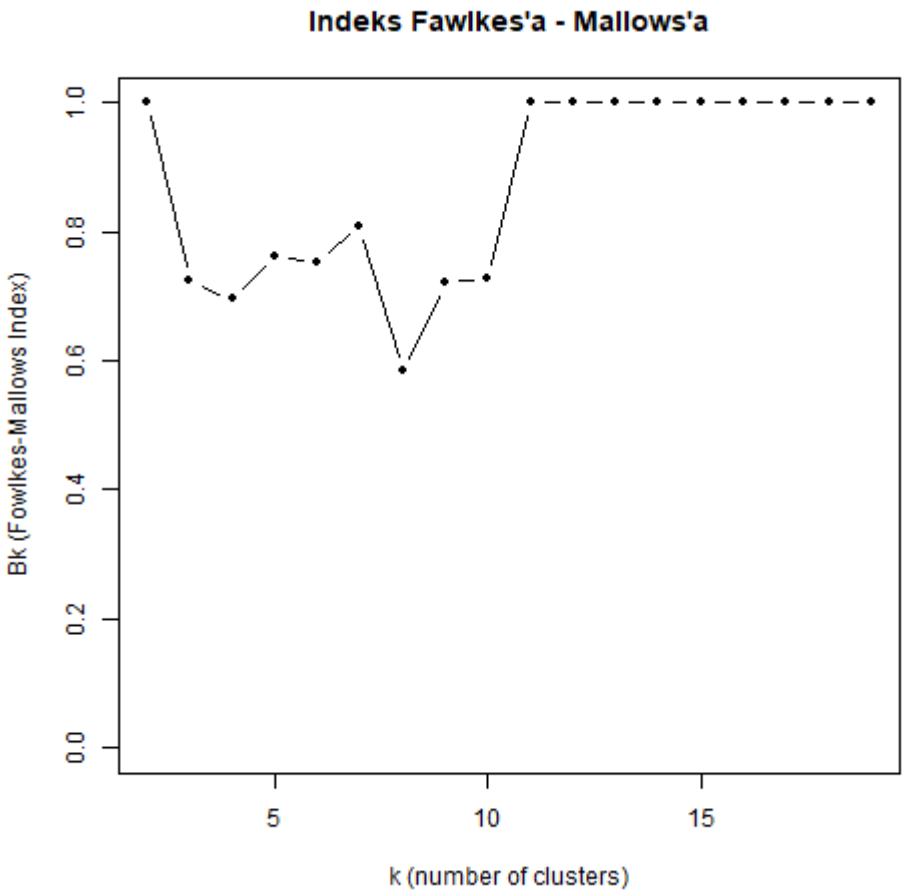
Rys. 9. Dendrogram dla Exp_complete_all



Rys. 10. Wykres dla Exp_complete_all

Porównanie wyników eksperymentów 1 i 2

Pierwsza para eksperymentów wykorzystywała macierz całościową drm_tfidf_all, lecz wykorzystywała różne metody analizy skupień (ward i complete). W tym porównaniu zdecydowanie lepiej sprawdziła się metoda "Ward", użyta w pierwszym eksperymencie. Dokumenty zostały podzielone na bardzo zbliżoną liczbę skupień (6) do liczby zadanych tematów i większość dokumentów została bardzo prawidłowo zgrupowana. Natomiast, podczas wykorzystania metody complete, dokumenty zostały podzielone na mniejszą ilość głównych skupień (3), przez co niektóre cykle zostały w dziwny sposób ze sobą powiązane.



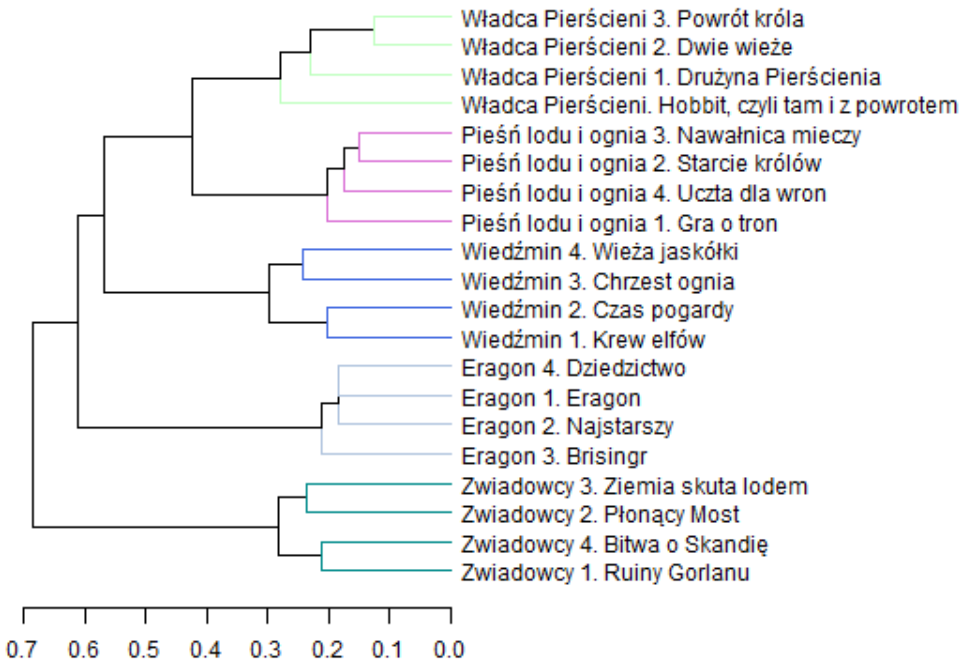
Rys. 11. Wykres FM dla Exp_ward_all i Exp_complete_all

3. Exp_ward_bounds

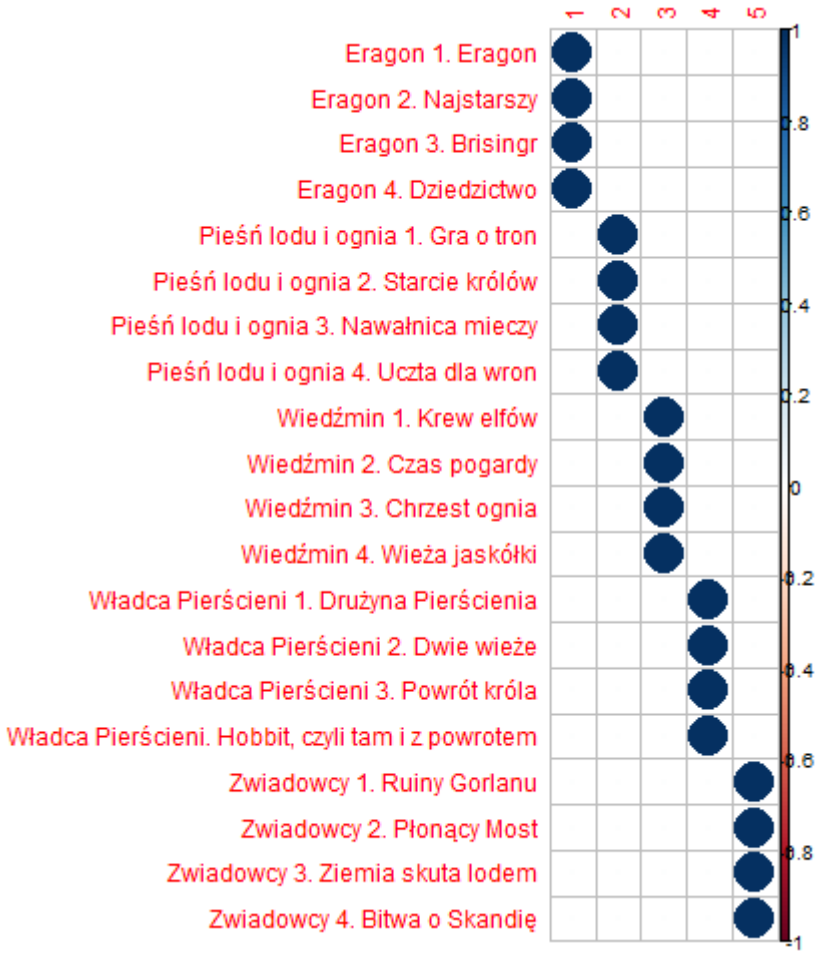
Trzeci eksperyment badał metodę "Ward" w przypadku użycia ograniczonej macierzy częstości dtm_tfidf_410.

Wyniki tego eksperymentu są zdecydowanie bardziej zadowalające. Dokumenty zostały podzielone dokładnie na 5 głównych skupień, czyli tyle samo ile cykli książek. Dodatkowo wszystkie książki zostały poprawnie zgrupowane. Można zauważyć, że w przypadku cyklu "Eragon" oraz "Pieśni lodu i ognia" rozrzut klastrów był bardzo miały, wręcz minimalny. Natomiast w przypadku sagi "Wiedźmin" i "Zwiadowcy" zostały bardzo wyraźnie wyróżnione jeszcze dwie podgrupy tematyczne. Natomiast w przypadku książek J.R.R. Tolkien'a wyraźnie widać, że "Hobbit" odstaje nieco od pozostałych książek, choć bliżej mu do pierwszej części trylogii, niż do 2 pozostałych części.

Wyniki zostały przedstawione na poniższych wykresach.



Rys. 12. Dendrogram dla Exp_ward_bounds



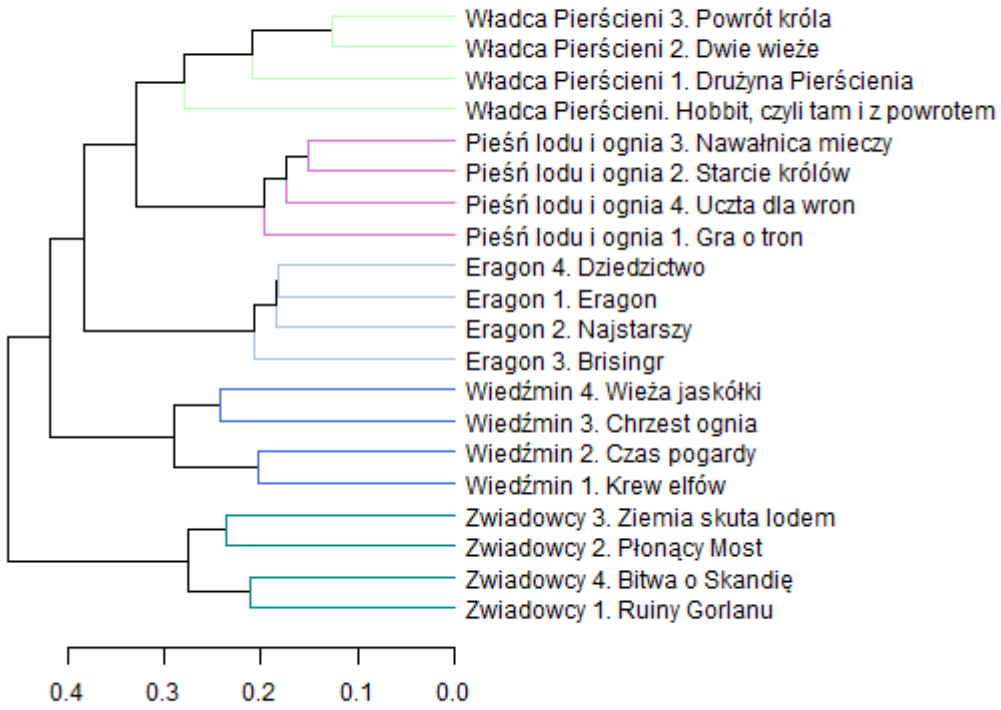
Rys. 13. Wykres dla Exp_ward_bounds

4. Exp_complete_bounds

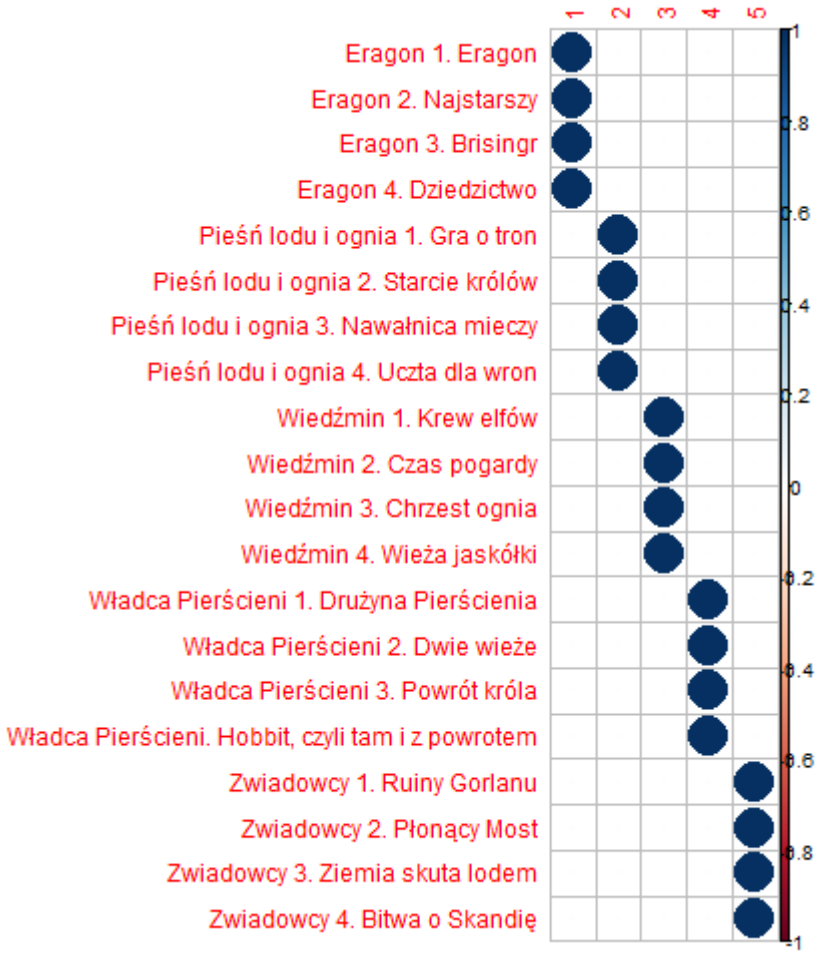
Czwarty eksperyment badał metodę “Complete” w przypadku użycia ograniczonej macierzy częstości dtm_tfidf_410.

Wyniki tego eksperymentu są niezmiernie zbliżone do eksperymentu numer 3. Dokumenty również zostały podzielone dokładnie na 5 głównych skupień oraz wszystkie zostały prawidłowo sklasyfikowane. Dendrogram wygląda niemal identycznie. Jedynymi różnicami są niektóre długości linii, czyli stopień w jakim dokumenty są do siebie podobne. Tak na przykład w przypadku książek J. R. R. Tolkien’a metoda ta znalazła więcej różnic pomiędzy “Hobbitem”, “Drużyną” pierścienia” i pozostałymi częściami trylogii, jednak w dalszym ciągu zakwalifikowała je jako jedno skupienie.

Wyniki zostały przedstawione na poniższych wykresach.



Rys. 14. Dendrogram dla Exp_complete_bounds

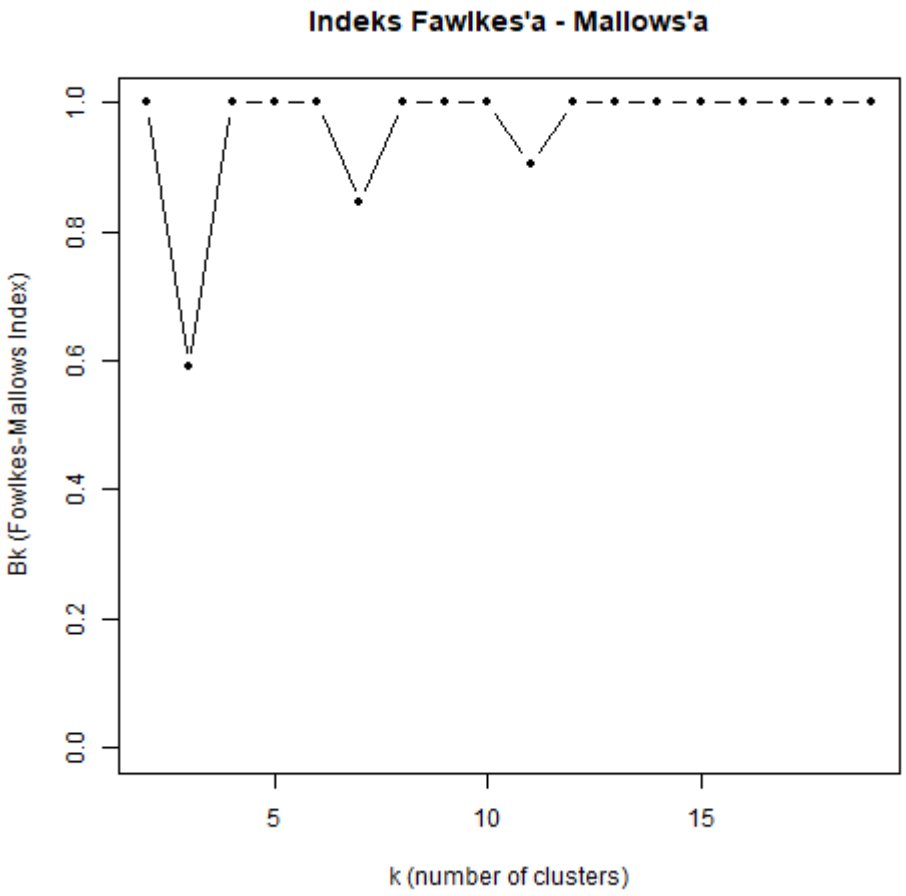


Rys. 15. Wykres dla Exp_ward_bounds

Porównanie wyników eksperymentów 3 i 4

W drugiej parze eksperymentów została wykorzystana macierz ograniczona dtm_tfidf_410, oraz zostały porównane metody “Ward” i “Complete”. W tym zestawieniu obie te metody osiągnęły bardzo dobry, niemal identyczny, wynik.

W porównaniu wykorzystywanych macierzy algorytm działał zdecydowanie lepiej z wykorzystaniem macierzy organicznej dla obu metod.



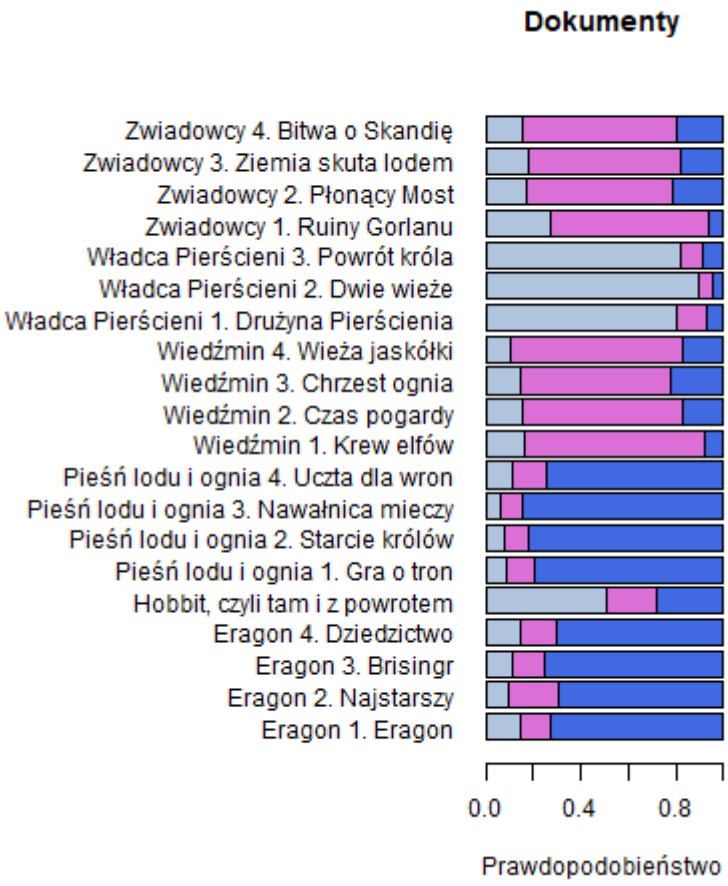
Rys. 16. Wykres FM dla Exp_ward_bounds i Exp_complete_bounds

Metoda ukrytej alokacji Dirichleta

Wynikiem metody ukrytej alokacji Dirichleta jest podział dokumentów na zadaną liczbę tematów, przy pomocy odnalezionych kluczowych słów dla każdego tematu. Dla opisanych wcześniej eksperymentów wyniki przedstawiają się następująco:

1. Exp_lda_3_all

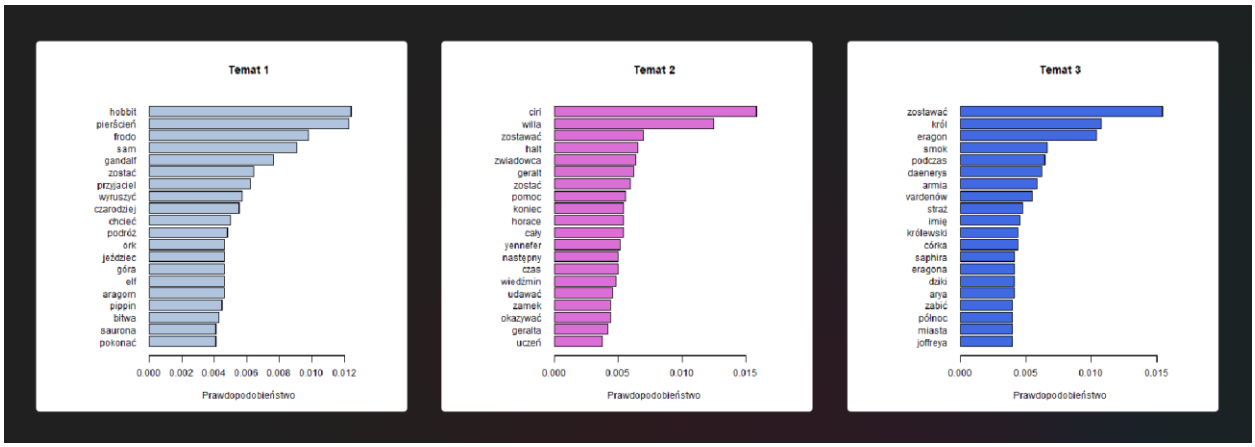
W pierwszym eksperymencie metody LDA przyjęto podział dokumentów na 3 tematy, przy użyciu macierzy całościowej dtm_tf_all.



Rys. 17. Wykres prawdopodobieństwa dla Exp_lda_3_all

Na poniższym wykresie można zauważyć, że do pierwszego tematu (kolor szary) przypisano trylogię "Władcy pierścieni". Cykl "Zwiadowcy" oraz "Wiedźmin" zostały w dużym prawdopodobieństwem przydzielone do tematu drugiego, oznaczonego kolorem różowym. Jest to dość ciekawe, ponieważ są to cykle diametralnie różne w tonie oraz świecie przedstawionym, jednak pewnym punktem wspólnym mogły tu być takie elementy, jak wspólna podróż bohaterów oraz wyprawa na ratunek/pomoc porwanym bohaterom. Trzeci temat (kolor niebieski) został przypisany z dużym prawdopodobieństwem do serii "Eragon" oraz książek z cyklu "Pieśń lodu i ognia". Znów jest to dość ciekawe połączenie, ponieważ książki te są bardzo różne, jednak jest ono zrozumiałe, gdyż w obu tych seriach dość dużą rolę odgrywają smoki. Największy problem z przypisaniem do tematu pojawił się w przypadku książki "Hobbit, czyli tam i z powrotem", bo, choć jest to książka osadzona w świecie wykreowanym przez J. R. Tolkiena, jest ona niemal całkiem odrębną historią od tej opowiedzianej w trylogii "Władcy pierścieni". Dla tej książki zostały rozpoznane 2 tematy (1 i 3) niemal z takim samym prawdopodobieństwem. Pierwszy temat ma lekką przewagę, zapewne dzięki światotwórczym elementom, takim jak "hobbici", "elfy", itp. lub też powtarzającym się postaciom np. "Gandalf". Nie mniej jednak temat 3 również jest rozpoznany w dość dużej mierze, co jest najprawdopodobniej spowodowane tym, iż duża część historii "Hobbita" ma związek z wyprawą na smoka.

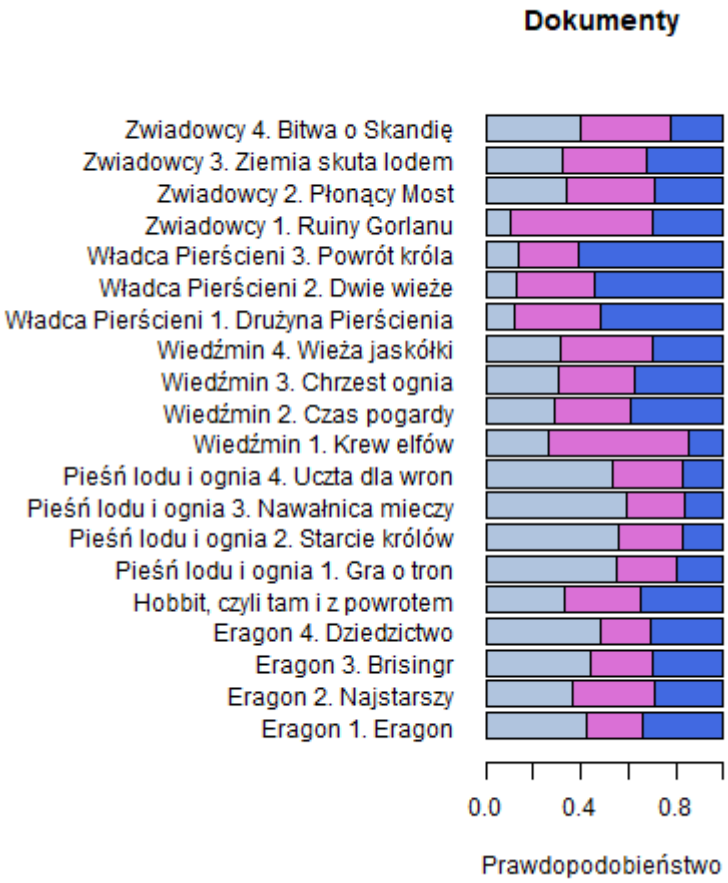
Na poniższej grafice można zaobserwować jakie słowa najczęściej występowały w wyodrębnionych przez metodę LDA tematach.



Rys. 18. Podział tematów dla Exp_lda_3_all

2. Exp_lda_3_bounds

Drugi eksperyment metody LDA polegał na wyodrębnieniu 3 tematów przy użyciu ograniczonej macierzy częstości dtm_tf_612.

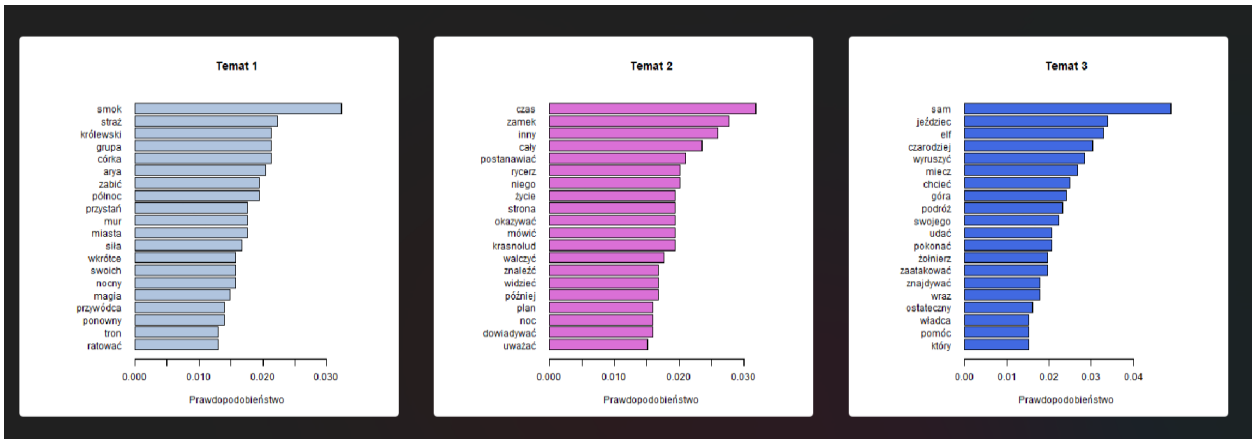


Rys. 19. Wykres prawdopodobieństwa dla Exp_lda_3_bounds

Na powyższym wykresie można zauważyć, że prawdopodobieństwo przypisania do tematu stało się zdecydowanie mniej wyraźne niż w przypadku pierwszego eksperymentu. Niemniej jednak wciąż

można zauważyć pewne prawidłowości. Cykl “Pień lodu i ognia” oraz “Eragon” w dalszym ciągu zostały z największym prawdopodobieństwem przydzielone do jednego tematu (tym razem oznaczonego kolorem szarym). Wyróżnia się też trylogia “Władcy pierścieni”, która jest przypisana do tematu niebieskiego z największym prawdopodobieństwem. W pozostałych przypadkach ciężko jest jednak jednoznacznie przydzielić dane książki do konkretnych tematów. Dzieje się tak, ponieważ została ograniczona liczba słów w macierzy częstości, a z racji tego, iż wszystkie te książki należą do gatunku fantasy i są osadzone w nieco podobnych realiach, pod względem rozwoju technologicznego przypominających średniowiecze. Pojawia się więc tam wiele elementów z każdego z wyznaczonych tematów.

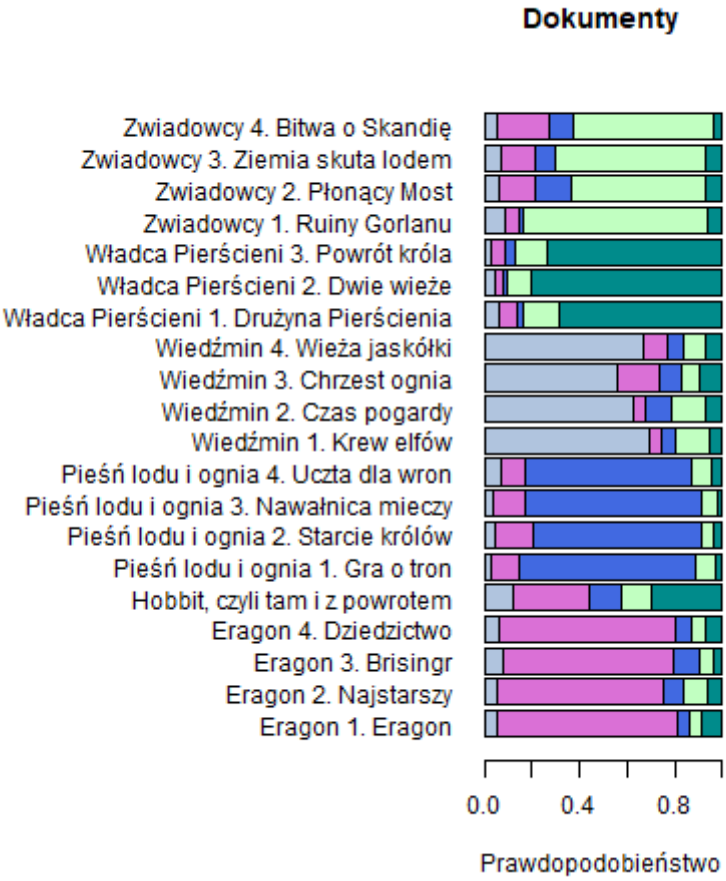
Na poniższej grafice można zaobserwować jakie słowa najczęściej występowały w wyodrębnionych przez metodę LDA tematach, w tym eksperymencie.



Rys. 20. Podział tematów dla Exp_lda_3_bounds

3. Exp_lda_5_all

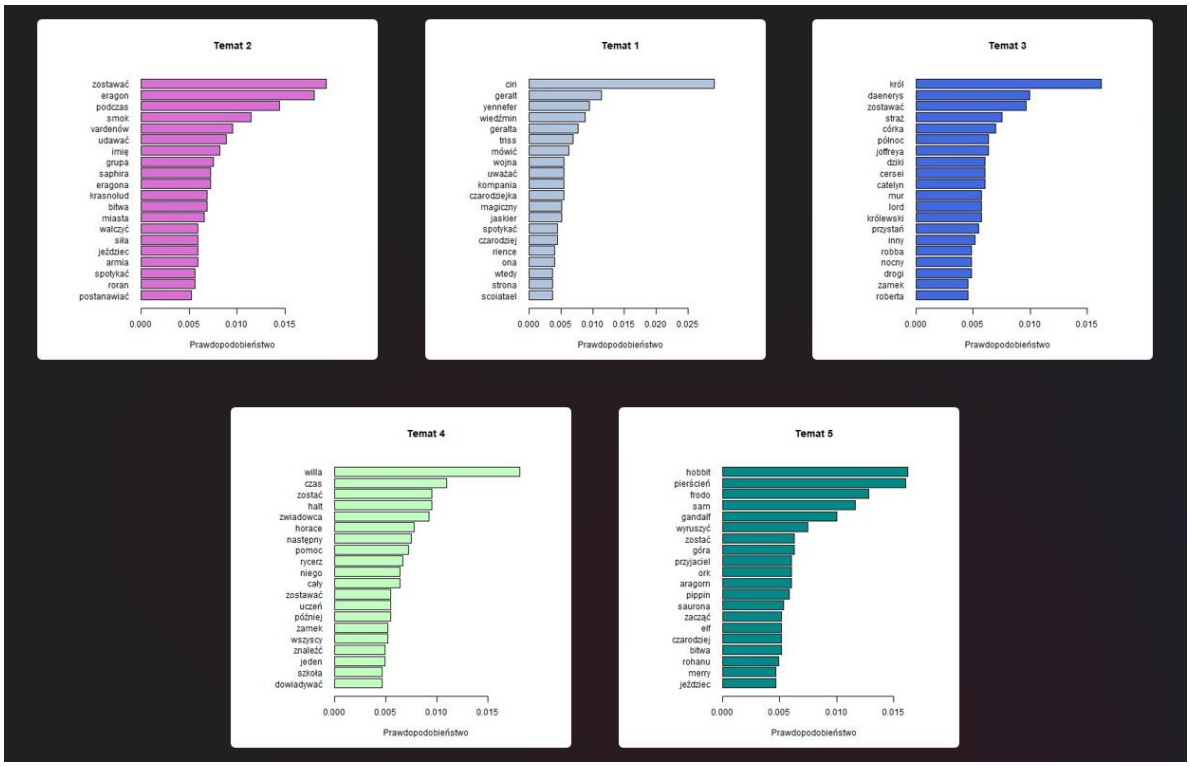
W trzecim eksperymencie LDA, wyznaczono podział na 5 tematów przy pomocy całościowej macierzy: dtm_tf_all.



Rys. 21. Wykres prawdopodobieństwa dla Exp_lda_5_all

Na powyższym wykresie można zauważyć, iż metoda LDA z zadanymi parametrami oraz dla wybranych dokumentów całkiem dobrze wyodrębniła kolejne cykle i przypisała je do odrębnych tematów. Wyjątek jednak znów stanowi książka "Hobbit, czyli tam i z powrotem". Pomimo że przynależność do tematu ciemnozielonego wyodrębnionego dla "Władcy pierścieni" również jest widoczna, widać także duże prawdopodobieństwo przynależności do tematu oznaczonego kolorem różowym, który cechuje cykl "Eragon". Prawdopodobnie jest to związane z tym, iż historia zawarta w książce "Hobbit" opowiada o wyprawie na smoka.

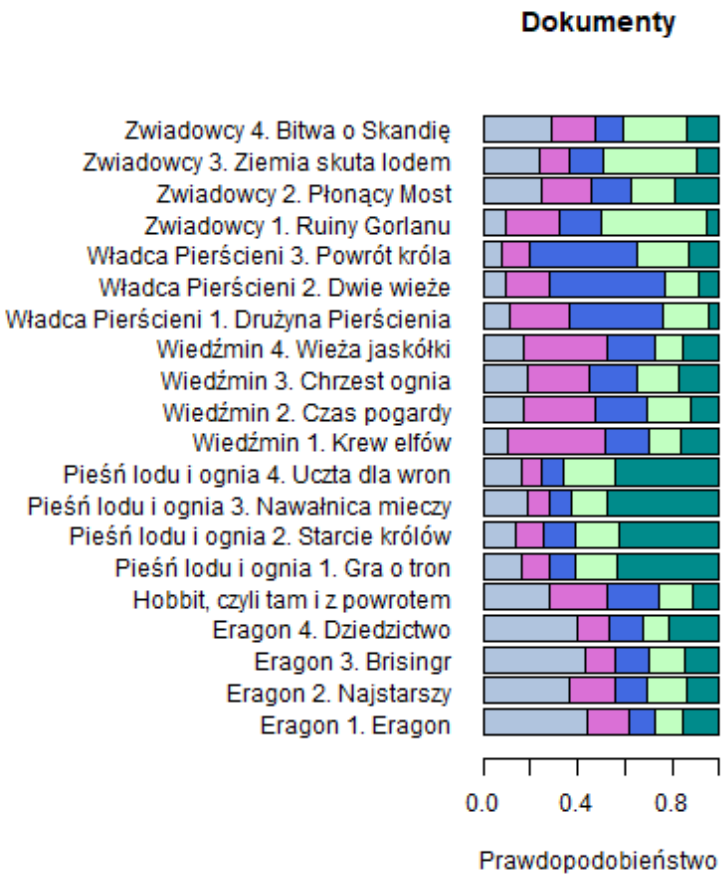
Na poniższej grafice można zaobserwować jakie słowa najczęściej występowały w wyodrębnionych przez metodę LDA tematach, w tym eksperymencie.



Rys. 22. Podział tematów dla Exp_lda_5_all

4. Exp_lda_5_bounds

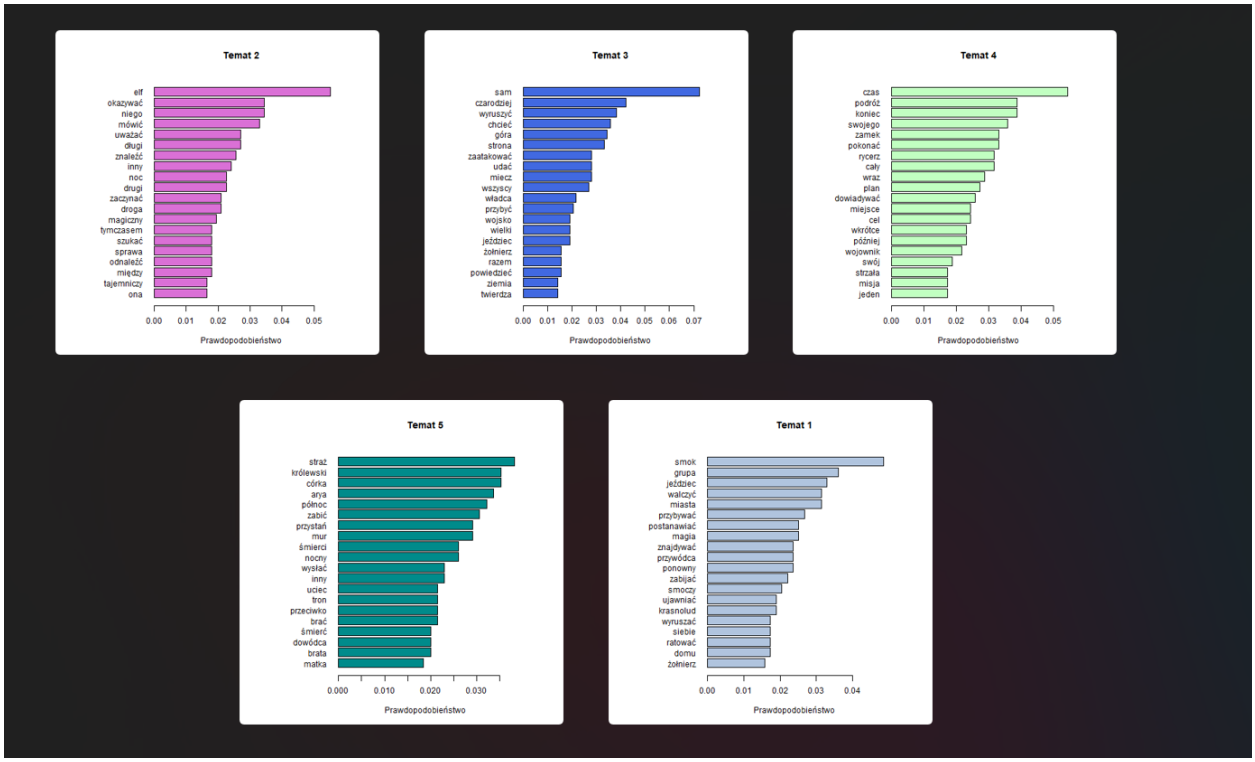
Kolejny eksperyment metody LDA polegał na wyodrębnieniu 5 tematów przy użyciu ograniczonej macierzy częstości dtm_tf_612.



Rys. 23. Wykres prawdopodobieństwa dla Exp_lda_5_bounds

Na powyższym wykresie można zauważyć że ograniczenie macierzy, także w tym przypadku, zmniejszyło wyrazistość podziału na tematy, jednak wciąż można zaobserwować pewne dominujące elementy. W przypadku cykli: “Eragon”, “Pieśń lodu i ognia” i “Władcy pierścieni”, wciąż da się zauważyć jaki temat jest tym głównym (najbardziej prawdopodobnym) dla każdego z tych cykli. W przypadku “Wiedźmina” oraz “Zwiadowców” jest już trochę ciężej, ponieważ różnice pomiędzy tematami są coraz mniejsze i można w nich odnaleźć każdy z nich niemal w równym stopniu. Problemem jest też “Hobbit”, który znów bardziej wyłapał temat spójny z cyklem Eragon, aniżeli z “Władcy pierścieni”, można wręcz powiedzieć, iż kolor niebieski i różowy, czyli tematy “Władcy pierścieni” i “Wiedźmina”, są odnalezione w “Hobbicie” na równi. Nie jest to wynik w pełni zadowalający.

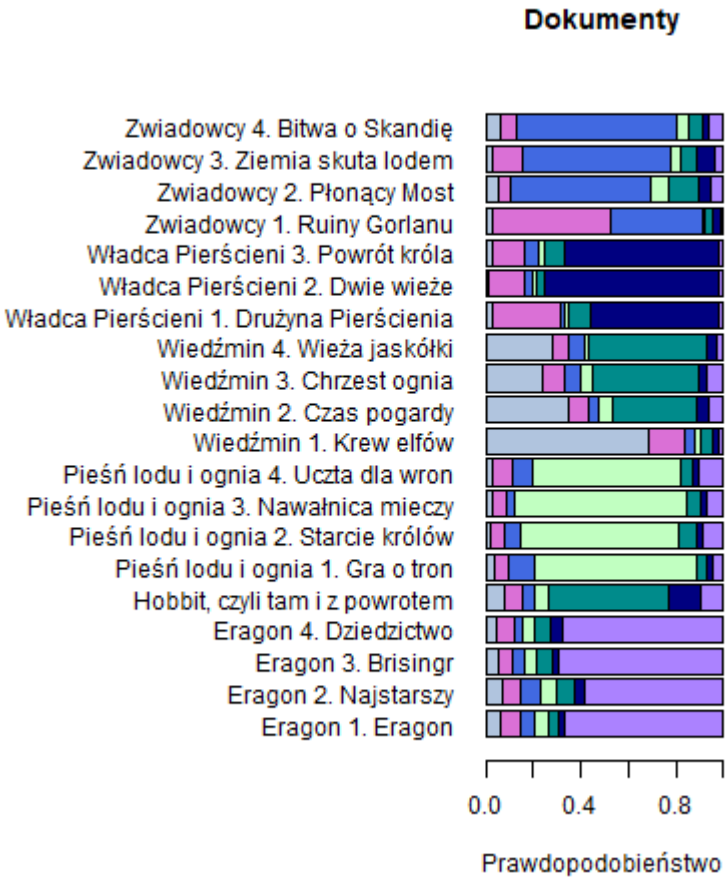
Na poniższej grafice można zaobserwować jakie słowa najczęściej występowały w wyodrębnionych przez metodę LDA tematach, w tym eksperymencie.



Rys. 24. Podział tematów dla Exp_lda_5_bounds

5. Exp_lda_7_all

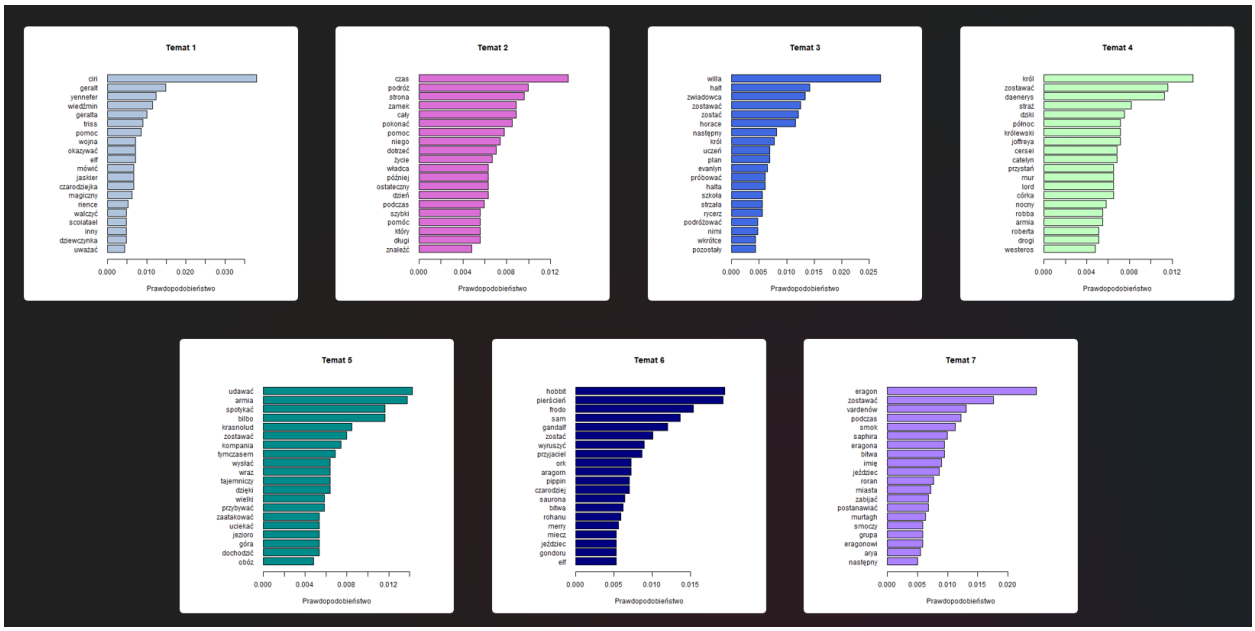
W piątym eksperymencie LDA, wyznaczono podział na 7 tematów przy pomocy całościowej macierzy: dtm_tf_all.



Rys. 25. Wykres prawdopodobieństwa dla Exp_lda_7_all

Na powyższym wykresie można zaobserwować wyniki dla tego eksperymentu. Wyraźnie rozdzielone zostały cykl "Eragon", "Pieśń lodu i ognia" oraz "Władcy pierścieni". Dość mocno odznacza się również seria zwiadowcy kolorem niebieskim, jednak w przypadku pierwszej książki z cyklu został wyróżniony także inny temat, oznaczony kolorem różowym. Może to być wiązane z tym, iż pierwsza książka miała nieco inny przebieg niż pozostałe 3 z tej serii. Większy nacisk położony tu został na wprowadzeniu głównego bohatera Will'a do roli Zwiadowcy oraz jego czas spędzony na nauce i ćwiczeniach. Metoda lda zlokalizowała także osobny temat, oznaczony kolorem ciemnozielonym, dla książki "Hobbit", natomiast połączenie tematyczne z "Władcą pierścieni" jest na drugim miejscu a z serią "Eragon" na trzecim, co jest dość spójne ze stanem faktycznym jak i poprzednimi wynikami. Co ciekawe, temat ciemnozielony został także bardzo wyraźnie przydzielony do 2, 3 i 4 części "Wiedźmina". Bardzo prawdopodobnie jest to spowodowane tym, iż w tych częściach sagi "Wiedźmin" duża część fabuły skupia się na wędrówce kompani Geralta, podobnie jak "Hobbit" skupia się na wyprawie krasnoludów. Na podstawie tych wyników można stwierdzić, iż wyróżnione tutaj 2 dodatkowe tematy, odnoszą się bardziej szczegółowo do elementów fabularnych wybranych książek, a nie tylko do cechujących ich nazw.

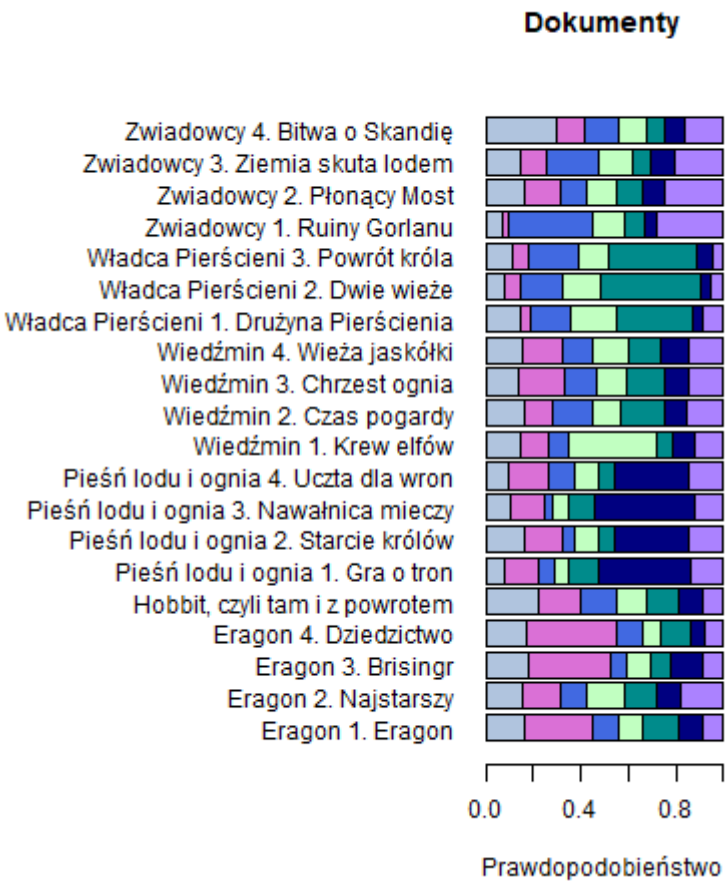
Na poniższej grafice można zaobserwować jakie słowa najczęściej występowały w wyodrębnionych przez metodę LDA tematach, w tym eksperymencie.



Rys. 26. Podział tematów dla Exp_lda_7_all

6. Exp_lda_7_bounds

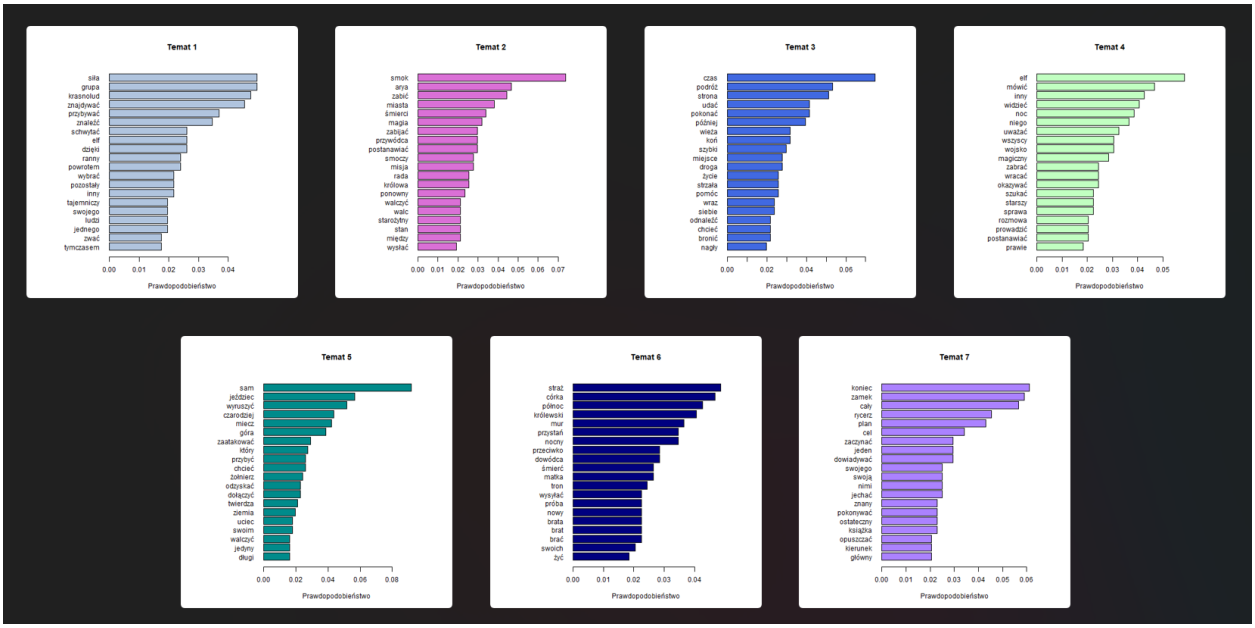
Ostatni eksperyment metody LDA polegał na wyodrębnieniu 7 tematów przy użyciu ograniczonej macierzy częstości dtm_tf_612.



Rys. 27. Wykres prawdopodobieństwa dla Exp_lda_7_bounds

Na powyższym wykresie przedstawiono wyniki tego eksperymentu. Można zauważyć, iż tak samo jak w poprzednich eksperymentach wykorzystujących macierz ograniczoną, tematyki są zakreślone mniej wyraźnie. W dalszym ciągu najbardziej można wyróżnić trzy serie czyli “Eragon”, “Pieśń lodu i ognia” oraz “Władcy pierścieni”, jednak w przypadku pozostałych dwóch cykli oraz “Hobbita” jest to już niezmiernie trudne, ponieważ tematy bardzo się zrównują.

Na poniższej grafice można zaobserwować jakie słowa najczęściej występowały w wyodrębnionych przez metodę LDA tematach, w tym eksperymencie.



Rys. 28. Podział tematów dla Exp_lda_7_bounds

Podsumowanie wyników LDA

Podsumowując, na podstawie przeprowadzonych eksperymentów, można stwierdzić, iż w przypadku tak dobranej próby dokumentów najlepsze wyniki daje macierz całościowa wraz z podziałem na liczbę tematów, równą faktycznej lub nieznacznie większą (Exp_lda_5_all i Exp_lda_7_all). W przypadku macierzy ograniczonej, określenie tematu dominującego stawało się trudniejsze, jednak nie niemożliwy w przypadku niektórych cykli. Natomiast w przypadku podziału na mniejszą liczbę tematów, do wspólnej grupy zostały zaliczone książki mające pewne wspólne elementy, jednak bardzo od siebie różne w szerszej perspektywie, przez to wyniki tych eksperymentów prezentowały się najgorzej spośród pozostałych eksperymentów.

Słowa i frazy kluczowe

1. Exp_cloud_tags

Dla każdego dokumentu została wygenerowana chmura słów kluczowych, Zaliczone do nich zostały słowa o największej wadze, bardzo często były to nazwy własne. Przeglądając je można zauważyć, iż lematyzacja nie poradziła sobie z częścią nazw własnych i ich odmian np. “Joffreya”, “Stannisa”,” Eragona”, “Eragonowi”. Ciekawy przypadek stanowi imię głównego bohatera serii “Zwiadowcy” - “Will”, to imię nie występuje w słowach kluczowych w swojej podstawowej formie ani razu, poniewaz zistało ono rozpoznane i przekształcone jako polskie słowo “willa” - czyli duży bogaty dom. Przykłady chmur można zobaczyć poniżej.

Rys. 29. Przykładowe chmury słów

Rys. 30. Przykładowe chmury słów

2. Exp_keywords_tf_all

W tym eksperymencie wykorzystano macierz całościową dtm_tf_all do wyznaczenia słów kluczowych. Można zaobserwować, iż w uzyskanych wynikach dominują nazwy własne. Co w przypadku znajomości uniwersów tych książek może dać jasną informację czego dotyczy dany dokument. Wyniki eksperymentu zostały przedstawione na grafice poniżej

```
[1] "Eragon 1. Eragon"
   eragon  zostawać   saphira   eragona eragonowi  jeździec
       12        11         7         6         6         6
[1] "Eragon 2. Najstarszy"
  eragon vardenów    roran    bitwa  eragona zostawać
      10       10        8        7        7        7
[1] "Eragon 3. Brisingr"
   eragon  vardenów     roran  zostawać   saphira eragonowi
       18         9         7         7         6         5
[1] "Eragon 4. Dziedzictwo"
  eragon     smok zostawać     arya  eragona  podczas
      15        8        8        6        6        6
[1] "Pieśń lodu i ognia 1. Gra o tron"
   drogi daenerys     król      ned  roberta westeros
      13        9        7        7        7        7
[1] "Pieśń lodu i ognia 2. Starcie królów"
joffreya daenerys    dziki     król     siła stannisa
       9        8        8        7        6        6
[1] "Pieśń lodu i ognia 3. Nawałnica mieczy"
zostawać     król daenerys    straż     lord  catelyn
      17       12       10       10        9        8
[1] "Pieśń lodu i ognia 4. Uczta dla wron"
  cersei zostawać  żelazny daenerys     król  braavos
      11        8        7        6        6        5
[1] "Wiedźmin 1. Krew elfów"
    ciri yennefer   geralt    triss  geralta wiedźmin
      59       20       19       19       16       15
[1] "Wiedźmin 2. Czas pogardy"
      ciri czarodziej     geralt nilfgaardu    thanedd   zostawać
         8          4          4          4          4          4
[1] "Wiedźmin 3. Chrzest ognia"
kompania  jaskier     ciri   geralt     obóz    regis
       7        5        4        4        4        4
[1] "Wiedźmin 4. Wieża jaskółki"
    ciri   udawać belhaven   geralt    hanza kompania
       9        5        4        4        4        4
[1] "Władca Pierścieni 1. Drużyna Pierścienia"
pierścień    hobbit     frodo    zostać   gandalf   drużyna
       53        29        26        15        13        12
```

Rys. 31. Wynik dla Exp_keywords_tf_all

3. Exp_keywords_tf_612

W tym eksperymencie wykorzystano macierz ograniczoną dtm_tf_612. Z tego powodu część nazw własnych charakteryzujących dane książki, nie została uwzględniona w macierzy. Wynikiem są więc słowa bardziej ogólne, jednak mogą one być również wartościowe. Ciężej jest na ich podstawie bezpośrednio zidentyfikować dany dokument, jednak dają one ogólne rozeznanie o tym czego on może dotyczyć. Dla osób niezaznajomionych z tymi cyklami będą one wręcz bardziej pomocne, ponieważ zarysują bardzo ogólnie czego można się spodziewać sięgając po daną książkę. Jest to jednak w

większości zbiór, który można by było określić ogólnie jako słowa kluczowe gatunku “Fantastyka”. Wyniki dla tego eksperymentu przedstawione zostały na poniższej grafice.

```
[1] "Eragon 1. Eragon"
jeździec      elf    grupa     smok    magia   miasta
       6        4        4        4        3        3
[1] "Eragon 2. Najstarszy"
   jeździec    okazywać   krasnolud     książka       magia postanawiać
          5           4           3           3           3           3
[1] "Eragon 3. Brisingr"
 żołnierz     misja      smok   zabijać   dowódca krasnolud
        5         4         4         4         3         3
[1] "Eragon 4. Dziedzictwo"
     smok      arya  jeździec znajdywać     magia  atakować
        8         6         5         5         4         3
[1] "Pieśń lodu i ognia 1. Gra o tron"
   północ królewski       mur  przystań     straż      brać
        6         5         5         5         4         3
[1] "Pieśń lodu i ognia 2. Starcie królów"
  siła    cel północ   smok  straż   tron
     6      4      4      4      4      4
[1] "Pieśń lodu i ognia 3. Nawałnica mieczy"
    straż królewski    północ       mur  przystań     nocny
       10         8         8         7         7         6
[1] "Pieśń lodu i ognia 4. Uczta dla wron"
 brata  córka  straż   arya   brat koniec
     5      5      5      4      4      4
[1] "Wiedźmin 1. Krew elfów"
     mówić     uważać       inny        elf   okazywać czarodziej
        10         10          7          6          6          5
[1] "Wiedźmin 2. Czas pogardy"
czarodziej       czas   dołączać      droga        elf   okazywać
         4          2          2          2          2          2
[1] "Wiedźmin 3. Chrzest ognia"
     elfi    strona tymczasem    wojsko  atakować      cały
        3         3         3         3         2         2
[1] "Wiedźmin 4. Wieża jaskółki"
       elf   odnaleźć        ona tajemniczy  tymczasem dziewczyna
         3         3         3         3         3         2
[1] "Władca Pierścieni 1. Drużyna Pierścienia"
       elf   wyruszyć       góra czarodziej     podróż        sam
        11        10         9         7         7         7
```

Rys. 32. Wynik dla Exp_keywords_tf_612

4. Spis Grafik