Imię	Nazwisko	Numer albumu	Grupa dziekańska	Wkład (Zadania)	Wkład (Procentowo)
Michal	Antos	220576	ZZISN2-2411IS	100%	100%
Jakub	Abramczyk	220585	ZZISN2-2411IS	100%	100%

__/100 pkt

Wstęp

Zbiór dokumentów składa się z 20 tekstów w języku polskim zawierających około 3 tysiące słów każdy. Dokumenty zostały podzielone na 5 różnych grup tematycznych. Wybrane grupy tematyczne to:

- Grupa 1 (fragment książki)
  - chlopi-czesc-czwarta-lato.txt
  - chlopi-czesc-czwarta-wiosna.txt
  - chlopi-czesc-czwarta-jesien.txt
  - chlopi-czesc-czwarta-zima.txt
- Grupa 2 (artykuł)
  - Jan_Pawel_II.txt
  - Jozef_Pilsudski.txt
  - Robert_Lewandowski.txt
  - Mikolaj_Kopernik.txt
- Grupa 3 (fragment książki)
  - Wiedzmin-01.txt
  - Wiedzmin-02.txt
  - Wiedzmin-03.txt
  - Wiedzmin-04.txt
- Grupa 4 (fragment książki)
  - George R.R. Martin 01 - Gra o tron.txt
  - George R.R. Martin 03 - Nawalnica mieczy 01 - Stal i snieg.txt
  - George R.R. Martin 04 - Nawalnica mieczy 02 - Krew i zloto.txt
  - George R.R. Martin 05 - Uczta dla wron 01 - Cienie smierci.txt
- Grupa5 (fragment książki)
  - Hobbit.txt
  - Wladca_pierscien_Powrot_Krola.txt
  - Wladca_pierscieni_Druzyna_Pierscienia.txt
  - Wladca_pierscieni_Dwie_Wieze.txt

Ponadto Grupa 5 i Grupa 3 są do siebie zbliżone jako że obie to gatunek fantastyki. Grupa 2 została wybrana w taki sposób aby w jak największym stopniu wyróżniać się od pozostałych.

Preprocessing

Projekt rozpoczęliśmy od utworzenia korpus dokumentów na którym następnie możliwe było przeprowadzenie eksperymentów. Na gotowym korpusie zostało przeprowadzone wstępne przetwarzanie (ang. “preprocessing”) mające na celu odpowiednio przygotować nasz zbiór danych. Preprocessing miał na celu przeprowadzenie takich działań jak konwersja znaków na małe litery, usunięcie znaków specjalnych, usunięcie cyfr, usunięcie ciągów znaków zawartych w pliku stopwords pl.txt itd. W kolejnym kroku korpus został poddany lematyzacji, tzn procesowi który sprowadza słowa do ich form podstawowych. Tak przetworzony korpus został zapisany do plików tekstowych w folderze “new_dataset_przetworzone”, na których możliwe było przeprowadzenie dalszych eksperymentów.

Macierze częstotliwości

W tym kroku należało przygotować macierze częstotliwości zmieniając wartości parametrów takich jak:

- waga
- max liczba dokumentów w których mogą wystąpić słowa
- liczba dokumentów w których mogą wystąpić słowa

W przypadku wagi do wyboru mieliśmy trzy dostępne parametry:

- weightTfIdf
- weightTf (wartość domyślna)
- weightBin

Max i Min liczba dokumentów musiał znajdować się w zakresie <1,20>

W naszym przypadku postanowiliśmy wykorzystać trzy kombinacje:

1. waga=WeightTf, Min = 6, Max = 14
2. waga=WeightTfIdf, Min=1, Max = 20
3. waga=WeightTfIdf, Min =4, Max = 18

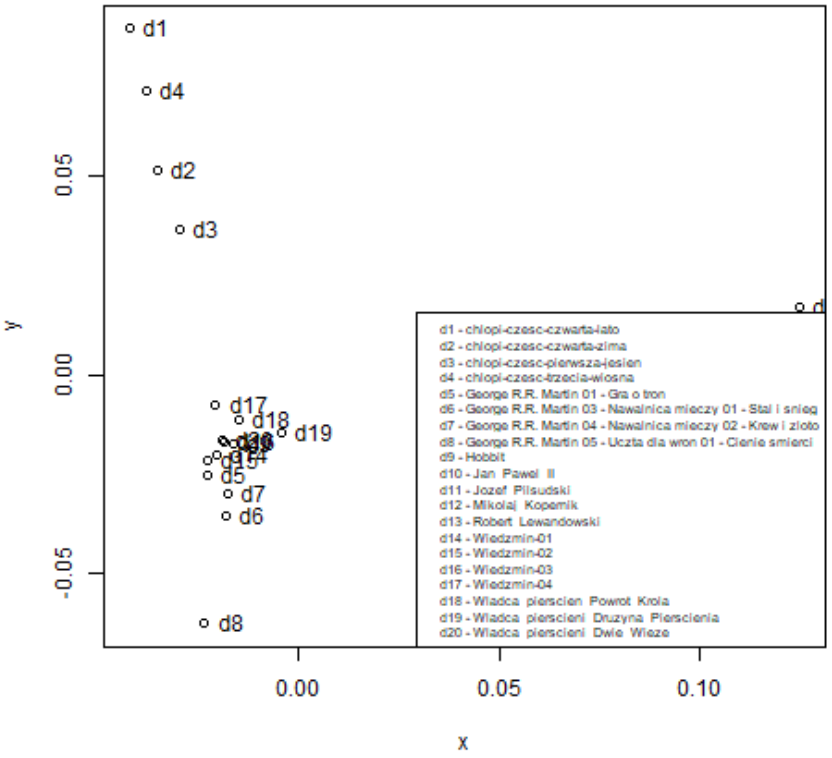
Macierze wygenerowane przez skrypt to odpowiednio:

- dtm_tf_6_14_m.csv oraz tdm_tf_6_14_m.csv
- dtm_tfidf_1_20_m.csv oraz tdm_tfidf_1_20_m.csv
- dtm_tfidf_4_18_m.csv oraz tdm_tfidf_4_18_m.csv

Analiza głównych składowych

Analiza składowych głównych (PCA) to najbardziej popularny algorytm redukcji wymiarów. W ogólnym skrócie polega on na rzutowaniu danych do przestrzeni o mniejszej liczbie wymiarów tak, aby jak najlepiej zachować strukturę danych.
Służy głównie do redukcji zmiennych opisujących dane zjawisko oraz odkrycia ewentualnych prawidłowości między cechami. Dokładna analiza składowych głównych umożliwia wskazanie tych zmiennych początkowych, które mają duży wpływ na wygląd poszczególnych składowych głównych czyli tych, które tworzą grupę jednorodną.

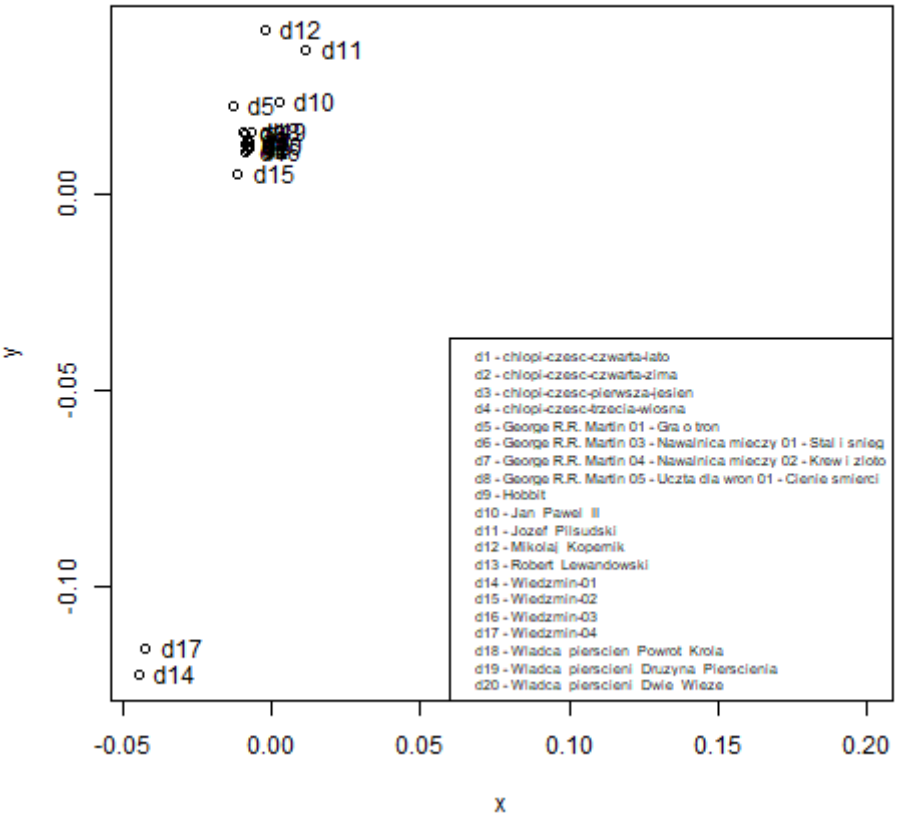
Jako pierwsza przeprowadzona została analiza głównych składowych dla macierzy częstotliwości o parametrach (WeightTfIdf, Min=4, Max-18).



PCA dla (WeightTfIdf, Min=4, Max-18)

Dzięki przedstawieniu danych w postaci graficznej możemy w łatwy i czytelny sposób zaobserwować podobieństwo między dokumentami. Jak wynika z wykresu najbardziej wyróżniające się dokumenty to zbiór dokumentów zawierających fragmenty powieści Chłopi. Moze to wynikac z faktu że powieść ta została napisana językiem staropolskim i zawiera słowa których nie używa się w mowie potocznej.

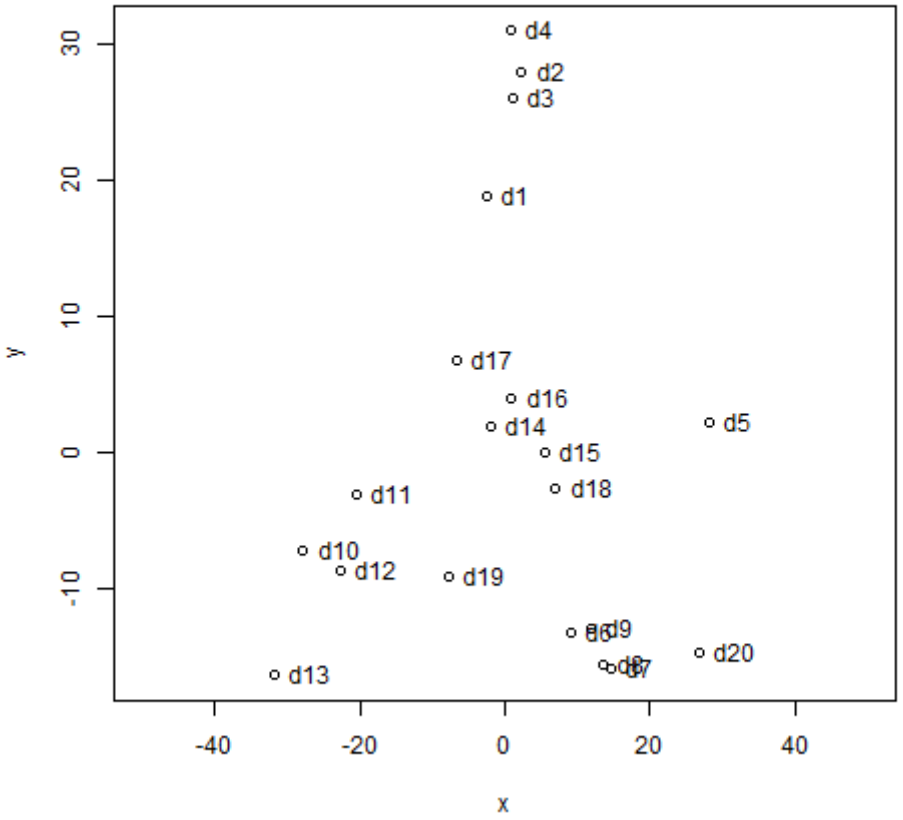
Analiza głównych składowych dla macierzy częstotliwości o parametrach (WeightTfIdf, Min=1, Max=20).



PCA dla (WeightTfIdf, Min=1, Max=20)

W przypadku analizy dla parametrów (WeightTfIdf, Min=1, Max=20) dokumenty które wyróżniały się od pozostałych to (d17 i d14) oraz (d12 i d11) co najprawdopodobniej wynika z faktu ze musialy one zawierać duza ilosc slow nie pojawiajacych sie w zadnym innym dokumencie.

Analiza głównych składowych dla macierzy częstotliwości o parametrach (WeightTf, Min=6, Max=14).



PCA dla (WeightTf, Min=6, Max=14)

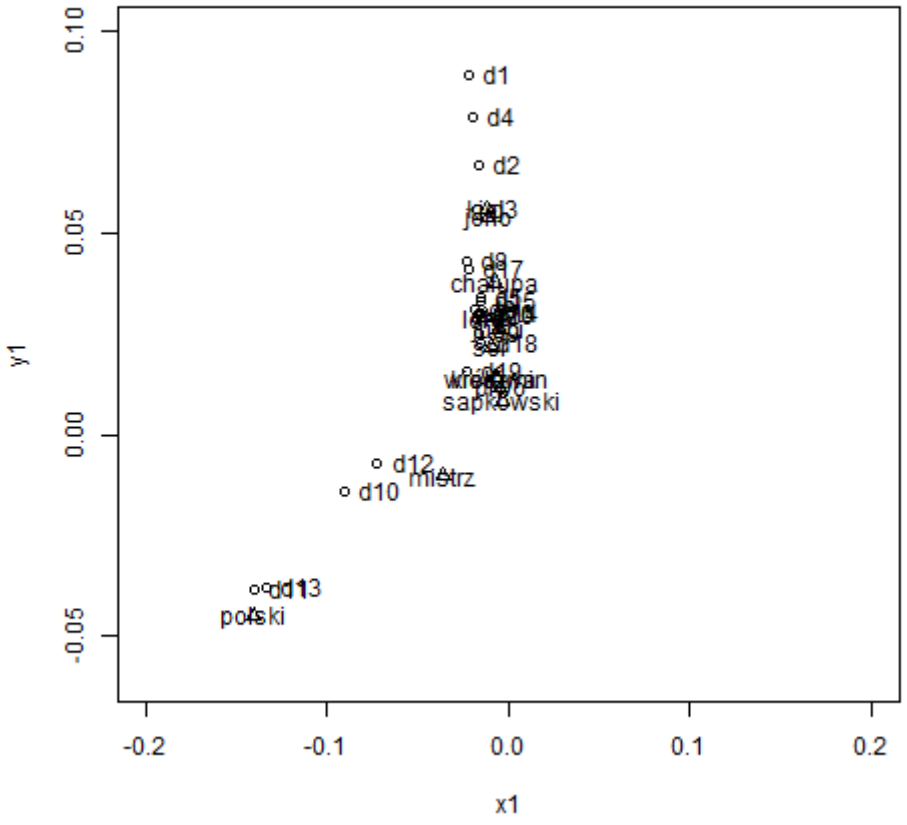
W przypadku analizy glownych skladowych dla parametrow (WeightTf, Min=6, Max=14), nie możemy jasno okreslic ktore teksty wyróżniają się od pozostalych poniewaz punkty na wykresie nie skupiają się wokół jednego miejsca. Możemy natomiast stwierdzić które teksty są do siebie podobne. Tak np. d14 i d18 co wynika prawdopodobnie z faktu że są to książki o podobnej tematyce. (Wiedźmin i Wladca pierscieni)

Dekompozycja według wartości osobliwych

Utajona analiza semantyczna (LSA) to technika przetwarzania języka naturalnego , w szczególności semantyki dystrybucyjnej , polegająca na analizowaniu relacji między zbiorem dokumentów a zawartymi w nich terminami poprzez tworzenie zbioru pojęć związanych z dokumentami i terminami. LSA zakłada, że słowa o zbliżonym znaczeniu wystąpią w podobnych fragmentach tekstu.

Dekompozycje rozpoczęliśmy od wykorzystania macierzy częstotliwości o parametrach (WeightTfIdf, Min=4, Max-18).

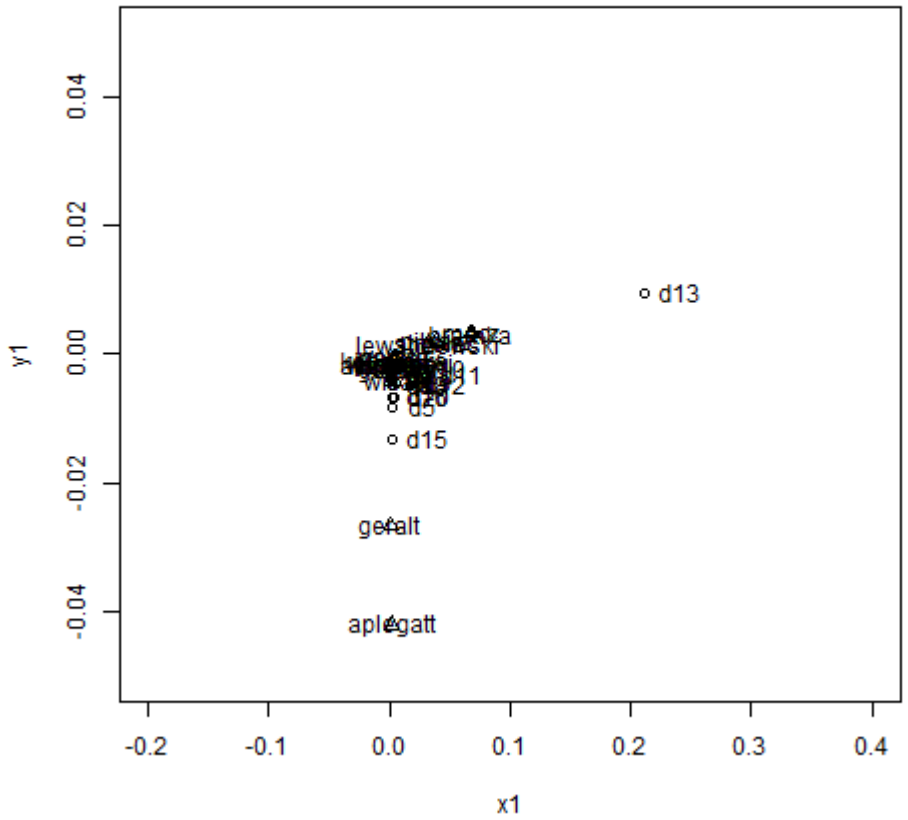
Own terms:
["sapkowski","piwo","juści","wiedźmin","ino","królowa","chałupa","ser","mistrz","jeno","kiej", "lord","polski"]



LSA dla (WeightTfIdf, Min=4, Max-18)

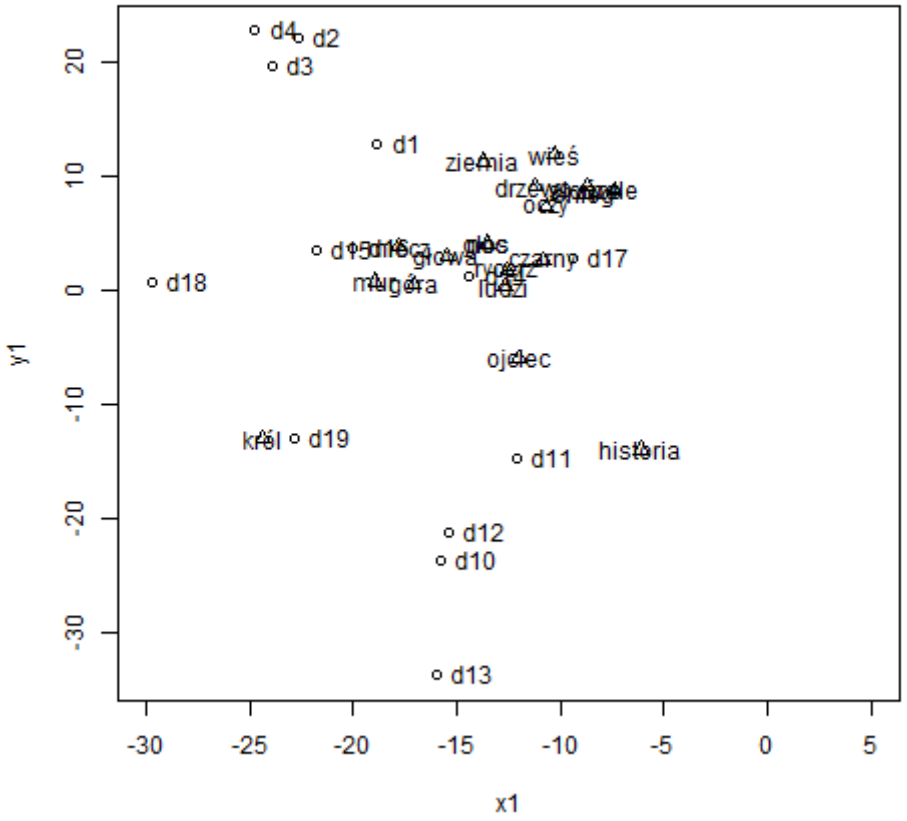
Dekompozycja dla macierzy częstotliwości o parametrach (WeightTfIdf, Min=1, Max-20).

Own terms:
["lewandowski","geralt","strzelec","piłkarz","kopernika","waymar","aplegatt","hobbit","royce",
"papież","wojtyła","astronom","gared","bramka","mecz","kopernik","piłsudski","willa"]



LSA dla (WeightTfIdf, Min=1, Max=20)

Dekompozycja dla macierzy częstotliwości o parametrach (WeightTf, Min=6, Max=14).
Own terms:
["oczy","noc","ludzi","czarny","śnieg","pole","głos","słońce","ojciec","drzewo","głowa","rycerz",
"historia","ziemia","góra","wieś","miecz","mur","król"]



LSA dla (WeightTf, Min=6, Max=14)

Analiza skupień dokumentów

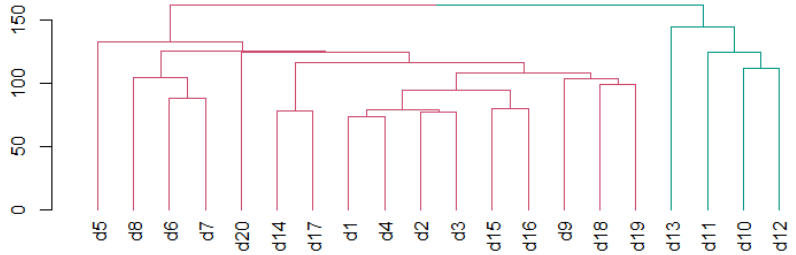
Metoda tzw. klasyfikacji bez nadzoru. Jest to metoda dokonująca grupowania elementów we względnie jednorodne klasy. Podstawą grupowania w większości algorytmów jest podobieństwo pomiędzy elementami – wyrażone przy pomocy funkcji podobieństwa.

Analiza Hierarchiczna

W początkowym etapie wykorzystaliśmy metodę hierarchiczną, która ma na celu zbudowanie hierarchii klastrów. Służy do dzielenia obserwacji na grupy bazując na podobieństwach między nimi. W przeciwieństwie do wielu algorytmów służących do klastrowania w tym wypadku nie jest konieczne wstępne określenie liczby tworzonych klastrów. W przypadku naszego projektu liczba klastrów w metodzie hierarchicznej dobraliśmy na podstawie długości wiązań.

Eksperyment pierwszy

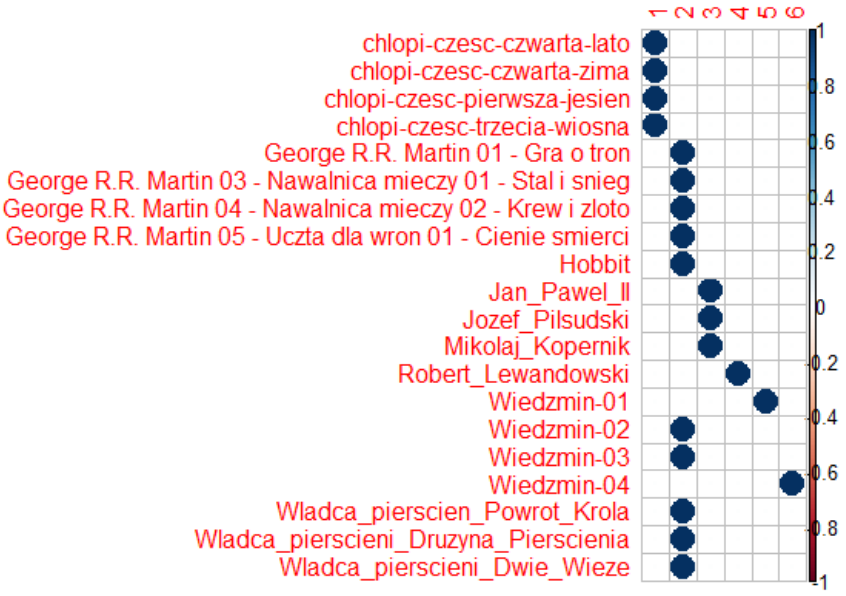
W pierwszym eksperymencie zastosowaliśmy euklidesową macierz odległości, oraz metodę Complete-linkage clustering. W metodzie tej na początku procesu każdy element znajduje się we własnym klastrze. Klastry są następnie kolejno łączone w większe klastry, aż wszystkie elementy znajdą się w tym samym klastrze.

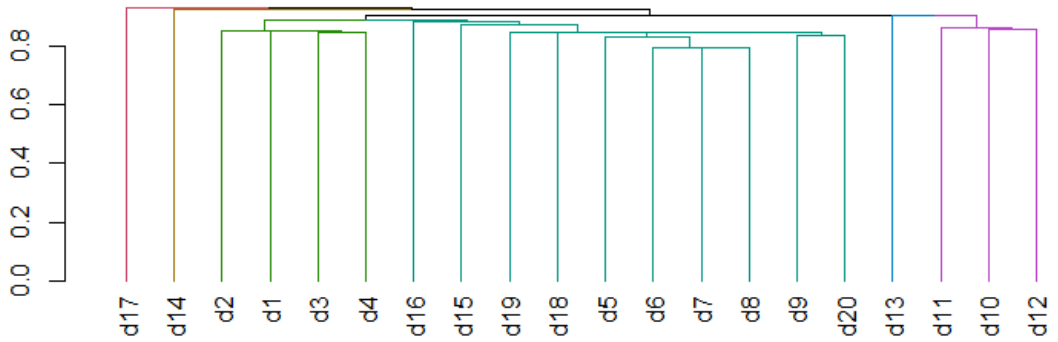


Metody użyte w tym eksperymencie podzieliły nasze dane na dwa główne klastry, odróżniając tym samym najbardziej odległy temat od innych. Jak możemy zauważyć na dendrogramie głębszy podział nie jest już taki dokładny.

Eksperyment drugi

W drugim eksperymencie zastosowaliśmy Indeks Jaccarda , znany również jako współczynnik podobieństwa Jaccarda, który jest metodą używaną do pomiaru podobieństwa i różnorodności zestawów próbek**.** Oraz metodę single-linkage clustering, która polega ona na grupowaniu klastrów w sposób oddolny (grupowanie aglomeracyjne), łącząc na każdym kroku dwa klastry zawierające najbliższą parę elementów, które nie należą jeszcze do tego samego klastra.

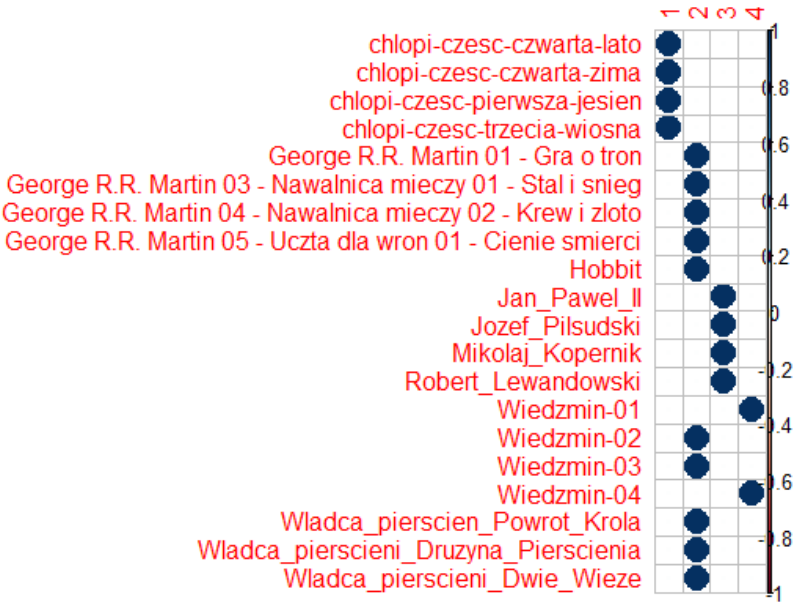


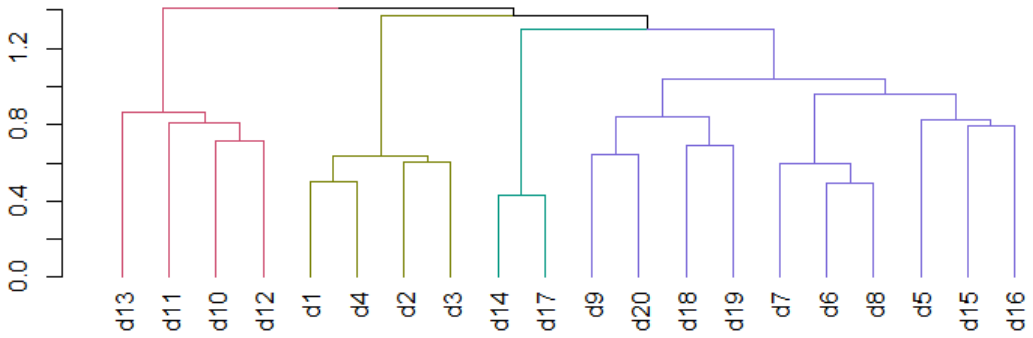


Metody użyte w tym eksperymencie podzieliły nasze dane na sześć różnych klastrów i oddzieliły od siebie trzy najbardziej odległe tematy, zauważamy też, że do niektórych klastrów został przydzielony tylko jeden dokument co nie jest optymalnym wynikiem.

Eksperyment trzeci

W trzecim eksperymencie zastosowaliśmy cosinusową macierz odległości, która używana jest do określenia, jak podobne są dwie jednostki, niezależnie od ich wielkości. Mierzy cosinus kąta między dwoma wektorami w przestrzeni wielowymiarowej. Oraz metodę Warda, w której łączenie dokonywane jest w taki sposób, aby w najmniejszym stopniu zwiększyć wariancję wewnątrzgrupową.



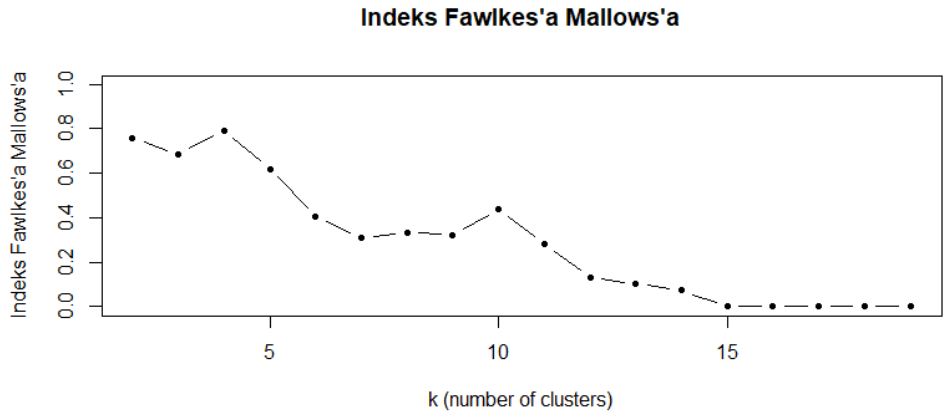


Metody użyte w tym eksperymencie podzieliły nasze dane na cztery klastry i podobnie jak w eksperymencie drugim oddzieliły od siebie trzy najbardziej rozbieżne tematyki. Możemy zaobserwować, że klastry nie mają przydzielonych do siebie pojedynczych dokumentów, tak jak miało to miejsce w eksperymencie drugim. Zastosowanie powyższych metod okazało się najbardziej skuteczne w przypadku naszego zbioru danych.

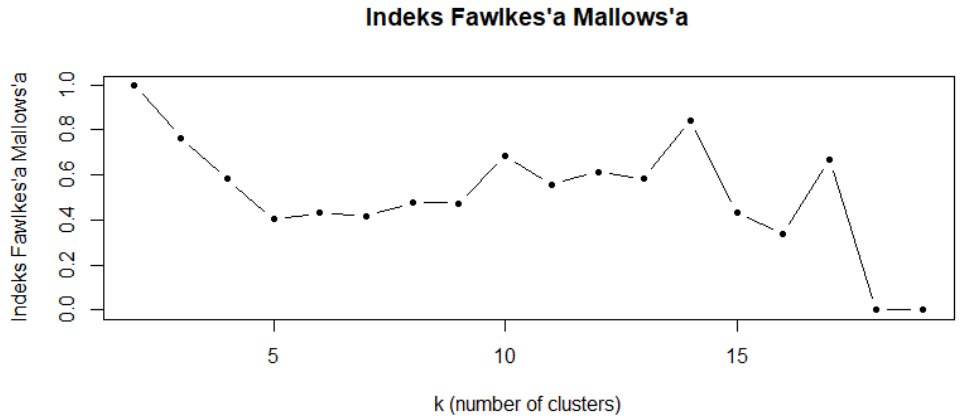
Porównanie eksperymentów

Indeks Fowlkesa-Mallowsa jest metodą oceny, która służy do określenia podobieństwa między dwoma klastrami. Wyższa wartość wskaźnika Fowlkesa-Mallowsa wskazuje na większe podobieństwo między klastrami. . Chociaż technicznie jest używany do określenia podobieństwa między dwoma klasteryzacjami, jest zwykle używany do oceny wydajności klasteryzacji algorytmu grupowania, zakładając, że drugie grupowanie jest podstawową prawdą, tj. obserwowanymi danymi, i zakładając, że jest to idealne grupowanie.

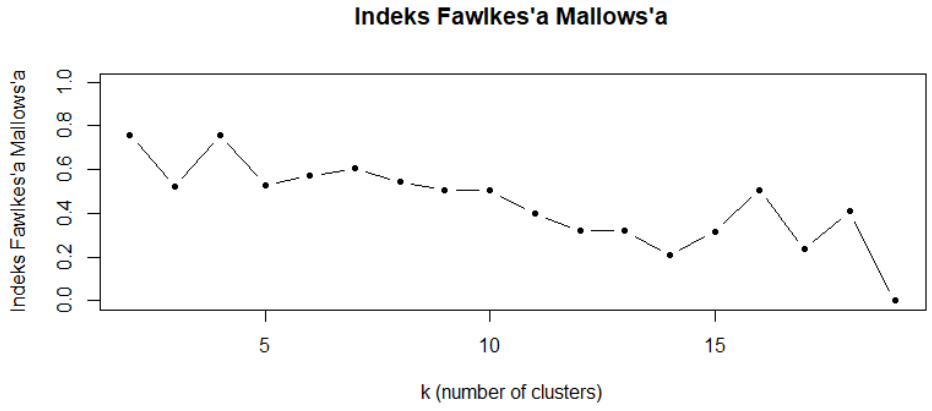
Porównanie eksperymentów 1 i 2.



Porównanie eksperymentów 1 i 3.



Porównanie eksperymentów 2 i 3.



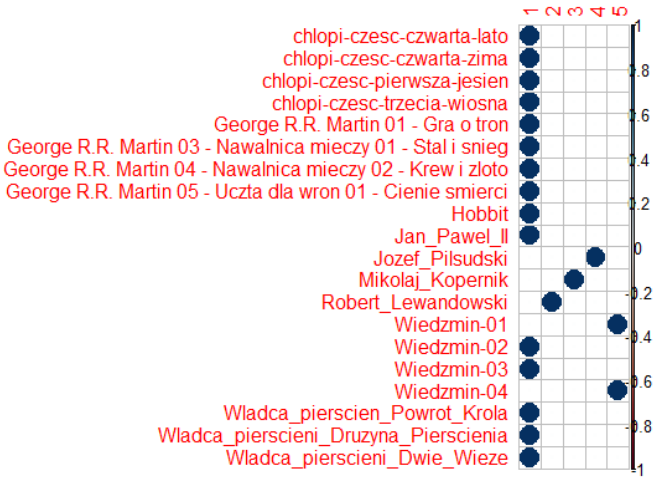
Dzięki porównaniu ze sobą różnych eksperymentów, możemy ocenić jakość klasteryzacji. W przypadku przeprowadzonych przez nas badań, widzimy, że najbardziej różnią się od siebie pierwszy i drugi eksperyment, a największe podobieństwa eksperymenty mają dla czterech klas.

Analiza nie hierarchiczna

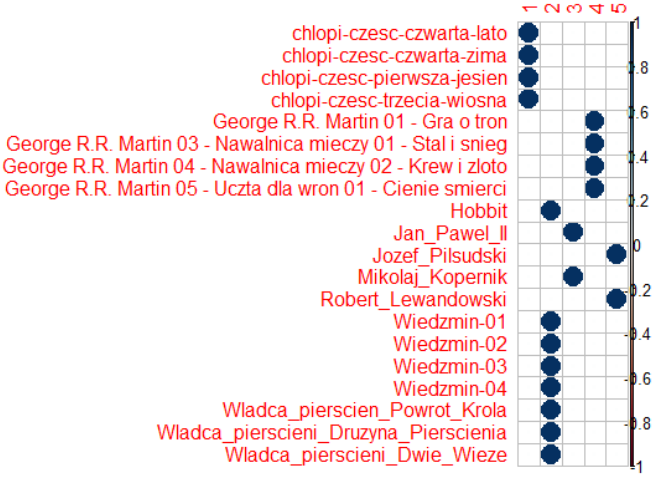
W analizie niehierarchicznej użyliśmy metoda k-średnich.Celem tej metody jest podzielenie obiektów na zadaną liczbę klas w taki sposób, aby suma odległości poszczególnych obiektów od środków klas do których należą była minimalna. Liczba klas w naszym przypadku była ustalona na podstawie liczby tematów.

- **waga=WeightTf, Min = 6, Max = 14**

- **waga=WeightTfIdf, Min=1, Max = 20**



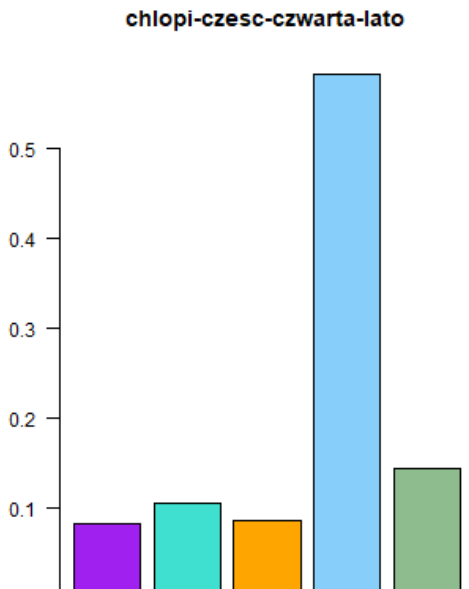
- **waga=WeightTfIdf, Min =4, Max = 18**

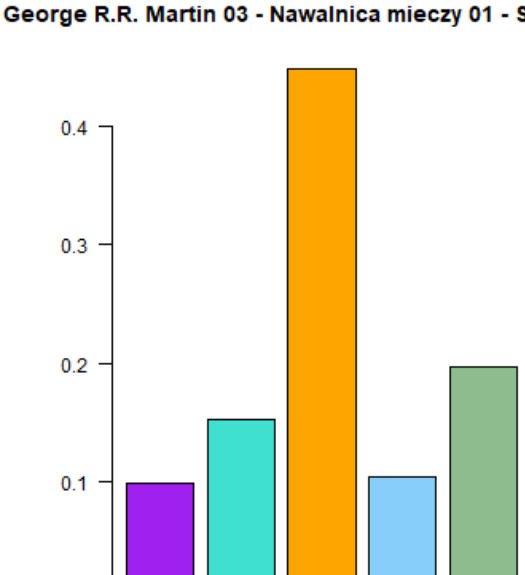


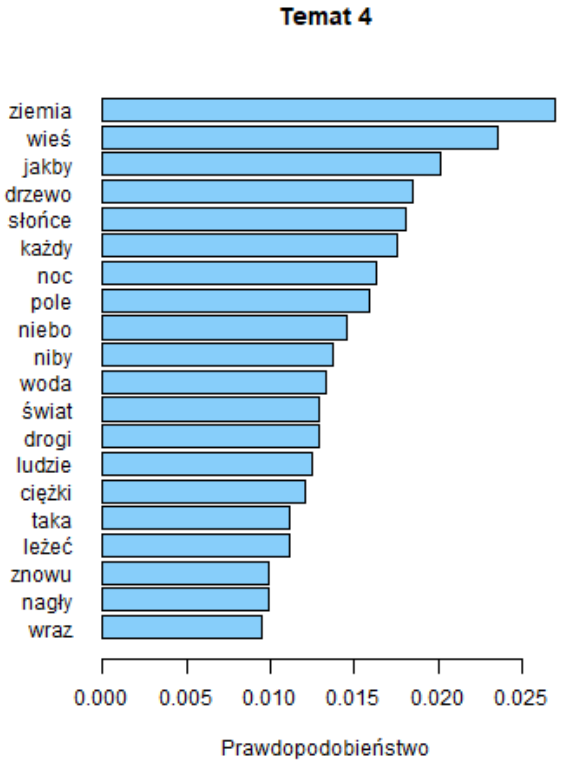
Po porównaniu wszystkich wyników z analizy nie hierarchicznej możemy dojść do wniosku, że przy żadnych wagach klasyfikacja nie była idealna, a najlepszy wynik możemy obserwować dla waga=WeightTfIdf, Min =4, Max = 18.

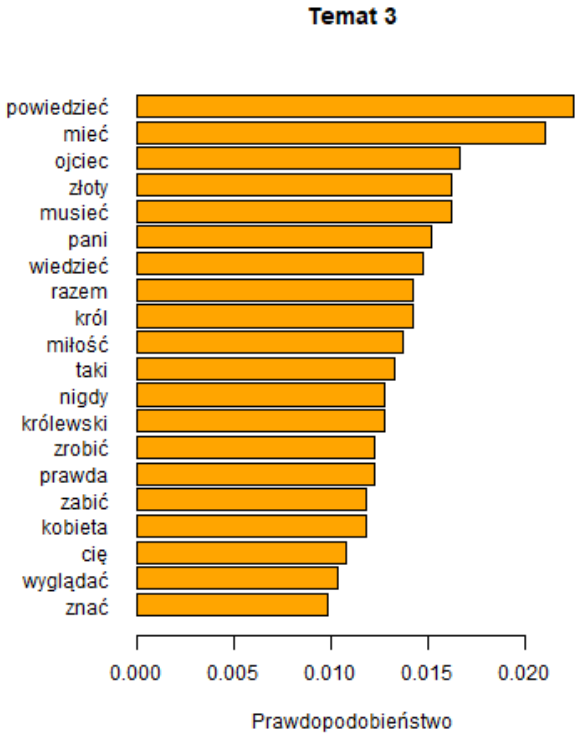
Analiza Tematów

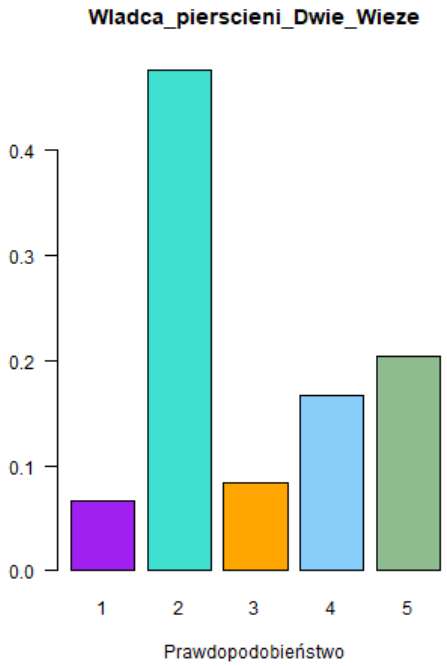
Przeprowadziliśmy analizę tematyk korzystając z metody ukrytej alokacji Dirichlet'a. Liczbę tematyk wyznaczyliśmy na podstawie liczby badanych tematów. Przeprowadziliśmy także eksperymenty dla większej i mniejszej liczby tematyk. Poniżej znajdują się przykładowe wykresy prawdopodobieństwa należenia danego dokumentu do odpowiedniego tematu, oraz wykresy prawdopodobieństwa wystąpienia danych słów w danym temacie.

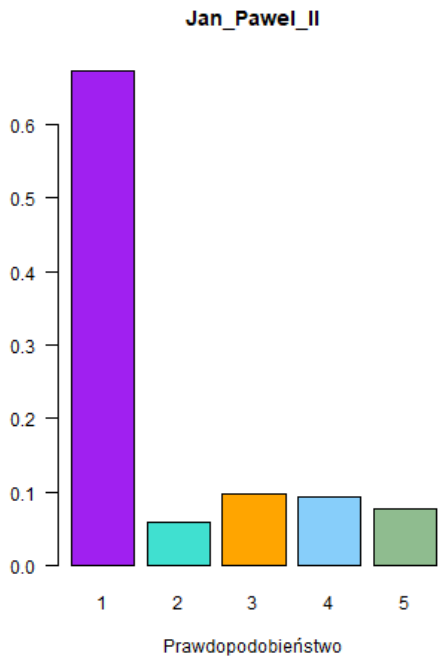


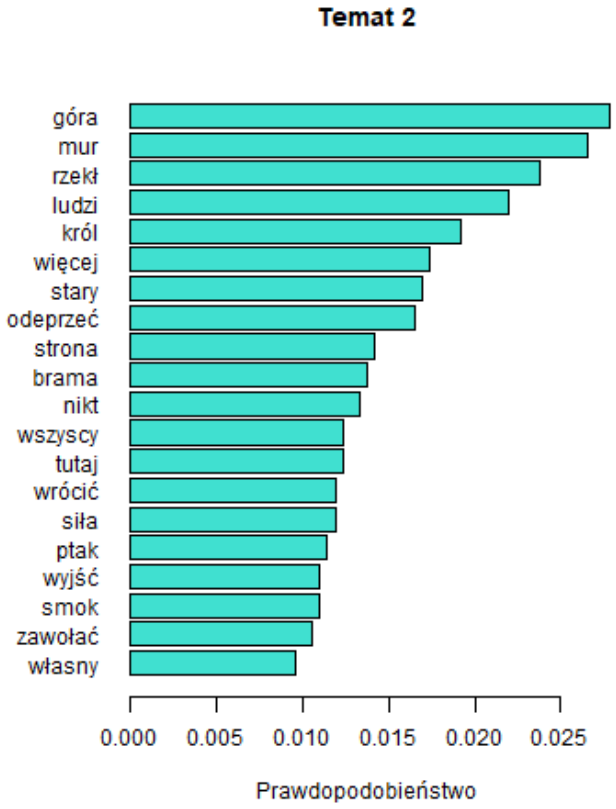


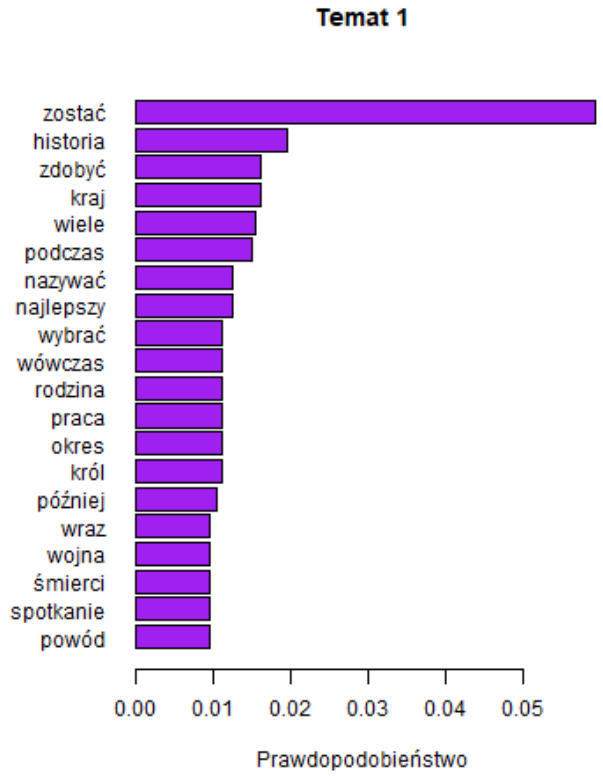

Temat 4
ziemia
wieś
jakby
drzewo
słońce
każdy
noc
pole
niebo
niby
woda
świat
drogi
ludzie
ciężki
taka
leżeć
znowu
nagły
wraz
0.000
0.005
0.010
0.015
0.020
0.025
Prawdopodobieństwo


Temat 3
powiedzieć
mieć
ojciec
złoty
musieć
pani
wiedzieć
razem
król
miłość
taki
nigdy
królewski
zrobić
prawda
zabić
kobieta
cię
wyglądać
znać
0.000
0.005
0.010
0.015
0.020
Prawdopodobieństwo


Wladca_pierscieni_Dwie_Wieze
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
Prawdopodobieństwo


Jan_Pawel_II
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
Prawdopodobieństwo





Na podstawie przeprowadzonych eksperymentów możemy stwierdzić, że przy odpowiedniej ilości tematyk (w naszym przypadku równemu liczbie tematów) możemy jednoznacznie dopasować dany dokument do odpowiedniej tematyki. Co więcej dopasowania te będą ewidentnie pokrywać się z wybranym przez nas tematem. Prawdopodobieństwo wystąpienia słów w danym temacie, także jasno koreluje z dopasowanymi do nich dokumentami .

Frazy Kluczowe

Przeprowadziliśmy analizę fraz kluczowych korzystając z metody ukrytej alokacji Dirichlet'a. Przy użyciu LDA można wyznaczyć ważność słowa w zidentyfikowanym temacie lub w rozpatrywanym dokumencie.

Analiza fraz kluczowych na przykładzie wybranego dokumentu.

waga tfidf 1_14 jako miara istotności słów
chlopi-czesc-czwarta-lato
jeno kiej pacierz jaże jambroż chałupa

waga tfidf 1_20 jako miara istotności słów
"chlopi-czesc-czwarta-lato"
jeno kiej pacierz jaże jambroż chałupa

waga tfidf 4_18 jako miara istotności słów
"chlopi-czesc-czwarta-lato"
jeno kiej chałupa juści izba żałosny

prawdopodobieństwo w LDA jako miara istotności słów

"chlopi-czesc-czwarta-lato"
ziemia droga jakby kiej wieś oczy

Na podstawie przykładu można wywnioskować, że miary istotności słow w zależnośći od wag nie różnią się znacznie od siebie. Zdecydwoaną różnice można jednak zauważyć przy użyciu prawdopodobieństwa metody LDA jako miary istotności słów. Dobrane słowa kluczowe są adekwatne do oczekiwanych w większości przypadków. Jednak z przeprowadzonych eksperymentów dokładnie widać, że najczęściej pojawiające się w dokumencie słowa niekoniecznie są frazami kluczowymi.