Projekt

Eksploracyjna analiza tekstu w R

Grupa projektowa: Projekt 60

Imię	Nazwisko	Numer albumu	Grupa dziekańska	Wkład w prace nad projektem (zadania)	Wkład w prace nad projektem (procentowo)
Konrad	Adamik	215941	ZZISS2-2411IS	2, 3, 4, 7	50
Weronika	Mirek	215916	ZZISS2-2412IS	1, 5, 6, 7	50

Zadania odpowiadają kolejnym rozdziałom w sprawozdaniu.

Kod oraz dokumenty do projektu dostępne także na platformie github:
https://github.com/konrad-adamik/pjn-projekt/

___/100 pkt

1. Charakterystyka zbioru dokumentów

Wybrany przez nas zbiór dokumentów składa się z pięciu tematyk:

- IT,
- Ekonomia,
- Zdrowie,
- COVID-19,
- Literatura.

Każdy z 20 tekstów ma ok. 14000 - 32000 znaków. Wszystkie dokumenty zapisane są w formacie .txt z kodowaniem UTF-8. Został z nich utworzony korpus dokumentów, z którego można wyciągnąć poniższe informacje:

- podstawowe informacje o korpusie:

```
> corpusPP
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 20
```

- podsumowanie korpusu:

```
> summary(corpusPP)
                                                                                                  Length Class             Mode
Efektywne zastosowanie IT w przedsiębiorstwie.txt                                                 2      PlainTextDocument list
Ekonomia Behawioralna - Hybryda Teorii i Eksperymentu.txt                                         2      PlainTextDocument list
Ekonomia i społeczne imaginarium.txt                                                              2      PlainTextDocument list
Ekonomia społeczna jako aktor rynku pracy.txt                                                     2      PlainTextDocument list
Ekonomia w dobie finansyzacji gospodarki.txt                                                      2      PlainTextDocument list
Folklor i literatura.txt                                                                          2      PlainTextDocument list
Informacje i tymczasowe wytyczne dla farmaceutów i pracowników aptek.txt                          2      PlainTextDocument list
Jakość życia a zdrowie.txt                                                                        2      PlainTextDocument list
Koronawirus 2019-nCoV - transmisja zakażenia, objawy i leczenie.txt                               2      PlainTextDocument list
Kreowanie wartości poprzez efektywne zarządzanie usługami IT.txt                                  2      PlainTextDocument list
Literatura i pisarze wobec cenzury PRL - wprowadzenie.txt                                         2      PlainTextDocument list
Literatura jako trop rzeczywistości - Wprowadzenie.txt                                            2      PlainTextDocument list
Literatura nowoczesna, cztery dyskursy.txt                                                        2      PlainTextDocument list
Metody i narzędzia rozwiązywania problemów komunikacji w relacji IT-Biznes w projektach informatycznych.txt 2      PlainTextDocument list
Podręcznik prewencji i leczenia COVID-19.txt                                                      2      PlainTextDocument list
Religia a zdrowie - czy religia może sprzyjać trosce o zdrowie.txt                                2      PlainTextDocument list
Sprężystość psychiczna i zmienne pośredniczące w jej wpływie na zdrowie.txt                       2      PlainTextDocument list
Styl życia młodzieży i jego wpływ na zdrowie.txt                                                  2      PlainTextDocument list
Zarządanie projektami w przedsiębiorstwach branży IT - studium literaturowe.txt                   2      PlainTextDocument list
Zrozumieć COVID-19.txt                                                                            2      PlainTextDocument list
```

- szczegółowe informacje o korpusie:

```
> inspect(corpusPP)
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 20

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 21654

[[2]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 20749

[[3]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 31921

[[4]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 31397

[[5]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 29183

[[6]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 20168

[[7]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 14070

[[8]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 18230

[[9]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 15058
```

```
[[10]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 24833

[[11]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 18944

[[12]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 20210

[[13]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 22424

[[14]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 22616

[[15]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 18283

[[16]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 20824

[[17]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 20555

[[18]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 17213

[[19]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 23123

[[20]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 22436
```

- informacje o pojedynczych dokumentach w korpusie:

```
> corpusPP[1]
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1

...
> corpusPP[20]
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1
```

- zawartość pojedynczych dokumentów w korpusie:

```
> corpusPP[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 21654

...
> corpusPP[[20]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 22436
```

- podsumowanie pojedynczych dokumentów w korpusie:

```
> summary(corpusPP[1])
                                                 Length Class             Mode
Efektywne zastosowanie IT w przedsiębiorstwie.txt 2      PlainTextDocument list

...
> summary(corpusPP[20])
                         Length Class             Mode
Zrozumieć COVID-19.txt 2      PlainTextDocument list
```

- podsumowanie zawartości pojedynczych dokumentów w korpusie:

```
> summary(corpusPP[[1]])
        Length Class            Mode
content 482    -none-           character
meta      7    TextDocumentMeta list

...
> summary(corpusPP[[20]])
        Length Class            Mode
content 318    -none-           character
meta      7    TextDocumentMeta list
```

- informacje szczegółowe o pojedynczych dokumentach w korpusie:

```
> inspect(corpusPP[1])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 21654

...
```

```
> inspect(corpusPP[20])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 22436
```

- informacje szczegółowe o zawartości pojedynczych dokumentów w korpusie (w tym zawartość dokumentów):

```
> inspect(corpusPP[[1]])
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 21654

Technologie informacyjne (IT) odgrywają we współczesnych przedsiębiorstwach bardzo istotną rolę, przenikając
już prawie każdy aspekt ich działalności. Świadczy o tym
m.in. wysokość wydatków na inwestycje w IT, które
pochłaniają często ponad połowę wszystkich wydatków
inwestycyjnych przedsiębiorstw. Jednocześnie najczęściej
nie przynoszą zakładanych efektów. Pojawia się zatem

...
```

```
> inspect(corpusPP[[20]])
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 22436

2. Informacje podstawowe
2.1 Mechanizm powstawania nowych chorób zakaźnych
SARS-COV-2 jest typowym wirusem pochodzenia zwierzęcego, jest przyczyną choroby
COVID-19 zaliczanej do grupy zoonoz. To w tej grupie klasyfikowana jest zdecydowana większość chorób ludzkich. Przenoszenie patogenów między zwierzętami i ludźmi (ale też ludźmi
i zwierzętami) odbywa się jako element naturalnego cyklu biologicznego.
Przez ostatnie 70 lat obserwuje się na świecie postępującą i przyspieszającą presję antropogeniczną. Masowo i masywnie przekształcane jest naturalne środowisko. Prowadzi to do po
wstawania nowych warunków obiegu patogenów (wirusów, bakterii, pasożytów) – w ramach
nisz socjo-ekologicznych. W rezultacie powstaje zupełnie nowy ekosystem, którego atrybuty
```

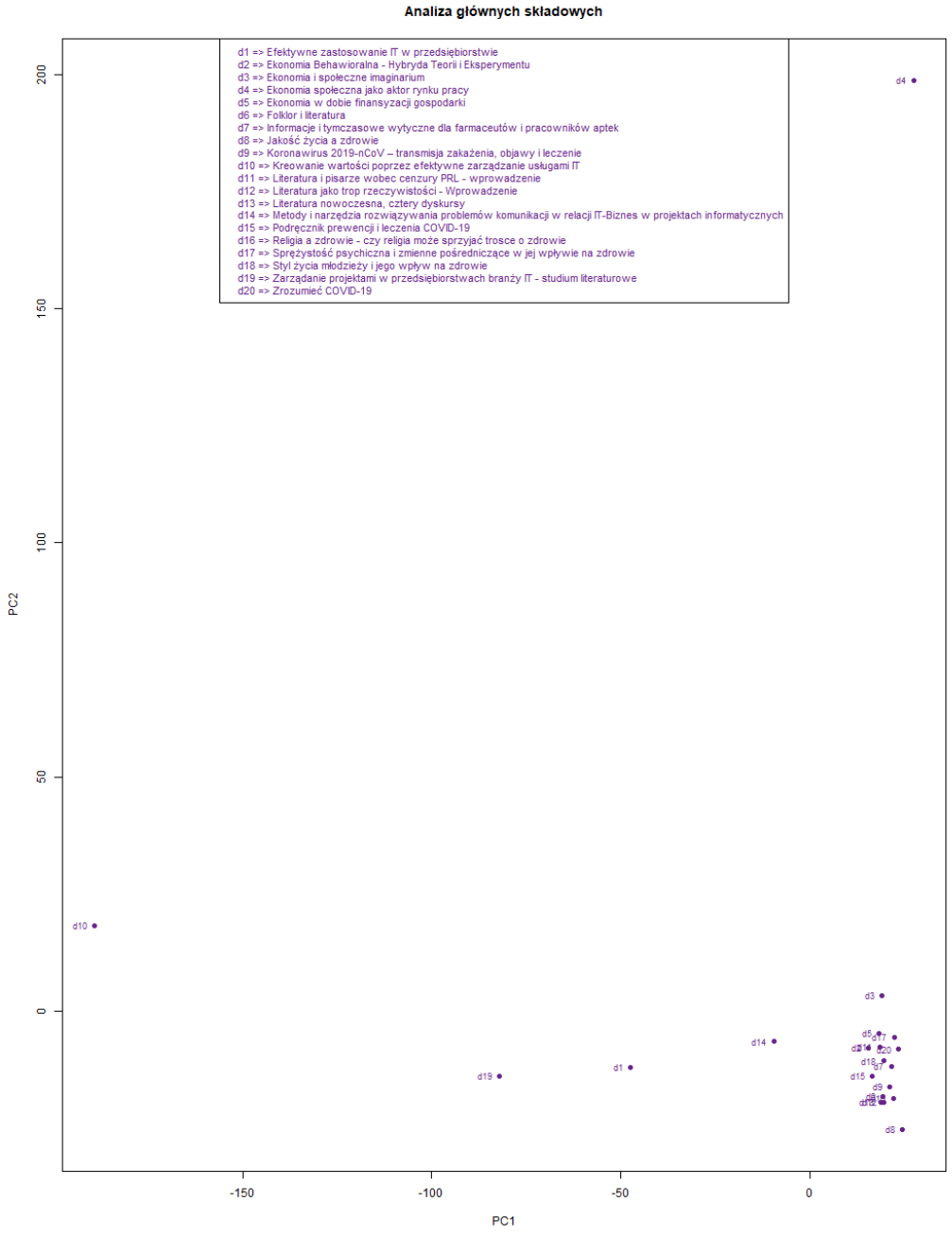
2. Analiza głównych składowych

Analiza głównych składowych została przeprowadzona na macierzach DTM pod kątem porównania dwóch zmiennych:

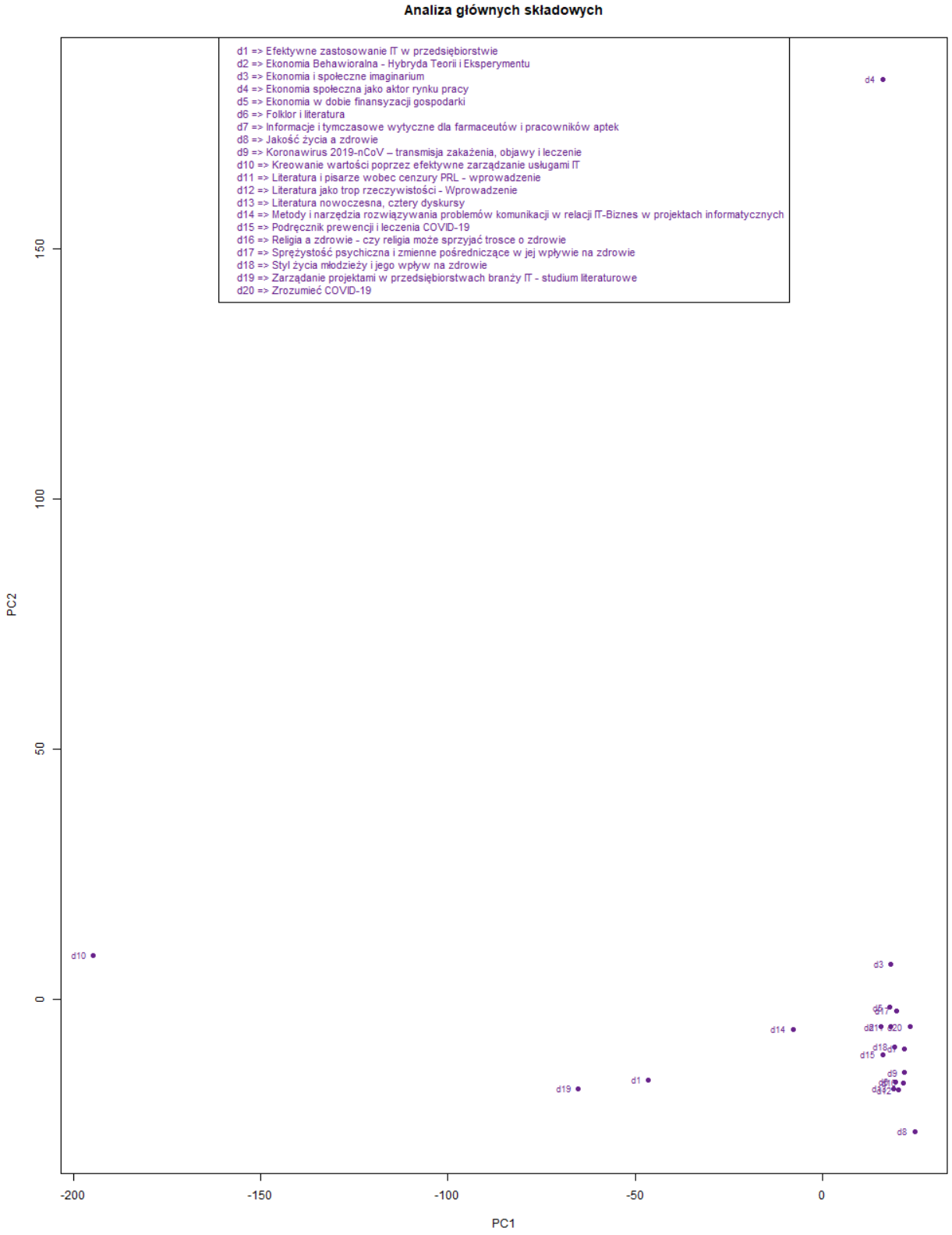
- Metody ważenia, Tf oraz TfIdf
- Ograniczeń w występowanych dokumentach

Analiza głównych składowych macierzy częstości DTM z wykorzystanie wagi Tf:

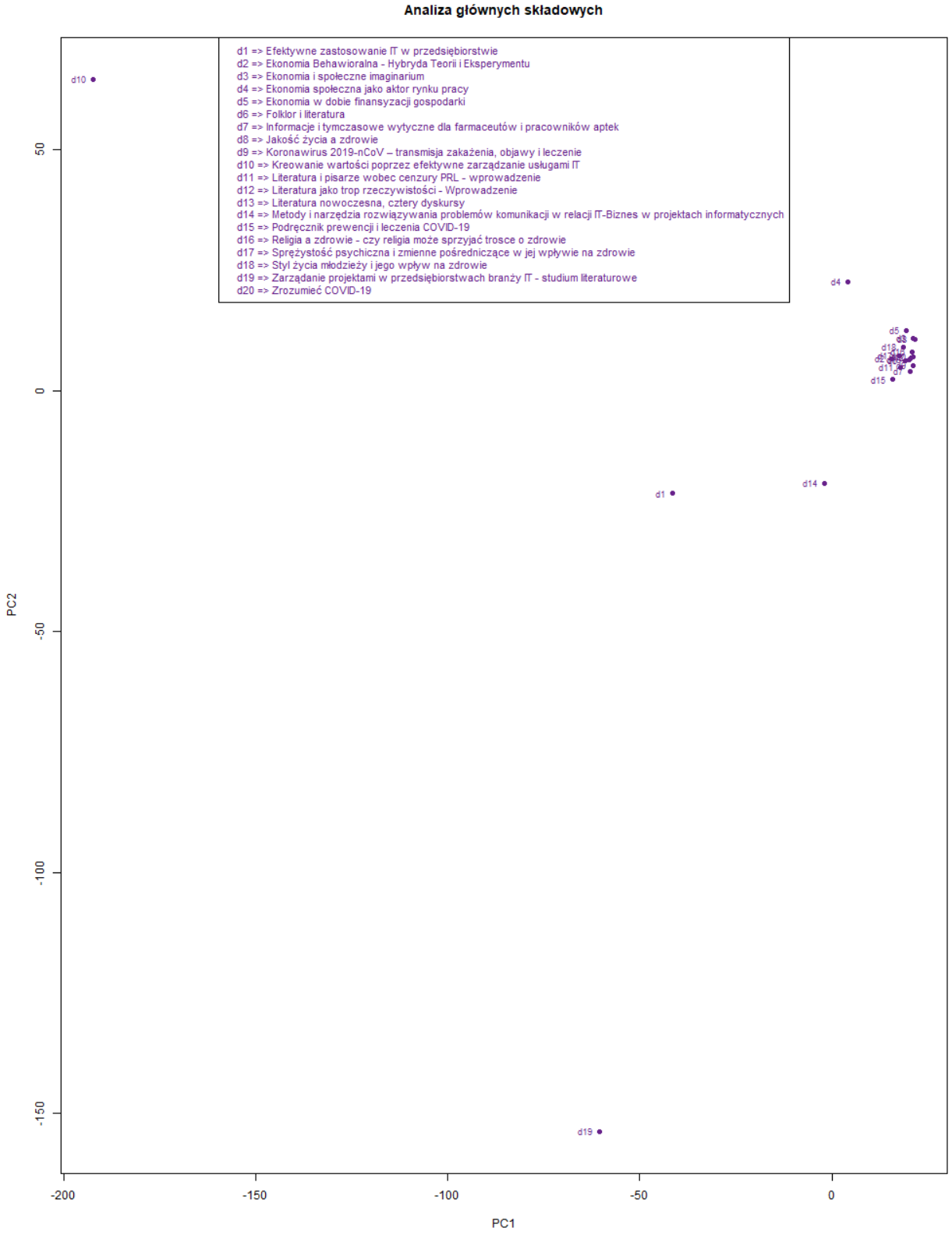
1. Bez granic



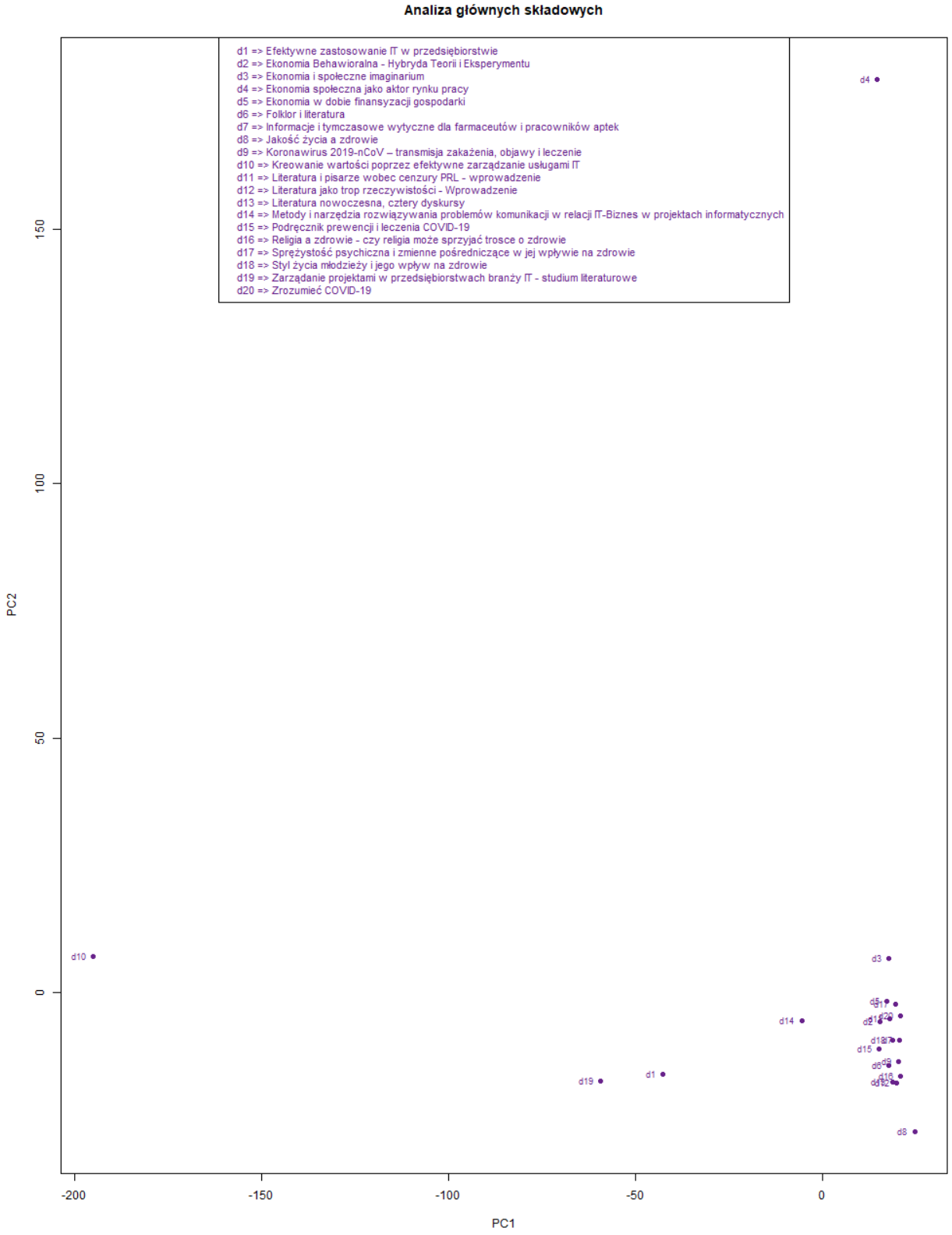
2. Granice między 2 a 18 dokumentów



3. Granice między 3 a 14 dokumentów



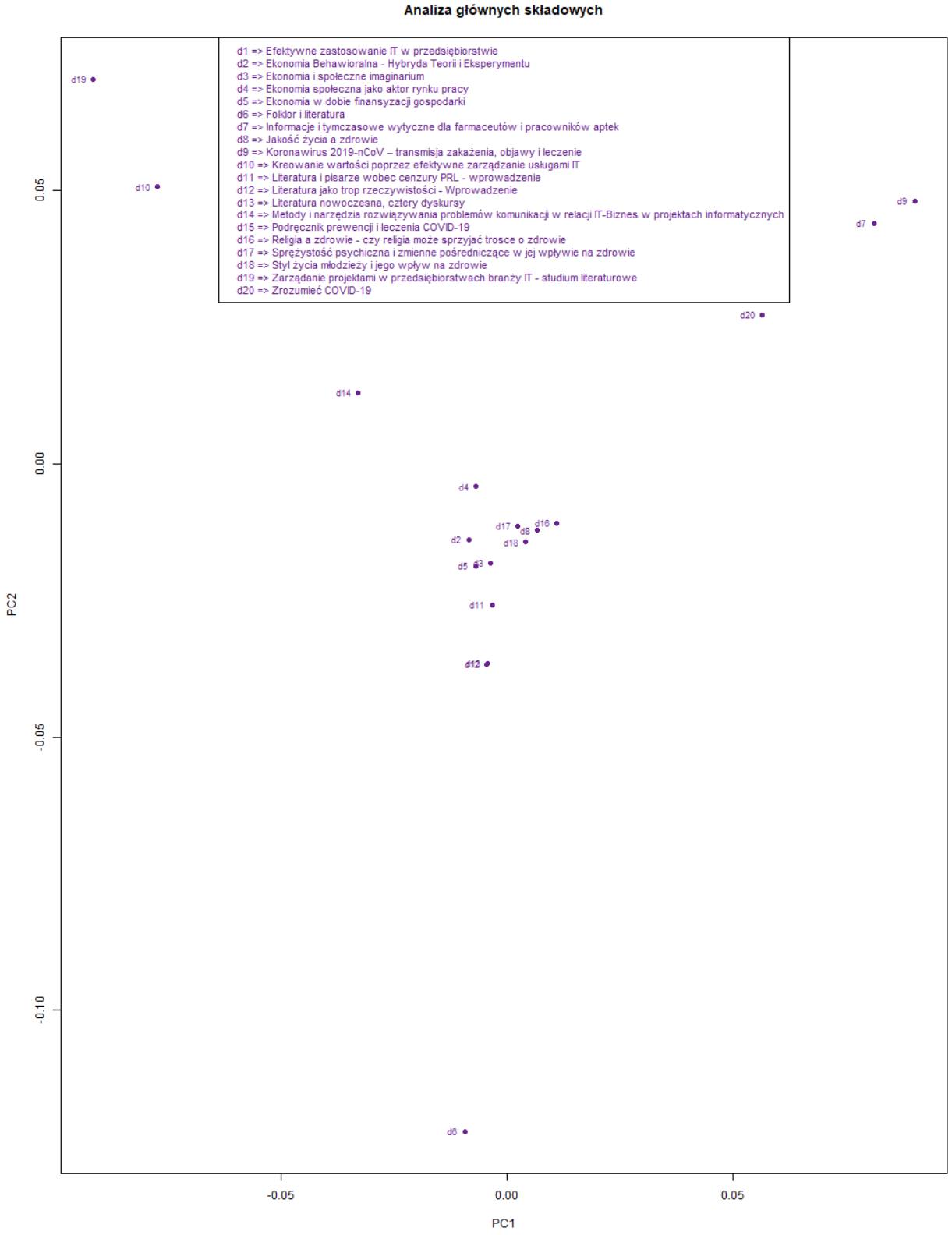
4. Granice między 4 a 20 dokumentów



Analiza głównych składowych macierzy częstości DTM z wykorzystanie wagi TfIdf:

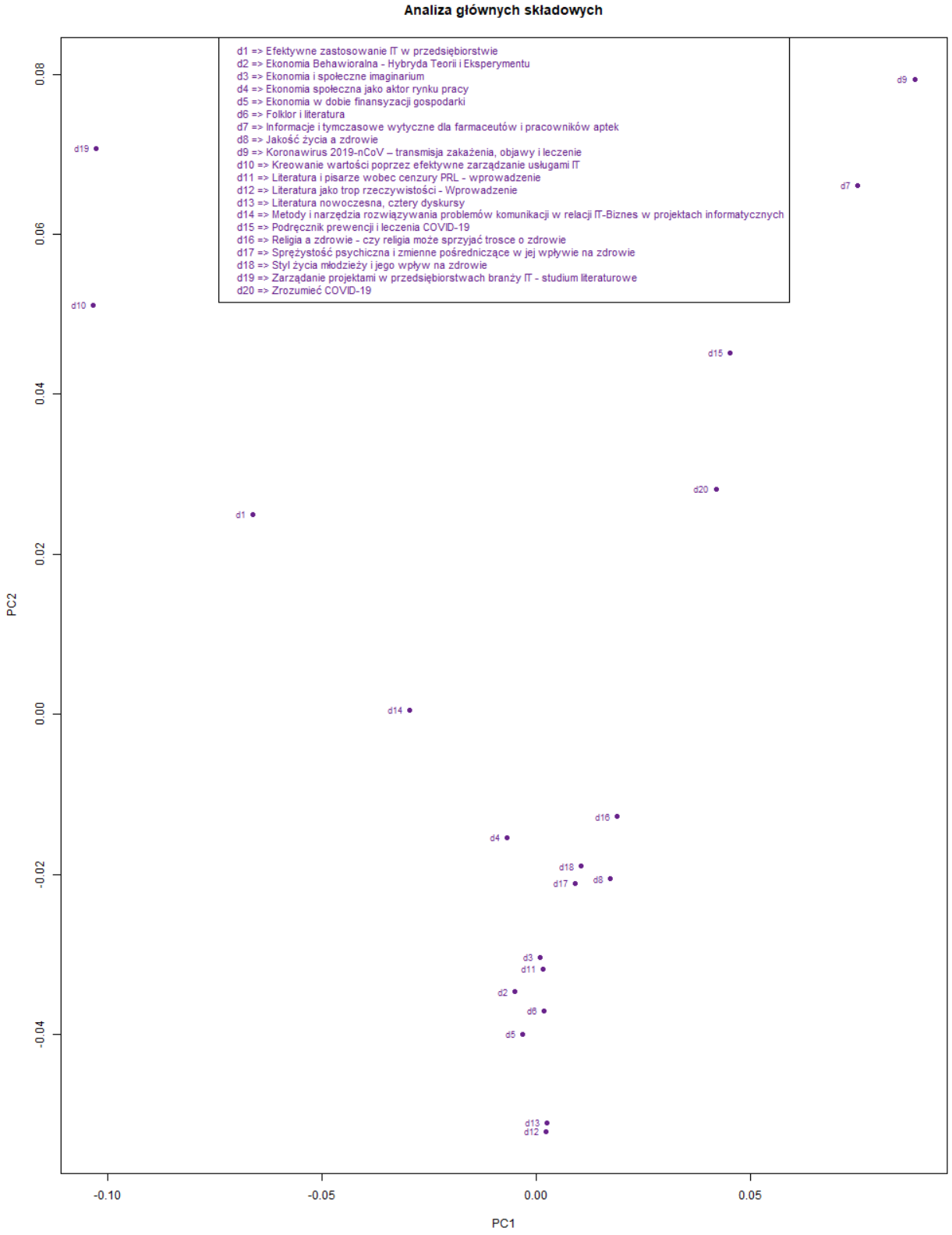
1. Bez granic

2. Granice między 2 a 18 dokumentów



3. Granice między 3 a 14 dokumentów

4. Granice między 4 a 20 dokumentów



Analiza głównych składowych dla macierzy z wagą Tf niewiele mówi nam o zróżnicowaniu dokumentów. Niezależnie od ograniczeń nałożonych na takie macierze, skupienie dokumentów jest bardzo duże, jedynym tematem odstającym od głównych skupień jest temat IT (d1, d10, d14, d19) oraz wyjątkowy przypadek dokumentu d4 z kategorii ekonomia, który na każdym wykresie znacznie odstaje od reszty dokumentów.

Inaczej sytuacja ma się w przypadku rozważania macierzy z wagą TfIdf, dla której nałożenie ograniczeń na dokumenty przynosi znaczną poprawę przejrzystości wykresu. W przypadku braku takich ograniczeń, dokumenty wciąż skupione są w jednym miejscu, z wyjątkiem dokumentu d15 na temat prewencji przeciwko COVID, oraz d17 na temat wpływu zdrowia psychicznego na ogólne zdrowie.

Po nałożeniu ograniczeń możemy znacznie jaśniej określić skupiska dokumentów podobnych do siebie, najlepszym przykładem jest tutaj para liczb 4, 20. Po lewej stronie wykresu możemy wyróżnić tematy IT wcześniej wspomniane, na prawą stronę wyróżniamy zaś tematy związane z koronawirusem oraz zdrowiem, co zapewne spowodowane jest skorelowaniem tych dwóch tematów. Na dole wykresu pozostają tematy o tematyce literatury oraz ekonomii, z wyjątkiem wspomnianego wcześniej przykładu d4, odstającego od głównego skupiska. Podobieństwo tematyki literatury i ekonomii jest tutaj dość ciekawe, co zostanie poddane dalszej analizie w kolejnych podpunktach.

3. Dekompozycja według wartości osobliwych

Dekompozycja według wartości osobliwych została przeprowadzona na macierzach TDM pod kątem porównania dwóch zmiennych:

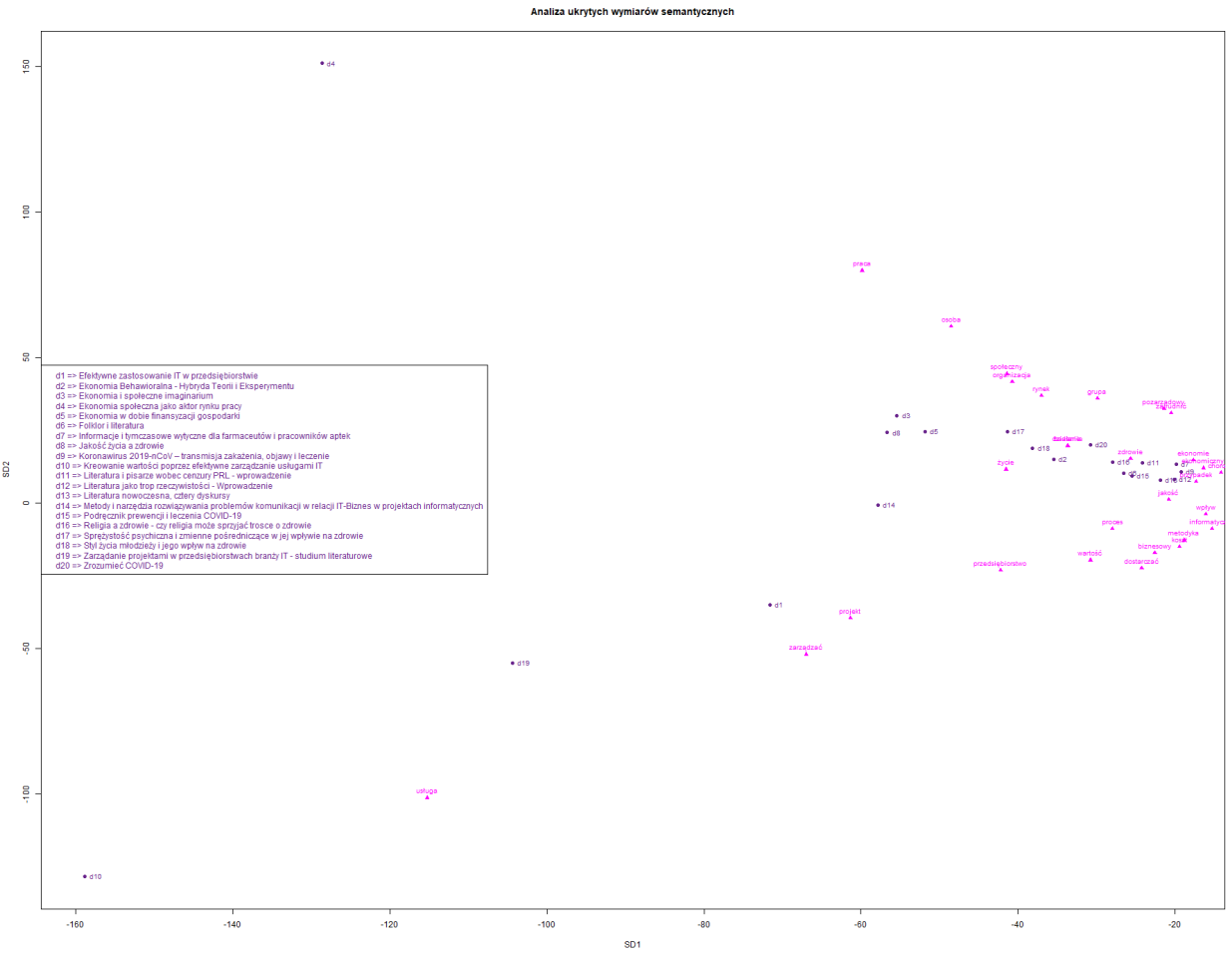
- Metody ważenia, Tf oraz TfIdf
- Ograniczeń w występowanych dokumentach

Przy przeprowadzania dekompozycji według wartości osobliwych następujące słowa zostały użyte w zmiennej opisujące własne warunki:

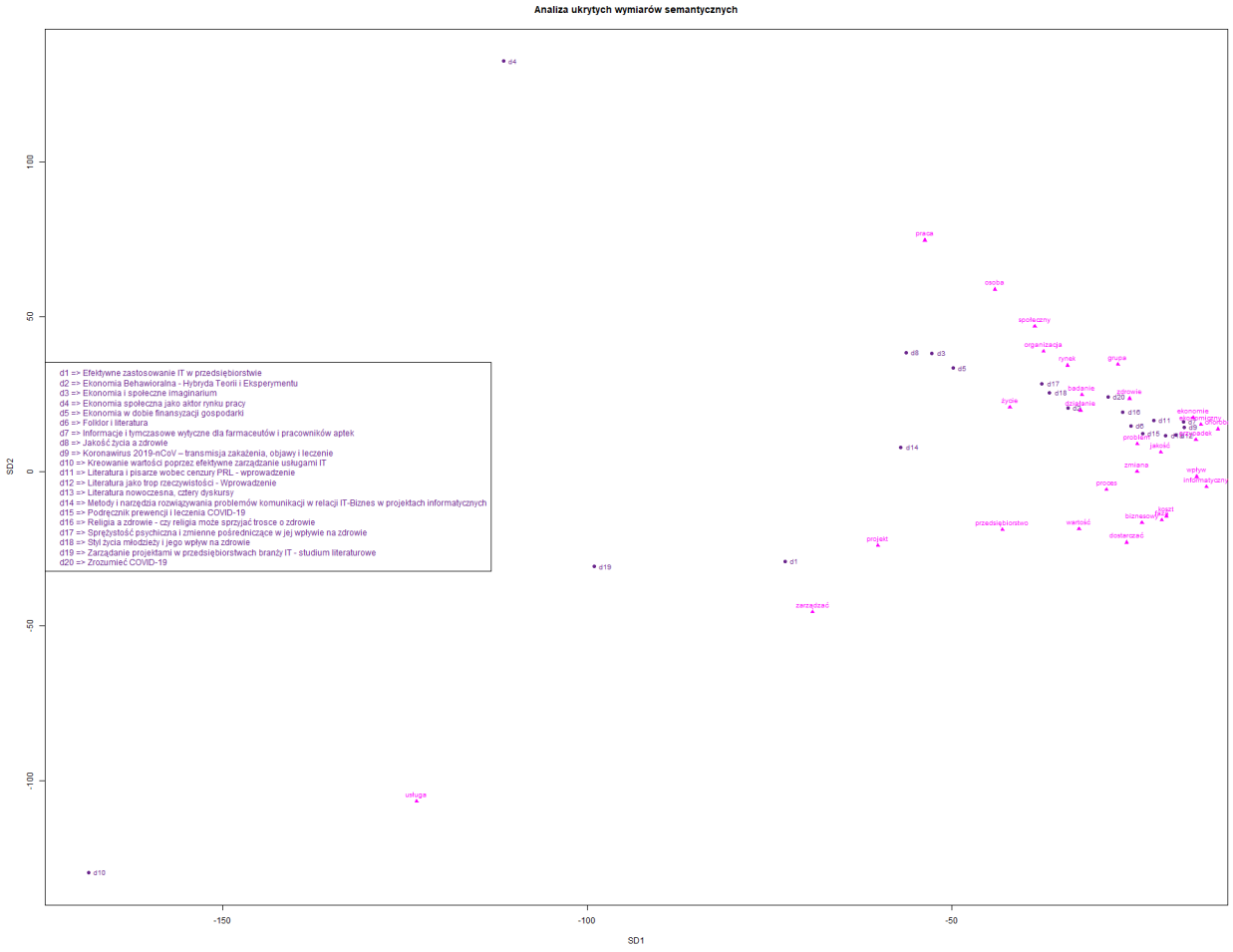
"zdrowie", "covid", "przedsiębiorstwo", "oprogramowanie", "zakażenie", "pacjent", "ekonomia", "literatura", "biznes", "zarządzać"”

Analiza głównych składowych macierzy częstości TDM z wykorzystanie wagi Tf:

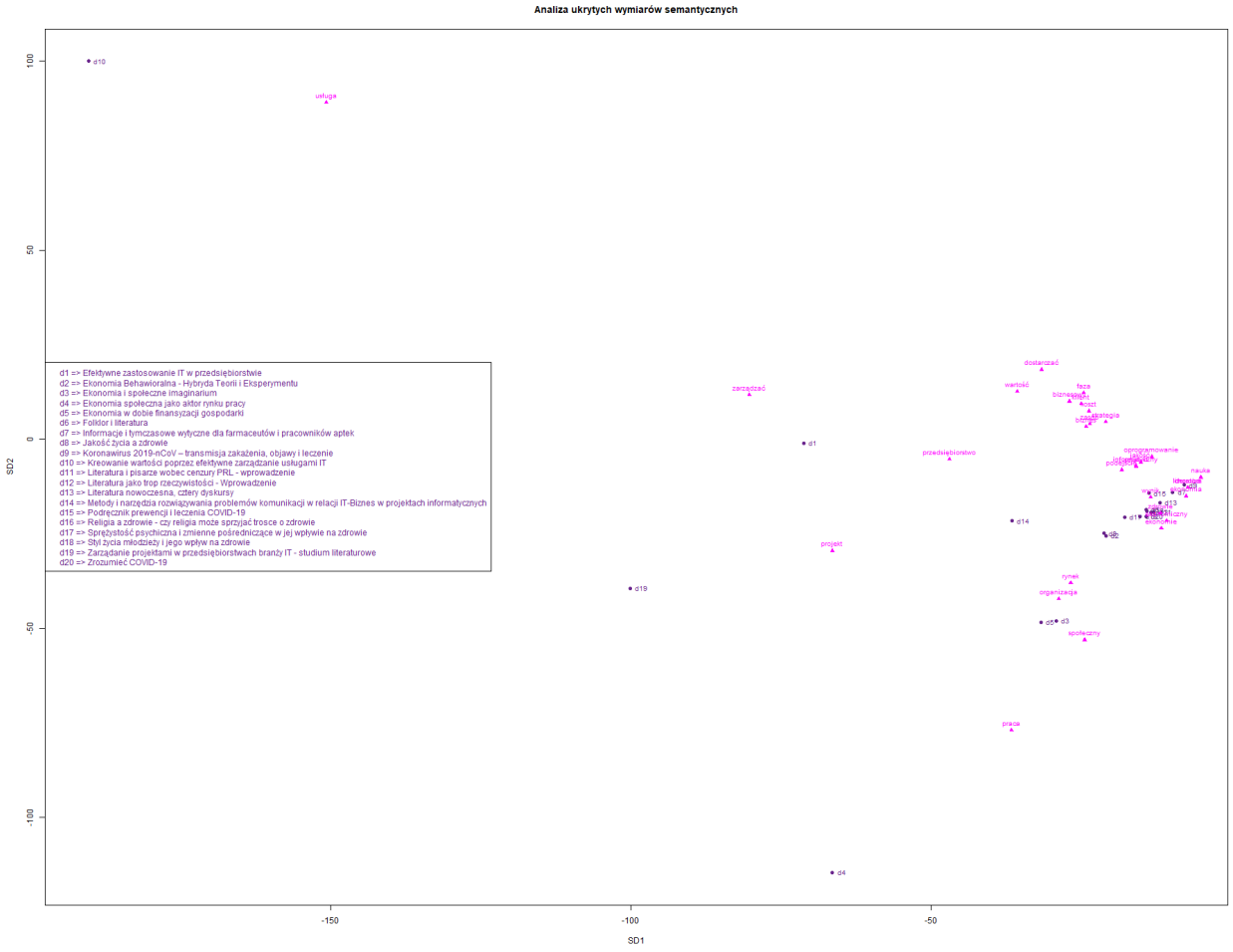
1. Bez granic



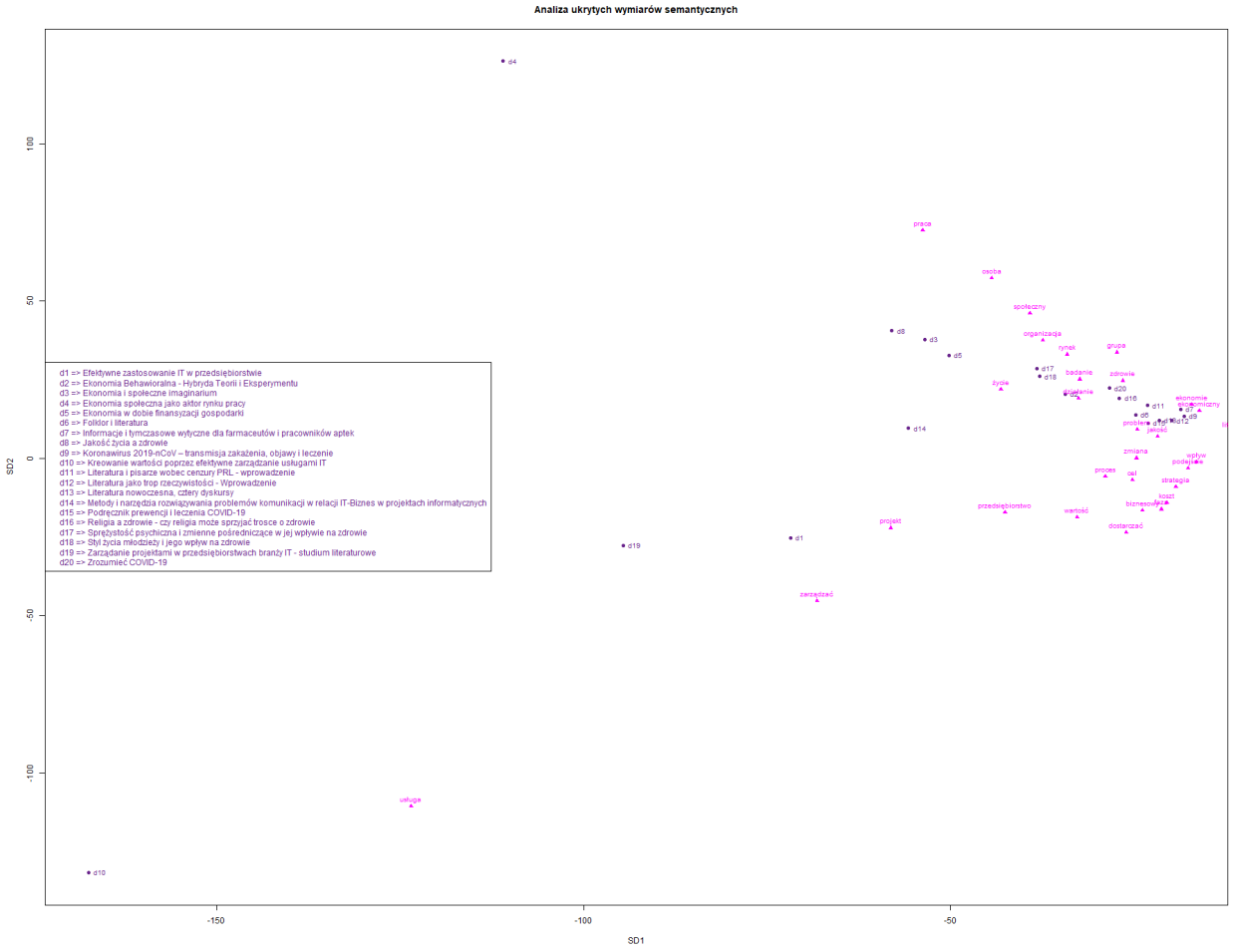
2. Granice między 2 a 18 dokumentów



3. Granice między 3 a 14 dokumentów

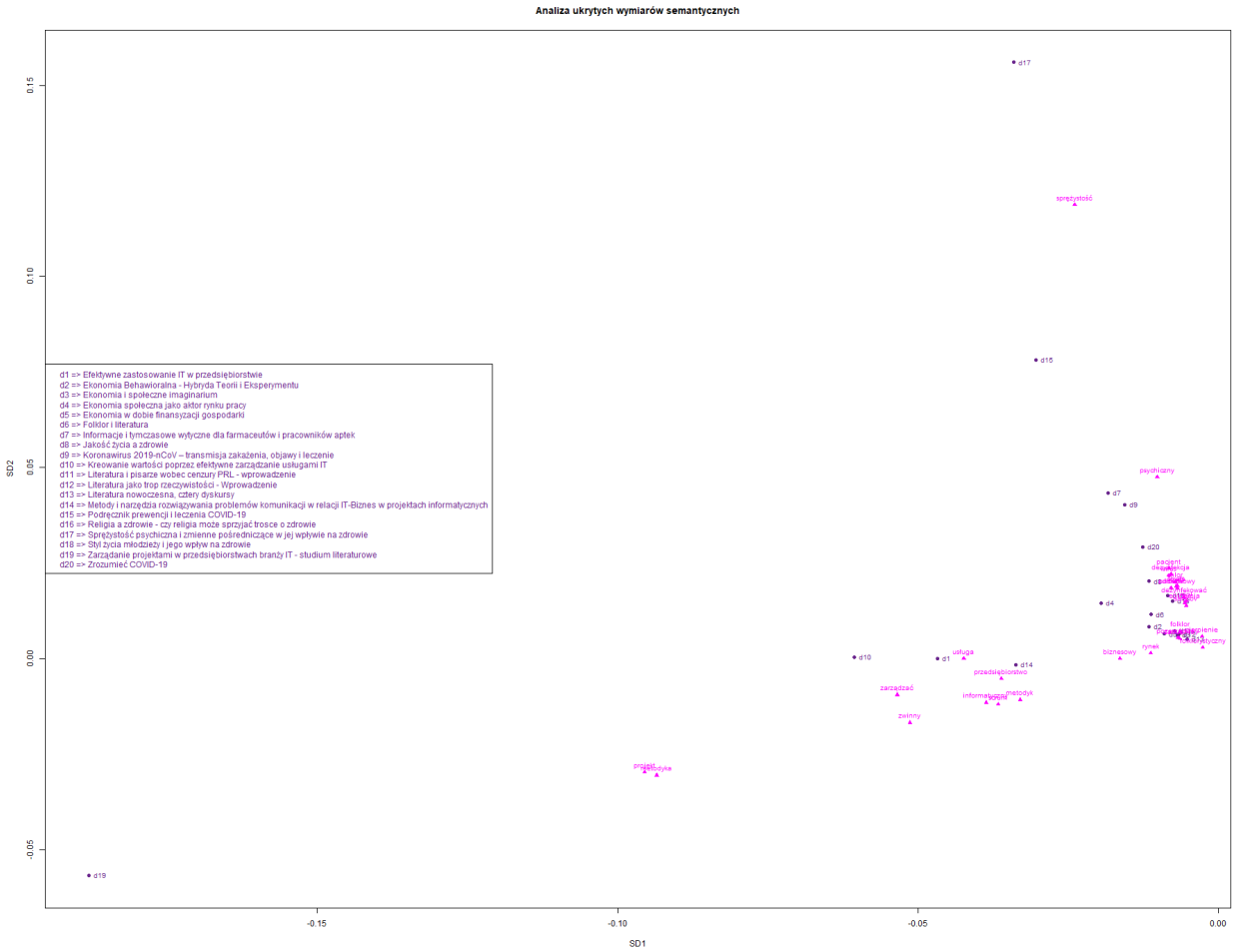


4. Granice między 4 a 20 dokumentów

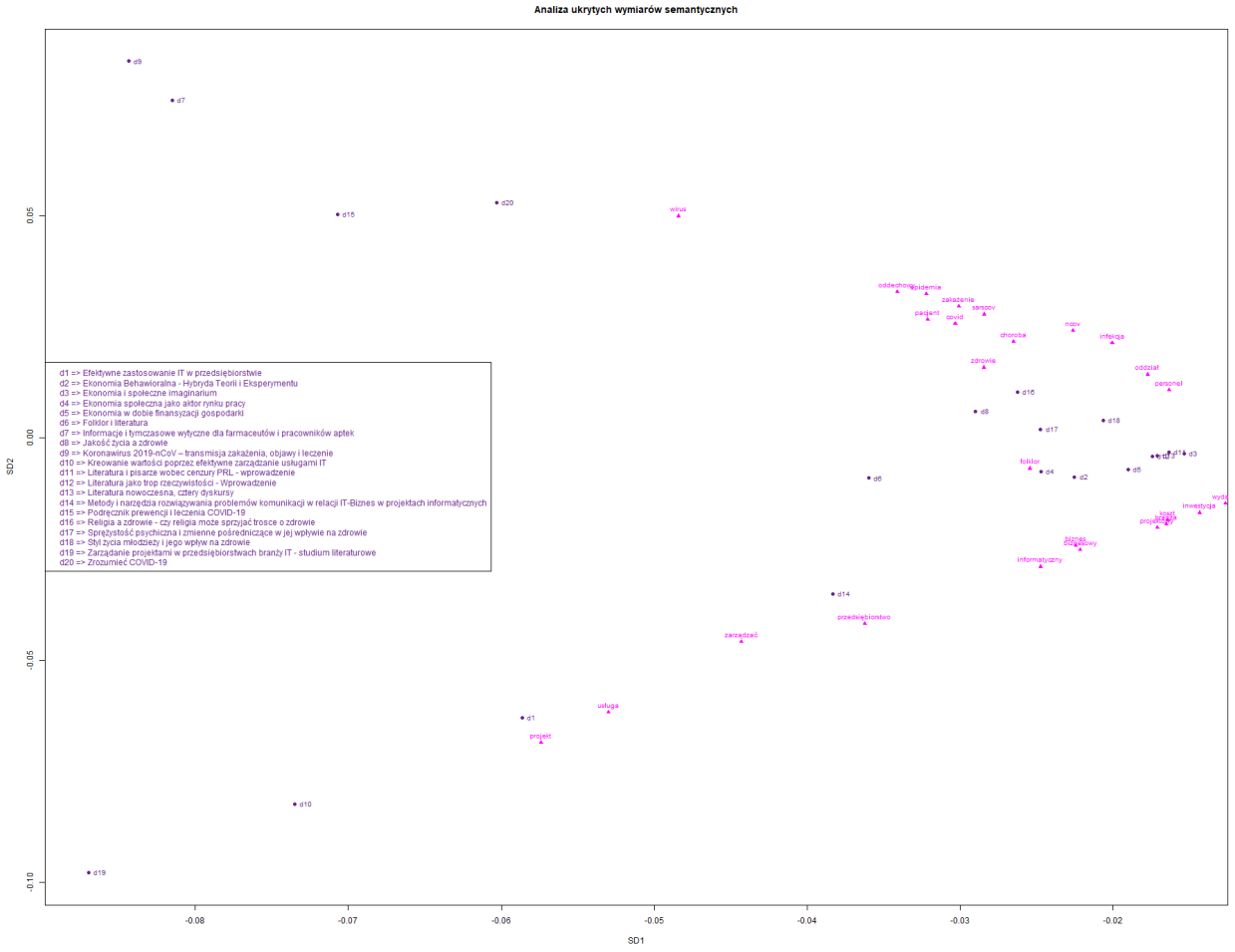


Analiza głównych składowych macierzy częstości TDM z wykorzystanie wagi TfIdf:

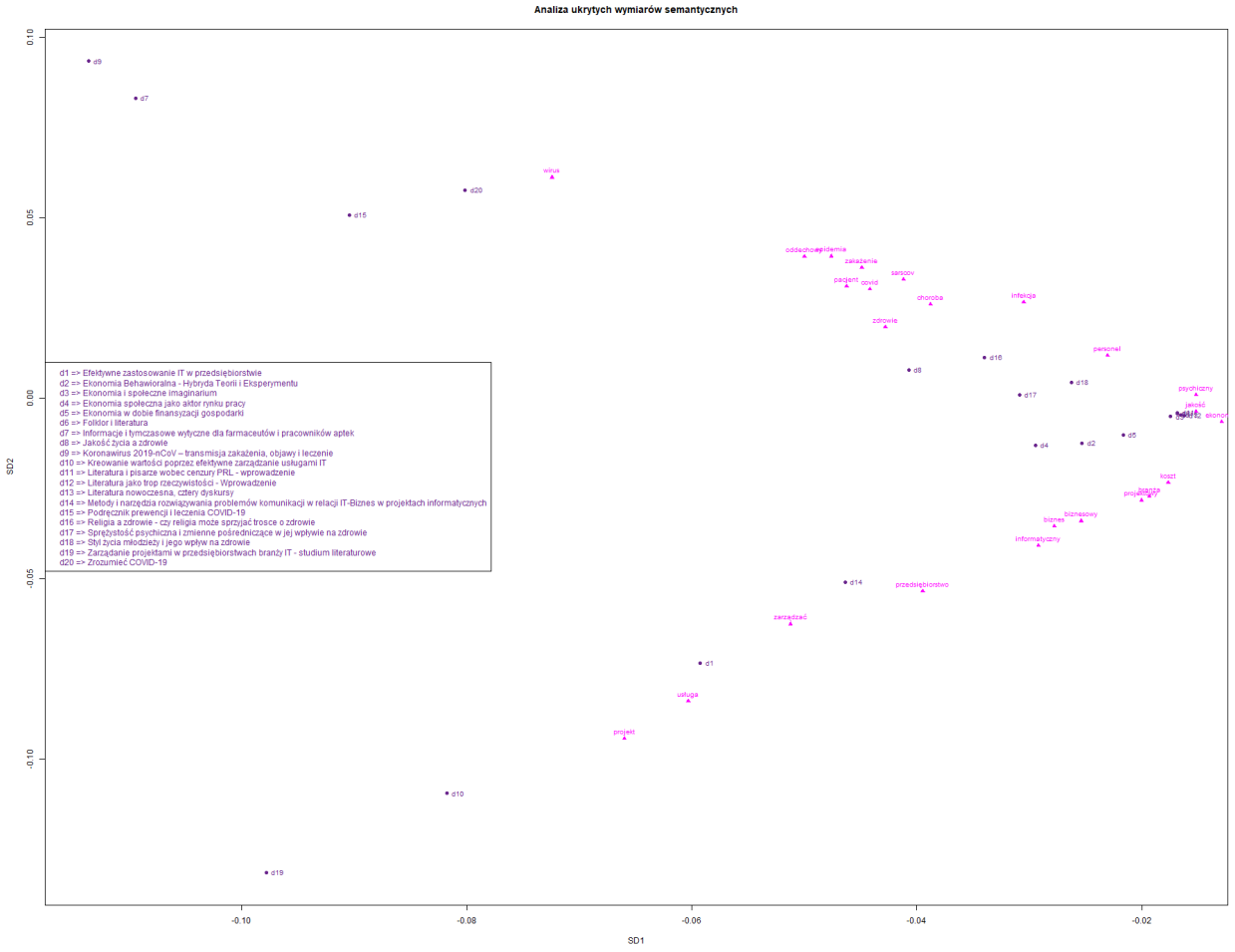
1. Bez granic



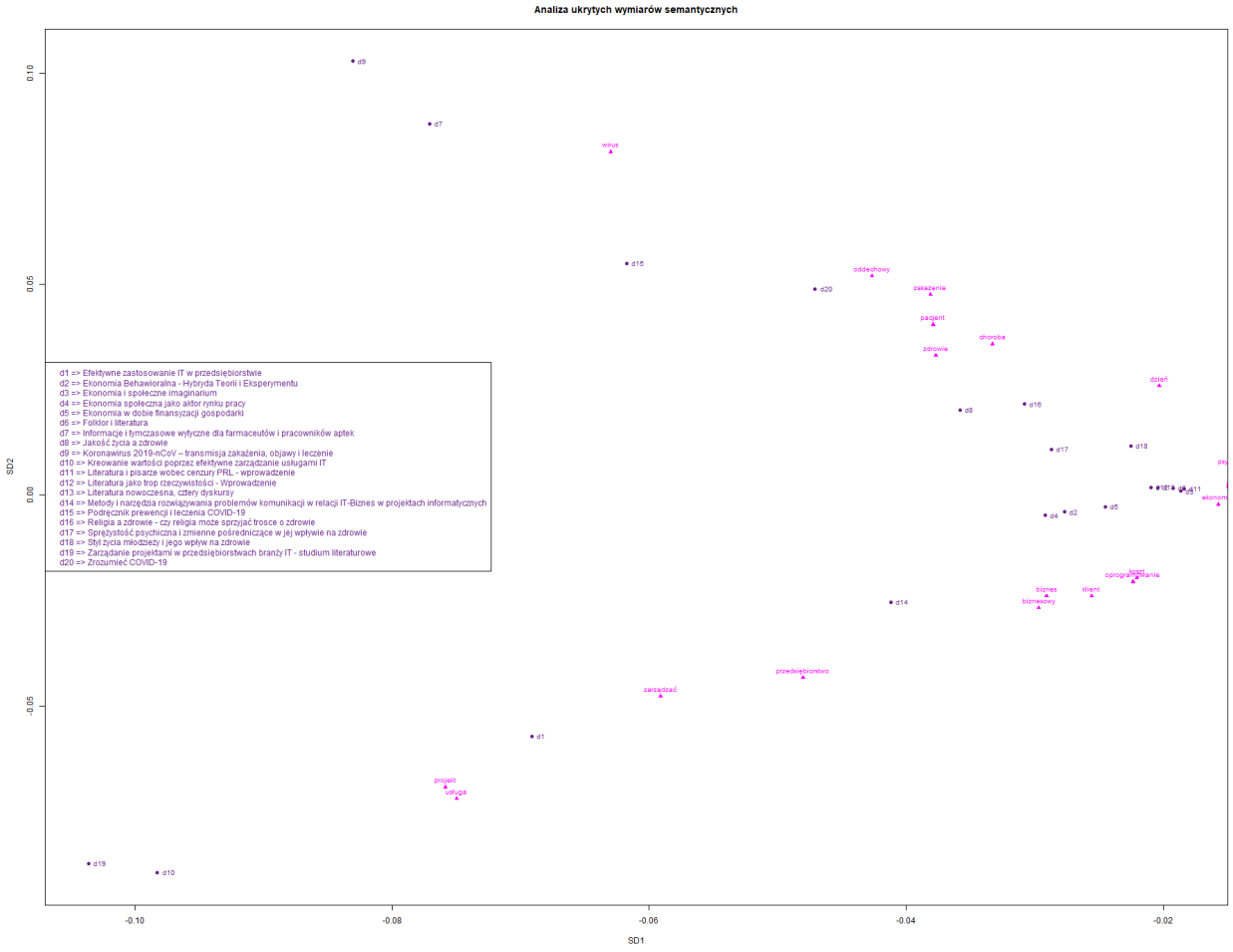
2. Granice między 2 a 18 dokumentów



3. Granice między 3 a 14 dokumentów



4. Granice między 4 a 20 dokumentów



Dekompozycja według wartości osobliwych dla macierzy z wagą Tf ma się podobnie jak w przypadku metody PCA, niezależnie od nałożonych ograniczeń skupienie zarówno dokumentów jak i wartości osobliwych kumuluje się praktycznie w jednym punkcie i nawet przy zwiększeniu rozdzielczości wykresu bardzo ciężko odczytać z niego zróżnicowanie dokumentów oraz słów. Na wykresie można zauważyć także wyróżniające się dokumenty z tematyki IT (d1, d10, d14,d19) oraz jeden wyjątek którym jest także dokument z dziedziny ekonomii - d4.

Macierz z wagą TfIdf ale bez ograniczeń także posiada znaczne skupienie tematów w jednym miejscu, gdy jednak nałożymy dodatkowe ograniczenia, rozkład dokumentów oraz wartości osobliwych staje się znacznie czytelniejszy, co tak samo jak w przypadku metody PCA najlepiej widać dla pary liczb 4, 20.

W tym przypadku możemy jasno rozgraniczyć dokumenty, które kierują się w stronę lewej dolnej krawędzi wykresu. Są one związane z tematyką IT a słowom z nimi powiązane to przede wszystkim projekt oraz usługa, choć można tutaj zauważyć te wpływ słów kojarzących się z ekonomią takich jak np zarządzać czy przedsiębiorstwo. Na górze wykresu wyróżniają się tematy związane z COVID-19 oraz zdrowiem, co wynika z ich wzajemnego powiązania. Wyróżniające się słowa zdecydowanie oscylują wokół tych tematyk. Znajdziemy tutaj więc zarówno słowa takie jak wirus oraz oddechowy związany z obecną pandemią ale także zakażenie, pacjent, choroba czy zdrowie, które blisko związane są z obiema tematykami.

Po prawej stronie wykresu pozostają tematy związane z literaturą oraz ekonomią. Choć w tym wypadku możemy wyraźniej odróżnić ich skupienie (temat literatury jest tutaj niemal w jednym punkcie), to ciekawym jest to że tematyki te są ze sobą tak powiązane pomimo występowania tutaj taki słów jak ekonomiczny czy psychologia.

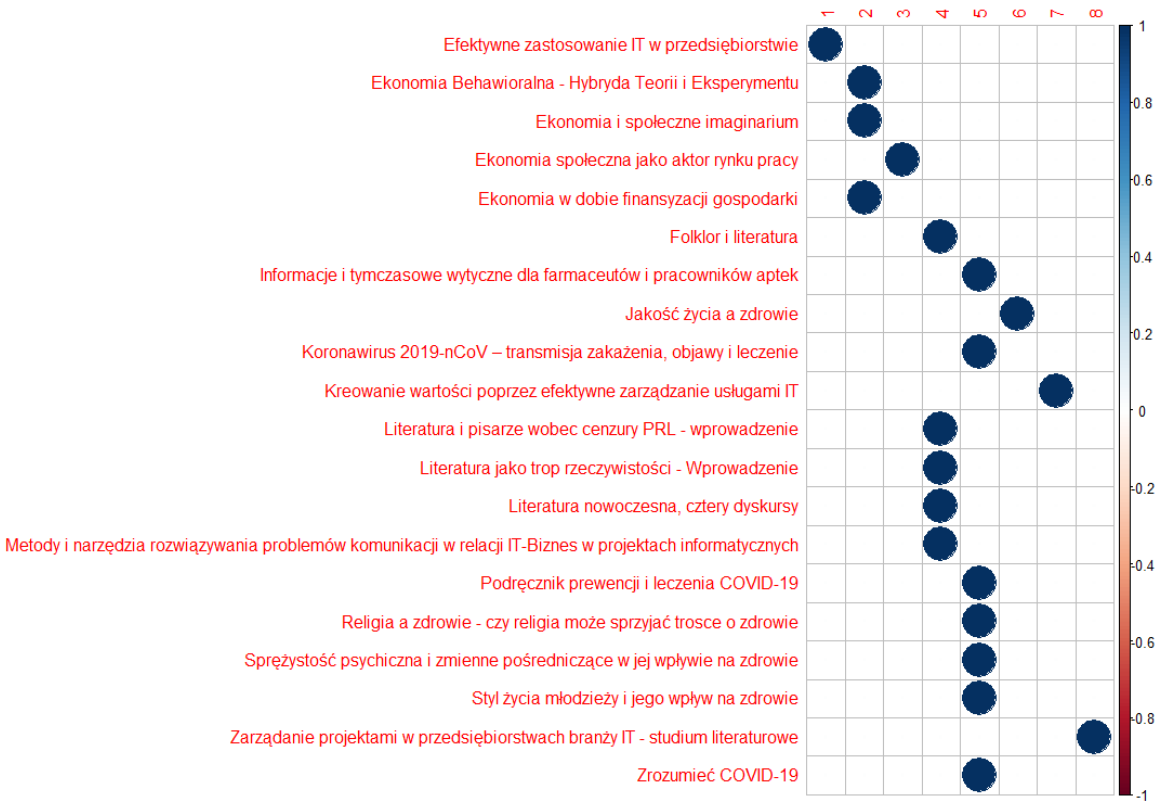
4. Analiza skupień

Eksperymenty do analizy skupień zostały wykonane na wybranych macierzach częstości DTM, zarówno z wagą Tf jak i TfIdf, oraz z różnymi granicami. Zaprezentowane zostaną wynikowe dendrogramy a przy porównaniu macierzy, Indeks Fawlkes'a Mallows'a.

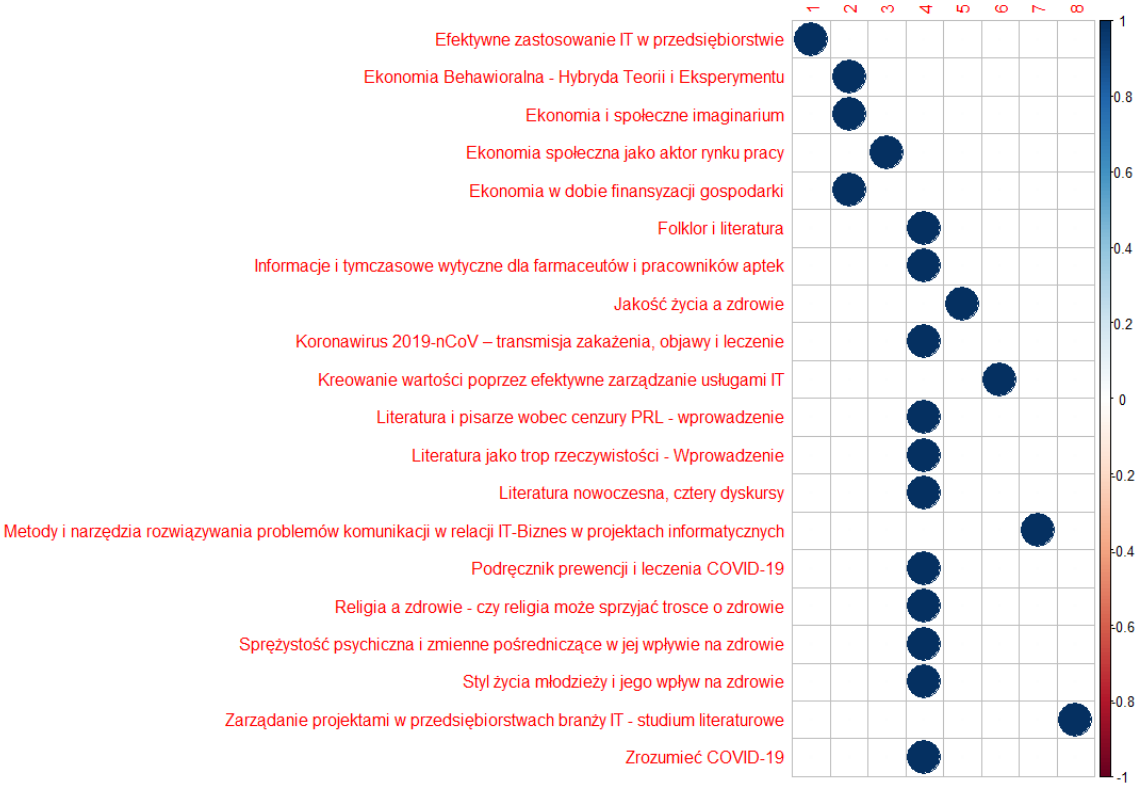
Przy porównywaniu macierzy metoda wyznaczania odległości została przyjęta jako “euclidean” a sposób wyznaczania odległości jako “single”. Liczba skupień została ustalona na stałą liczbę wynoszącą 8.

Dla jednej z wybranych macierzy zostały również poddane eksperymentom wyżej wymienione zmienne.

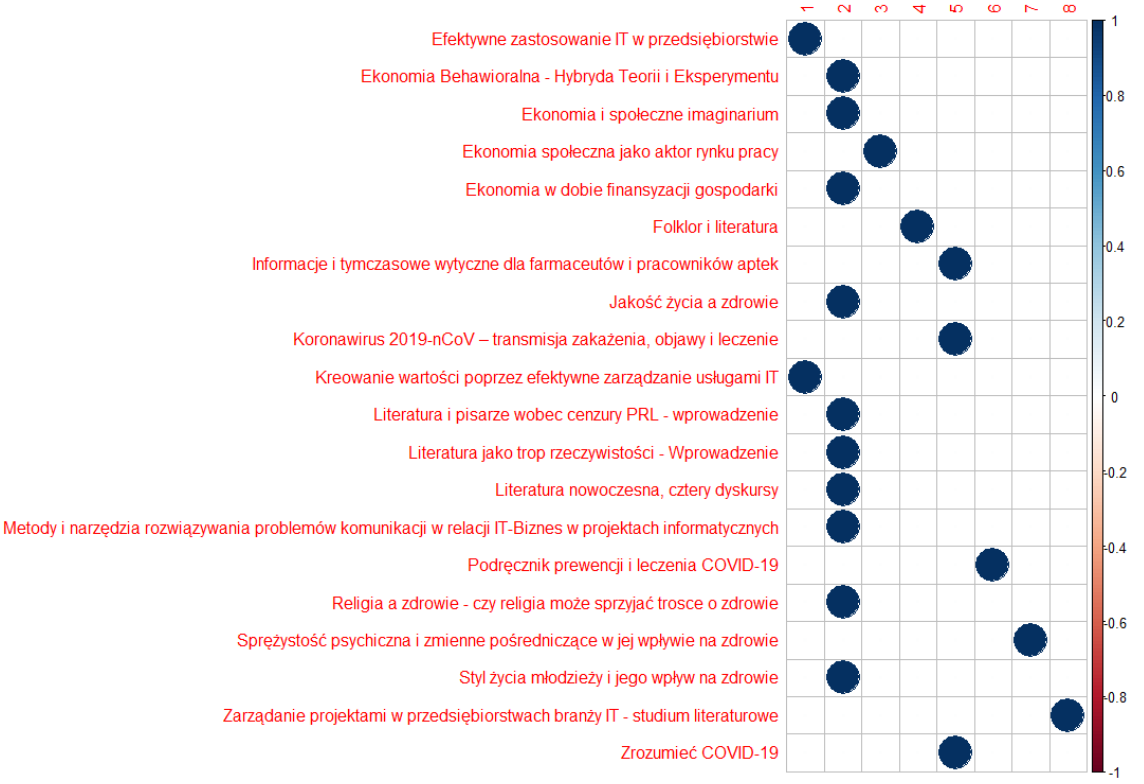
1. Macierz częstości z wagą Tf oraz bez ograniczeń



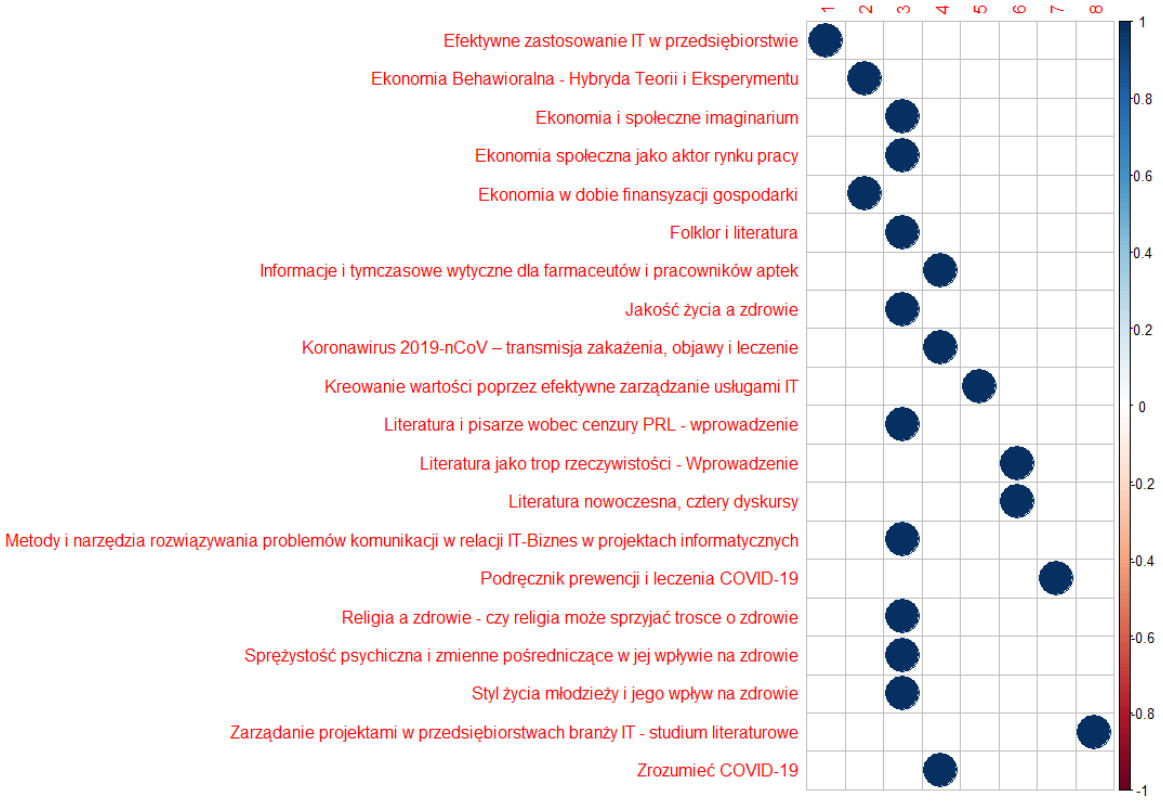
2. Macierz częstości z wagą Tf oraz parą ograniczeń 4-20



3. Macierz częstości z wagą TfIdf oraz bez ograniczeń



4. Macierz częstości z wagą TfIdf oraz parą ograniczeń 4-20



Macierz z ważeniem Tf bez ograniczeń, bardzo dobrze grupuje skupienia literatury pod jedną kategorię (z wyjątkiem jednego dokumentu z tematyki IT), a także skupienia dotyczące podobnych tematyk jakim jest zdrowie oraz COVID. Także tematyka ekonomii leży w podobnym skupieniu, najbardziej zróżnicowany jest temat IT, każdy dokument posiada własne oznaczenie skupienia.

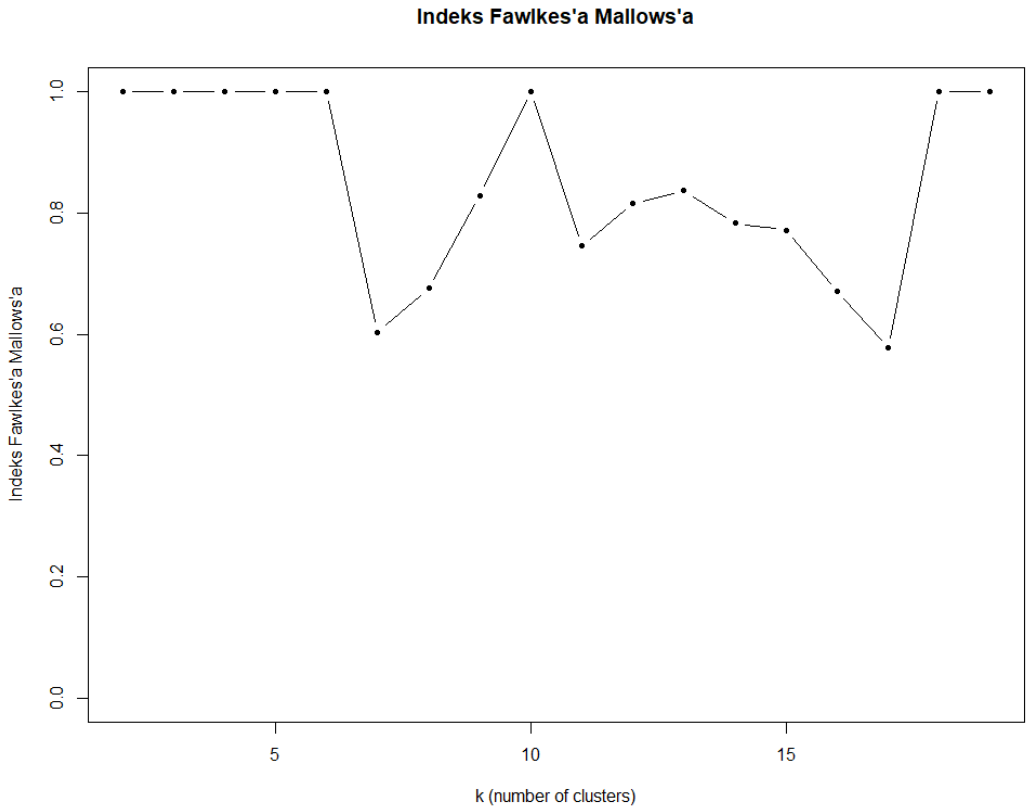
Dla macierzy z ważeniem TF z ograniczeniami 4-20 można zauważyć, że skupienia dla literatury jaki i zdrowia mocno zbiegają się, tematy ekonomiczne posiadają własne skupienie, najbardziej zróżnicowany jest temat IT, każdy dokument posiada własne oznaczenie skupienia.

Dla macierzy z ważeniem TfIdf bez ograniczeń można wyróżnić jedno silne skupienie pod numerem 2 do którego można zakwalifikować zarówno tematy z ekonomii, zdrowia jak i literatury. Pozostałe dokumenty posiadają różne skupienia.

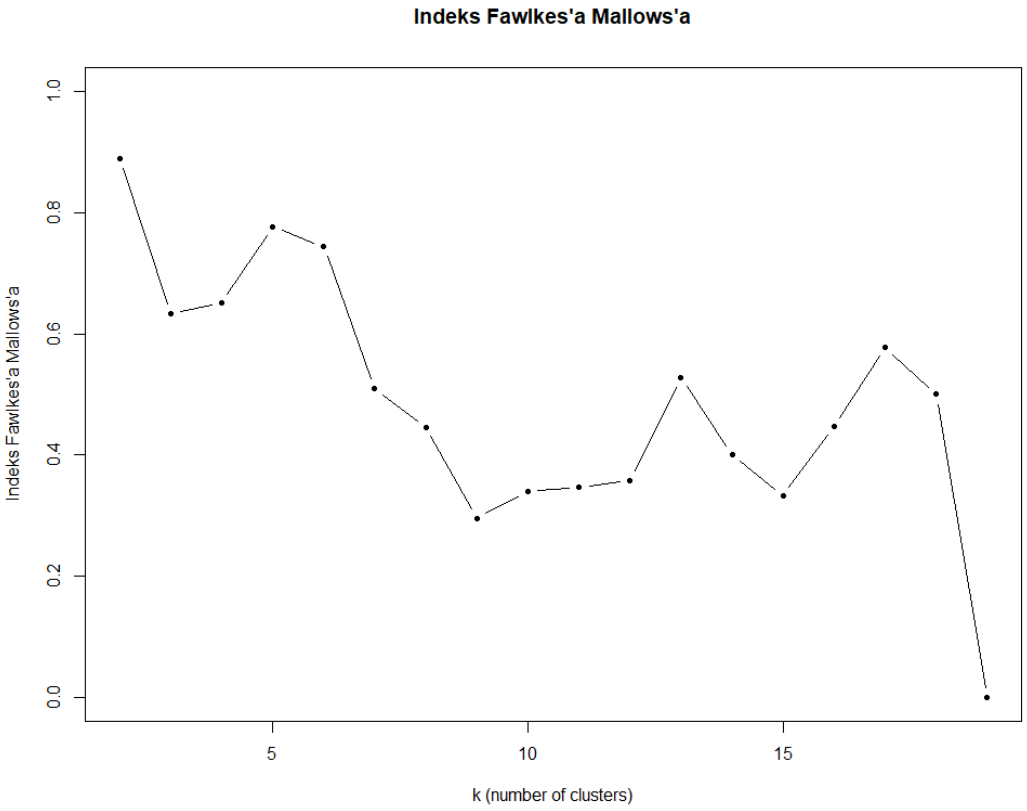
Dla macierzy z ważeniem TfIdf z ograniczeniami 4-20 także można wyróżnić jedno silne skupienie do którego zalicza się co najmniej jeden dokument z każdej kategorii. Pozostałe tematyki posiadają różne skupienia.

Poniżej przedstawione zostanie porównanie dendrogramów przy użyciu Indeks Fawlkes'a Mallows'a:

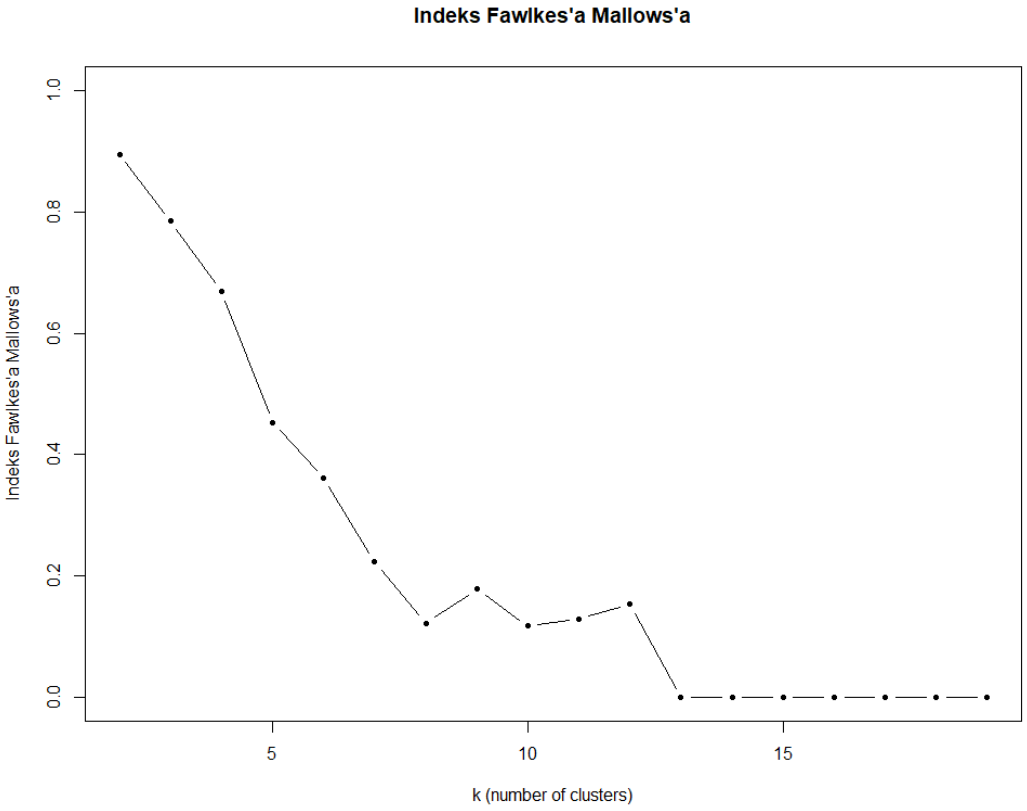
1. Porównanie macierzy z ważeniem Tf (bez ograniczeń kontra ograniczenia 4-20)



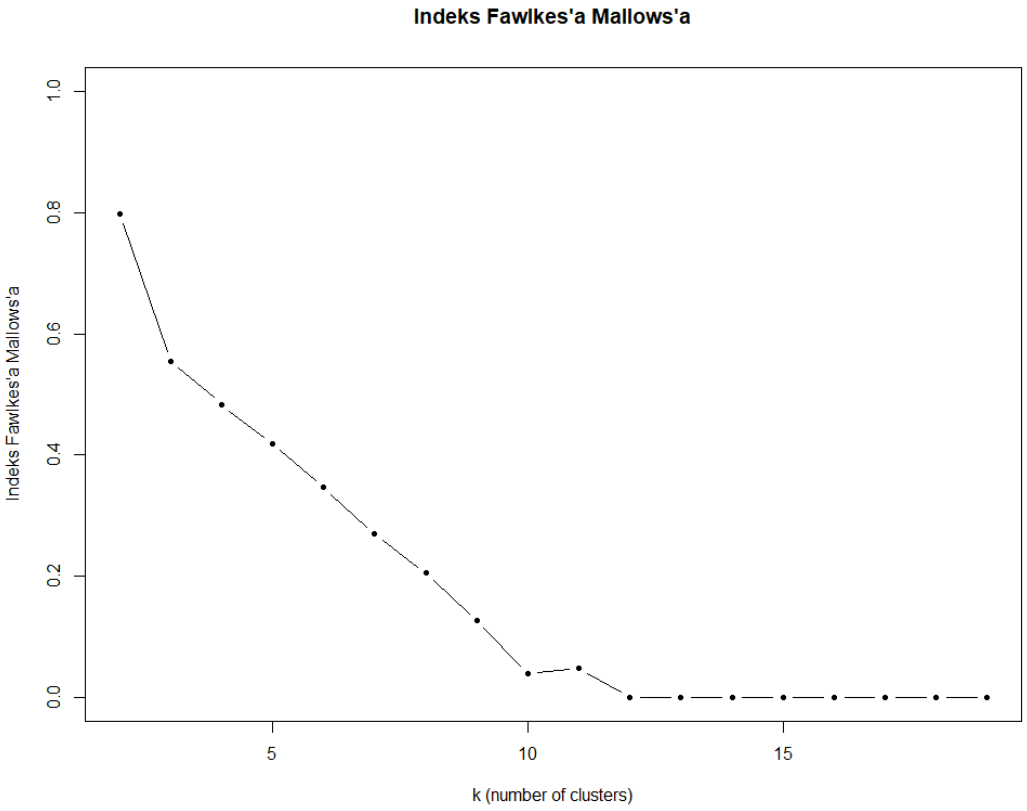
2. Porównanie macierzy z ważeniem TfIdf (bez ograniczeń kontra ograniczenia 4-20)



3. Porównanie macierzy z ważeniem Tf oraz TfIdf bez ograniczeń



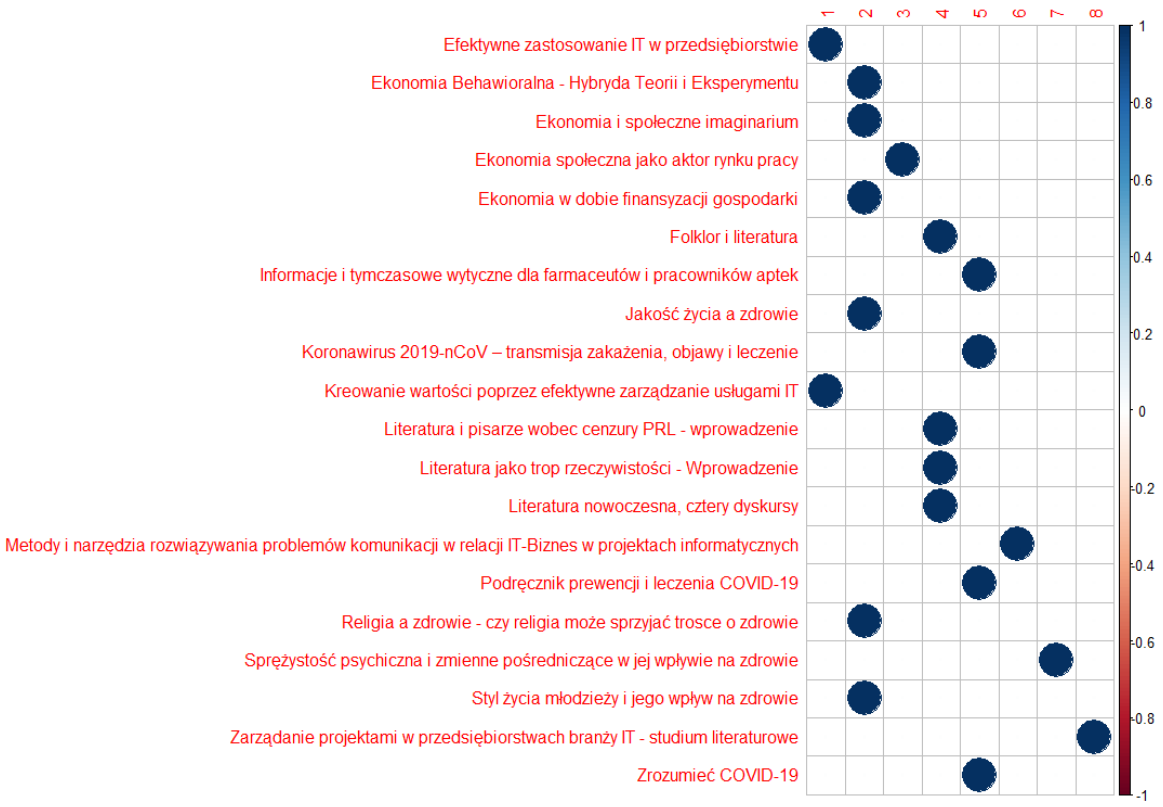
4. Porównanie macierzy z ważeniem Tf oraz TfIdf z ograniczeniami 4-20



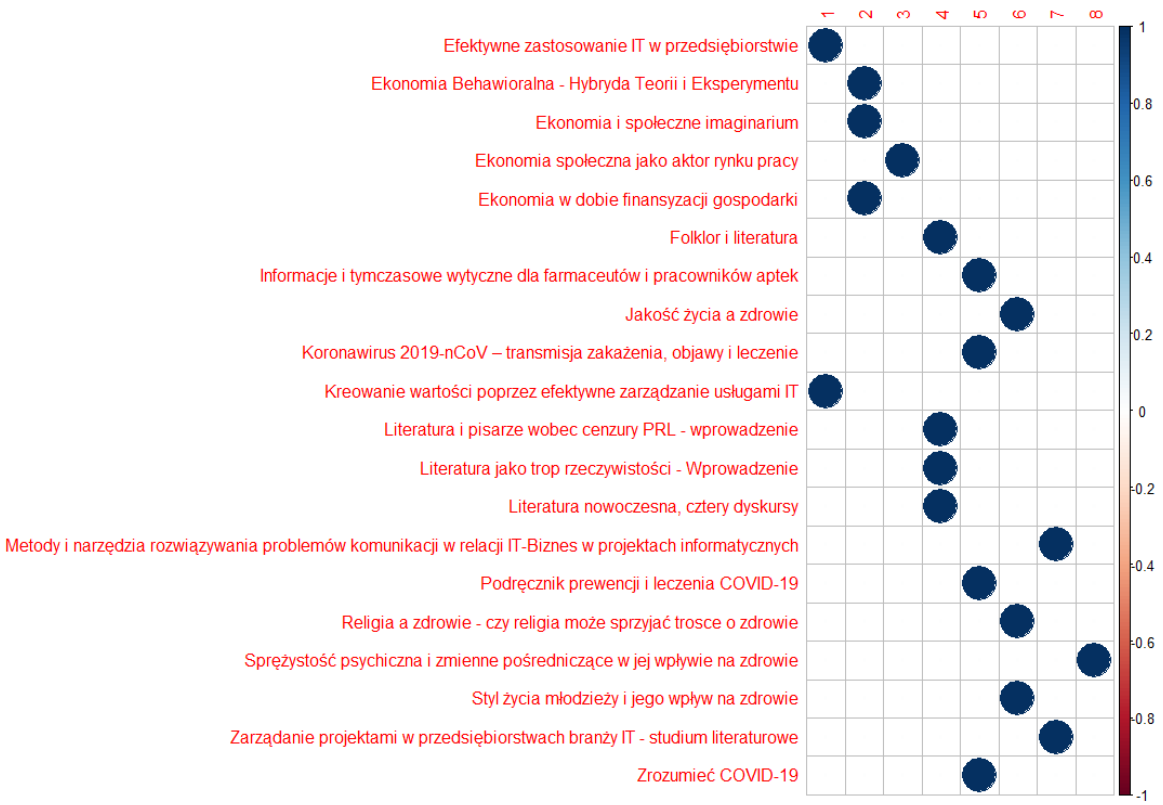
Z porównania można wywnioskować, że analiza skupień pomiędzy macierzami z ważeniem Tf nie różni się zbyt wiele w zależności od ilości skupień, analiza ta jest bardzo podobna do siebie a więc nałożenie ograniczeń może nie mieć większych skutków przy próbie eksperymentów. Zupełnie inaczej sprawa ma się dla macierzy z wagą TfIdf gdzie można zauważyć znaczną różnicę w zależności od ilości skupień.

Porównując macierze z ważeniem Tf oraz TfIdf można zauważyć ogromną różnice wraz ze wzrostem ilości skupień. Zmiana metody ważenia powoduje więc znaczną różnice pomiędzy analizą skupień dla dokumentów.

Macierz z ważeniem Tf bez ograniczeń została poddana kolejnym eksperymentom, na początku zmianie metody wyznaczania odległości. Ustawienie jej na wartość “jaccard” nie przyniosła zmian, ustawienie jej natomiast na cosine przyniosła znacznie lepsze efekty w grupowaniu skupień, przedstawione poniżej



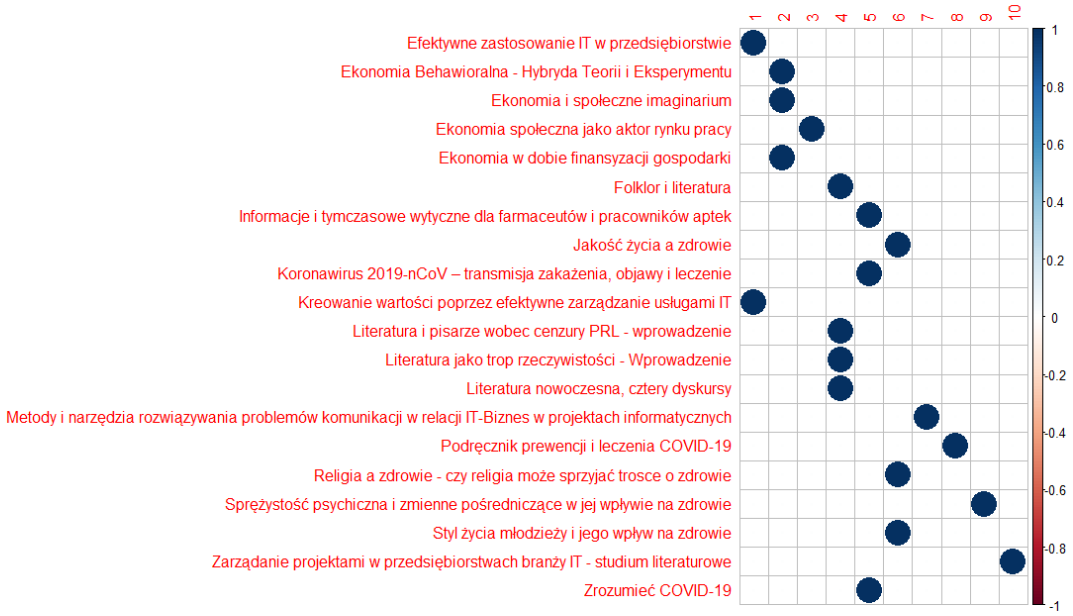
Następnie eksperymentom został poddany sposób wyznaczania odległości pomiędzy skupieniami. Zarówno wartość “complete” jak i “ward.D2” przynosiły bardzo podobne wyniki, dlatego też przyjęliśmy wartość “complete, dla której wyniki zaprezentowane są poniżej.



Można zauważyć że eksperymenty przyniosły ciekawe rezultaty, można zdecydowanie wyróżnić skupienia dla podobnych dokumentów z tematów literatura, oraz ekonomia. Zdrowie oraz COVID także mają podobnie przeplatające się skupienia ze względu na podobną tematykę. Tematyka IT dalej zostaje “porozrzucana” po skupieniach, choć częściowo udało się ją zaklasyfikować do podobnych skupień.

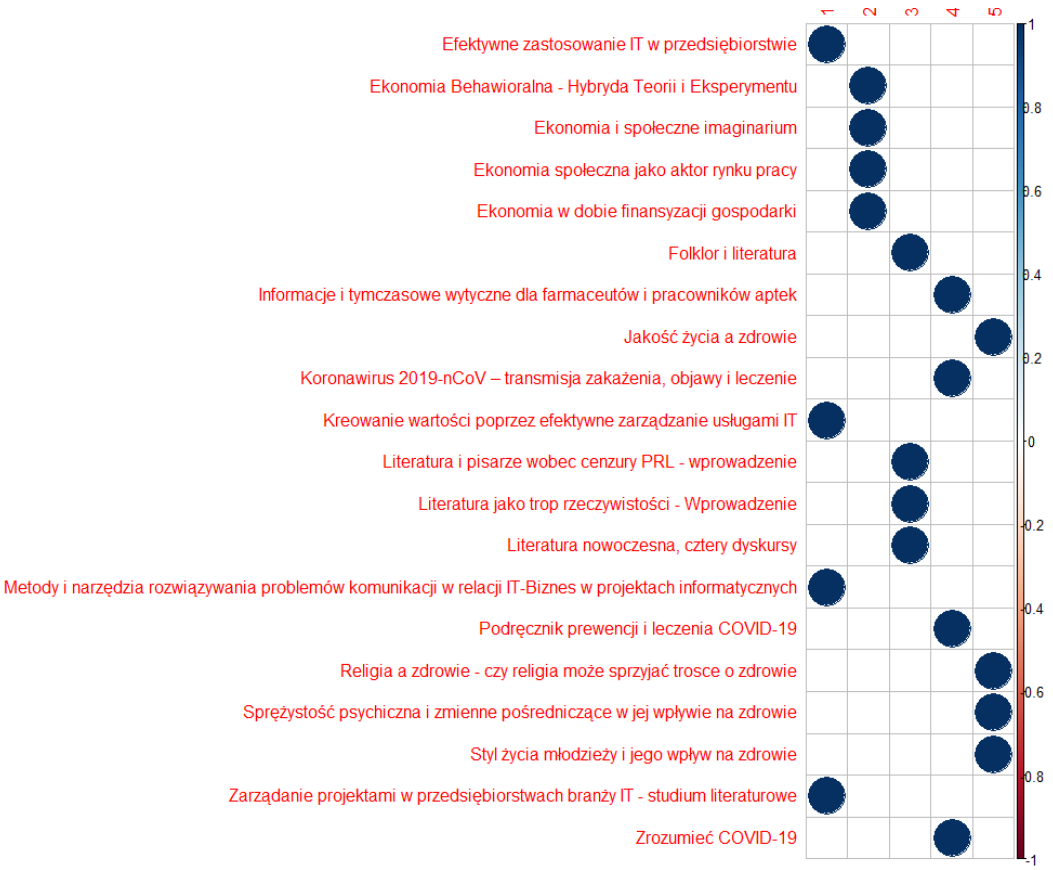
Następnie wykonane zostały eksperymenty przyjmując różną liczbę skupień, aby zobaczyć wpływ na dendrogram.

- Dendrogram dla liczby skupień wynoszącej 10



- Dendrogram dla liczby skupień wynoszącej 12

- Dendrogram dla liczby skupień wynoszącej 5



Zmniejszając liczbę skupień na 5 możemy niemal bezbłędnie zaklasyfikować tematyki do tych samych skupień, jedynym wyjątkiem pozostają tematyki na temat COVIDa oraz zdrowia, które są problematyczne z racji tego że są podobne.

Nawet dla większej ilości skupień, grupowanie dla literatury jest także niemal niezmienne.

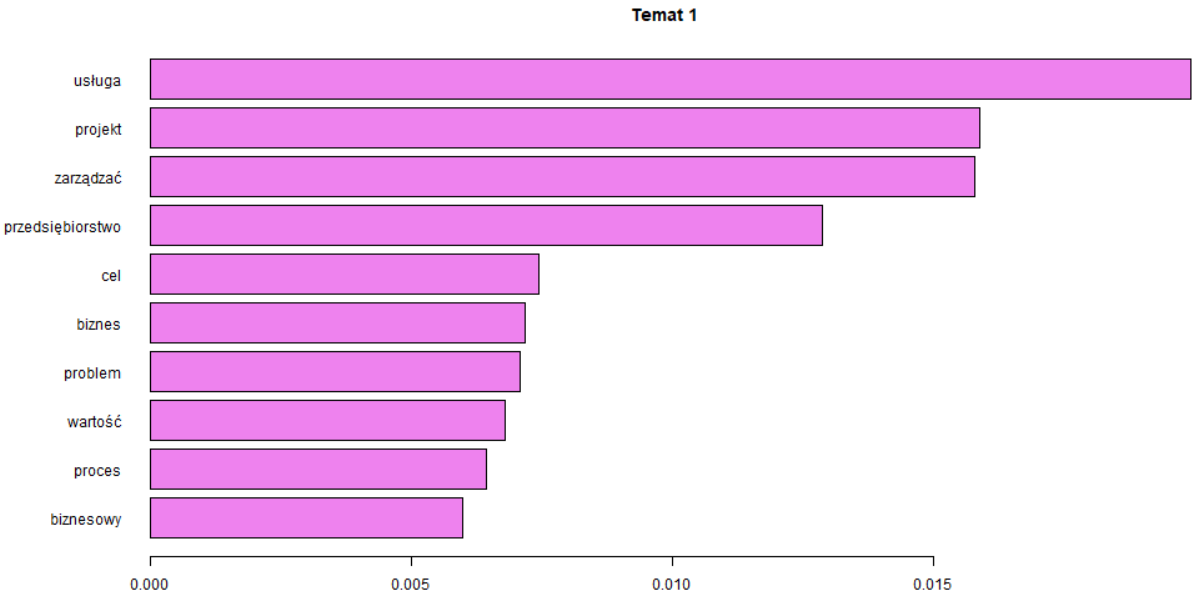
5. Metoda ukrytej alokacji Dirichlet'a

Eksperymenty z LDA zostały przeprowadzone na dokumentach wstępnie przetworzonych, używając macierzy częstości DTM z wagą Tf, ponieważ metoda LDA wymaga "term frequency weighting".

Eksperymenty przeprowadzone zostały dla różnych macierzy DTM (o różnych granicach) i dla różnej liczby tematów. Liczba słów w tematach ustawiona została na stałą liczbę wynoszącą 10. Przy każdym eksperymencie sprawdzony został udział niektórych słów w tematach.

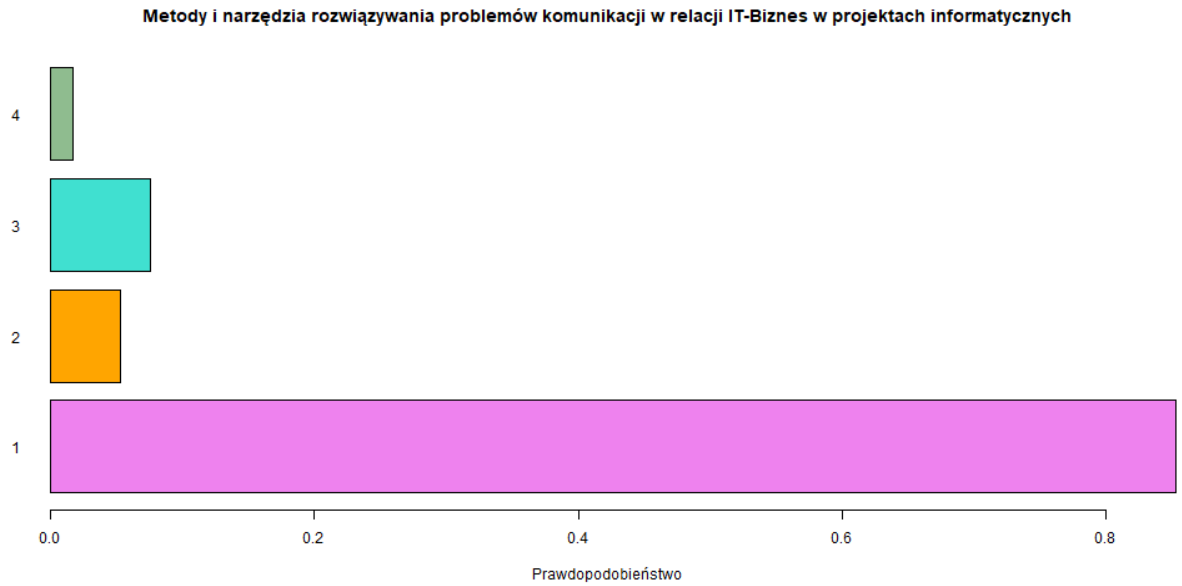
Poniższe zrzuty ekranu i komentarze przedstawiają niektóre z wyników uzyskanych w eksperymentach, które uznane zostały za najbardziej interesujące.

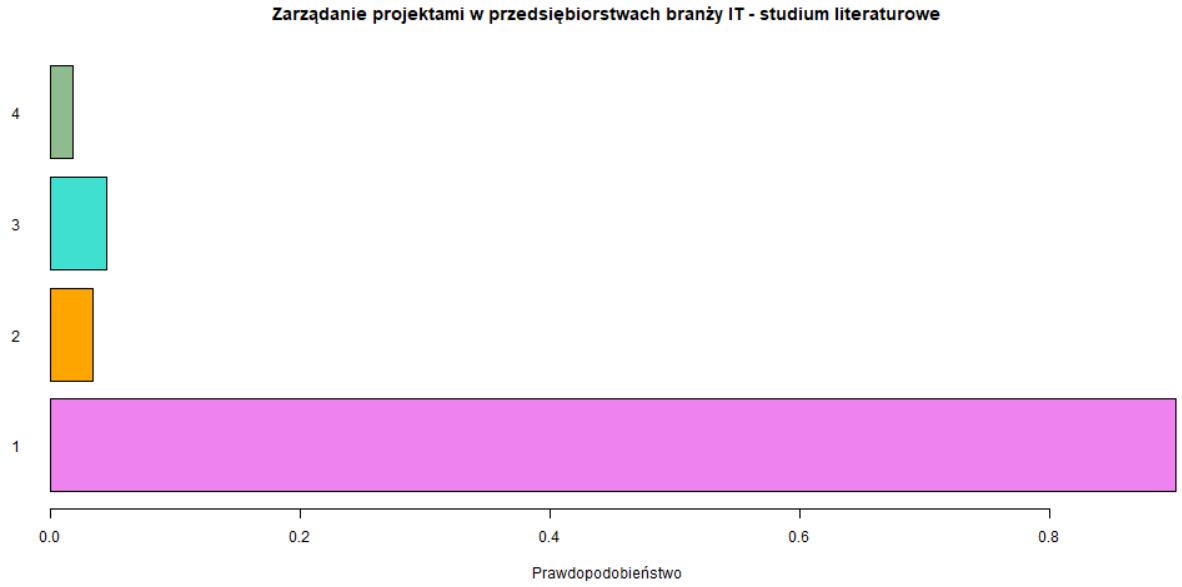
- Eksperyment pierwszy:
  - DTM_Tf_NoBounds,
  - 4 tematy.

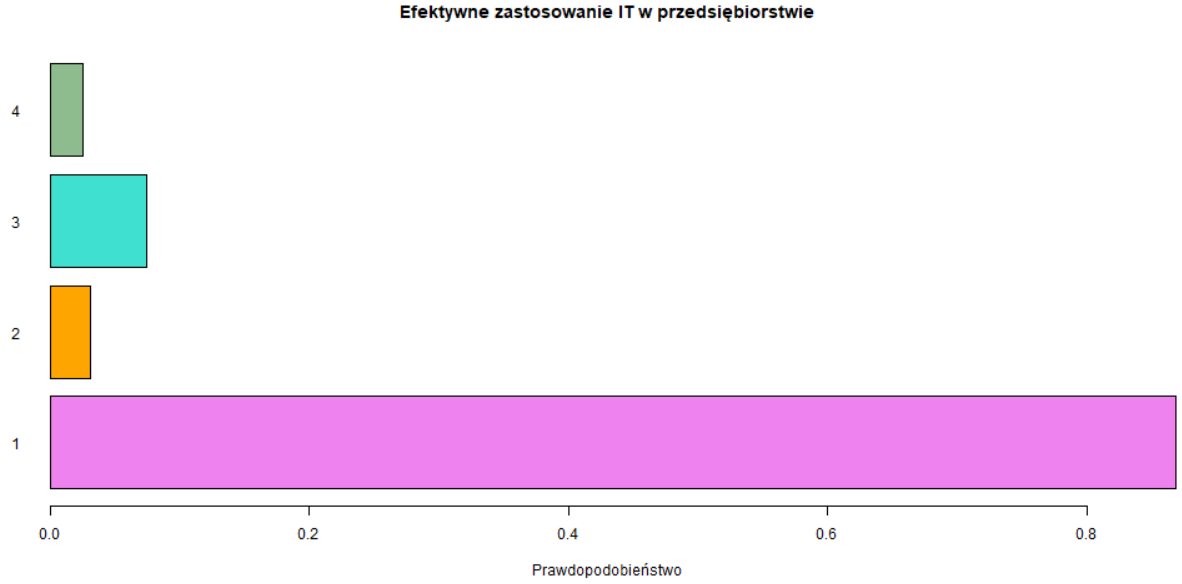


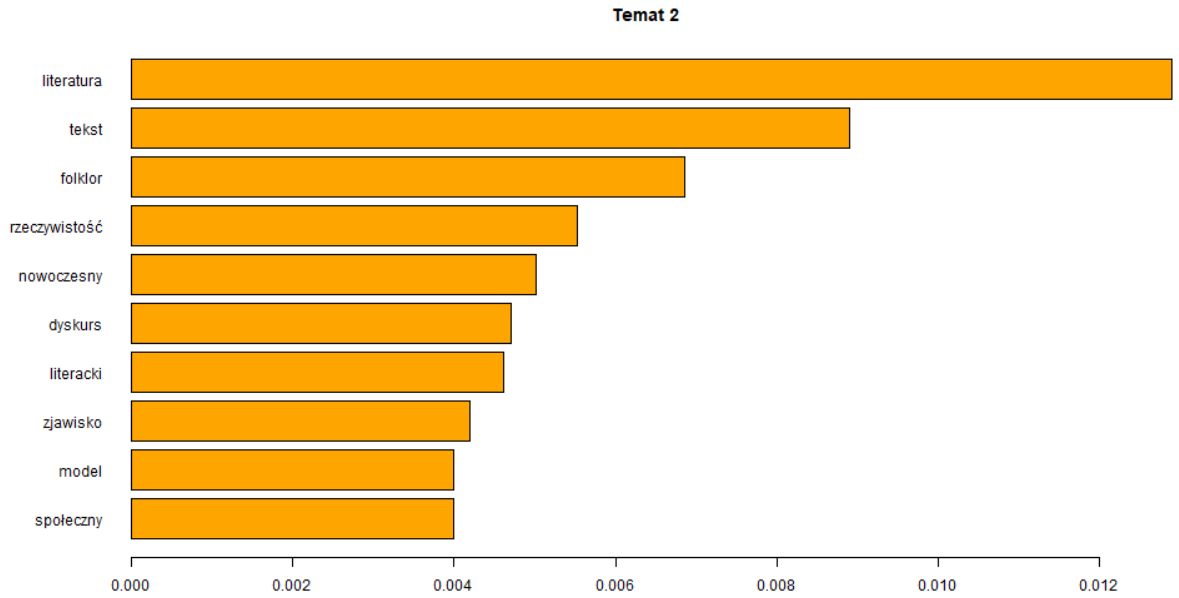
Z wybranych przez nas tematyk, najbardziej do Tematu 1 pasowałyby: IT oraz Ekonomia.

Poniżej dokumenty, których dopasowanie do Tematu 1 okazało się najlepsze. Rzeczywiście, największe prawdopodobieństwo bycia w Temacie 1 mają teksty o tematyce Ekonomii i IT:

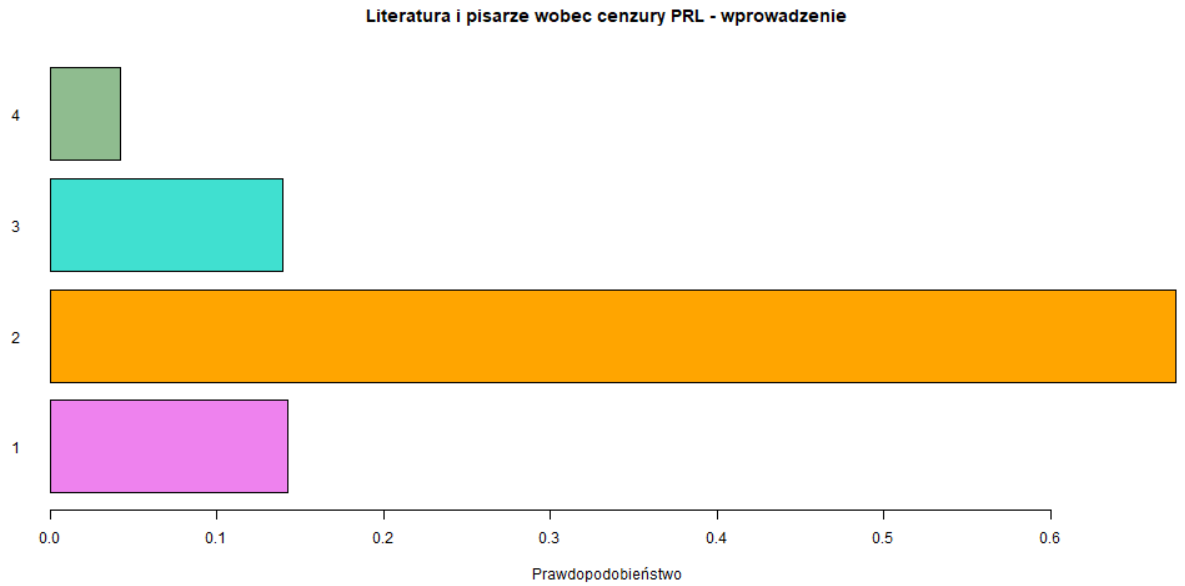


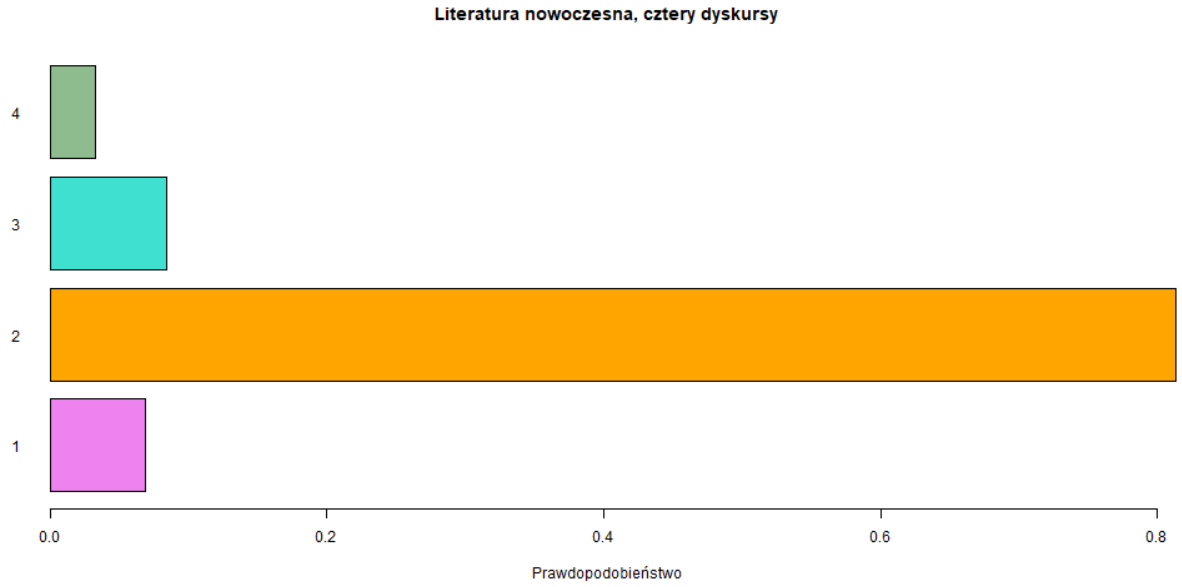

Zarządanie projektami w przedsiębiorstwach branży IT - studium literaturowe
4
3
2
1
0.0
0.2
0.4
0.6
0.8
Prawdopodobieństwo

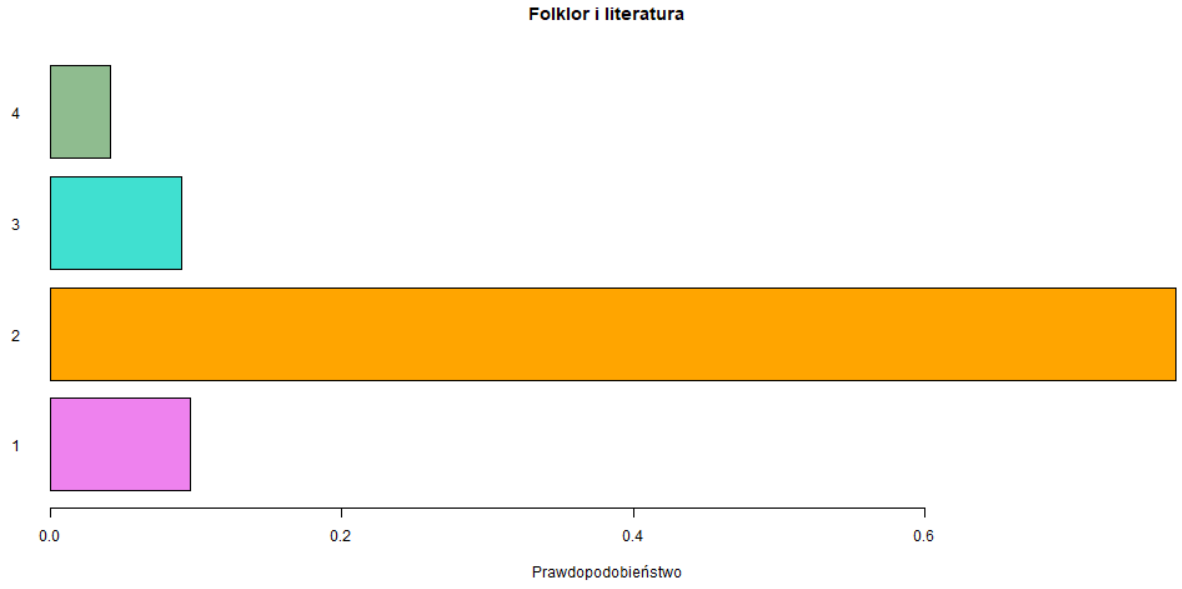

Efektywne zastosowanie IT w przedsiębiorstwie
4
3
2
1
0.0
0.2
0.4
0.6
0.8
Prawdopodobieństwo



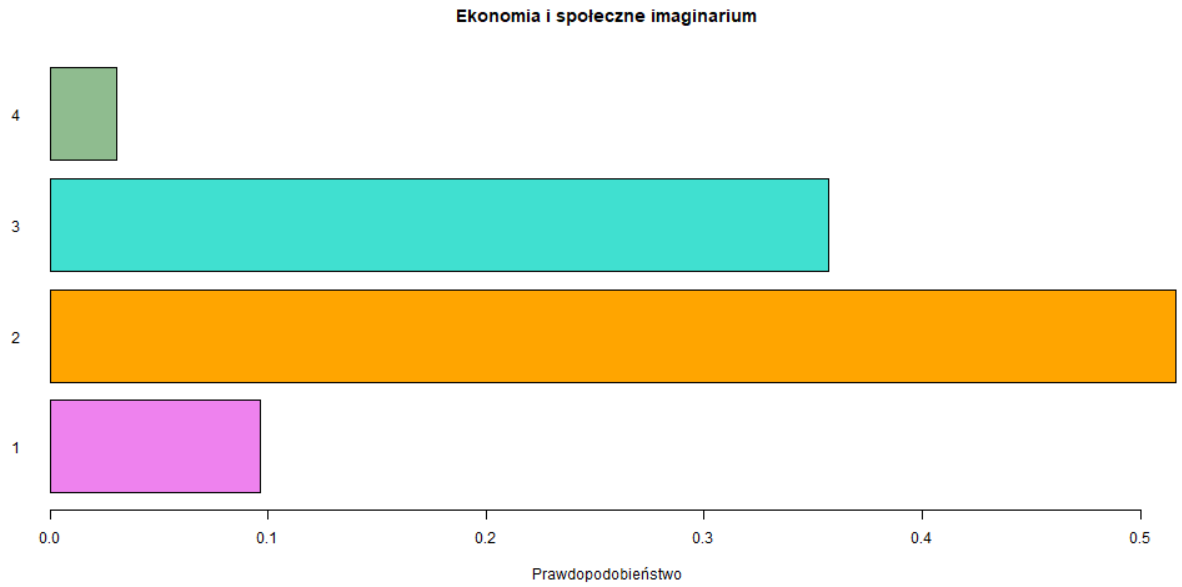
Temat 2 można by przyporządkować do wybranej przez nas tematyki Literatura. Rzeczywiście, wszystkie cztery teksty z tematyki Literatura wykazują wysokie prawdopodobieństwo przynależności do Tematu 2:




Literatura nowoczesna, cztery dyskursy
4
3
2
1
0.0
0.2
0.4
0.6
0.8
Prawdopodobieństwo


Folklor i literatura
4
3
2
1
0.0
0.2
0.4
0.6
Prawdopodobieństwo

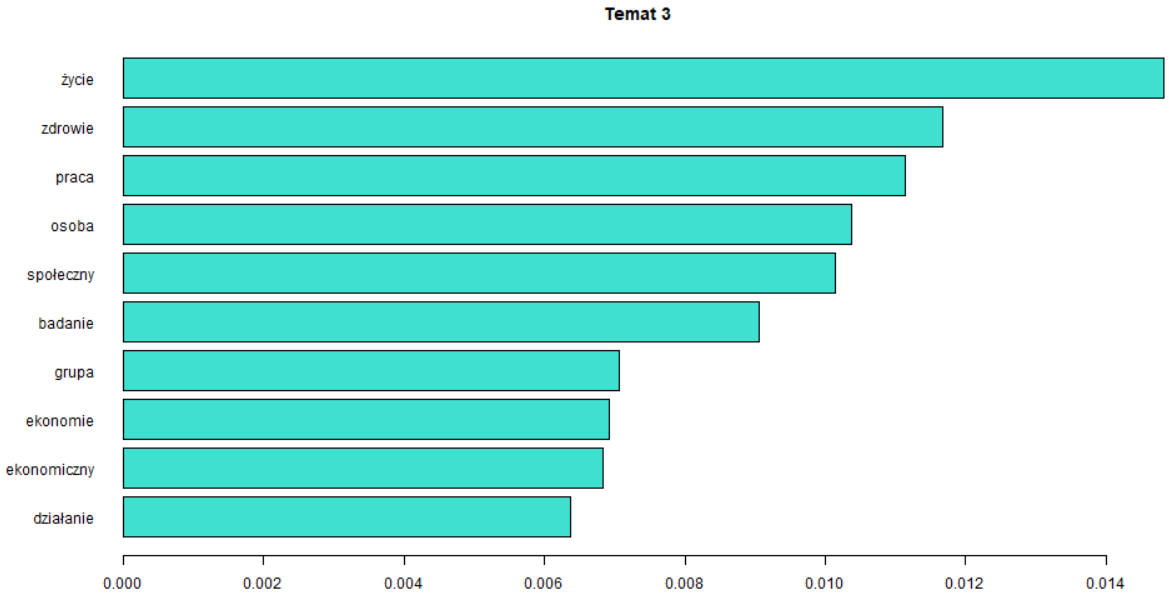
W kontekście Tematu 2, warto zwrócić uwagę na poniższy wynik:



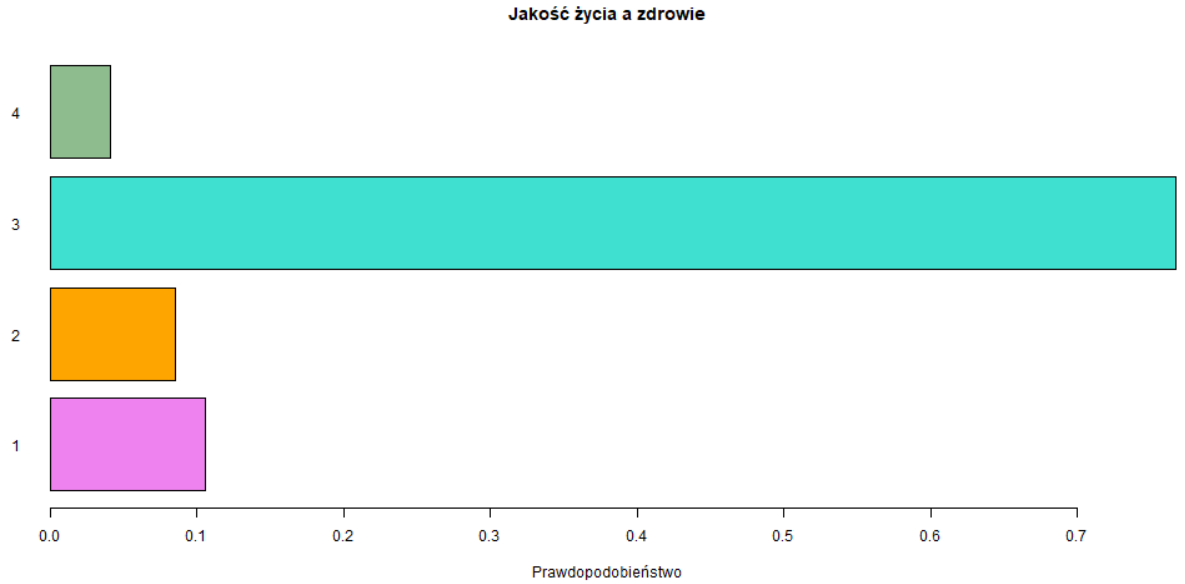
Pomimo że tekst “Ekonomia i społeczne imaginarium” pochodzi z wybranej przez nas tematyki Ekonomia, według eksperymentu został przyporządkowany do Tematu 2, który oscyluje bardziej w tematyce Literatury. Może to wynikać z wysokiego prawdopodobieństwa występowania słowa “społeczny” w Temacie 2, ponieważ, zgodnie z tytułem, tekst dotyczy ekonomii i społecznego imaginarium:

```
> words1 <- c("społeczny")
> round(experiment1$terms[,words1],4)
     1      2      3      4
0.0000 0.0040 0.0101 0.0000
```

Temat 3 może wskazywać na powiązanie z tematyką Ekonomii i również w nieco mniejszym stopniu z IT:



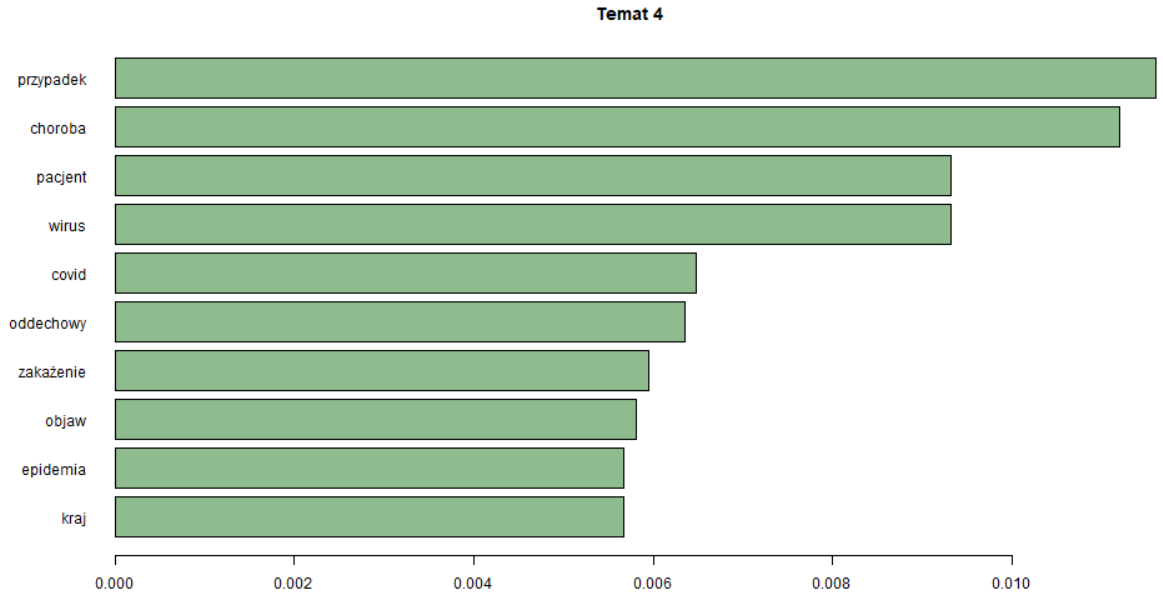
Jednak, występujące w Temacie 3 słowo “zdrowie” zahacza również o wybraną przez nas tematykę Zdrowie, dlatego także tekst o takiej tematyce został przyporządkowany do tego Tematu z bardzo wysokim prawdopodobieństwem:



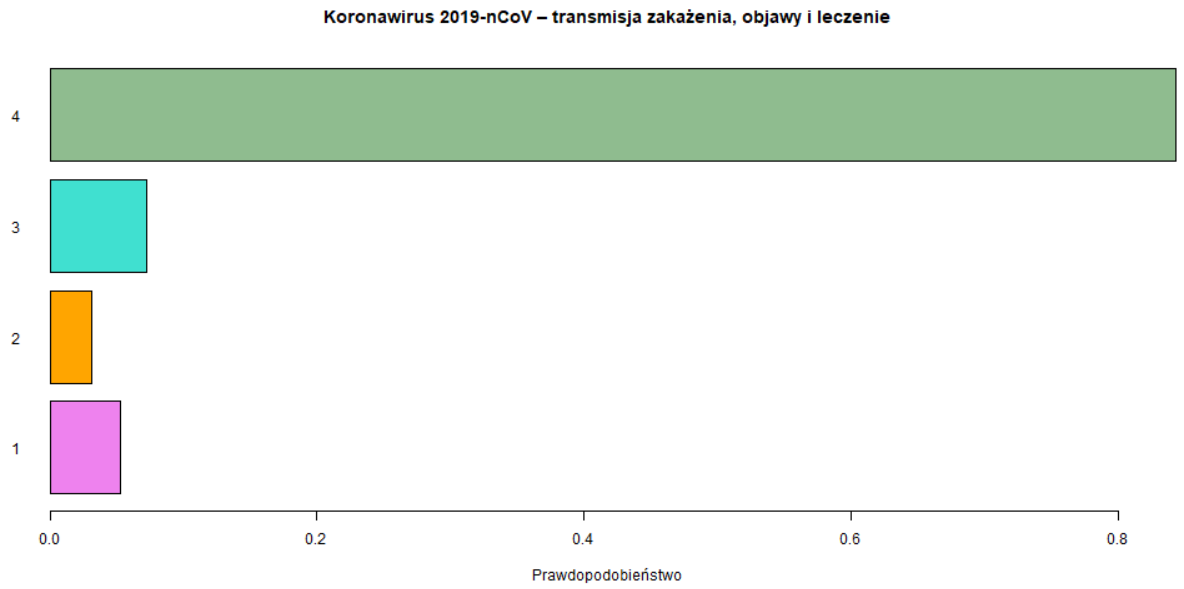
Prawdopodobieństwo wystąpienia słowa “zdrowie” w Tematach 3 i 4 jest znaczące, stąd podobieństwo:

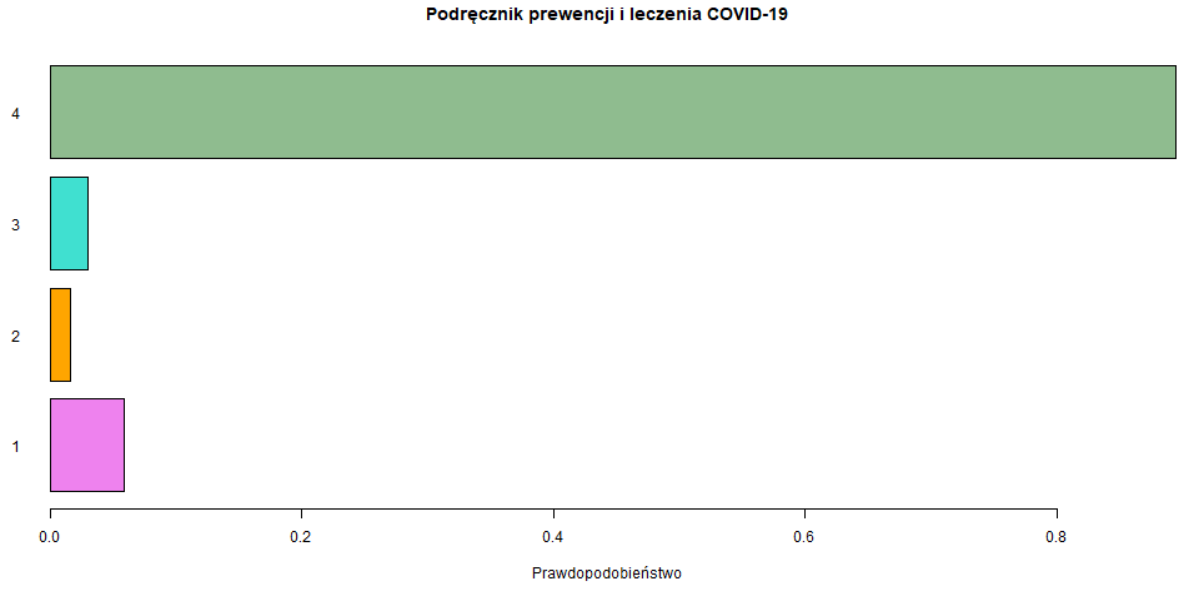
```
> words2 <- c("zdrowie")
> round(experiment1$terms[,words2],4)
     1      2      3      4
0.0000 0.0000 0.0117 0.0024
```

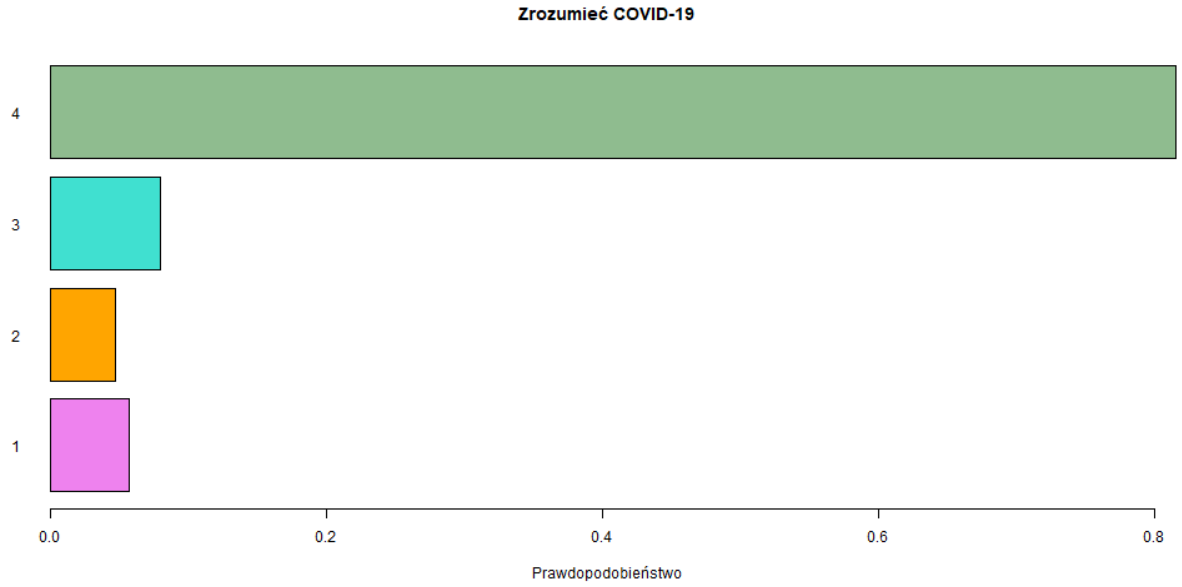
Temat 4 najbardziej wskazuje na tematykę pandemii COVID-19, jednak jest także mocno powiązany z tematyką Zdrowie.

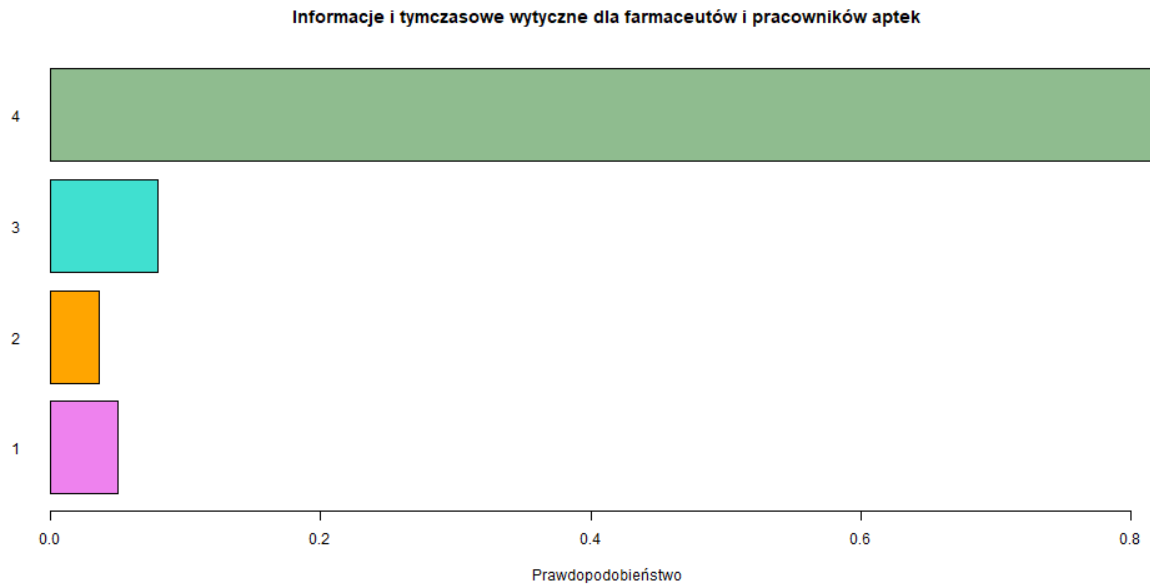


Rzeczywiście, wszystkie 4 teksty o tematyce COVID-19 mają największe prawdopodobieństwo przynależności do Tematu 4:

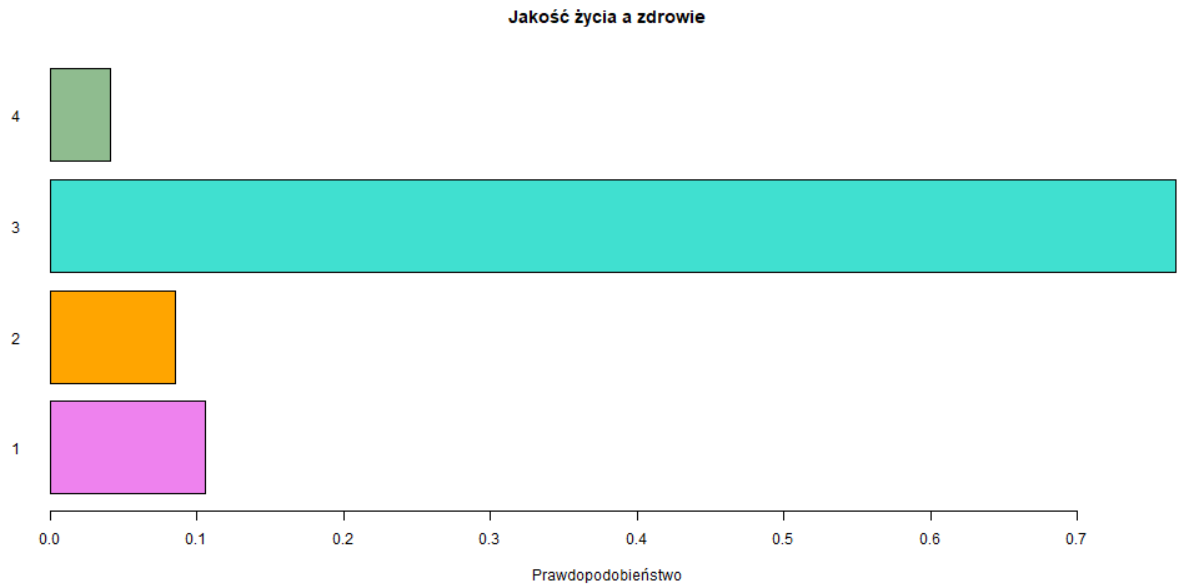





Zrozumieć COVID-19
4
3
2
1
0.0
0.2
0.4
0.6
0.8
Prawdopodobieństwo


Informacje i tymczasowe wytyczne dla farmaceutów i pracowników aptek
4
3
2
1
0.0
0.2
0.4
0.6
0.8
Prawdopodobieństwo

Co ciekawe, mimo bliskości tematyki Zdrowie do Tematu 4, tekst pod tytułem “Jakość życia i zdrowie” ma bardzo niskie prawdopodobieństwo przynależności do Tematu 4:



Analizując udział niektórych słów w tematach, udało się znaleźć kilka takich, które są wspólne dla wielu tematów:

- słowo “proces” ma prawdopodobieństwo występowania we wszystkich czterech tematach, jednak największe w temacie pierwszym, który najbardziej pasuje do tematyki Ekonomia i IT

```
> words3 <- c("proces")
> round(experiment1$terms[,words3],4)
     1      2      3      4
0.0064 0.0030 0.0015 0.0001
```

- słowo “życie” ma najwyższe prawdopodobieństwo występowania w Temacie 3 (tematyka Ekonomia)

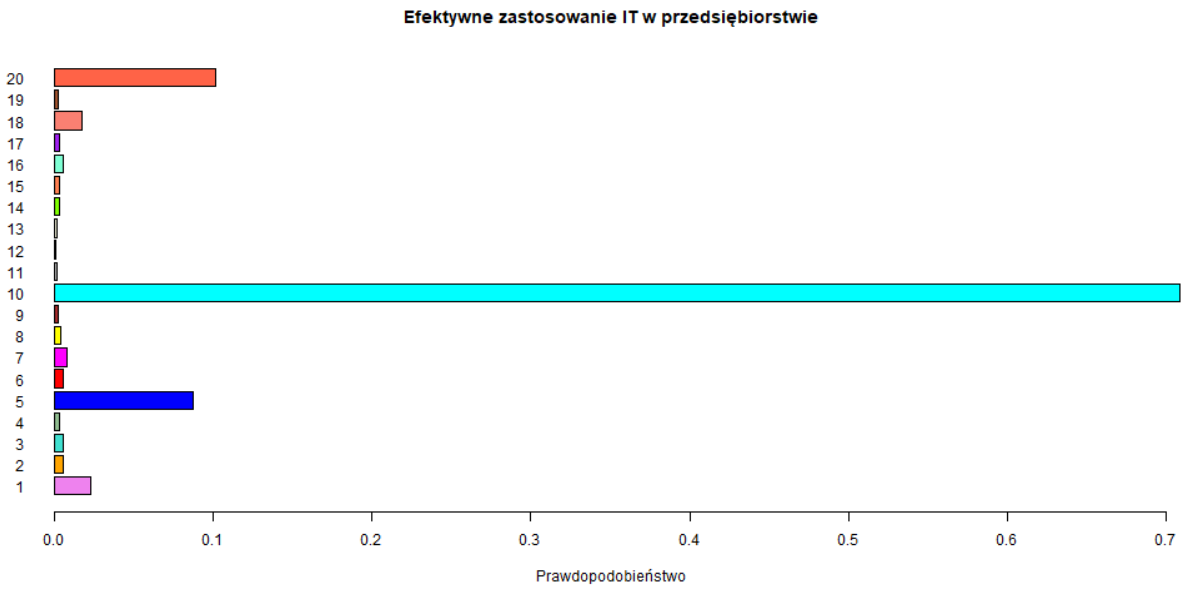
```
> words4 <- c("życie")
> round(experiment1$terms[,words4],4)
     1      2      3      4
0.0013 0.0000 0.0148 0.0000
```

- słowo “praca” występuje w Temacie 1, 3 i 4, z najwyższym prawdopodobieństwem występowania w Temacie 3 (Ekonomia)

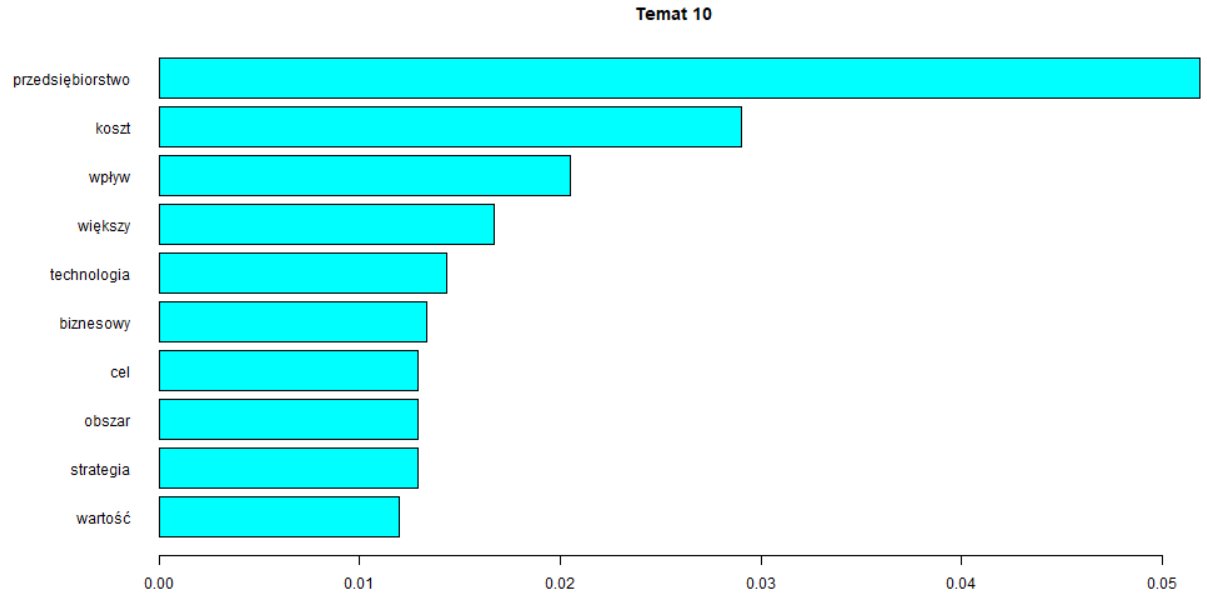
```
> words5 <- c("praca")
> round(experiment1$terms[,words5],4)
     1      2      3      4
0.0011 0.0000 0.0111 0.0011
```

a. Eksperyment drugi:
    - DTM_Tf_2_18,
    - 20 tematów.

Drugi eksperyment zawierał aż 20 tematów, czyli liczbę równą liczbie analizowanych tekstów. Wyniki wyszły dosyć interesujące, dlatego przedstawimy je z perspektywy każdego z dokumentów.



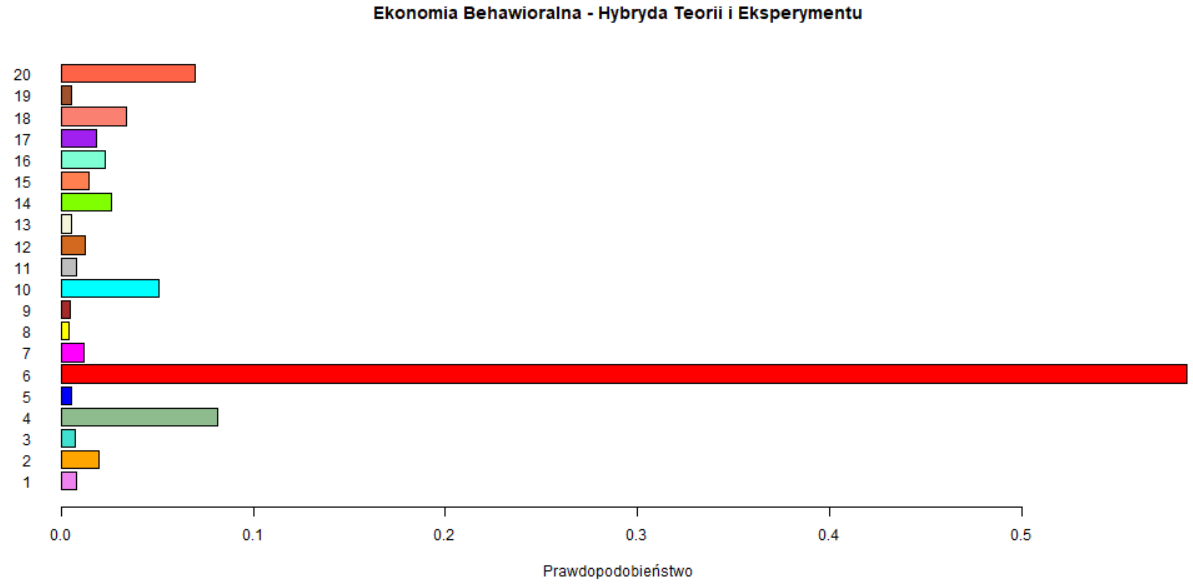
Tekst "Efektywne zastosowanie IT w przedsiębiorstwie", należący do tematyki IT wykazał największe prawdopodobieństwo przynależności do Tematu 10:



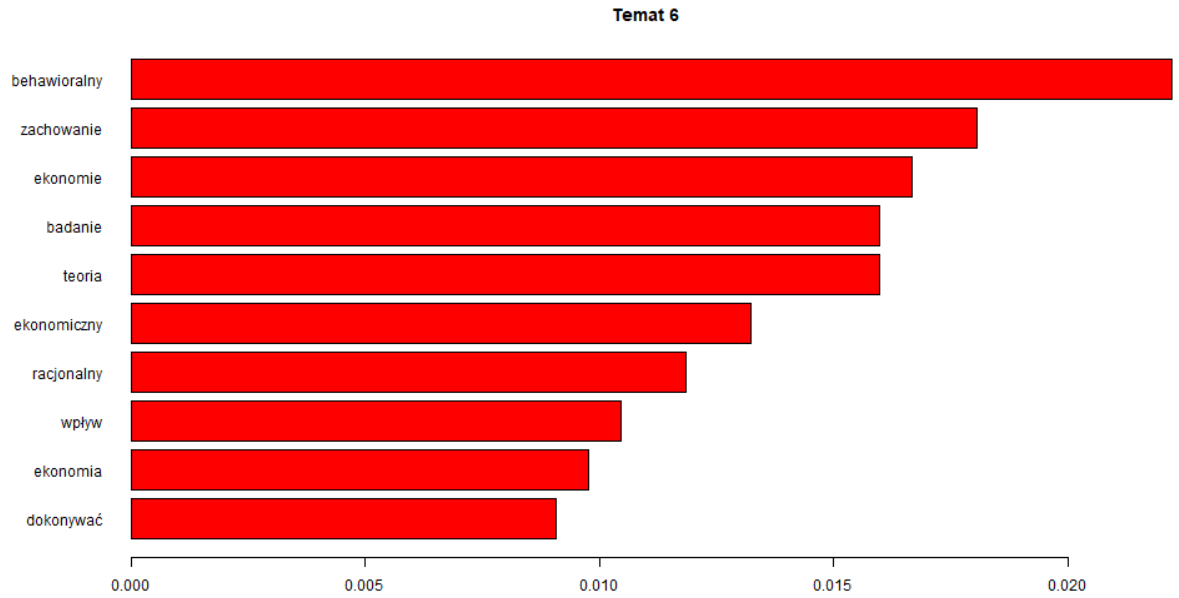
Rzeczywiście słowa występujące w Temacie 10 wskazują na tematykę Ekonomia (z której pochodzi tekst), jednak można także zauważyć słowa pasujące do tematyki IT, jak np. “technologia”.

Słowo “przedsiębiorstwo”, które ma najwyższe prawdopodobieństwo występowania w Temacie 10, ma także znaczne prawdopodobieństwo występowania w Temacie 1, dla którego to rozważany dokument prawdopodobieństwo przynależności ma na poziomie ok. 0.025.

```
> words1 <- c("przedsiębiorstwo")
> round(experiment2$terms[,words1],20)
            1             2             3             4             5             6             7             8
0.01554203540 0.00007304602 0.00008025682 0.00004662005 0.00004568296 0.00006920415 0.00007062147 0.00006648936
            9            10            11            12            13            14            15            16
0.00004403347 0.05187826914 0.00006038647 0.00007911392 0.00005707763 0.00149038462 0.00006476684 0.00006600660
           17            18            19            20
0.00005117707 0.00004757374 0.00004995005 0.00003321156
```

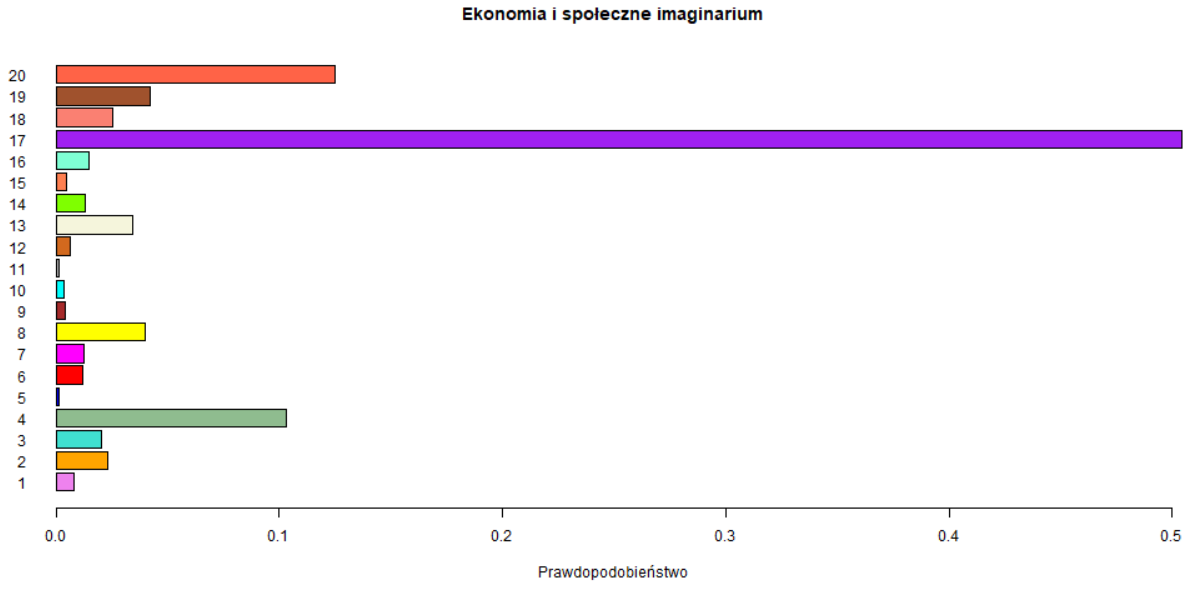


Tekst “Ekonomia Behawioralna - Hybryda Teorii i Eksperymentu” wykazał bardzo wysokie prawdopodobieństwo przynależności do Tematu 6.

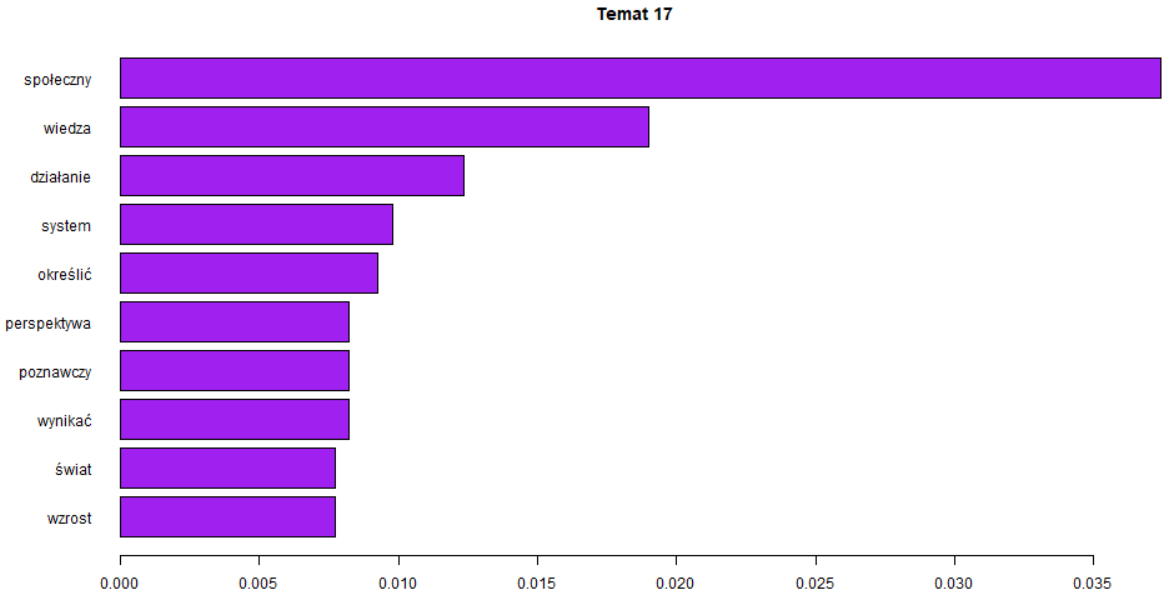


Temat 6 jest bardzo zbliżony do tematyki Ekonomii, a szczególnie do tekstu o ekonomii behawioralnej, warto zwrócić uwagę na słowo o największym prawdopodobieństwie wystąpienia, jakim jest “behawioralny”. Ma ono bardzo niskie prawdopodobieństwo wystąpienia w pozostałych dziewiętnastu tematach:

```
> words2 <- c("behawioralny")
> round(experiment2$terms[,words2],20)
                1             2             3             4             5             6             7             8
0.00005530973 0.00007304602 0.00008025682 0.00004662005 0.00004568296 0.02221453287 0.00007062147 0.00006648936
                9            10            11            12            13            14            15            16
0.00004403347 0.00004755112 0.00006038647 0.00007911392 0.00005707763 0.00004807692 0.00006476684 0.00006600660
               17            18            19            20
0.00005117707 0.00004757374 0.00004995005 0.00069744271
```



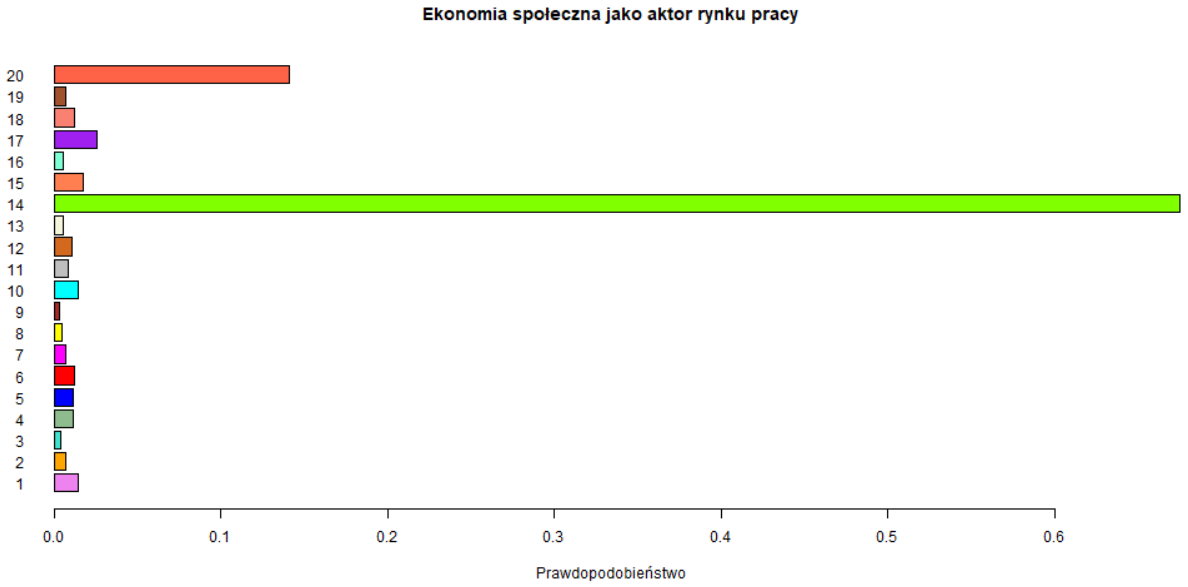
Dokument “Ekonomia i społeczne imaginarium” wykazuje wysokie prawdopodobieństwo przynależności do Tematu 17.



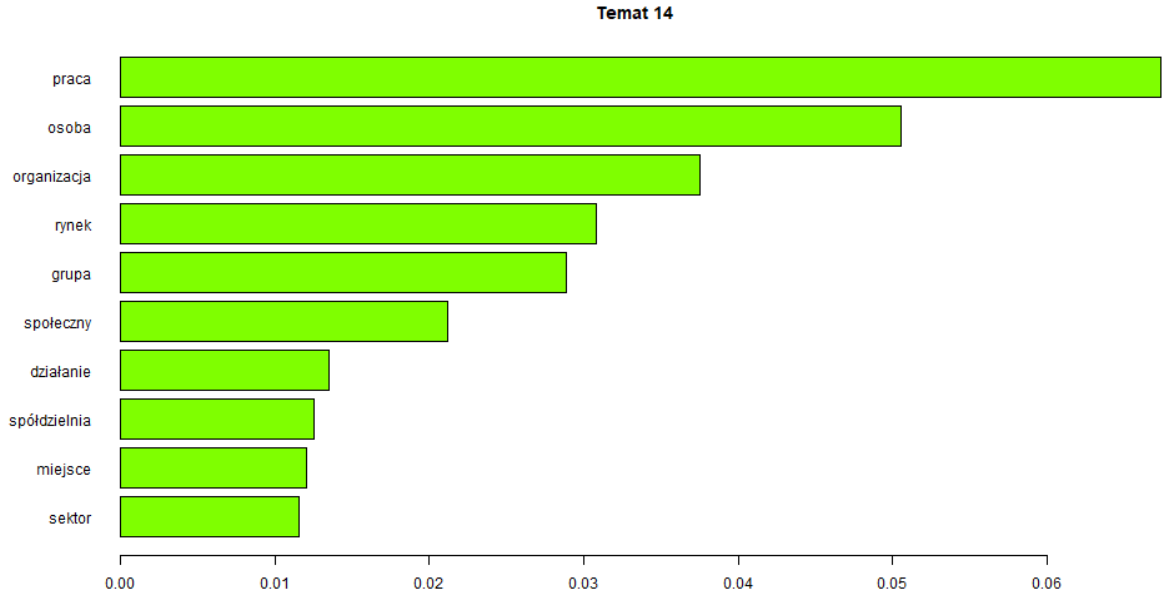
Słowa występujące w Temacie 17 dosyć jasno wskazują na powiązanie z tematyką Ekonomii, szczególnie pasują one konkretnie do wyżej wspomnianego dokumentu o ekonomii i społecznym imaginarium.

Mimo że słowo “społeczny” ma bardzo wysokie prawdopodobieństwo występowania w Temacie 17, nie występuje aż tak często w Temacie 20, do którego prawdopodobieństwo przynależności omawianego dokumentu również jest znaczne (ok. 0.125).

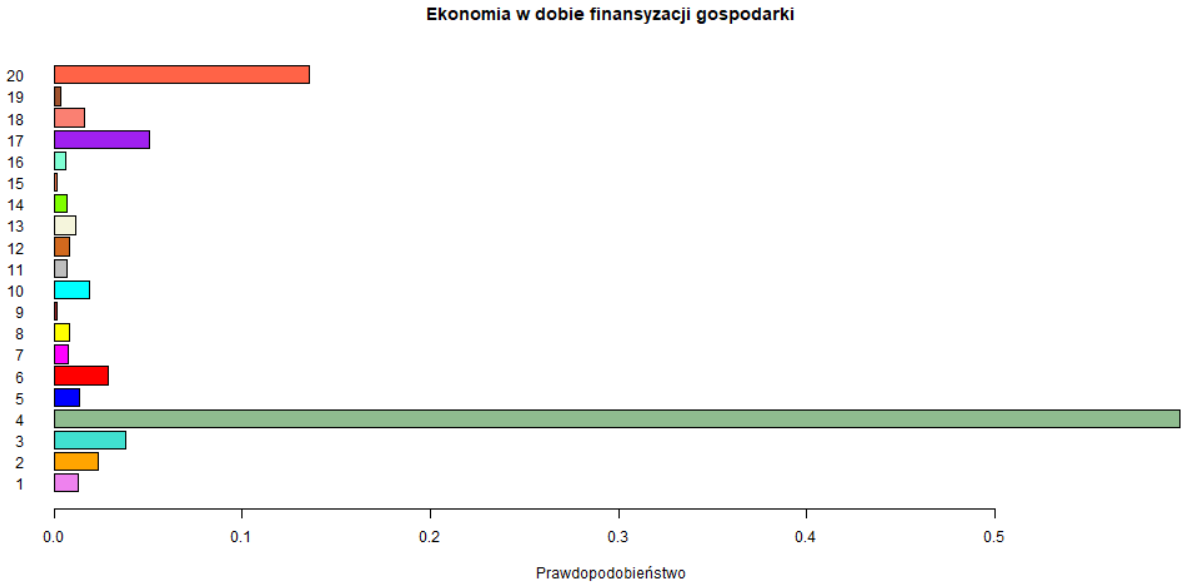
```
> words3 <- c("społeczny")
> round(experiment2$terms[,words3],20)
            1             2             3             4             5             6             7             8
0.00005530973 0.01906501096 0.00008025682 0.00797202797 0.00004568296 0.00006920415 0.00007062147 0.00405585106
            9            10            11            12            13            14            15            16
0.00004403347 0.00004755112 0.00006038647 0.00007911392 0.00005707763 0.02120192308 0.00006476684 0.00336633663
           17            18            19            20
0.03741044012 0.00004757374 0.00004995005 0.00003321156
```



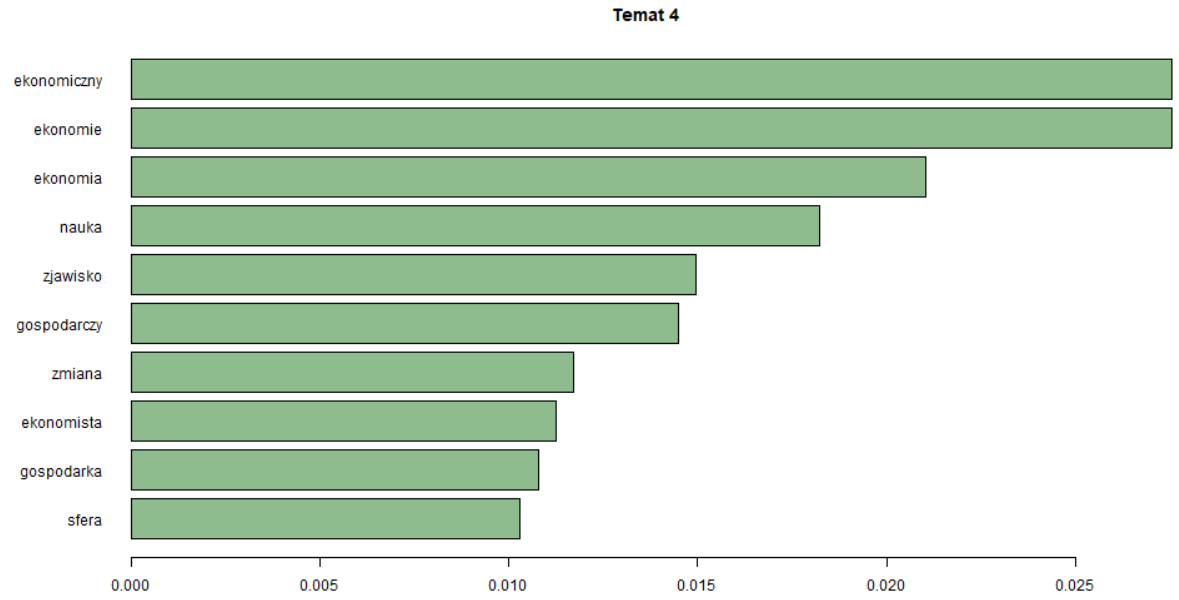
Tekst "Ekonomia społeczna jako aktor rynku pracy" wykazał wysokie prawdopodobieństwo przynależności do Tematu 14.



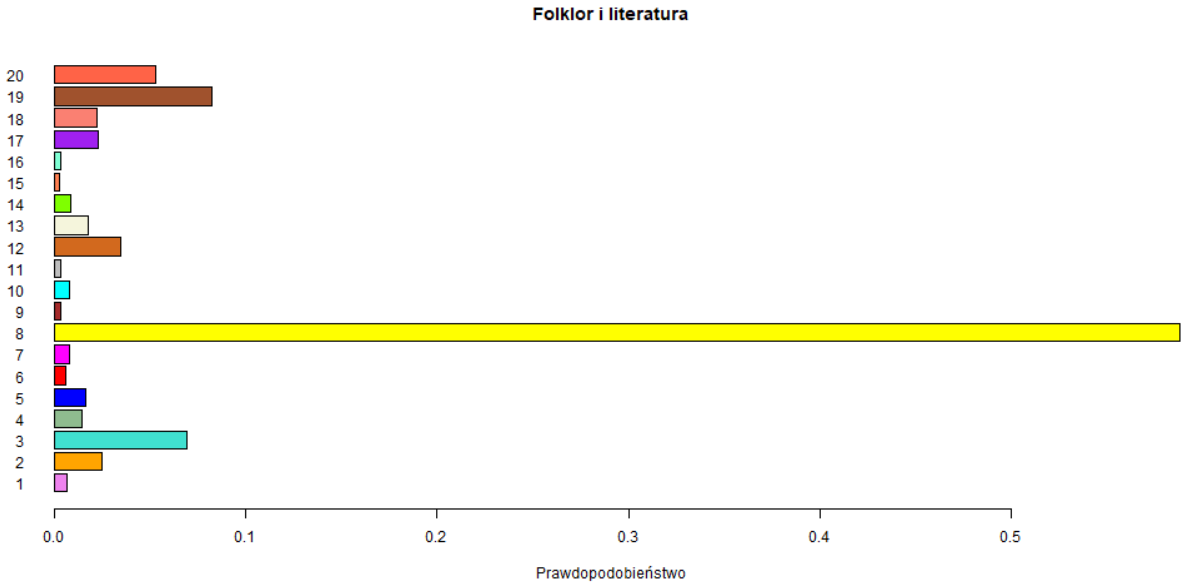
Ze słów występujących w Temacie 14 można wywnioskować, że rzeczywiście dotyczy on tematyki omawianego dokumentu (Ekonomia).



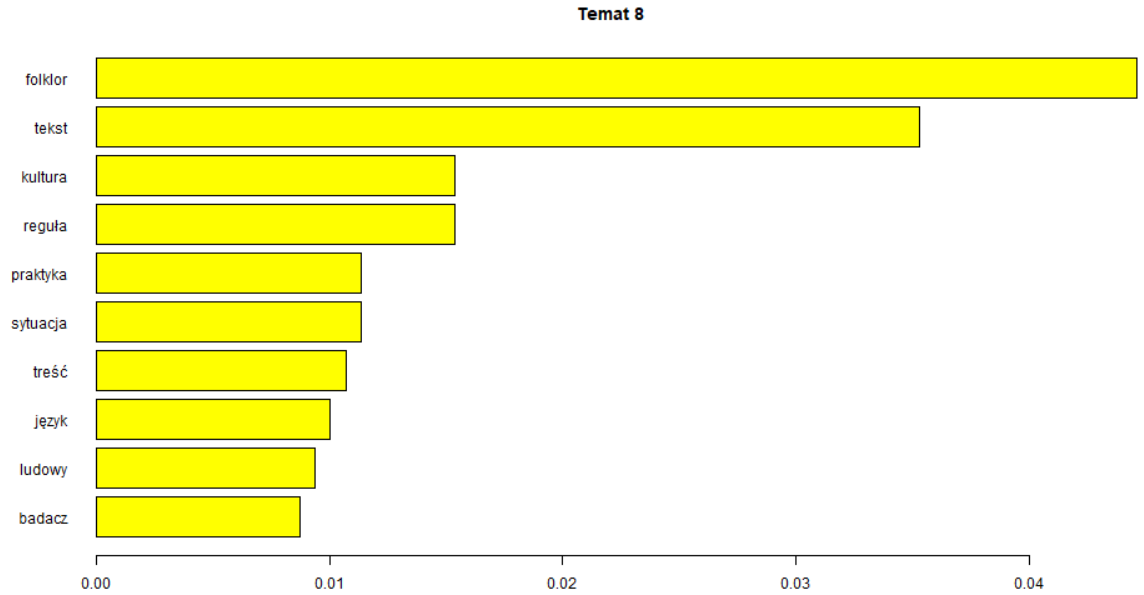
Tekst "Ekonomia w dobie finansyzacji gospodarki" wykazał największe prawdopodobieństwo przynależności do Tematu 4.



Nietrudno zauważyć, że w rzeczy samej słowa występujące w Temacie 4 wpasowują się w tematykę tekstu.

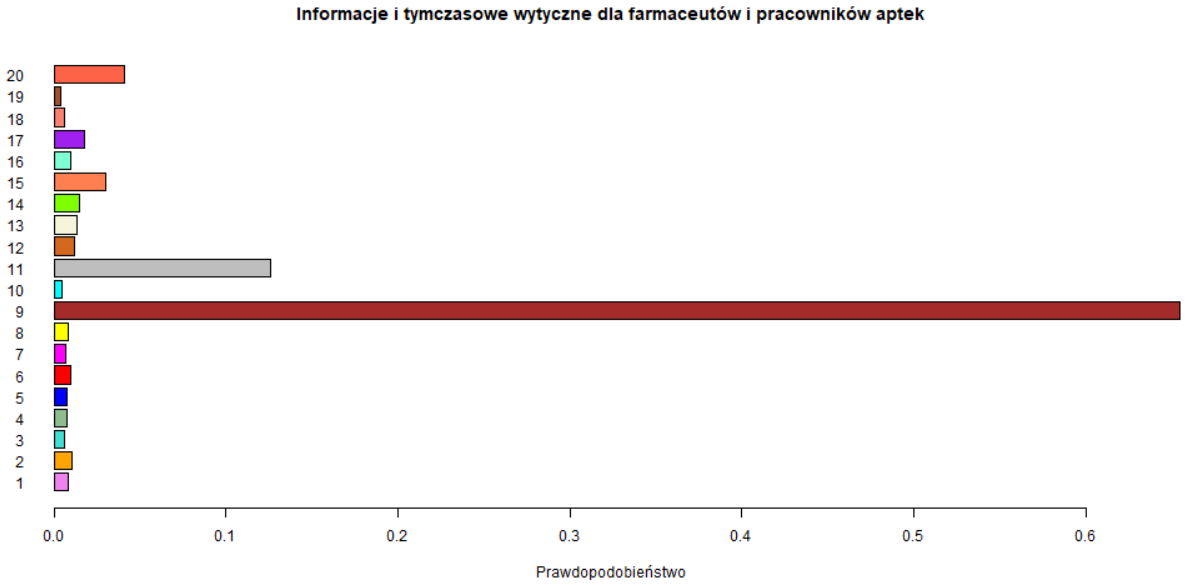


Tekst “Folklor i literatura” ma największe prawdopodobieństwo przynależności do Tematu 8

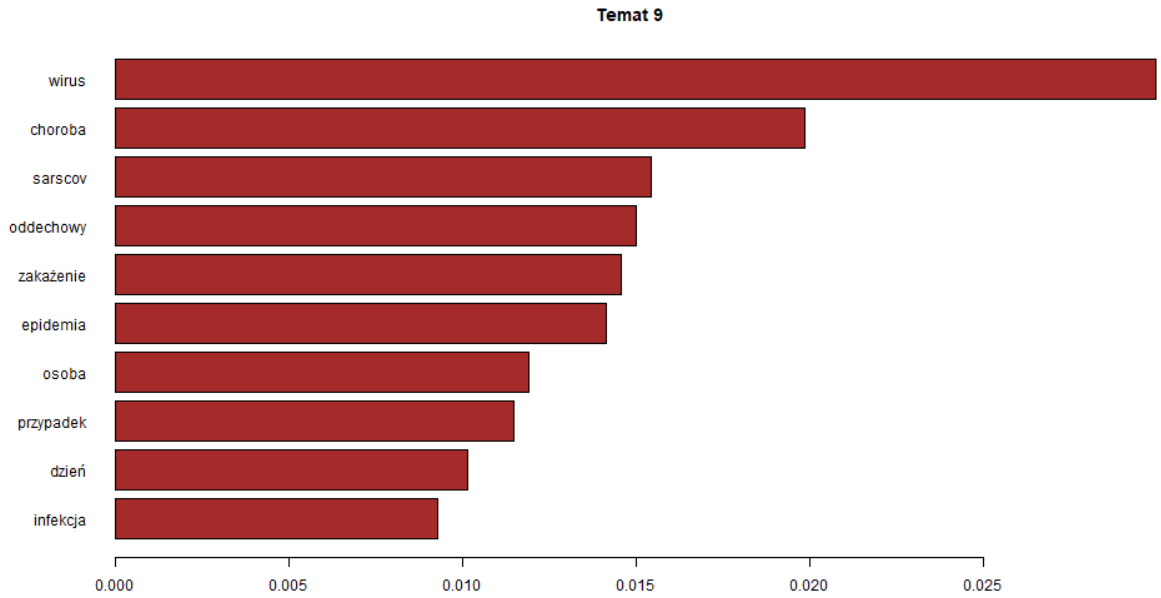


Temat ten nie wskazuje może jednoznacznie na powiązanie z ogólnie pojętą tematyką literatury, jednak zdecydowanie wpasowuje się w tematykę omawianego tekstu. Słowo “folklor” ma bardzo niskie prawdopodobieństwo wystąpienia w innych Tematach:

```
> words5 <- c("folklor")
> round(experiment2$terms[,words5],20)
            1             2             3             4             5             6             7             8
0.00005530973 0.00007304602 0.00008025682 0.00004662005 0.00004568296 0.00006920415 0.00007062147 0.04461436170
            9            10            11            12            13            14            15            16
0.00004403347 0.00004755112 0.00006038647 0.00007911392 0.00005707763 0.00004807692 0.00006476684 0.00006600660
           17            18            19            20
0.00005117707 0.00004757374 0.00004995005 0.00003321156
```

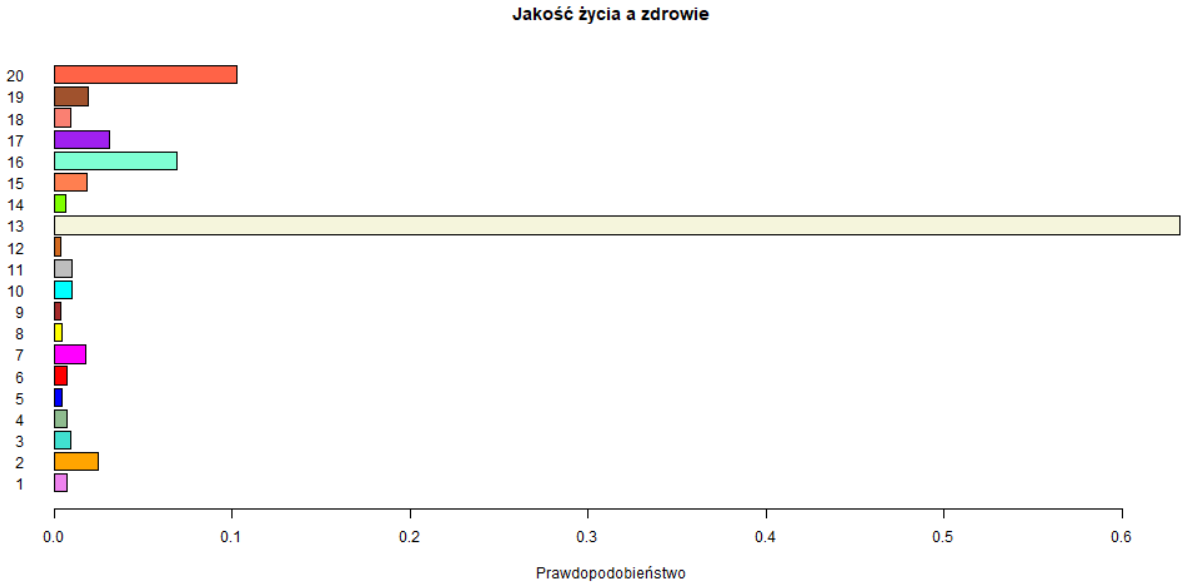


Teksty “Informacje i tymczasowe wytyczne dla farmaceutów i pracowników aptek” oraz “Koronawirus 2019-nCoV - transmisja zakażenia, objawy i leczenie” przyporządkowane przez nas do tematyki COVID-19, wykazały bardzo wysokie prawdopodobieństwo przynależności do Tematu 9.

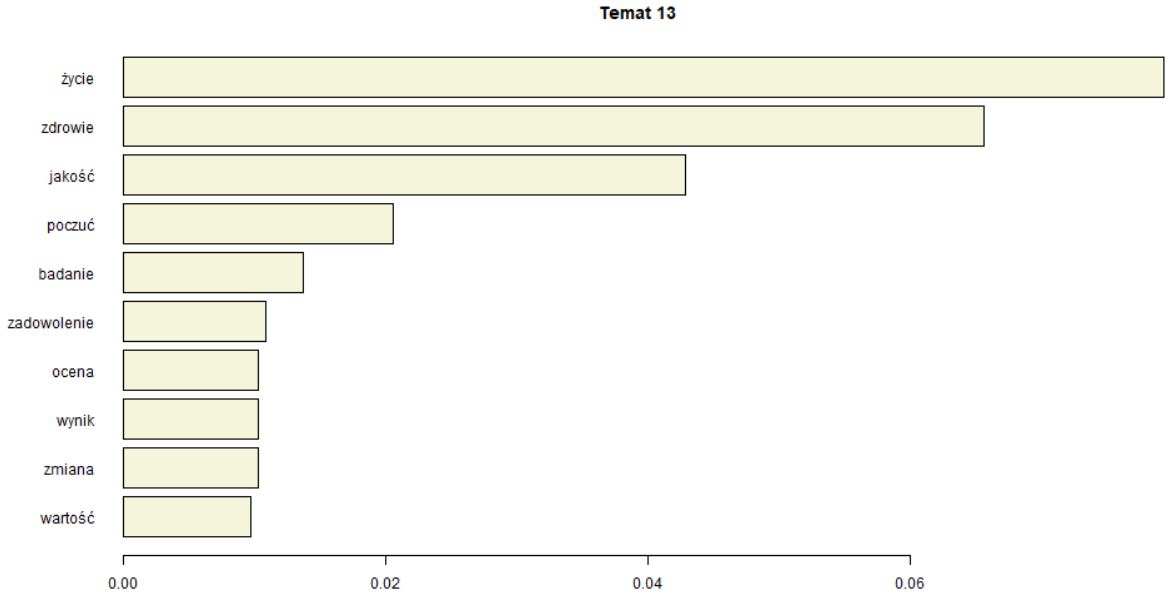


Rzeczywiście, choć temat ten mógłby łączyć w sobie tematyki Zdrowie oraz COVID-19, występujące w nim słowa “wirus”, “sarscov” czy “epidemia” uświadczają w przekonaniu, że temat odnosi się do pandemii COVID-19. Patrząc na prawdopodobieństwo występowania tych słów w innych tematach, można rzeczywiście stwierdzić, że w Temacie 9 jest ono zdecydowanie najwyższe:

```
> words6 <- c("wirus","sarscov","epidemia")
> words7 <- c("sarscov")
> words8 <- c("epidemia")
> round(experiment2$terms[,words6],20)
           wirus       sarscov      epidemia
1  0.00005530973 0.00005530973 0.00005530973
2  0.00007304602 0.00007304602 0.00007304602
3  0.00008025682 0.00008025682 0.00008025682
4  0.00004662005 0.00004662005 0.00004662005
5  0.00004568296 0.00004568296 0.00004568296
6  0.00006920415 0.00006920415 0.00006920415
7  0.00007062147 0.00007062147 0.00007062147
8  0.00006648936 0.00006648936 0.00006648936
9  0.02998678996 0.01545574637 0.01413474240
10 0.00004755112 0.00004755112 0.00004755112
11 0.00006038647 0.00006038647 0.00006038647
12 0.00007911392 0.00007911392 0.00007911392
13 0.00005707763 0.00005707763 0.00005707763
14 0.00052884615 0.00004807692 0.00004807692
15 0.00006476684 0.00006476684 0.00654145078
16 0.00006600660 0.00006600660 0.00006600660
17 0.00005117707 0.00005117707 0.00005117707
18 0.00004757374 0.00004757374 0.00004757374
19 0.00004995005 0.00004995005 0.00004995005
20 0.00003321156 0.00003321156 0.00003321156
```

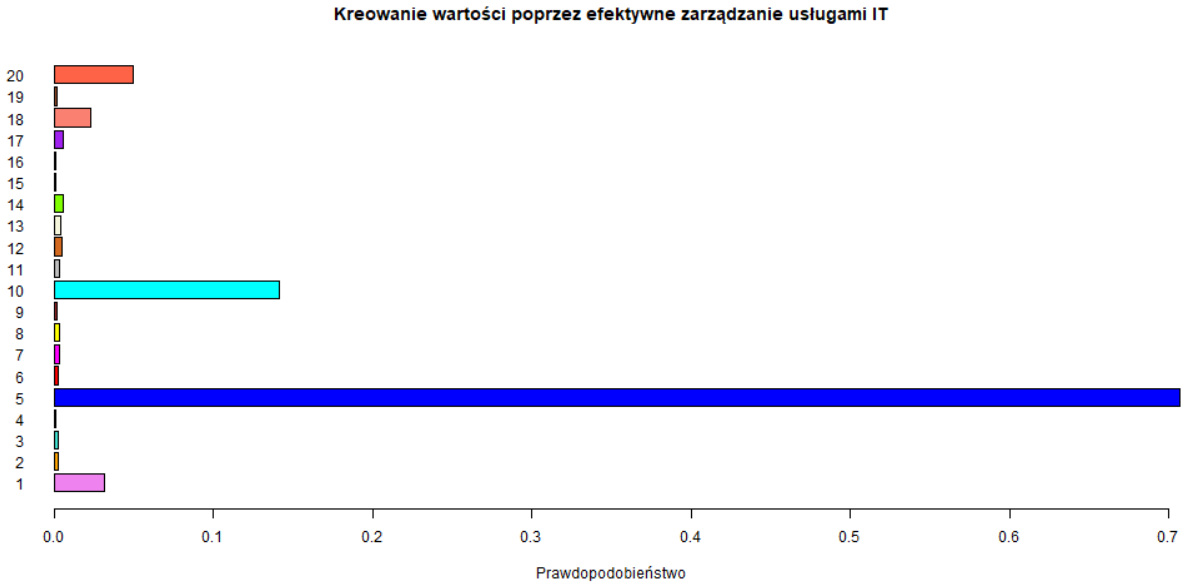


Tekst "Jakość życia a zdrowie" uzyskał w eksperymencie najwyższe prawdopodobieństwo przyporządkowania do Tematu 13.

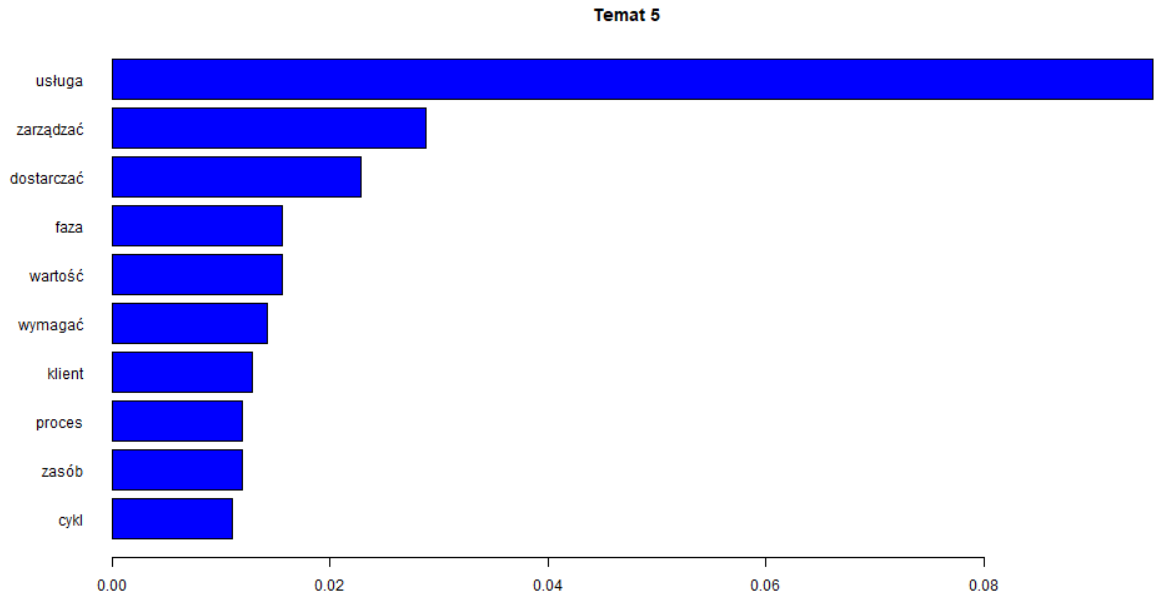


Rzeczywiście, temat ten najbardziej oscyluje w okolicach tematyki Zdrowie, co jak najbardziej zgadza się z tematyką omawianego tekstu. Co ciekawe, najczęściej występujące w temacie słowo, "życie", występuje z jedynie dwa razy mniejszym prawdopodobieństwem w Temacie 2, dla którego omawiany tekst ma nieznaczne prawdopodobieństwo przynależności.

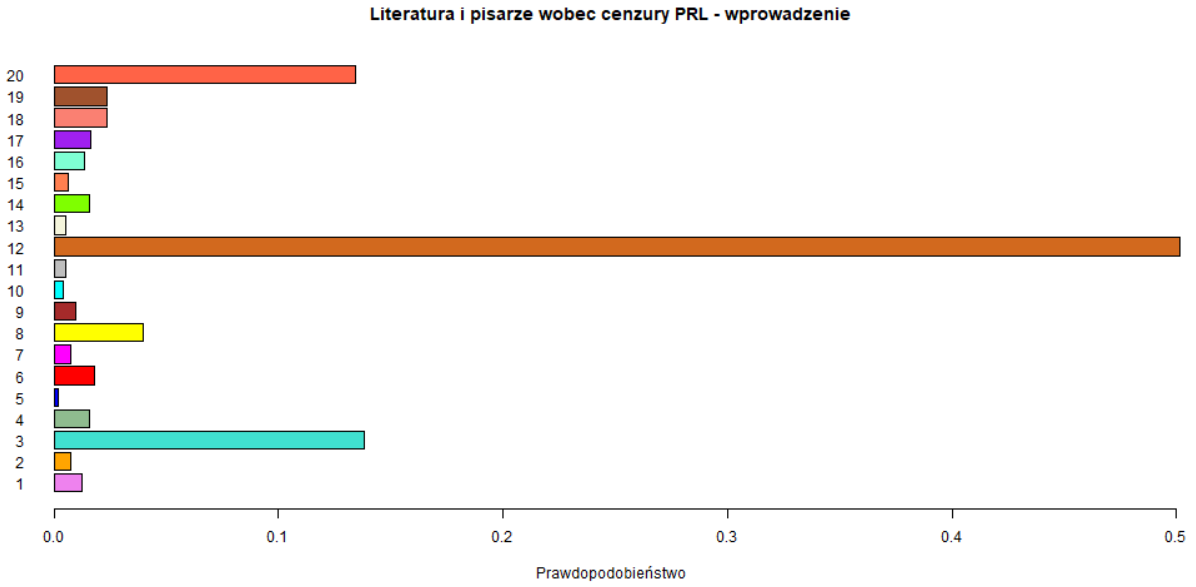
```
> words9 <- c("życie")
> round(experiment2$terms[,words9],20)
            1             2             3             4             5             6             7             8
0.00005530973 0.03659605551 0.00008025682 0.00004662005 0.00689812700 0.00006920415 0.00148305085 0.00006648936
            9            10            11            12            13            14            15            16
0.00004403347 0.00004755112 0.00066425121 0.00007911392 0.07939497717 0.00004807692 0.00006476684 0.00006600660
           17            18            19            20
0.00005117707 0.00004757374 0.00004995005 0.00003321156
```



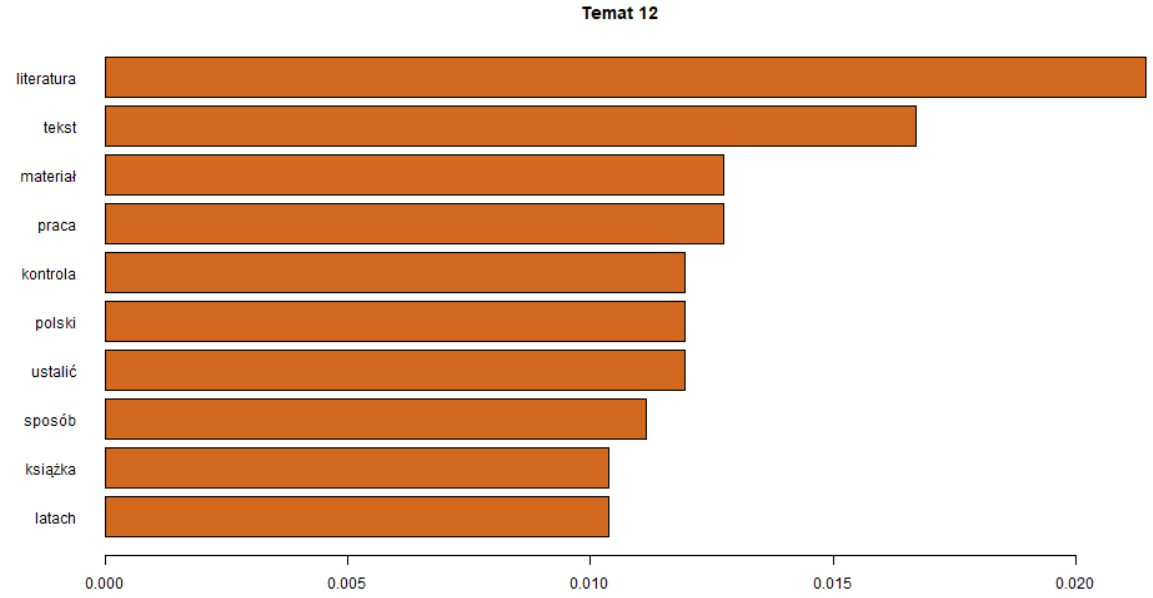
Dokument “Kreowanie wartości poprzez efektywne zarządzanie usługami IT” z tematyki IT uzyskał największe prawdopodobieństwo dopasowania do Tematu 5.



Choć dokument w naszym zamyśle dotyczy tematyki IT, Temat 5 nie wskazuje na nią jednoznacznie - słowa pasują bowiem także do tematyki Ekonomii, jednak wynika to z bliskości tematyki tekstu do obu tych tematyk.

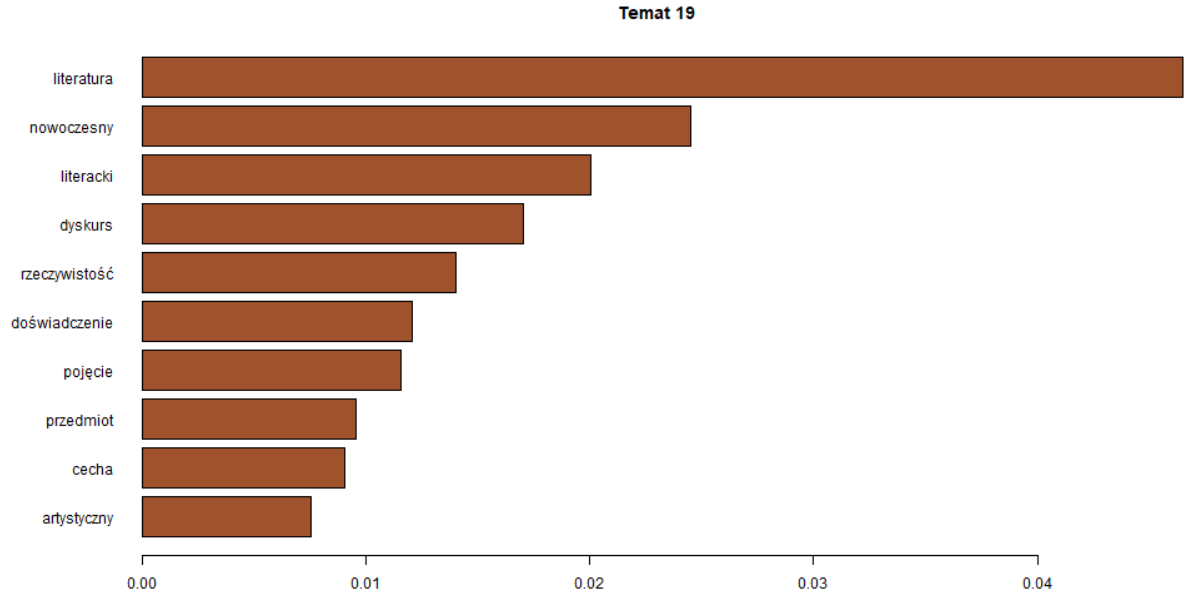


Tekst “Literatura i pisarze wobec cenzury PRL - wprowadzenie” wykazał największe prawdopodobieństwo przynależności do Tematu 12.



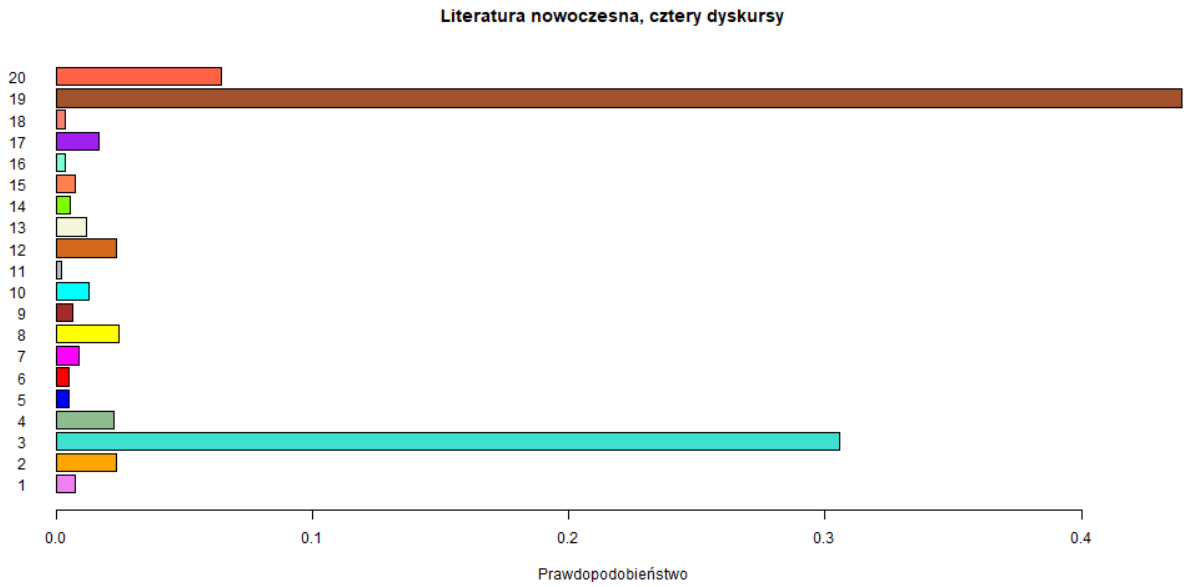
Można zdecydowanie stwierdzić, że rzeczywiście tematyka omawianego tekstu, która przez nas została określona jako Literatura zgadza się z dopasowanym Tematem 12. Warto jednak zwrócić także uwagę na dosyć duże prawdopodobieństwo przynależności tekstu do Tematu 3, który to z kolei opowiadał się raczej po stronie tematyki Ekonomii.

Tekst "Literatura jako trop rzeczywistości - wprowadzenie" uzyskał największe prawdopodobieństwo przynależności do Tematu 19.

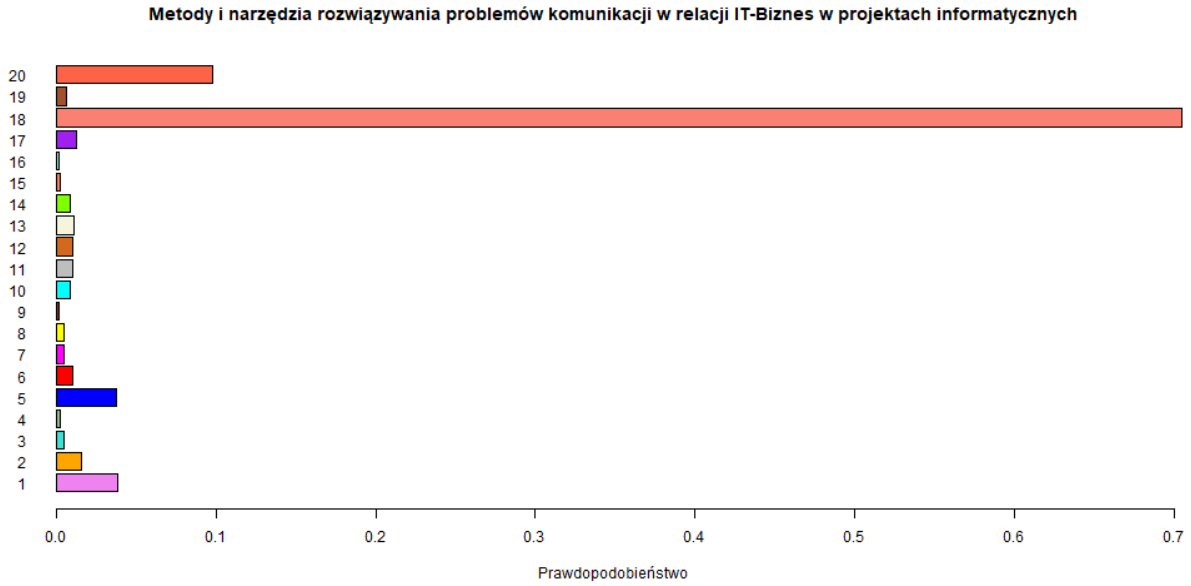


Choć temat ten nieco podobny jest do Tematu 12, tamten miał jednak więcej słów wskazujących na powiązanie konkretnie z folklorem, natomiast Temat 19 w nieco ogólniejszym stopniu dotyczy tematyki Literatury. Podobieństwo może wynikać z dużego prawdopodobieństwa występowania słowa "literatura" w obu Tematach:

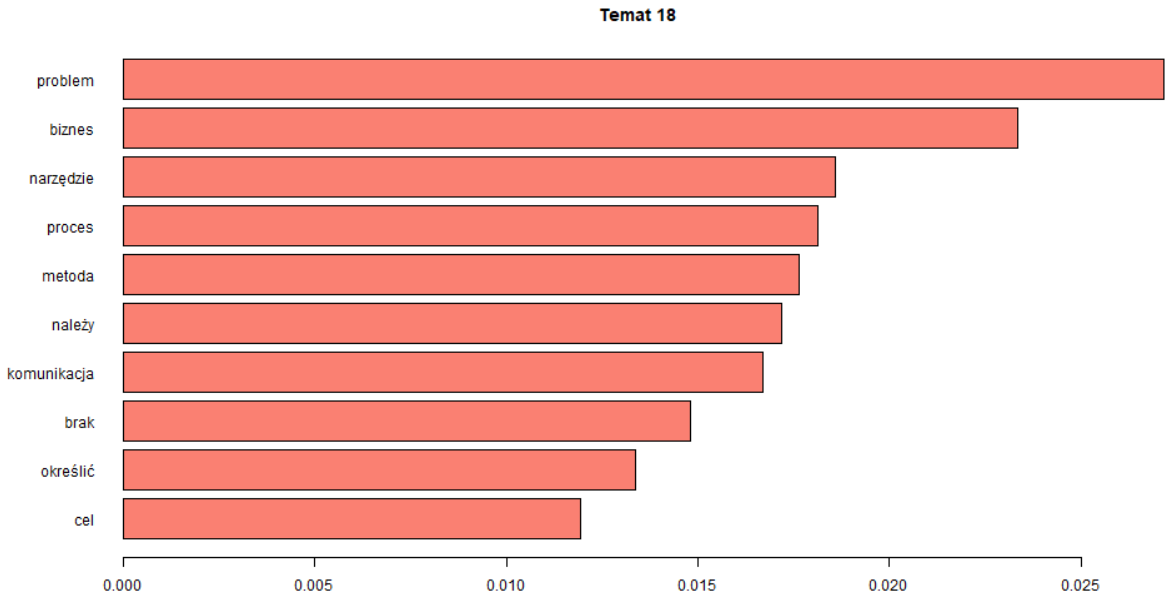
```
> words10 <- c("literatura")
> round(experiment2$terms[,words10],20)
           1            2            3            4            5            6            7            8
0.00005530973 0.00007304602 0.00088282504 0.00004662005 0.00004568296 0.00076124567 0.00007062147 0.00272606383
           9           10           11           12           13           14           15           16
0.00004403347 0.00004755112 0.00006038647 0.02143987342 0.00005707763 0.00004807692 0.00006476684 0.00006600660
          17           18           19           20
0.00005117707 0.00004757374 0.04650349650 0.00003321156
```



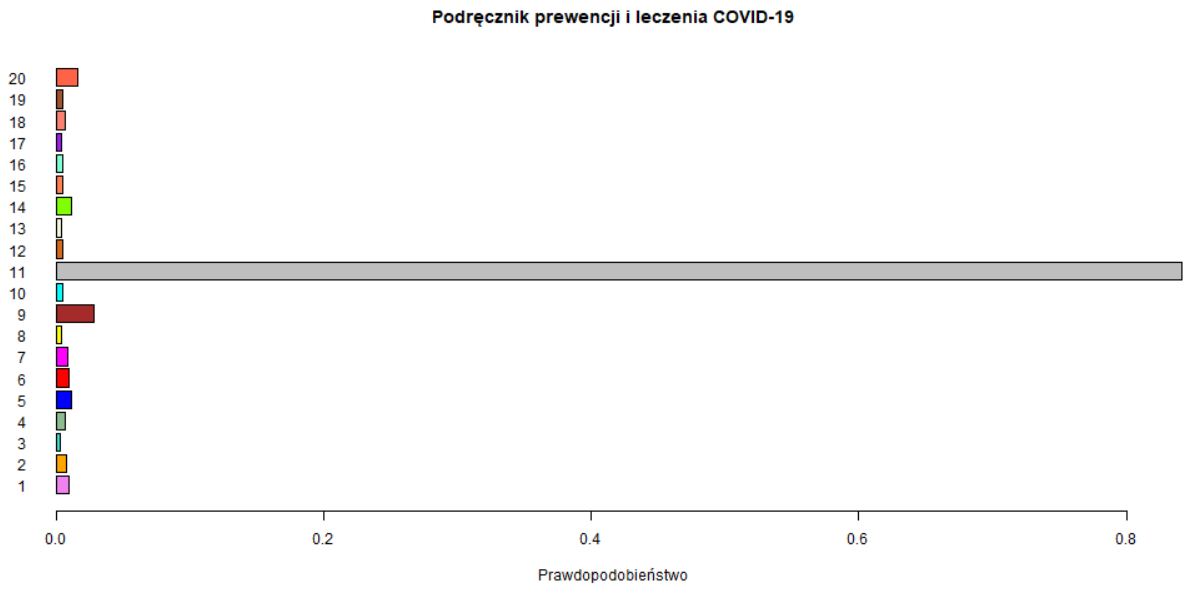
Dokument “Literatura nowoczesna, cztery dyskursy” również wykazuje bardzo wysokie prawdopodobieństwo przynależności do Tematu 19. Jednak ciekawe jest również prawdopodobieństwo na poziomie ok. 0.3 dopasowania do tematu 3, który mogłoby się zdawać dotyczy raczej tematyki Ekonomii, jednak już nie po raz pierwszy zdarza się, że tekst w innej tematyce jest do niego przyporządkowywany.



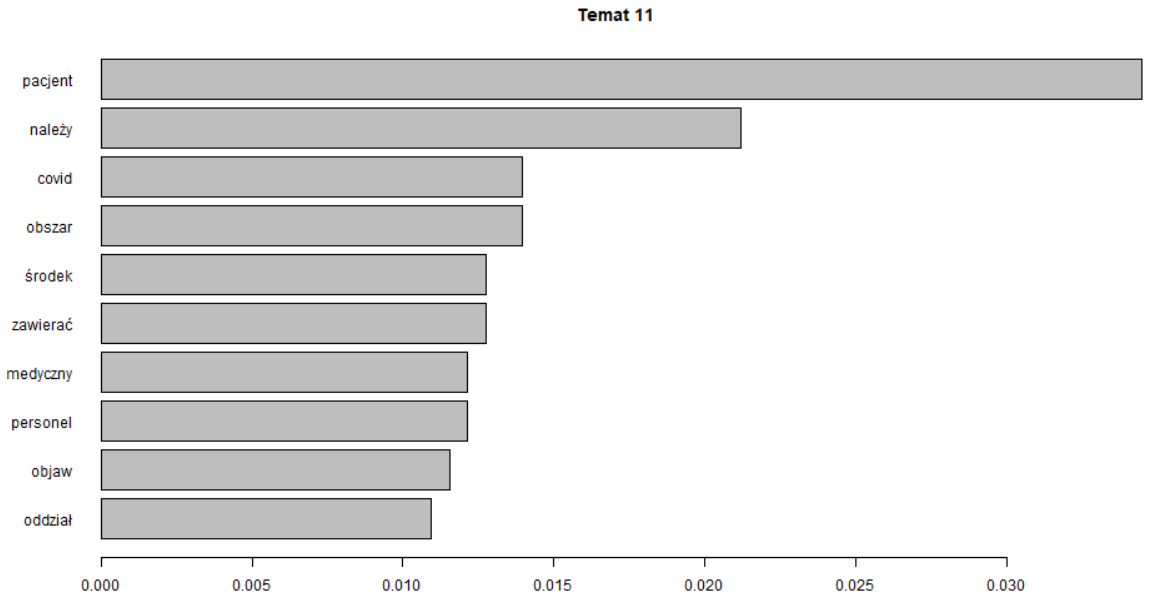
Tekst “Metody i narzędzia rozwiązywania problemów komunikacji w relacji IT-Biznes w projektach informatycznych” został przyporządkowany z największym prawdopodobieństwem do Tematu 18.



Zgodnie z naszym zamierzeniem, tekst należy do tematyki IT, jednak zahacza w pewnym stopniu także o tematykę Ekonomii. Temat 18 obejmuje słowa z obu tych tematyk.

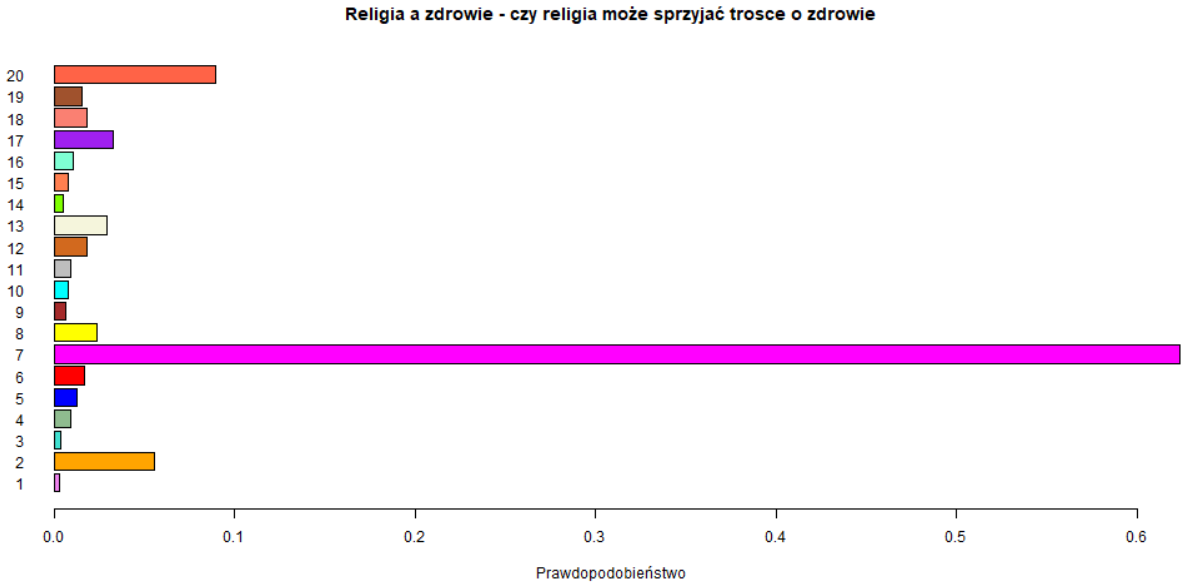


Tekst “Podręcznik prewencji i leczenia COVID-19” uzyskał największe prawdopodobieństwo przyporządkowania do Tematu 11.

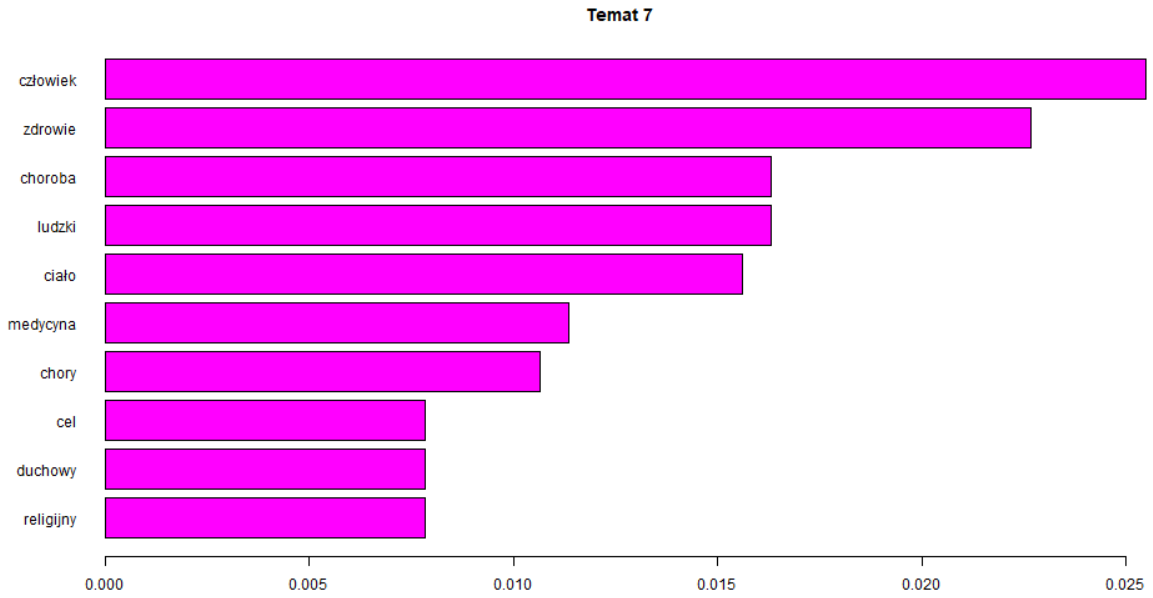


Interesującym jest fakt, że dopasowanie do Tematu 9, który wykazywał bardzo wyraźne ślady odniesienia do tematyki COVID-19 ma bardzo niskie prawdopodobieństwo w przypadku tego tekstu. Wprawdzie w Temacie 11 dosyć często pojawia się słowo “covid”, które nie było w pierwszych dziesięciu słowach w Temacie 9, jednak reszta słów jest zbliżona raczej do tematyki Zdrowia. Co ciekawe, słowo “covid” ma jeszcze większe prawdopodobieństwo występowania w Temacie 15, mimo to omawiany dokument ma prawie zerowe prawdopodobieństwo przyporządkowania do tego tematu.

```
> words11 <- c("covid")
> round(experiment2$terms[,words11],20)
            1             2             3             4             5             6             7             8
0.00005530973 0.00007304602 0.00008025682 0.00004662005 0.00004568296 0.00006920415 0.00007062147 0.00006648936
            9            10            11            12            13            14            15            16
0.00004403347 0.00004755112 0.01394927536 0.00007911392 0.00005707763 0.00004807692 0.01625647668 0.00006600660
           17            18            19            20
0.00005117707 0.00004757374 0.00004995005 0.00003321156
```

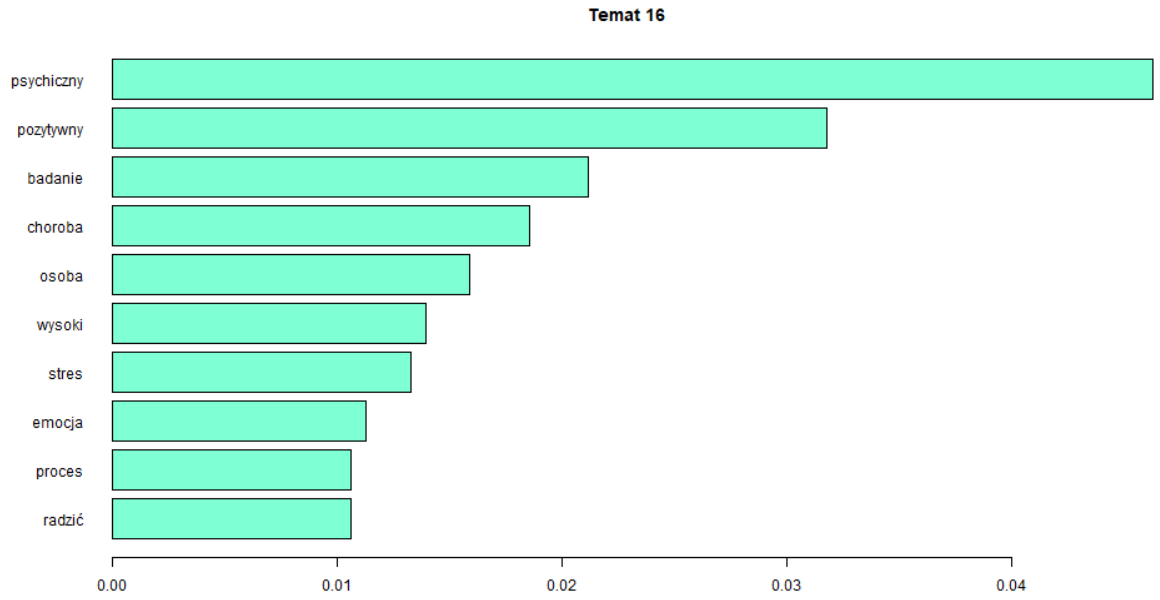


Tekst "Religia a zdrowie - czy religia może sprzyjać trosce o zdrowie", przyporządkowany przez nas do tematyki Zdrowie, uzyskał w eksperymencie największe prawdopodobieństwo dopasowania do Tematu 7.

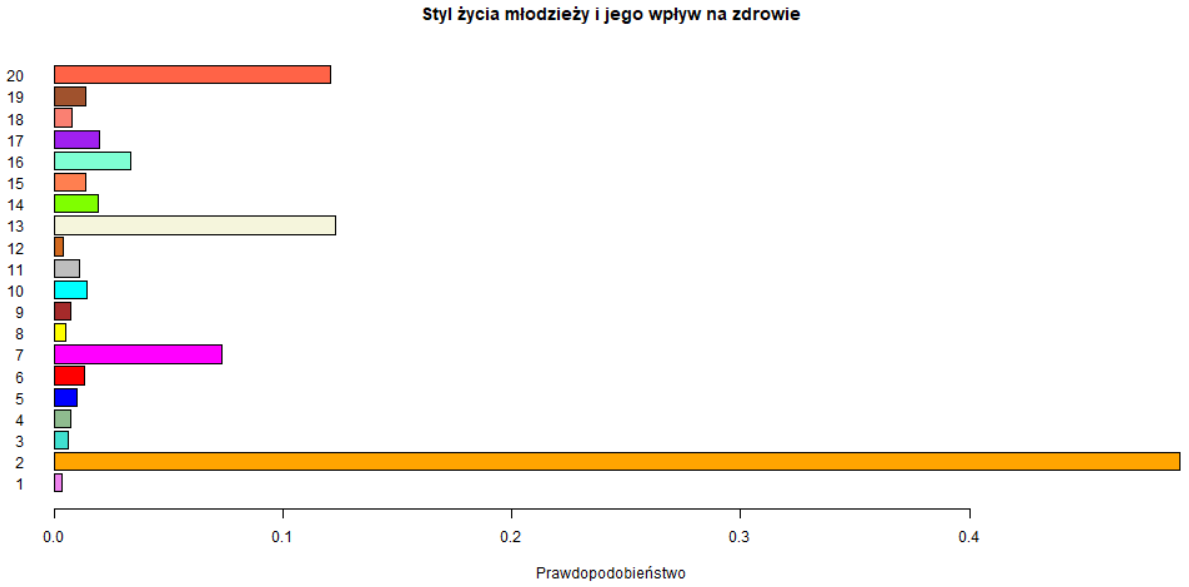


Rzeczywiście, słowa występujące najczęściej w Temacie 7 odnoszą się do tematyki Zdrowia. Dodatkowo pojawiają się takie słowa, jak "duchowy" czy "religijny" co dosyć jednoznacznie wskazuje na tekst o religii i zdrowiu.

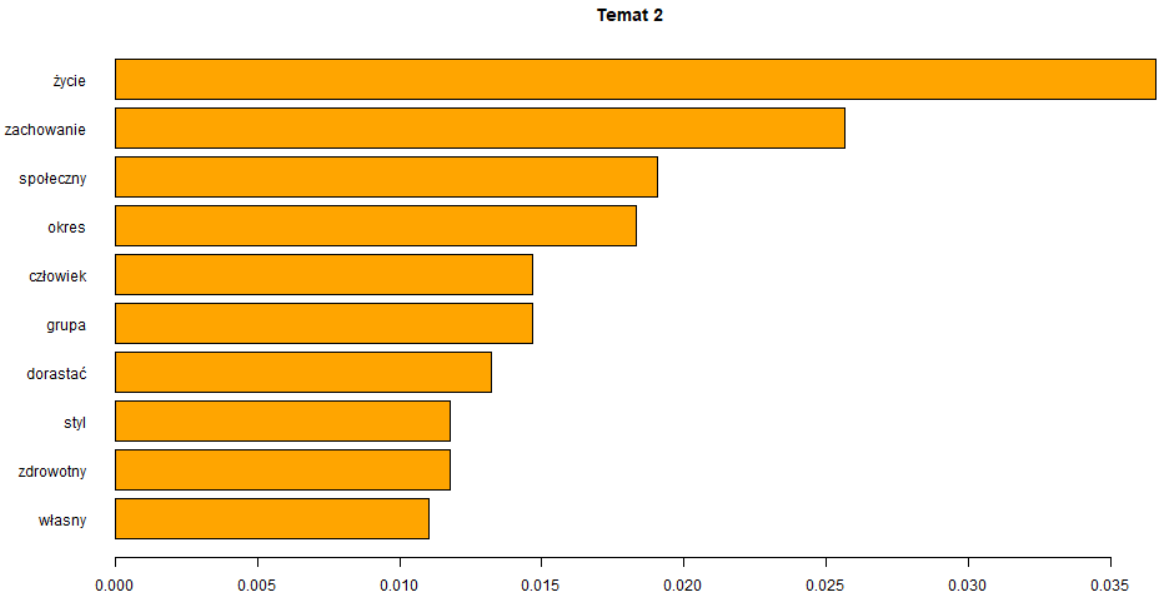
Dokument “Sprężystość psychiczna i zmienne pośredniczące w jej wpływie na zdrowie” z tematyki Zdrowie z największym prawdopodobieństwem przynależy do Tematu 16.



To już kolejny Temat dotyczący tematyki Zdrowie. Rzeczywiście, słowa w nich występujące pasują do omawianego dokumentu, jednak nie dotyczą jego tematyki bardzo szczegółowo.

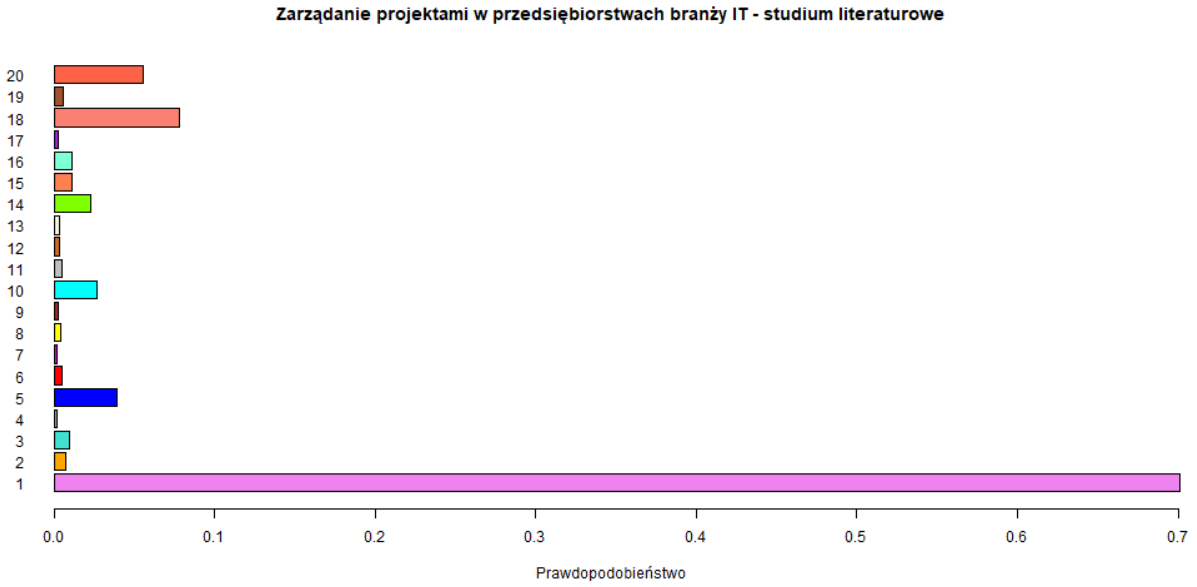


Tekst “Styl życia młodzieży i jego wpływ na zdrowie” przyporządkowany przez nas do tematyki Zdrowie uzyskał największe prawdopodobieństwo przyporządkowania do Tematu 2.

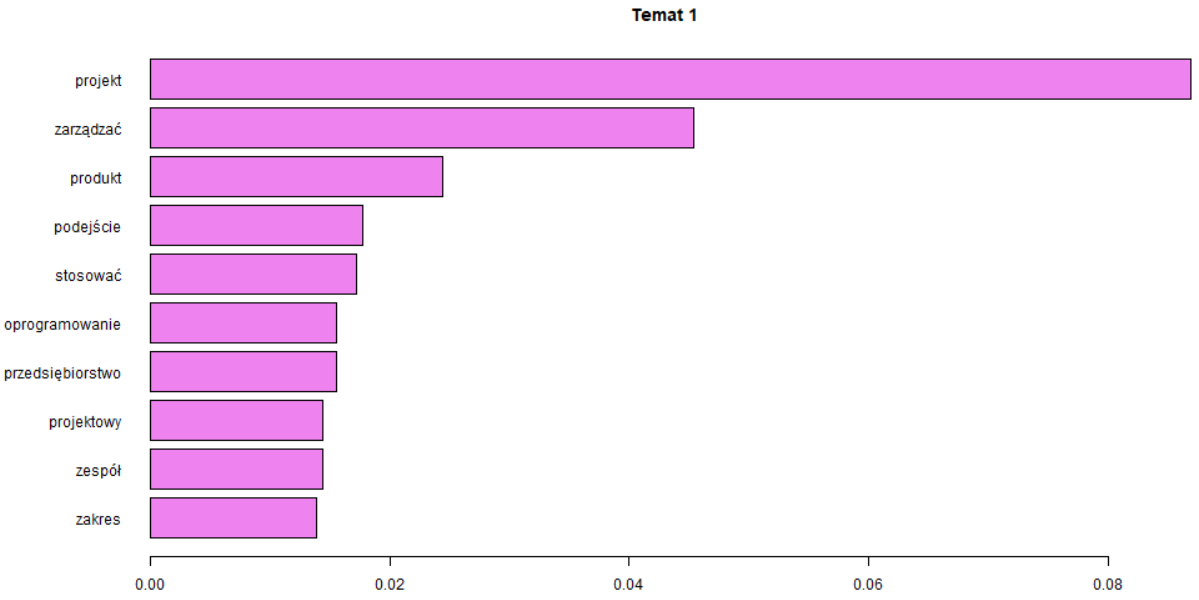


To kolejny temat, z którego częstości występowania słów można wywnioskować, że odnosi się do tematyki Zdrowie. Tym razem jest nieco bardziej ukierunkowany na temat społeczny i można zauważyć też kilka słów jak np. “dorastać”, które mogą bezpośrednio dotyczyć zdrowia młodzieży. Słowo to rzeczywiście występuje praktycznie tylko w Temacie 2:

```
> words12 <- c("dorastać")
> round(experiment2$terms[,words12],20)
            1             2             3             4             5             6             7             8
0.00005530973 0.01322132944 0.00008025682 0.00004662005 0.00004568296 0.00006920415 0.00007062147 0.00006648936
            9            10            11            12            13            14            15            16
0.00004403347 0.00004755112 0.00006038647 0.00007911392 0.00005707763 0.00004807692 0.00006476684 0.00006600660
           17            18            19            20
0.00005117707 0.00004757374 0.00004995005 0.00003321156
```

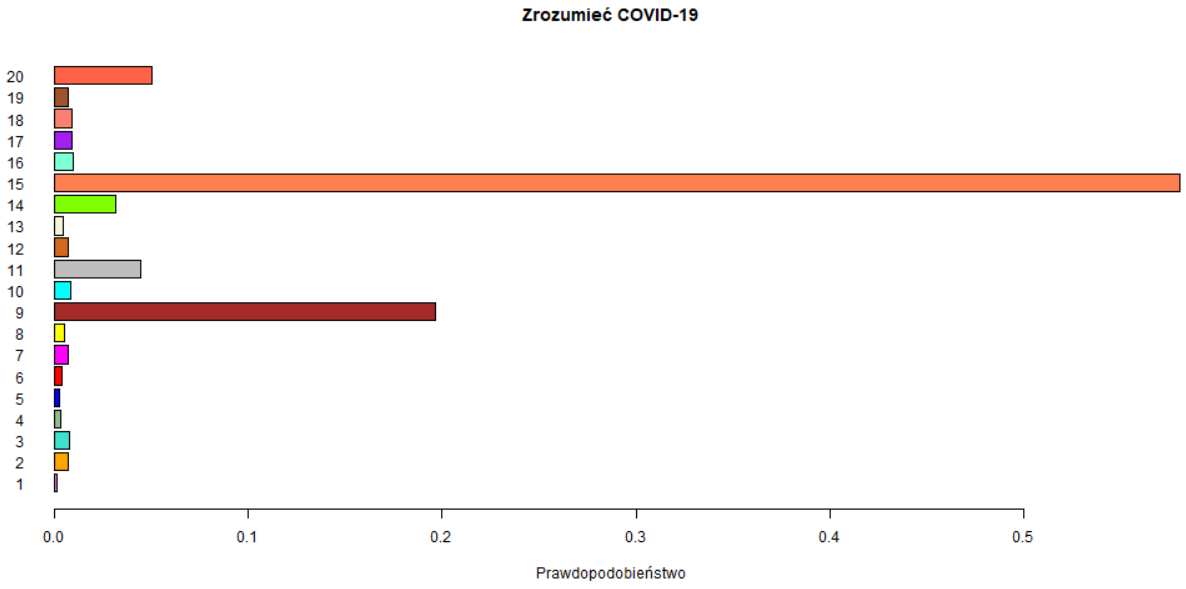


"Zarządzanie projektami w przedsiębiorstwach branży IT - studium literaturowe" - tekst ten wywodzi się z wybranej przez nas tematyki IT, a największe prawdopodobieństwo przyporządkowania wykazał do Tematu 1.

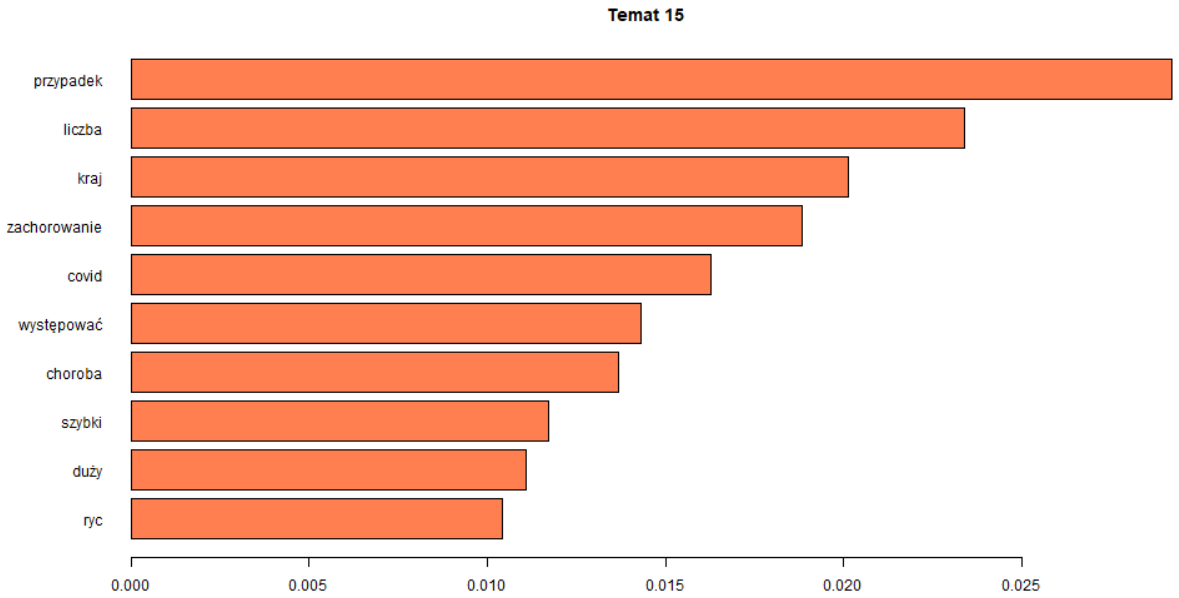


Temat 1 nie bardzo odróżnia się od innych tematów dotyczących tematyk IT i Ekonomii. Z dwóch najczęściej występujących słów: “projekt” oraz “zarządzać”, to drugie występuje także dosyć często w Temacie 5:

```
> words13 <- c("projekt", "zarządzać")
> round(experiment2$terms[,words13],20)
          projekt     zarządzać
1  0.08689159292 0.04540929204
2  0.00007304602 0.00007304602
3  0.00008025682 0.00008025682
4  0.00004662005 0.00004662005
5  0.00004568296 0.02882594792
6  0.00006920415 0.00006920415
7  0.00148305085 0.00007062147
8  0.00006648936 0.00006648936
9  0.00004403347 0.00004403347
10 0.00004755112 0.00955777461
11 0.00006038647 0.00489130435
12 0.00007911392 0.00007911392
13 0.00005707763 0.00005707763
14 0.00004807692 0.00004807692
15 0.00006476684 0.00006476684
16 0.00006600660 0.00006600660
17 0.00005117707 0.00005117707
18 0.00670789724 0.00004757374
19 0.00004995005 0.00004995005
20 0.00003321156 0.00003321156
```

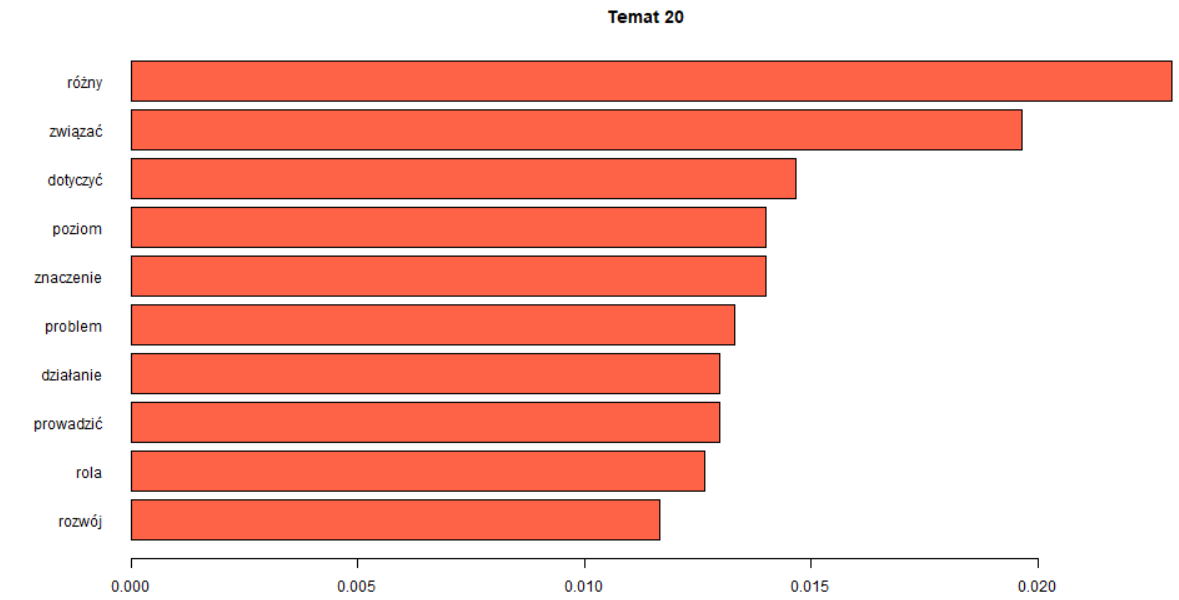


Ostatni omawiany dokument noszący tytuł “Zrozumieć COVID-19” został z największym prawdopodobieństwem przyporządkowany do Tematu 15.



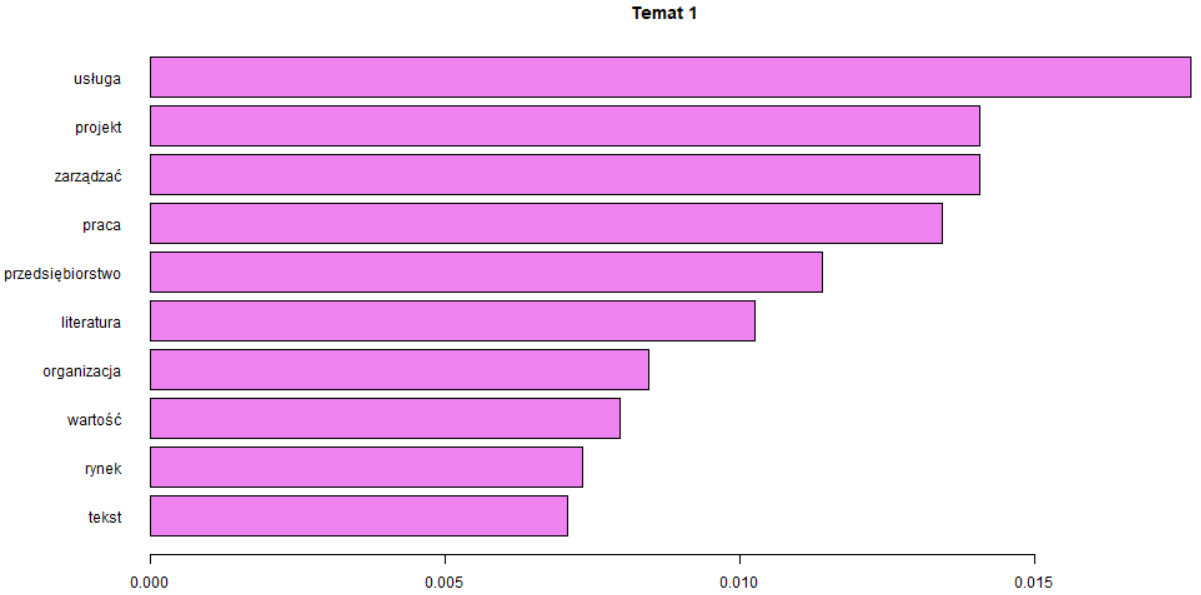
Słowa występujące w tym temacie dosyć jednoznacznie wskazują na powiązanie z tematyką COVID-19, co zgadza się jak najbardziej z tematyką omawianego tekstu.
Warto również zwrócić uwagę na znaczne prawdopodobieństwo dopasowania dokumentu do Tematu 9, który to również dotyczył pandemii.

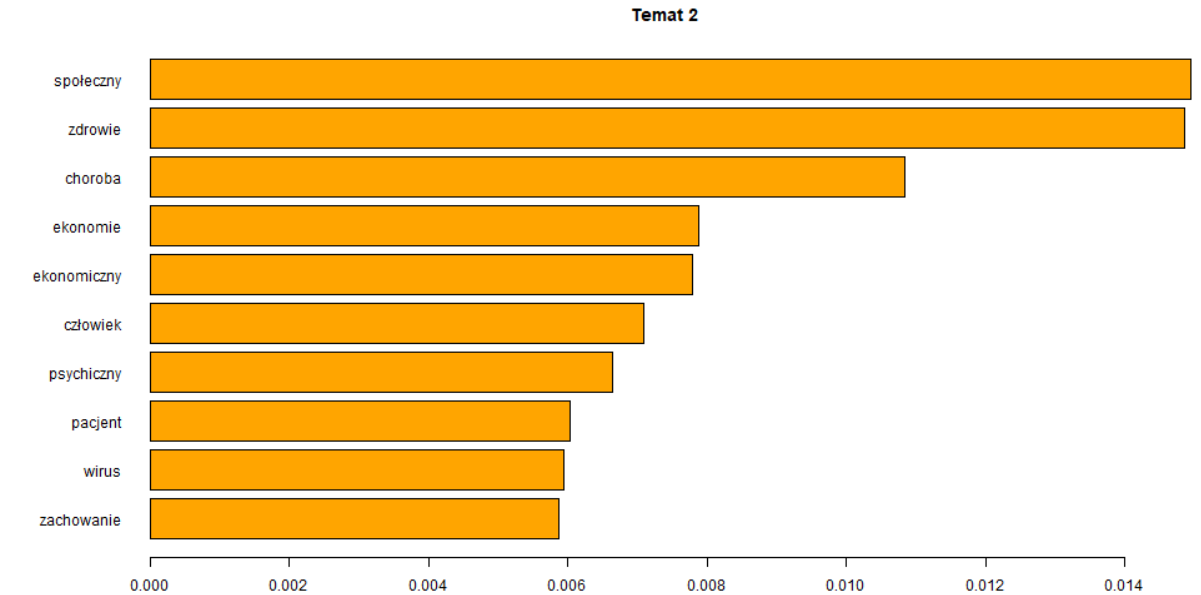
Podsumowując ten eksperyment, warto zauważyć, że większość tekstów wykazała znaczne prawdopodobieństwo przynależności do Tematu 20, co wskazuje na neutralność słów w nim występujących:



b. Eksperyment trzeci:
    - DTM_Tf_3_14,
    - 2 tematy.

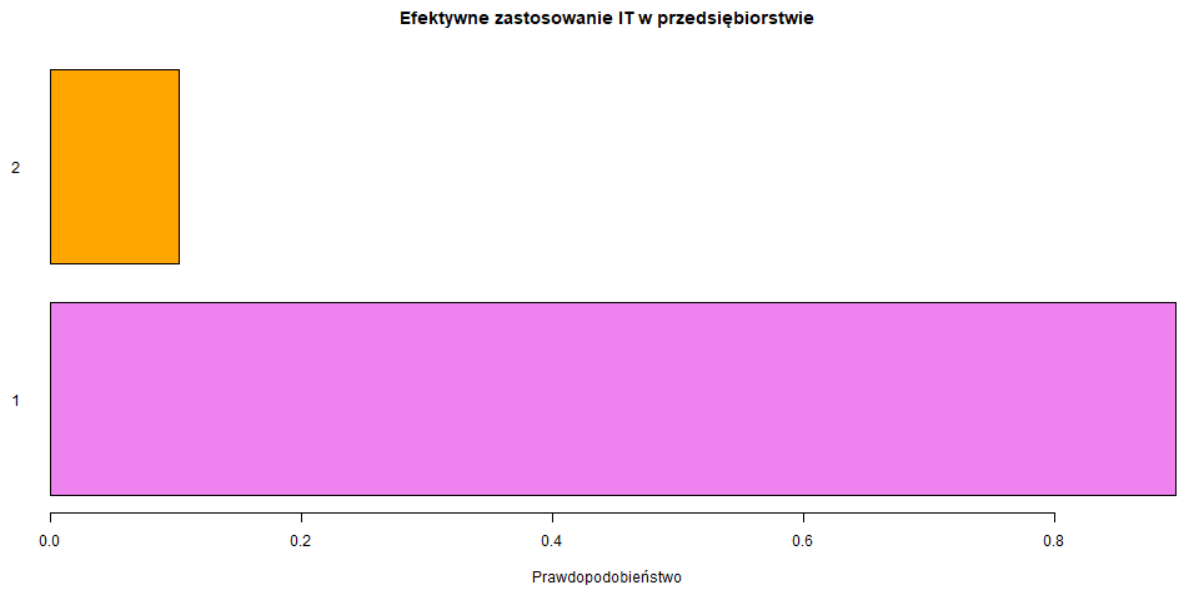
Trzeci eksperyment zawierał jedynie 2 tematy:



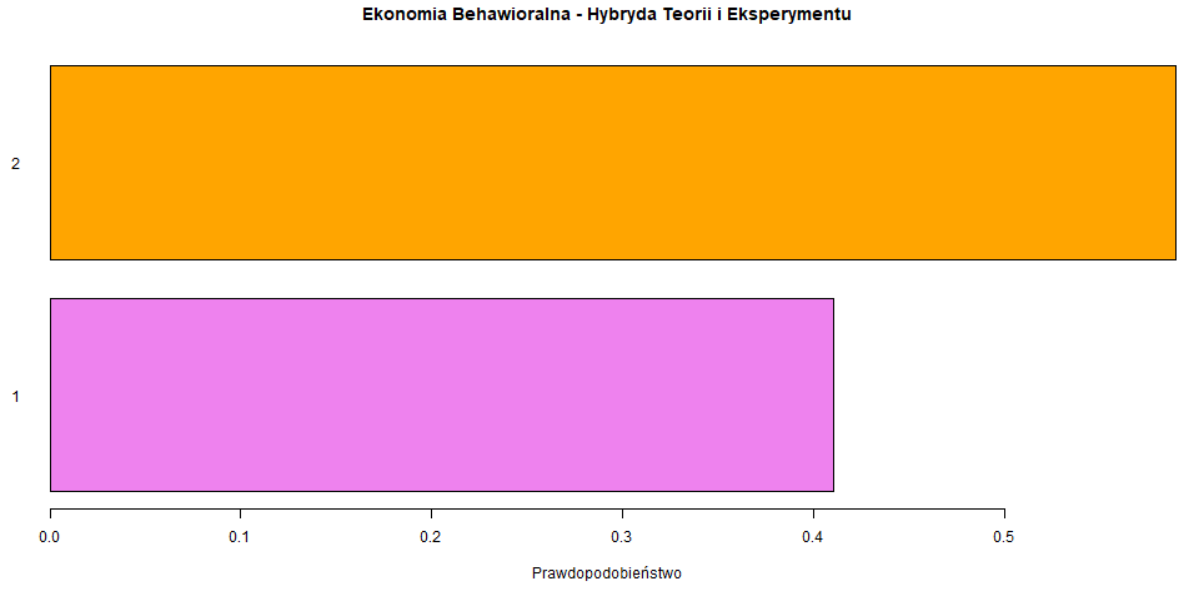


Temat 1 jest zbliżony do tematyk IT oraz Ekonomii, natomiast Temat 2 do tematów Zdrowia, COVID-19 ale także Ekonomii. Ciężko jednoznacznie powiedzieć, do którego z tematów pasowałaby tematyka Literatura, jednak występowanie w Temacie 1 słowa “literatura” mogłoby sugerować właśnie ten temat.

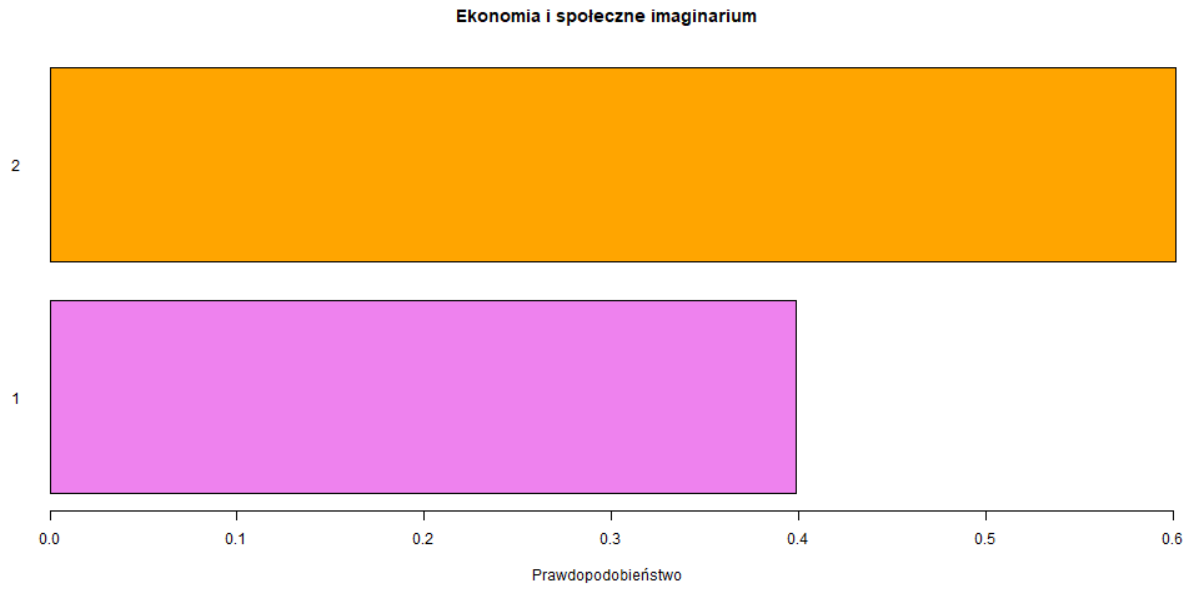
Przeanalizujemy teraz każdy dokument i jego przynależność do Tematu 1 i 2.



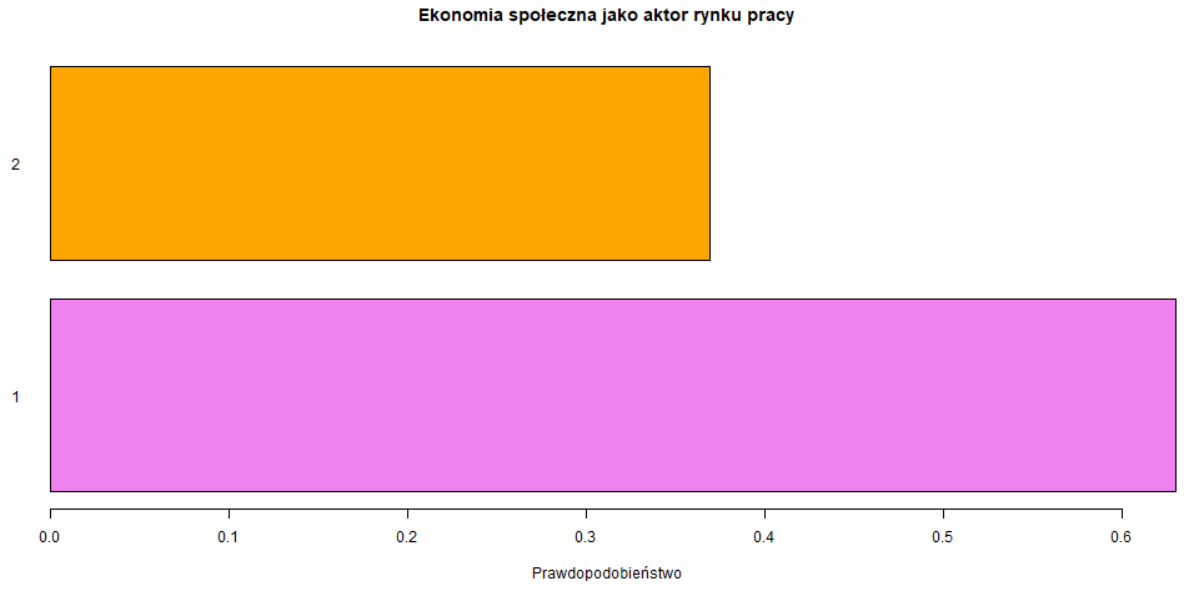
Zgodnie z przewidywaniami, tekst o efektywnym zastosowaniu IT w przedsiębiorstwie został przyporządkowany z niemal pewnym prawdopodobieństwem do Tematu 1.



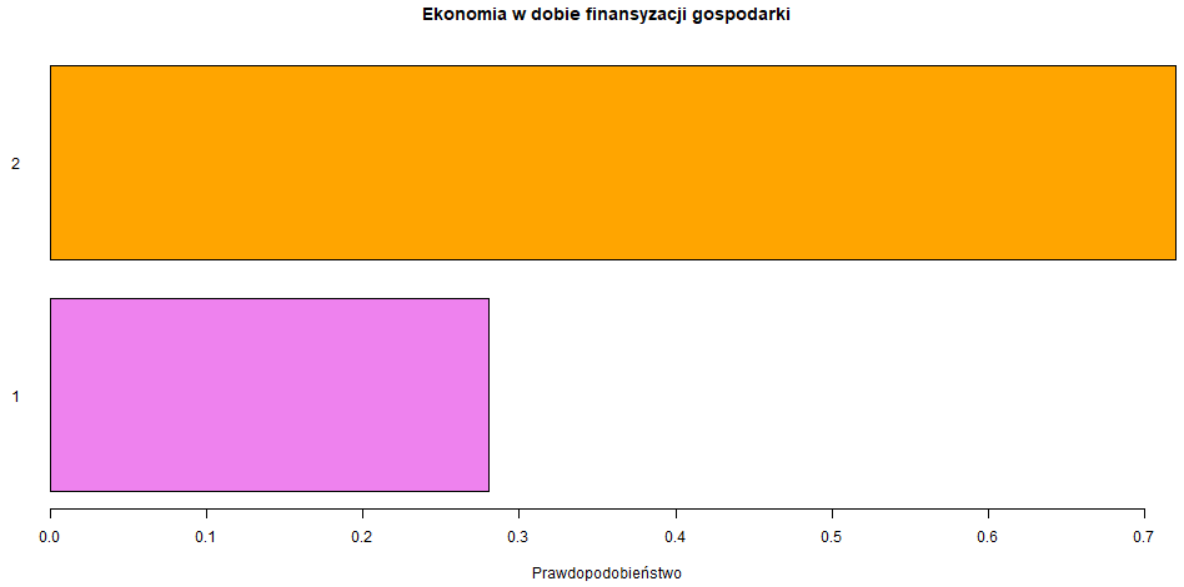
Dokument o ekonomii behawioralnej, mimo że wywodził się z tematyki Ekonomia, został z większym prawdopodobieństwem przyporządkowany do Tematu 2. Może to wynikać z nieco “społecznego” charakteru tego tekstu. Jednak prawdopodobieństwo przyporządkowania do Tematu 1 jest wciąż wysokie.



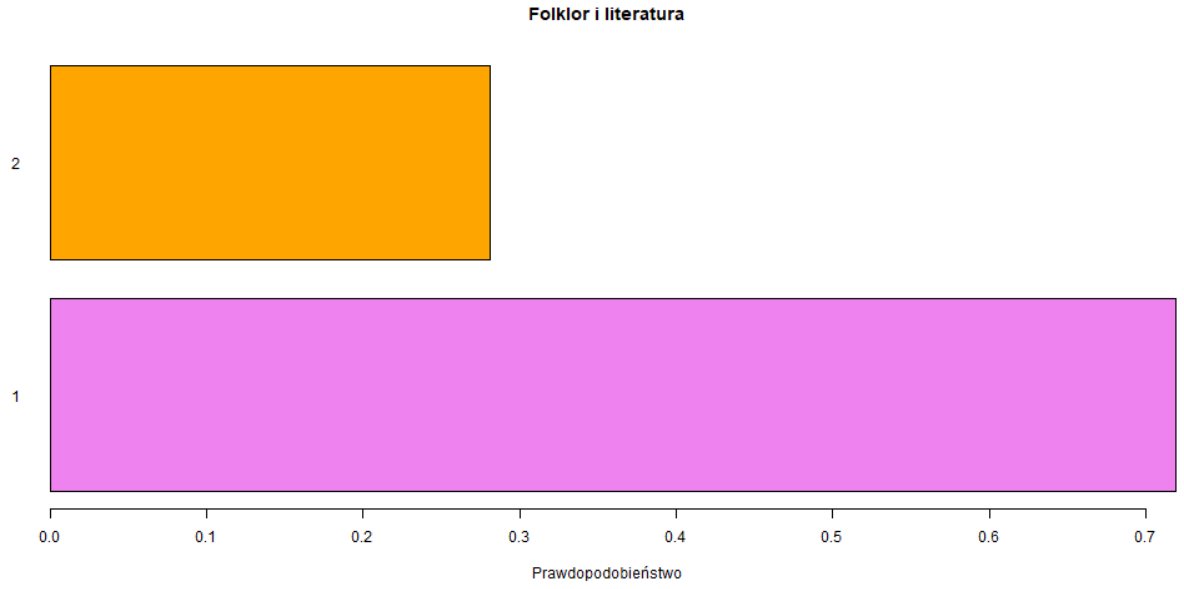
Bardzo podobny wynik uzyskał dokument “Ekonomia i społeczne imaginarium”.



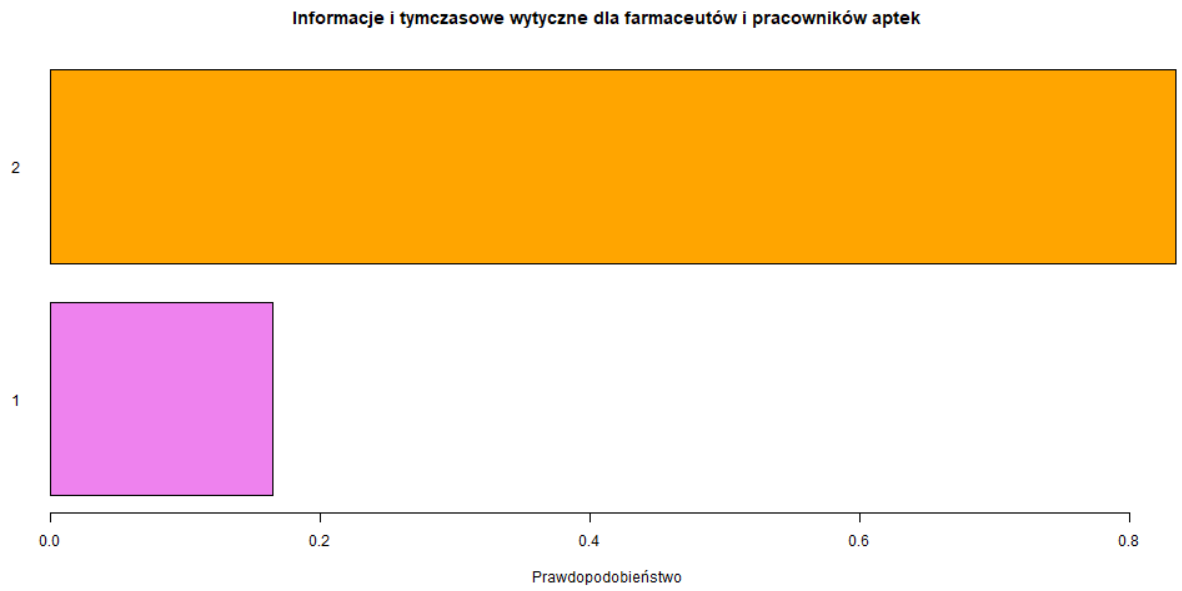
Następnie tekst “Ekonomia społeczna jako aktor rynku pracy”, mimo również “społecznego” charakteru skłania się w stronę Tematu 1.



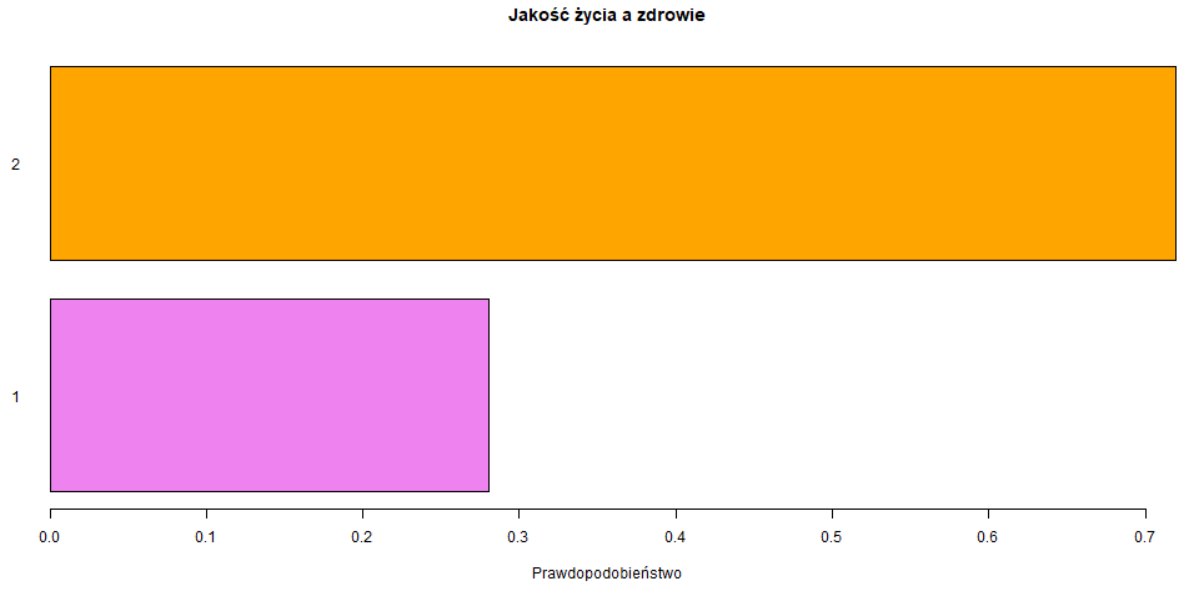
Z kolei "Ekonomia w dobie finansyzacji gospodarki" ponownie odwraca trend i wykazuje większe prawdopodobieństwo przynależności do Tematu 2.



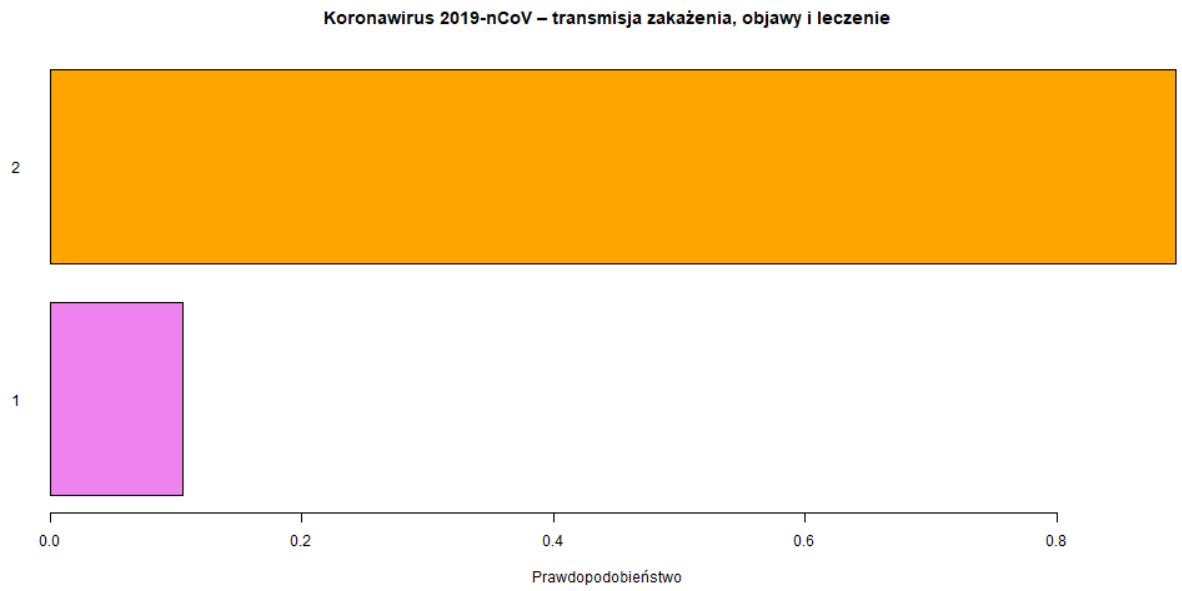
Temat 1 wykazywał większe powiązania z tematyką Literatura, prawdopodobnie w związku z tym tekst "Folklor i literatura" został przyporządkowany do niego z większym prawdopodobieństwem.



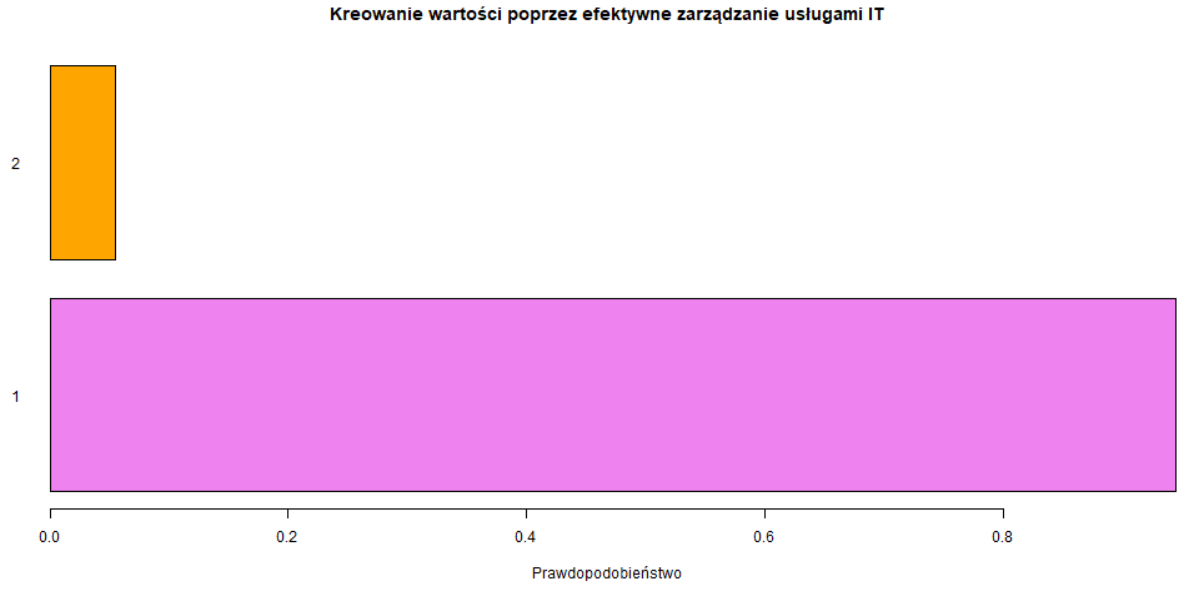
Dopasowana przez nas tematyka dokumentu “Informacje i tymczasowe wytyczne dla farmaceutów i pracowników aptek” to COVID-19. Tekst ten wykazuje bardzo wysokie prawdopodobieństwo do Tematu 2, co rzeczywiście zgadzałoby się z początkowymi spostrzeżeniami.



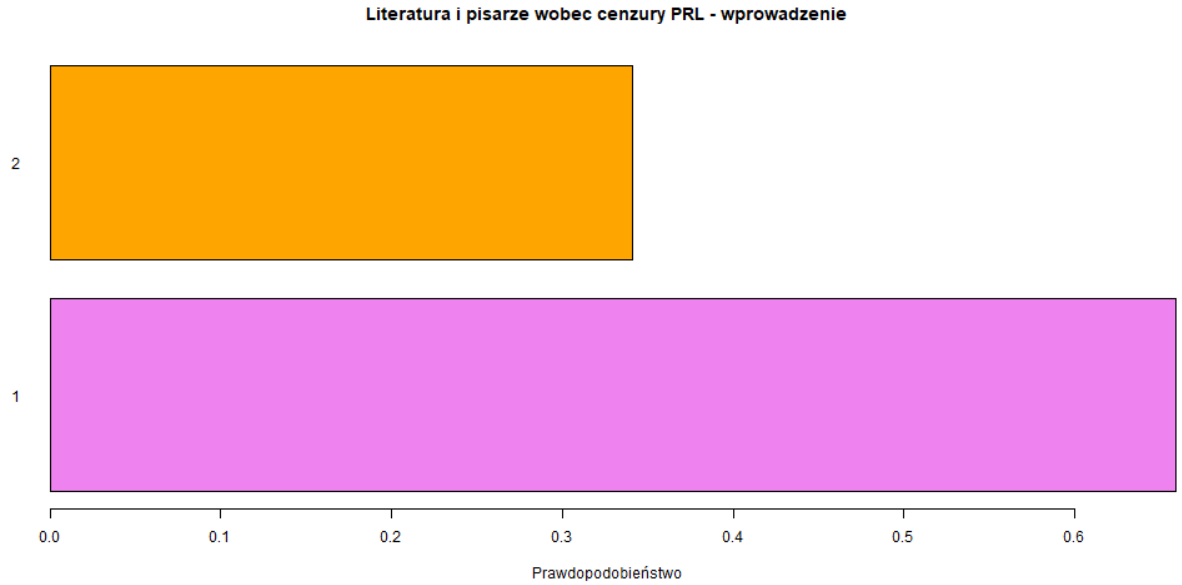
Podobnie jest w przypadku tekstu “Jakość życia i zdrowie” z tematyki Zdrowie. Prawdopodobieństwo przyporządkowania do Tematu 1 jest tu nieco wyższe niż w przypadku poprzedniego tekstu, jednak wciąż przeważa dopasowanie do Tematu 2.



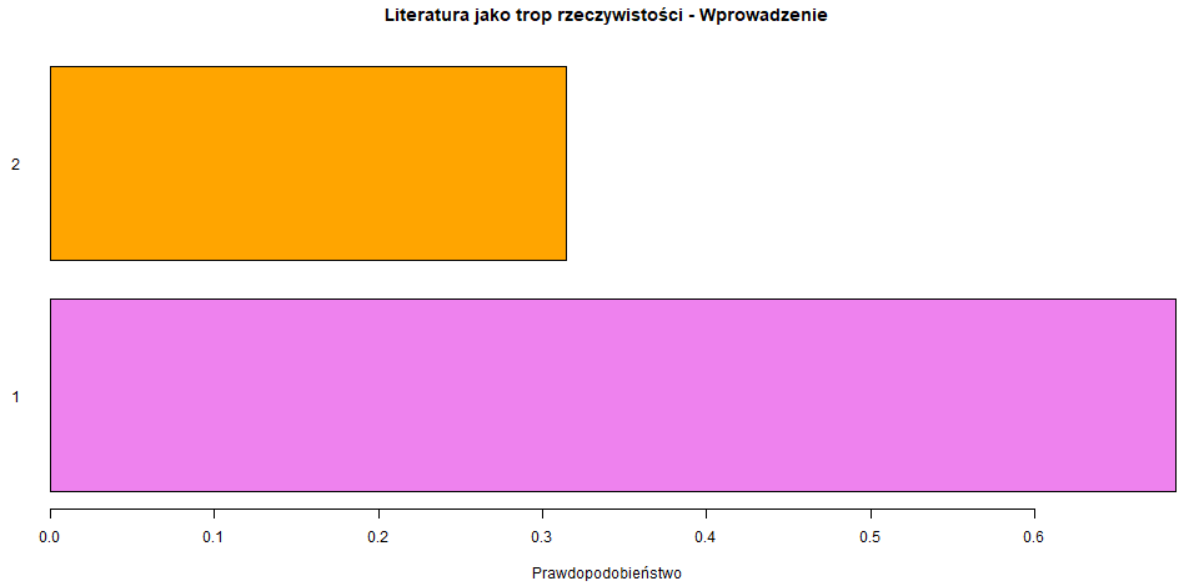
Kolejny tekst w tematyce COVID-19 również został dopasowany z większym prawdopodobieństwem do Tematu 2.

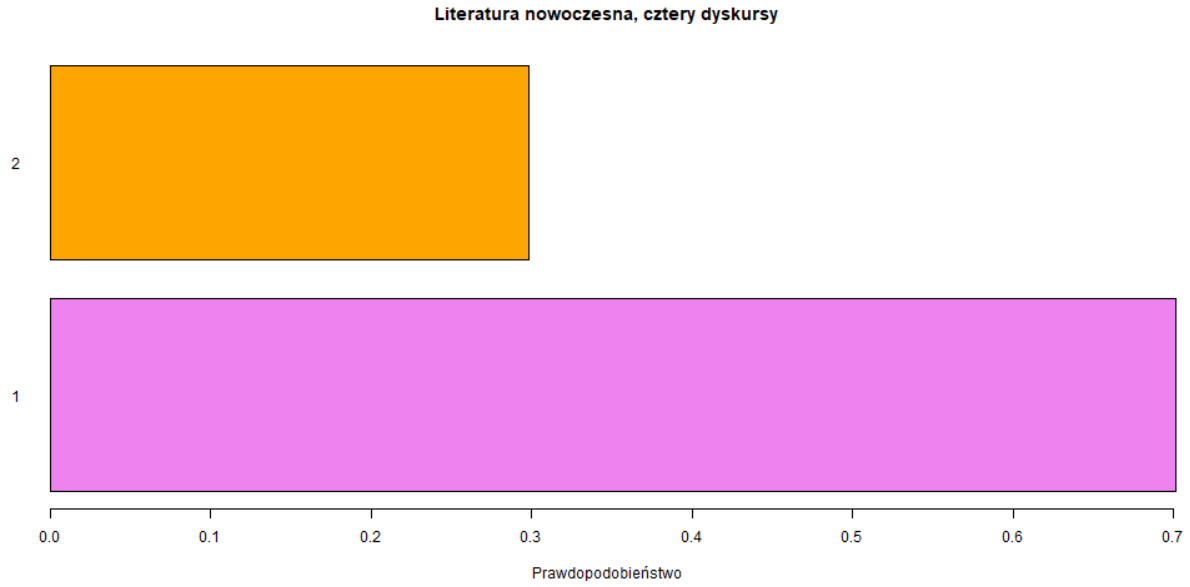


Przy tekście o tematyce IT eksperyment nie wykazał praktycznie żadnych wątpliwości co do jego przynależności do Tematu 1.

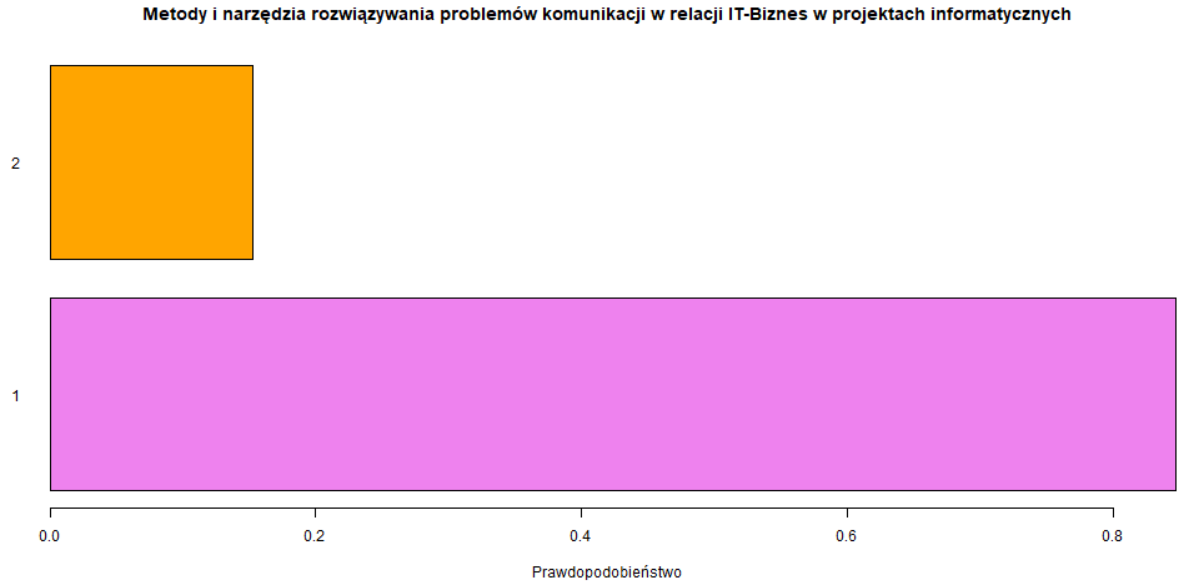


Dokument “Literatura i pisarze wobec cenzury PRL - wprowadzenie” został z większym prawdopodobieństwem przyporządkowany do Tematu 1, co rzeczywiście zgadza się z przewidywaniami, jednak wciąż prawdopodobieństwo dopasowania tekstu do Tematu 2 nie jest znikome.

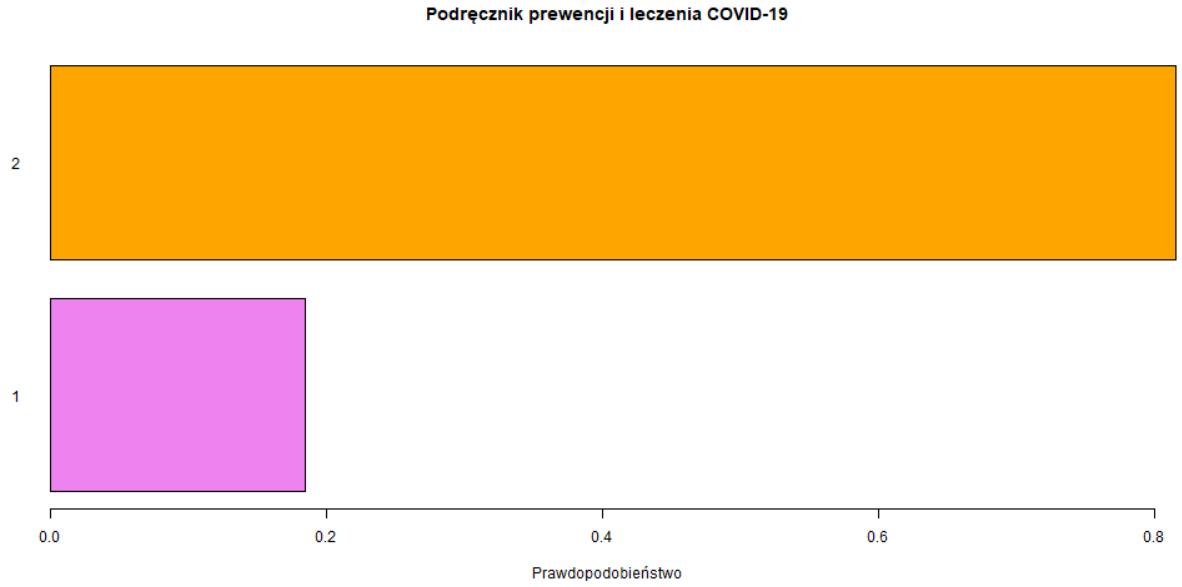




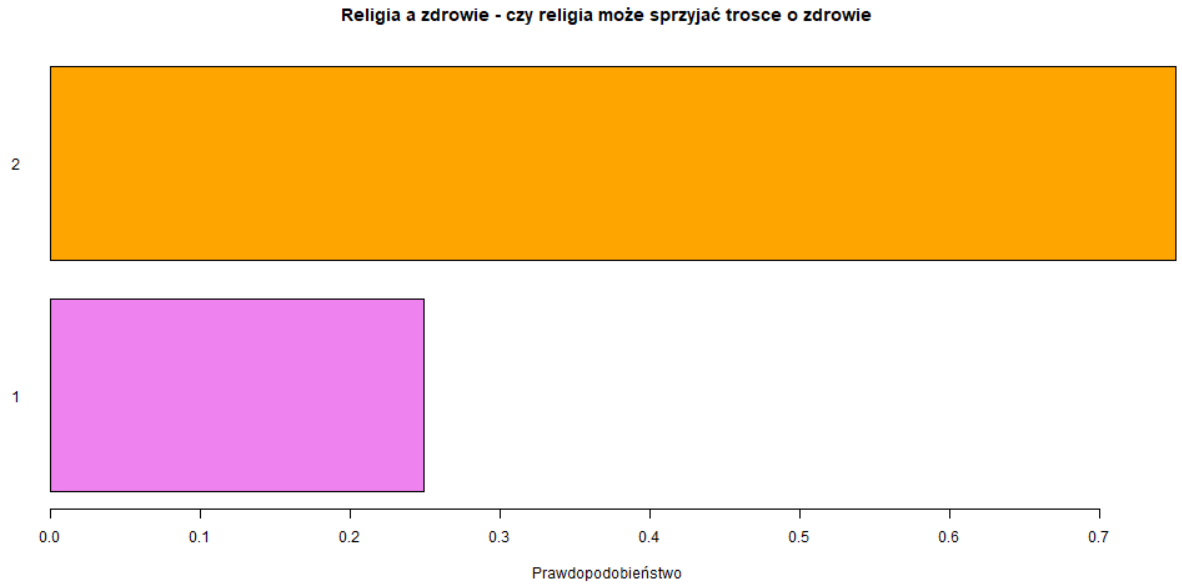
Bardzo podobnie jest w przypadku tekstów “Literatura jako trop rzeczywistości” oraz “Literatura nowoczesna - cztery dyskursy” - skłaniają się one w stronę Tematu 1, jednak wciąż nie wykluczają Tematu 2 jako dopasowania.

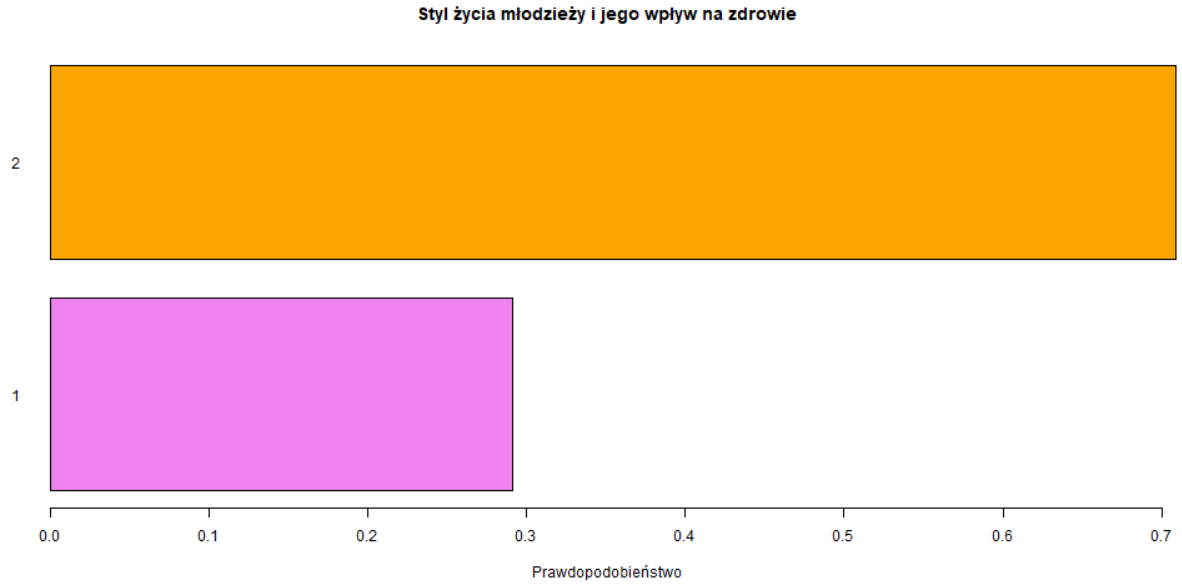


Tekst o metodach i narzędziach rozwiązywania problemów komunikacji w relacji IT-Biznes w projektach informatycznych zdecydowanie został dopasowany do Tematu 1 traktującego o IT i Ekonomii.

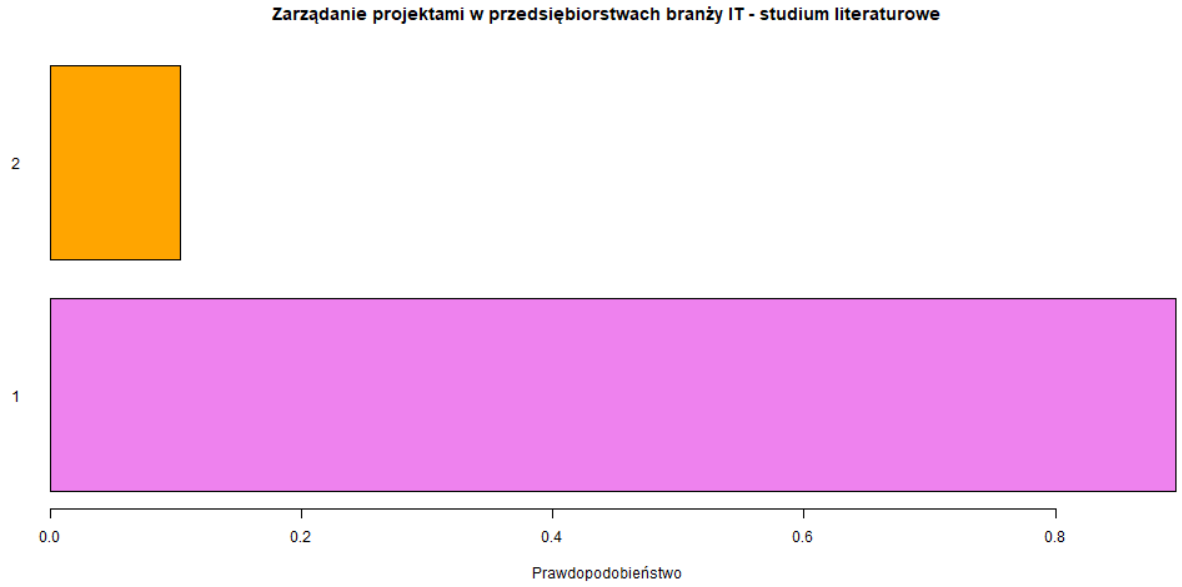


Również zgodnie z przewidywaniami, “Podręcznik prewencji i leczenia COVID-19” uzyskał wysokie prawdopodobieństwo przynależności do Tematu 2.

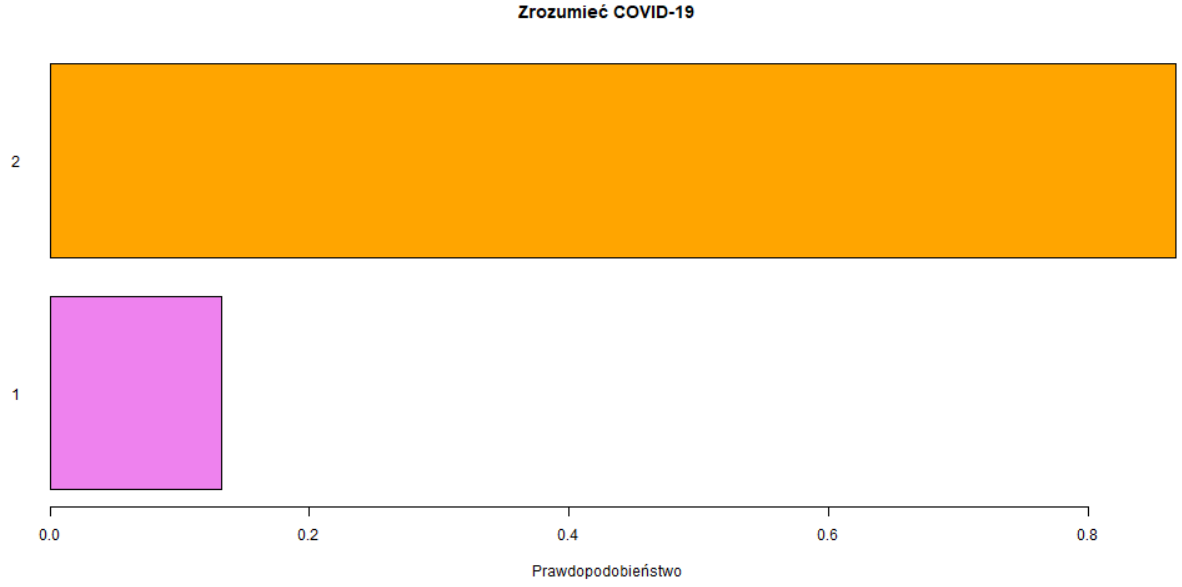




Tak samo w przypadku trzech powyższych tekstów o tematyce Zdrowie - wszystkie uzyskały wysoki poziom prawdopodobieństwa przynależności do Tematu 2.



Zdecydowaną przewagę jeżeli chodzi o przynależność tekstu o zarządzaniu projektami w przedsiębiorstwach branży IT do tematu ma Temat 1.



Ostatni tekst o tematyce COVID-19 został z bardzo dużym prawdopodobieństwem przyporządkowany do Tematu 2, czyli tematu, w którym występowały takie słowa jak “zdrowie”, “choroba” czy “wirus”.

6. Słowa/frazy kluczowe

Analiza słów kluczowych została przeprowadzona dla trzech różnych miar ważności słów:

- waga Tf,
- waga TdIdf,
- prawdopodobieństwo w modelu LDA (na podstawie eksperymentu 1 z poprzedniego podpunktu dla macierzy DTM_Tf_NoBounds i 4 tematów)

- Eksperyment 1
    - macierz DTM z wagami Tf, bez granic

```
> #eksperyment 1, macierz Tf bez granic
> weightExperiment(DTM_Tf_NoBounds_Matrix)
[1] "Dokument 1"
przedsiębiorstwo            koszt            wpływ        zarządzać        biznesowy        strategia
              92               38               29               26               23               23
[1] "Dokument 2"
behawioralny     ekonomie  ekonomiczny       teoria   zachowanie     ekonomia
          29           28           22           21           21           20
[1] "Dokument 3"
  społeczny      wiedza   działanie imaginarium    określić      zmiana
          58           25           21           21           16           16
[1] "Dokument 4"
      praca       osoba organizacja       rynek       grupa pozarządowy
         115           79           68           61           50           47
[1] "Dokument 5"
   ekonomie ekonomiczny    ekonomia       nauka       sfera   społeczny
         48           43           30           30           22           22
[1] "Dokument 6"
       folklor           tekst folklorystyczny      literatura         kultura          reguła
            66              35              21              18              17              15
[1] "Dokument 7"
  choroba   sarscov     wirus oddechowy     osoba  epidemia
        22        22        21        17        17        16
[1] "Dokument 8"
      życie     zdrowie      jakość     badanie      poczuć zadowolenie
         93           83           61           22           21           18
[1] "Dokument 9"
    wirus przypadek zakażenie     dzień  infekcja      ncov
       31        28        19        18        18        17
[1] "Dokument 10"
    usługa  zarządzać dostarczać    wartość       faza     klient
       185         56         37         35         29         24
```

```
[1] "Dokument 11"
   cenzura literatura      tekst      praca   materiał    gukppiw
         23         22         21         15         14         13
[1] "Dokument 12"
   literatura rzeczywistość    nowoczesny     literacki doświadczenie         tekst
           39            30            24            18            16            13
[1] "Dokument 13"
literatura nowoczesny    dyskurs  pisarstwo  literacki      model
         38         22         21         18         17         17
[1] "Dokument 14"
     biznes     problem   narzędzie komunikacja      metoda      należy
         49          38          31          30          26          24
[1] "Dokument 15"
    pacjent      należy dezynfekcja       covid       chlor    personel
         39          31          22          21          20          19
[1] "Dokument 16"
   zdrowie cierpienie   człowiek      ciało    choroba     ludzki
         33         30         30         21         20         19
[1] "Dokument 17"
sprężystość  psychiczny   pozytywny     badanie     choroba       osoba
         62          60          41          30          22          18
[1] "Dokument 18"
     życie    zdrowie      okres  społeczny zachowanie   dorastać
         44         23         21         20         19         17
[1] "Dokument 19"
      projekt     zarządzać      metodyka informatyczny     podejście        zwinny
          142            68            51            30            28            28
[1] "Dokument 20"
      liczba         kraj      choroba        covid    przypadek zachorowanie
          31           27           26           26           26           22
```

Dokumenty 1, 10, 14 oraz 19 pochodzą z tematyki IT. W trzech z nich (1, 10 i 19) dosyć często (odpowiednio 26, 56, 68 razy) pojawia się słowo “zarządzać”. Rzeczywiście te teksty mają charakter nieco “ekonomiczny”, w przeciwieństwie do tekstu nr 14, który skupia się na komunikacji i biznesie (słowa te występują w tym dokumencie odpowiednio 30 i 49 razy).

Dokumenty 2, 3, 4 oraz 5 pochodzą z tematyki Ekonomia. W dokumencie 2 traktującym o ekonomii behawioralnej, słowem występującym najwięcej (29) razy, jest “behawioralny”. W dokumentach 2 i 4 często pojawiają się też słowa pokrewne do “ekonomia”.

Dokumenty 6, 11, 12 oraz 13 pochodzą z tematyki Literatura. We wszystkich z nich, jednym ze słów kluczowych (najczęściej występujących) jest “literatura” - pojawia się w nich odpowiednio 18, 22, 39 i 38 razy.

Dokumenty 7, 9, 15 oraz 20 pochodzą z tematyki COVID-19. We wszystkich z nich jako słowa kluczowe możemy znaleźć “wirus” czy “covid”.

Dokumenty 8, 16, 17 oraz 18 pochodzą z tematyki Zdrowie. W trzech z nich najważniejszymi słowami kluczowymi są “zdrowie” oraz “życie”.

- Eksperyment 2
    - macierz DTM z wagami Tf, z granicami 2-18

```
[1] "Dokument 1"
przedsiębiorstwo            koszt            wpływ        zarządzać        biznesowy        strategia
              92               38               29               26               23               23
[1] "Dokument 2"
behawioralny     ekonomie  ekonomiczny       teoria   zachowanie     ekonomia
          29           28           22           21           21           20
[1] "Dokument 3"
  społeczny      wiedza   działanie    określić      zmiana ekonomiczny
          58           25           21           16           16           15
[1] "Dokument 4"
      praca       osoba organizacja       rynek       grupa   społeczny
         115           79           68           61           50           44
[1] "Dokument 5"
   ekonomie ekonomiczny    ekonomia       nauka       sfera   społeczny
          48           43           30           30           22           22
[1] "Dokument 6"
   folklor      tekst literatura    kultura     reguła      treść
        66         35         18         17         15         13
[1] "Dokument 7"
  choroba   sarscov     wirus oddechowy     osoba  epidemia
        22         22         21         17         17         16
[1] "Dokument 8"
      życie     zdrowie      jakość     badanie      poczuć zadowolenie
          93           83           61           22           21           18
[1] "Dokument 9"
    wirus przypadek zakażenie     dzień  infekcja      ncov
        31         28         19         18         18         17
[1] "Dokument 10"
    usługa  zarządzać dostarczać    wartość       faza     klient
       185          56          37          35          29          24

[1] "Dokument 11"
literatura      tekst      praca   materiał    książka   kontrola
        22         21         15         14         13         12
[1] "Dokument 12"
   literatura rzeczywistość    nowoczesny     literacki doświadczenie         tekst
           39             30             24             18             16             13
[1] "Dokument 13"
literatura nowoczesny    dyskurs  literacki      model      forma
        38         22         21         17         17         11
[1] "Dokument 14"
     biznes     problem   narzędzie komunikacja      metoda      należy
         49          38          31          30          26          24
[1] "Dokument 15"
 pacjent   należy    covid personel medyczny zawierać
      39       31       21       19       17       17
[1] "Dokument 16"
 zdrowie człowiek    ciało  choroba   ludzki medycyna
      33       30       21       20       19       15
[1] "Dokument 17"
psychiczny  pozytywny    badanie    choroba      osoba      stres
        60         41         30         22         18         18
[1] "Dokument 18"
     życie    zdrowie      okres  społeczny zachowanie   dorastać
        44         23         21         20         19         17
[1] "Dokument 19"
         projekt        zarządzać    informatyczny        podejście          produkt przedsiębiorstwo
             142               68               30               28               27               27
[1] "Dokument 20"
      liczba         kraj      choroba        covid    przypadek zachorowanie
          31           27           26           26           26           22
```

Przeanalizujmy jak zmiana granic macierzy częstości wpłynęła na słowa kluczowe dokumentów.

W dokumencie 3 słowo kluczowe “imaginarium” z 21 wystąpieniami nie zostało znalezione, w efekcie pojawiło się kolejne, “ekonomiczny” z 15 wystąpieniami.

W dokumencie 4 zamiast słowa “pozarządowy” z 47 wystąpieniami zostało znalezione “społeczny” z 44 wystąpieniami.

W dokumencie 6 słowo kluczowe “folklorystyczny” z 21 wystąpieniami nie zostało znalezione, w efekcie pojawiło się kolejne, “treść” z 13 wystąpieniami.

W dokumencie 11 w eksperymencie drugim nie zostało w ogóle znalezione słowo kluczowe “cenzura” z największą liczbą wystąpień (23).

W dokumencie 13 pominięte zostało słowo “pisarstwo”, które w pierwszym eksperymencie miało 18 wystąpień.

W dokumencie 15 nie zostały znalezione słowa kluczowe “covid” oraz “chlor”, które według eksperymentu 1 występowały w dokumencie odpowiednio 21 i 20 razy.

W dokumencie 16 nie zostało znalezione słowo “cierpienie”, mające w eksperymencie 1 30 wystąpień, tyle samo co słowo “człowiek” znalezione w obu eksperymentach.

W dokumencie 17 słowo “sprężystość” z 62 wystąpieniami w eksperymencie 1 nie zostało znalezione w eksperymencie 2.

W dokumencie 19 nie zostało znalezione słowo “metodyka” z 51 wystąpieniami.

Podsumowując, eksperyment 2 (wykorzystujący ograniczoną macierz częstości) wykazał gorsze wyniki - nie zostały znalezione niektóre słowa, które miały większą częstotliwość występowania.

- Eksperyment 3
    - macierz DTM z wagami Tf, z granicami 3-14

```
[1] "Dokument 1"
przedsiębiorstwo            koszt        zarządzać        biznesowy        strategia          większy
              92               38               26               23               23               21
[1] "Dokument 2"
behawioralny     ekonomie  ekonomiczny       teoria   zachowanie     ekonomia
          29           28           22           21           21           20
[1] "Dokument 3"
  społeczny      wiedza ekonomiczny     dyskurs       świat      świata
          58           25           15           14           13           13
[1] "Dokument 4"
      praca organizacja       rynek   społeczny      sektor     miejsce
         115           68           61           44           22           21
[1] "Dokument 5"
   ekonomie ekonomiczny    ekonomia       nauka       sfera   społeczny
         48           43           30           30           22           22
[1] "Dokument 6"
     tekst literatura    kultura     reguła      treść   praktyka
           35          18          17          15          13          12
[1] "Dokument 7"
     choroba      sarscov        wirus    oddechowy     epidemia koronawirusa
           22           22           21           17           16           12
[1] "Dokument 8"
   zdrowie     jakość     poczuć obiektywny      ocena      wynik
         83          61          21          14          12          11
[1] "Dokument 9"
    wirus zakażenie     dzień  infekcja      mers      sars
           31        19        18        18        15        15
[1] "Dokument 10"
    usługa  zarządzać dostarczać    wartość       faza     klient
       185         56         37         35         29         24
```

```
[1] "Dokument 11"
literatura      tekst      praca   materiał   kontrola     latach
        22         21         15         14         12         10
[1] "Dokument 12"
   literatura rzeczywistość    nowoczesny     literacki doświadczenie         tekst
          39            30            24            18            16            13
[1] "Dokument 13"
literatura nowoczesny    dyskurs  literacki      model      forma
        38         22         21         17         17         11
[1] "Dokument 14"
     biznes   narzędzie komunikacja      metoda   biznesowy rozwiązanie
         49         31         30         26         18         16
[1] "Dokument 15"
 pacjent    covid personel medyczny zawierać    objaw
      39       21       19       17       17       15
[1] "Dokument 16"
 zdrowie człowiek    ciało  choroba   ludzki medycyna
      33       30       21       20       19       15
[1] "Dokument 17"
psychiczny  pozytywny    choroba     emocja     wysoki     radzić
        60         41         22         15         13         12
[1] "Dokument 18"
   zdrowie  społeczny zachowanie   człowiek  zdrowotny    zadanie
        23         20         19         13         12         11
[1] "Dokument 19"
         projekt        zarządzać    informatyczny        podejście          produkt przedsiębiorstwo
             142               68               30               28               27               27
[1] "Dokument 20"
  liczba     kraj  choroba    covid epidemia    wirus
      31       27       26       26       17       16
```

Porównując eksperymenty 3 i 2 można stwierdzić, że eksperyment 3 wykazał jeszcze gorsze wyniki, nie znajdując niektórych kluczowych słów.

- Eksperyment 4
    - macierz DTM z wagami Tf, z granicami 4-20

```
[1] "Dokument 1"
przedsiębiorstwo            koszt            wpływ        zarządzać        biznesowy        strategia
              92               38               29               26               23               23
[1] "Dokument 2"
behawioralny     ekonomie  ekonomiczny       teoria   zachowanie     ekonomia
          29           28           22           21           21           20
[1] "Dokument 3"
  społeczny      wiedza   działanie    określić      zmiana ekonomiczny
          58           25           21           16           16           15
[1] "Dokument 4"
      praca       osoba organizacja       rynek       grupa   społeczny
         115           79           68           61           50           44
[1] "Dokument 5"
   ekonomie ekonomiczny    ekonomia       nauka       sfera   społeczny
         48           43           30           30           22           22
[1] "Dokument 6"
     tekst literatura    kultura     reguła      treść   praktyka
         35         18         17         15         13         12
[1] "Dokument 7"
  choroba     wirus oddechowy     osoba   zdrowie       lek
        22         21         17         17         12         10
[1] "Dokument 8"
     życie    zdrowie     jakość    badanie     poczuć obiektywny
         93         83         61         22         21         14
[1] "Dokument 9"
    wirus przypadek zakażenie     dzień     chory oddechowy
       31        28        19        18        13        12
[1] "Dokument 10"
    usługa  zarządzać dostarczać    wartość       faza     klient
       185         56         37         35         29         24
```

```
[1] "Dokument 11"
literatura      tekst      praca   materiał   kontrola    badanie
        22         21         15         14         12         10
[1] "Dokument 12"
   literatura rzeczywistość    nowoczesny     literacki doświadczenie         tekst
           39            30            24            18            16            13
[1] "Dokument 13"
literatura nowoczesny    dyskurs  literacki      model      forma
        38         22         21         17         17         11
[1] "Dokument 14"
     biznes     problem   narzędzie komunikacja      metoda      należy
         49          38          31          30          26          24
[1] "Dokument 15"
 pacjent   należy medyczny zawierać    objaw   obszar
      39       31       17       17       15       14
[1] "Dokument 16"
 zdrowie człowiek  choroba   ludzki    życie    chory
      33       30       20       19       14       12
[1] "Dokument 17"
psychiczny  pozytywny    badanie    choroba      osoba     poziom
        60         41         30         22         18         16
[1] "Dokument 18"
     życie    zdrowie      okres  społeczny zachowanie     rozwój
        44         23         21         20         19         14
[1] "Dokument 19"
         projekt        zarządzać        podejście          produkt przedsiębiorstwo   oprogramowanie
             142               68               28               27               27               25
[1] "Dokument 20"
   liczba      kraj   choroba przypadek     wirus      duży
       31        27        26        26        16        15
```

Po raz kolejny eksperyment wykazał gorsze wyniki od poprzednich. Wnioskując z eksperymentów na macierzach DTM z wagami TF, im większe granice tym lepsze wyniki i większa trafność w znajdowaniu słów kluczowych.

- Eksperyment 5
    - macierz DTM, z wagami TfIdf, bez granic

```
[1] "Dokument 1"
przedsiębiorstwo          wydatek            koszt       inwestycja        biznesowy           nadzór
      0.07556439       0.03783107       0.03579430       0.03062515       0.02896114       0.02226382
[1] "Dokument 2"
    ekonomie behawioralny    konsument     ekonomia   racjonalny  eksperyment
  0.03501022   0.03123317   0.03025582   0.02500730   0.02210796   0.02146642
[1] "Dokument 3"
imaginarium   społeczny       blyth     hausner     dyskurs       tamże
 0.03406925  0.02177177  0.01460111  0.01460111  0.01220232  0.00997576
[1] "Dokument 4"
 pozarządowy    zatrudnić        rynek spółdzielnia   bezrobocie pełnosprawny
  0.07421652   0.07105837   0.03871206   0.03034278   0.02842335   0.02842335
[1] "Dokument 5"
    ekonomie     ekonomia  gospodarczy  ekonomiczny       kryzys finansyzacja
  0.04472414   0.02795259   0.02306432   0.02281016   0.01976941   0.01907753
[1] "Dokument 6"
        folklor folklorystyczny           tekst        nosiciel           treść           bajka
     0.12207531      0.05053479      0.02951562      0.02034588      0.01447661      0.01443851
[1] "Dokument 7"
   sarscov      wirus   epidemia  oddechowy     apteka farmaceuta
0.04759940 0.03854584 0.03461775 0.03120378 0.03074890 0.03074890
[1] "Dokument 8"
    zdrowie      jakość zadowolenie       życie      poczuć  obiektywny
 0.06739560  0.04316474  0.03672893  0.02370914  0.02240558  0.01719902
[1] "Dokument 9"
     wirus       ncov   infekcja  zakażenie       mers       sars
0.05250166 0.04119094 0.03593390 0.03217843 0.02994492 0.02994492
[1] "Dokument 10"
    usługa  zarządzać     klient  biznesowy    portfel       cykl
0.12800184 0.03381817 0.02545741 0.02439669 0.02171823 0.02015378
```

```
[1] "Dokument 11"
   cenzura    gukppiw    książka      tekst     cenzor     piękny
0.05938133 0.03356336 0.02579753 0.01900002 0.01807258 0.01587540
[1] "Dokument 12"
  epifaniczny rzeczywistość    literatura    nowoczesny       poetyka     literacki
   0.02913064    0.02784142    0.02752949    0.02554361    0.02383416    0.02205882
[1] "Dokument 13"
  pisarstwo     dyskurs  literatura  nowoczesny   literacki fikcjonalny
 0.04018322  0.02518620  0.02261163  0.01973825  0.01756198  0.01715872
[1] "Dokument 14"
       biznes   komunikacja       diagram     biznesowy informatyczny  reprezentant
   0.05506993    0.03371628    0.02894226    0.02022977    0.01987148    0.01768694
[1] "Dokument 15"
 dezynfekcja        chlor       strefa dezynfekować      pacjent          mgl
  0.05602971   0.05093610   0.04838930   0.04074888   0.03991848   0.03820208
[1] "Dokument 16"
cierpienie      ciało        bóg    zdrowie   medycyna    kościół
0.06989641 0.03098452 0.03028844 0.02351678 0.02213180 0.02096892
[1] "Dokument 17"
sprężystość  psychiczny       stres   pozytywny       block      emocja
 0.15121870  0.05881373  0.03374419  0.03058637  0.02926814  0.02316844
[1] "Dokument 18"
    dorastać         styl adolescencja     edukacja      zdrowie    rozwojowy
  0.03592416   0.03381097   0.02749318   0.02199455   0.01934119   0.01741072
[1] "Dokument 19"
     metodyka       projekt        zwinny     zarządzać         scrum informatyczny
   0.10705116    0.10445332    0.05877318    0.04365765    0.04198085    0.03987808
[1] "Dokument 20"
zachorowanie        covid    sanitarny       powiat     epidemia          ryc
  0.03616151   0.03521084   0.02993914   0.02780063   0.02302247   0.02301187
```

Po przeprowadzeniu eksperymentu 5 warto zauważyć, że znalezione słowa kluczowe różnią się od tych znalezionych w eksperymencie 1 (który to również został przeprowadzony na macierzy częstości bez granic, jednak z innymi wagami). Zdarza się nawet, przykładowo w dokumencie 2, w przypadku słów “ekonomie” oraz “behawioralny”, że zostały one zamienione miejscami, co oznacza, że eksperyment 5 wykazał większą częstotliwość występowania słowa kluczowego “ekonomie” niż eksperyment 1.

- Eksperyment 6
    - macierz DTM, z wagami TfIdf, z granicami 2-18

```
[1] "Dokument 1"
przedsiębiorstwo          wydatek            koszt       inwestycja        biznesowy           nadzór
      0.08496386       0.04253688       0.04024676       0.03443462       0.03256363       0.02503322
[1] "Dokument 2"
    ekonomie behawioralny     ekonomia   racjonalny  eksperyment    producent
  0.04410718   0.03934871   0.03150513   0.02785243   0.02704419   0.02253683
[1] "Dokument 3"
   społeczny      dyskurs        tamże    wyjaśniać  gospodarczy gospodarować
  0.02661771   0.01491831   0.01219616   0.01219616   0.01130458   0.01067164
[1] "Dokument 4"
       rynek spółdzielnia        praca     socjalny       sektor    społeczny
  0.04880465   0.03825343   0.03292085   0.02754247   0.02352944   0.02026716
[1] "Dokument 5"
   ekonomie    ekonomia gospodarczy ekonomiczny      kryzys  ekonomista
 0.05252241  0.03282650  0.02708590  0.02678742  0.02321648  0.02192668
[1] "Dokument 6"
   folklor      tekst   nosiciel      treść    przekaz    kultura
0.15162327 0.03665979 0.02527055 0.01798064 0.01605759 0.01554134
[1] "Dokument 7"
     sarscov        wirus     epidemia    oddechowy koronawirusa      choroba
  0.05707416   0.04621847   0.04150848   0.03741496   0.03113136   0.02402281
[1] "Dokument 8"
    zdrowie      jakość zadowolenie       życie      poczuć  obiektywny
 0.07630044  0.04886800  0.04158185  0.02684179  0.02536598  0.01947149
[1] "Dokument 9"
     wirus       ncov   infekcja  zakażenie       mers       sars
0.06226624 0.04885188 0.04261711 0.03816318 0.03551426 0.03551426
[1] "Dokument 10"
    usługa  zarządzać     klient  biznesowy       cykl     biznes
0.13768847 0.03637738 0.02738392 0.02624292 0.02167894 0.02053794

[1] "Dokument 11"
   książka      tekst     piękny literatura   materiał      utwór
0.03457571 0.02546520 0.02127736 0.02029149 0.01697680 0.01533928
[1] "Dokument 12"
 rzeczywistość     literatura     nowoczesny      literacki           trop modernistyczny
    0.03444821     0.03406226     0.03160514     0.02729340     0.02282534     0.02266668
[1] "Dokument 13"
    dyskurs  literatura  nowoczesny   literacki fikcjonalny   twórczość
 0.03919654  0.03518981  0.03071804  0.02733119  0.02670360  0.02136288
[1] "Dokument 14"
       biznes   komunikacja       diagram     biznesowy informatyczny  reprezentant
   0.06479184    0.03966848    0.03405165    0.02380109    0.02337955    0.02080934
[1] "Dokument 15"
         pacjent            covid          oddział         personel         izolacja zanieczyszczenie
      0.06037581       0.05122663       0.04737152       0.04634790       0.02664648       0.02664648
[1] "Dokument 16"
     ciało    zdrowie   medycyna    religia   człowiek    duchowy
0.04370820 0.03317386 0.03122014 0.02526181 0.01967673 0.01873208
[1] "Dokument 17"
psychiczny      stres  pozytywny     emocja  kaczmarek  adaptacja
0.07438825 0.04268002 0.03868598 0.02930370 0.01896890 0.01823070
[1] "Dokument 18"
  dorastać       styl    zdrowie  rozwojowy      ciało  społeczny
0.04418840 0.04158908 0.02379057 0.02141601 0.01927441 0.01564945
[1] "Dokument 19"
       projekt      zarządzać  informatyczny         branża     projektowy oprogramowanie
    0.12809374     0.05353848     0.04890349     0.03586256     0.03586256     0.03457308
[1] "Dokument 20"
zachorowanie        covid     epidemia          ryc         kraj        wirus
  0.04634269   0.04512435   0.02950438   0.02949080   0.02593118   0.02355793
```

- Eksperyment 7
    - macierz DTM, z wagami TfIdf, z granicami 3-14

```
[1] "Dokument 1"
przedsiębiorstwo            koszt        biznesowy           nadzór      technologia     informacyjny
      0.10809987       0.05120612       0.04143083       0.03184987       0.03103181       0.02882145
[1] "Dokument 2"
    ekonomie behawioralny     ekonomia   racjonalny       wybory       teoria
  0.05779021   0.05155556   0.04127872   0.03649287   0.02919430   0.02827203
[1] "Dokument 3"
  społeczny     dyskurs gospodarczy       aktor     jedność ekonomiczny
 0.03403756  0.01907687  0.01445580  0.01284960  0.01284960  0.01163669
[1] "Dokument 4"
      rynek       praca      sektor   społeczny organizacja    zawodowy
 0.06572885  0.04433695  0.03168884  0.02729529  0.02170656  0.02155044
[1] "Dokument 5"
   ekonomie    ekonomia gospodarczy ekonomiczny      kryzys  ekonomista
 0.06532975  0.04083109  0.03369067  0.03331941  0.02887771  0.02727340
[1] "Dokument 6"
     tekst      treść    przekaz    kultura      utwór literatura
0.04819096 0.02363636 0.02110844 0.02042980 0.01990520 0.01885096
[1] "Dokument 7"
     sarscov        wirus     epidemia    oddechowy koronawirusa      choroba
  0.07433734   0.06019814   0.05406352   0.04873182   0.04054764   0.03128897
[1] "Dokument 8"
    zdrowie      jakość      poczuć  obiektywny subiektywny     ujmować
 0.10570331  0.06769960  0.03514092  0.02697495  0.02636768  0.02109415
[1] "Dokument 9"
     wirus   infekcja  zakażenie       mers       sars      dzień
0.08651415 0.05921320 0.05302480 0.04934433 0.04934433 0.03757858
[1] "Dokument 10"
    usługa  zarządzać     klient  biznesowy       cykl     biznes
0.17095548 0.04516655 0.03400017 0.03258349 0.02691680 0.02550013

[1] "Dokument 11"
     tekst literatura   materiał      utwór   dokument        akt
0.03365718 0.02681912 0.02243812 0.02027382 0.01904762 0.01719947
[1] "Dokument 12"
rzeczywistość    literatura    nowoczesny     literacki          trop   artystyczny
   0.04195494    0.04148488    0.03849231    0.03324100    0.02779928    0.01929580
[1] "Dokument 13"
   dyskurs literatura nowoczesny  literacki      model     status
0.05143512 0.04617734 0.04030933 0.03586498 0.01321563 0.01265823
[1] "Dokument 14"
        biznes    komunikacja      biznesowy  informatyczny         metoda przedstawiciel
    0.08874764     0.05433529     0.03260117     0.03202378     0.02680977     0.02184087
[1] "Dokument 15"
   pacjent      covid   personel   medyczny  oddechowy      linia
0.08181360 0.06941579 0.06280477 0.03566234 0.03365113 0.03084687
[1] "Dokument 16"
     ciało    zdrowie   medycyna   człowiek    duchowy    choroba
0.05736155 0.04353655 0.04097254 0.02582325 0.02458352 0.02299407
[1] "Dokument 17"
psychiczny  pozytywny     emocja  adaptacja    choroba     radzić
0.09982561 0.05191480 0.03932422 0.02446476 0.02427593 0.02298851
[1] "Dokument 18"
   zdrowie  rozwojowy      ciało  społeczny   młodzież   fizyczny
0.03344813 0.03010963 0.02709867 0.02200220 0.02107674 0.01980198
[1] "Dokument 19"
       projekt      zarządzać  informatyczny         branża     projektowy oprogramowanie
    0.15860575     0.06629138     0.06055234     0.04440505     0.04440505     0.04280841
[1] "Dokument 20"
     covid   epidemia       kraj      wirus    patogen     liczba
0.06204107 0.04056531 0.03565255 0.03238958 0.02863434 0.02702703
```

- Eksperyment 8
    - macierz DTM, z wagami TfIdf, z granicami 4-20

```
[1] "Dokument 1"
przedsiębiorstwo            koszt        biznesowy      technologia     informacyjny        strategia
      0.10126507       0.04796853       0.03881130       0.02906977       0.02699916       0.02531627
[1] "Dokument 2"
    ekonomie behawioralny     ekonomia       teoria  ekonomiczny         zysk
  0.05728105   0.05110132   0.04091503   0.02802294   0.02562328   0.01432026
[1] "Dokument 3"
    społeczny       dyskurs   ekonomiczny         świat      ekonomia rzeczywistość
   0.03329506    0.01866073    0.01138285    0.01130278    0.01066327    0.01043334
[1] "Dokument 4"
      rynek       praca      sektor   społeczny organizacja    zawodowy
 0.05863581  0.03955239  0.02826919  0.02434975  0.01936413  0.01922485
[1] "Dokument 5"
   ekonomie    ekonomia ekonomiczny       nauka  gospodarka        nauk
 0.06427483  0.04017177  0.03278138  0.02620369  0.02276400  0.01602736
[1] "Dokument 6"
     tekst      treść    przekaz    kultura literatura     ludowy
0.04716198 0.02313167 0.02065772 0.01999357 0.01844845 0.01779359
[1] "Dokument 7"
     wirus  oddechowy    choroba        lek      ostry    pacjent
0.06652181 0.05385099 0.03457581 0.03167705 0.02455662 0.02369667
[1] "Dokument 8"
   zdrowie     jakość      życie     poczuć obiektywny     pomiar
0.08848390 0.05667112 0.03112781 0.02941635 0.02258065 0.01612903
[1] "Dokument 9"
     wirus  zakażenie      dzień  oddechowy      chory      płuco
0.09384586 0.05751843 0.04076321 0.03632743 0.02944009 0.02724557
[1] "Dokument 10"
    usługa  zarządzać     klient  biznesowy       cykl     biznes
0.15741350 0.04158875 0.03130690 0.03000244 0.02478463 0.02348017

[1] "Dokument 11"
     tekst literatura   materiał   dokument        akt     latach
0.03405357 0.02713498 0.02270238 0.01927195 0.01740203 0.01621599
[1] "Dokument 12"
rzeczywistość    literatura    nowoczesny     literacki   artystyczny       dyskurs
   0.04394313    0.04345079    0.04031642    0.03481625    0.02021021    0.02021021
[1] "Dokument 13"
   dyskurs literatura nowoczesny  literacki      model     status
0.05021678 0.04508354 0.03935452 0.03501545 0.01290259 0.01235839
[1] "Dokument 14"
       biznes    komunikacja      biznesowy         metoda przedstawiciel    rozwiązanie
    0.07803462     0.04777630     0.02866578     0.02357348     0.01920439     0.01906135
[1] "Dokument 15"
   pacjent   medyczny  oddechowy      linia    ochrona     skazić
0.08251115 0.03596640 0.03393805 0.03110988 0.02679659 0.02545354
[1] "Dokument 16"
   zdrowie   człowiek    choroba      chory     ludzki podkreślać
0.04415347 0.02618916 0.02331990 0.02109675 0.01658647 0.01645091
[1] "Dokument 17"
psychiczny  pozytywny  adaptacja    choroba     radzić      skala
0.09023198 0.04692559 0.02211360 0.02194292 0.02077922 0.01573583
[1] "Dokument 18"
   zdrowie  społeczny      życie   fizyczny  tożsamość  zdrowotny
0.02972077 0.01955034 0.01785108 0.01759531 0.01588807 0.01550649
[1] "Dokument 19"
        projekt        zarządzać   oprogramowanie przedsiębiorstwo          produkt         standard
      0.15220764       0.06361720       0.04108153       0.02894089       0.02525977       0.02464892
[1] "Dokument 20"
       kraj       wirus      liczba     choroba   zakażenie uszkodzenie
 0.03748256  0.03405211  0.02841430  0.02745379  0.02128257  0.01915431
```

Podsumowując eksperyment 6, 7 oraz 8 ponownie można stwierdzić, że ograniczenie granic macierzy częstości negatywnie wpłynęło na wyniki eksperymentu, zostało znalezione mniej słów kluczowych i z mniejszą liczbą wystąpień.

- Eksperyment 9
    - prawdopodobieństwo w modelu LDA (dla macierzy DTM z wagami Tf, bez granic, dla 4 tematów)

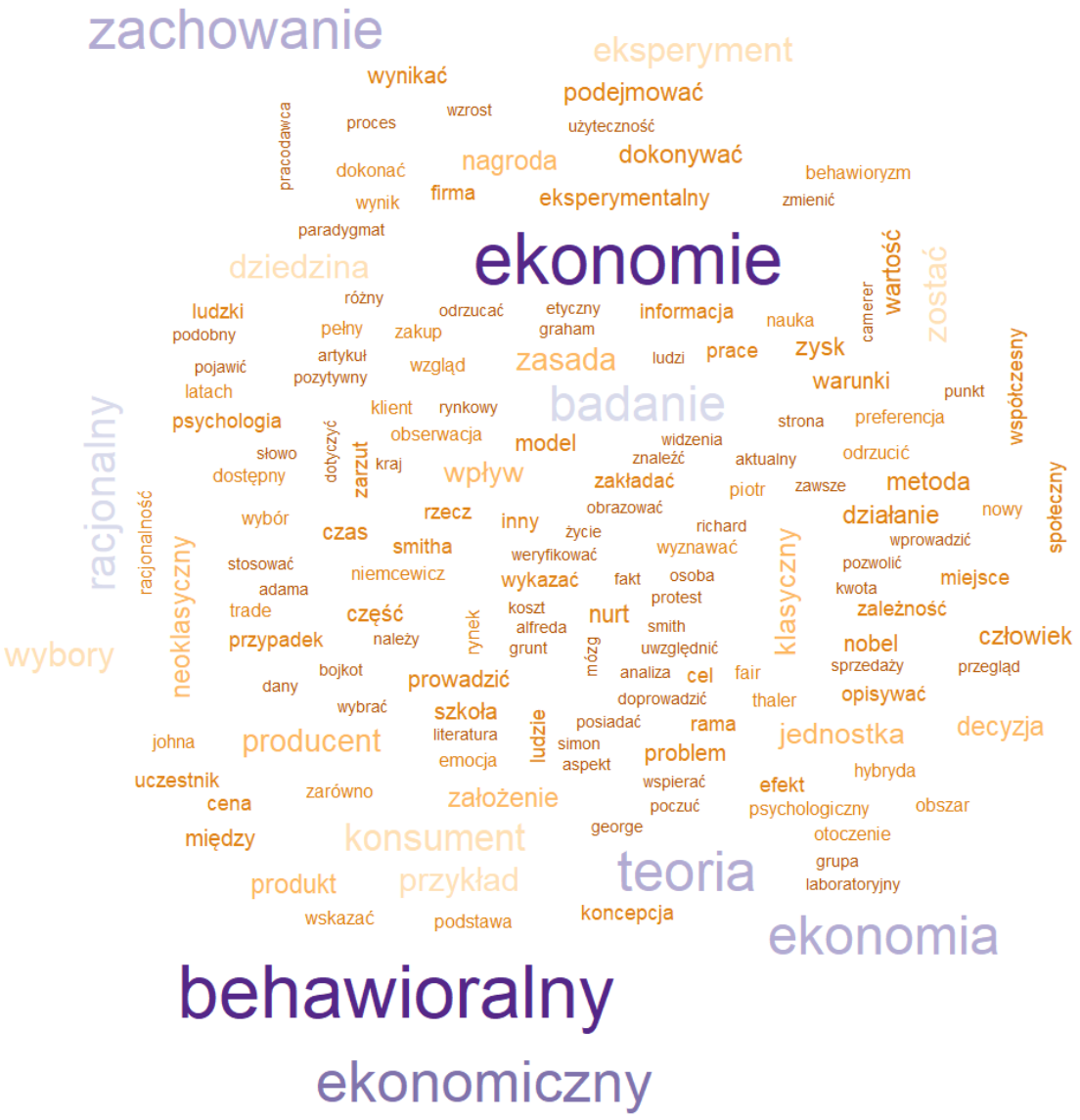
```
[1] "Dokument 1"
         usługa          projekt        zarządzać przedsiębiorstwo              cel           biznes
     0.017382552      0.013859843      0.013859843      0.011143380      0.006342419      0.006253993
[1] "Dokument 2"
      życie       praca     zdrowie       osoba     badanie   społeczny
0.009237954 0.007391090 0.006849363 0.006464949 0.006294474 0.006283233
[1] "Dokument 3"
 literatura   społeczny       tekst ekonomiczny    ekonomie       życie
0.006896748 0.005687968 0.004764785 0.004375453 0.004038312 0.003828757
[1] "Dokument 4"
      życie       praca     zdrowie       osoba     badanie   społeczny
0.014426768 0.011717721 0.010778221 0.010109423 0.009507370 0.007847034
[1] "Dokument 5"
 literatura       tekst   społeczny ekonomiczny    ekonomie     folklor
0.007789919 0.005381461 0.005350104 0.004795619 0.004300027 0.004146354
[1] "Dokument 6"
 literatura       tekst   społeczny ekonomiczny    ekonomie     folklor
0.008259844 0.005705969 0.005082188 0.005004489 0.004415907 0.004396289
[1] "Dokument 7"
  przypadek     choroba     pacjent       wirus       osoba       covid
0.010266236 0.008863712 0.007858470 0.007858470 0.005551454 0.005470654
[1] "Dokument 8"
      życie       praca     zdrowie       osoba     badanie   społeczny
0.012908053 0.010475634 0.009624001 0.009017467 0.008496633 0.007637260
[1] "Dokument 9"
  przypadek     choroba     pacjent       wirus       covid   oddechowy
0.010670948 0.009159155 0.008204291 0.008204291 0.005711275 0.005592560
[1] "Dokument 10"
         usługa          projekt        zarządzać przedsiębiorstwo              cel           biznes
     0.018193075      0.014506011      0.014506011      0.011659652      0.006599208      0.006545305

[1] "Dokument 11"
 literatura   społeczny       tekst ekonomiczny    ekonomie     folklor
0.007378170 0.005208912 0.005097243 0.004561638 0.004107385 0.003927538
[1] "Dokument 12"
 literatura       tekst ekonomiczny   społeczny     folklor    ekonomie
0.008846947 0.006111338 0.005340253 0.005298319 0.004708462 0.004694318
[1] "Dokument 13"
 literatura       tekst   społeczny ekonomiczny    ekonomie     folklor
0.008652776 0.005977279 0.005485890 0.005264426 0.004665126 0.004605229
[1] "Dokument 14"
         usługa          projekt        zarządzać przedsiębiorstwo              cel           biznes
     0.016540789      0.013188759      0.013188759      0.010605525      0.006055498      0.005951422
[1] "Dokument 15"
  przypadek     choroba     pacjent       wirus       covid   oddechowy
0.010902849 0.009212771 0.008360851 0.008360851 0.005820212 0.005699229
[1] "Dokument 16"
      życie       praca     zdrowie       osoba     badanie   społeczny
0.010758721 0.008825826 0.008316976 0.007959541 0.007327136 0.006927434
[1] "Dokument 17"
      życie       praca     zdrowie       osoba     badanie   społeczny
0.013211533 0.010699865 0.009887527 0.009310350 0.008815682 0.007308599
[1] "Dokument 18"
      życie       praca     zdrowie       osoba     badanie   społeczny
0.012497532 0.010190559 0.009441017 0.008909072 0.008305749 0.007492690
[1] "Dokument 19"
         usługa          projekt        zarządzać przedsiębiorstwo              cel           biznes
     0.017823600      0.014211463      0.014211463      0.011424161      0.006501544      0.006412530
[1] "Dokument 20"
  przypadek     choroba     pacjent       wirus       osoba       covid
0.009873728 0.008640702 0.007558259 0.007558259 0.005653373 0.005261766
```

W eksperymencie 9 na podstawie wyników analizy tematyk wyraźnie widać wyodrębnione 3 tematy oraz które dokumenty do nich przynależą wraz ze słowami kluczowymi danych tematyk.

Poniżej znajdują się zrzuty ekranów kilku chmur tagów stworzonych na podstawie niektórych dokumentów, po jednym z każdej z tematyk:

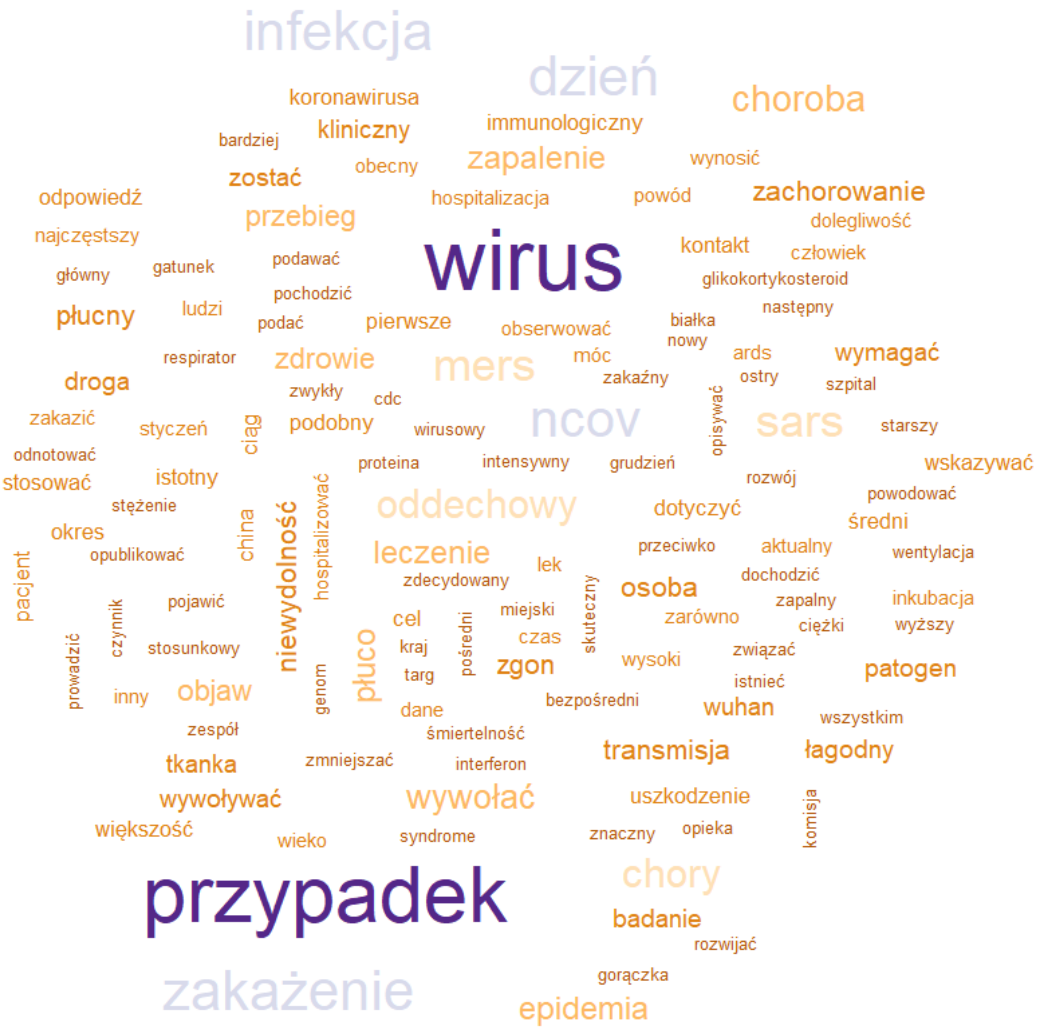
- IT ("Metody i narzędzia rozwiązywania problemów komunikacji w relacji IT-Biznes w projektach informatycznych")

- Ekonomia (“Ekonomia Behawioralna - Hybryda Teorii i Eksperymentu”)

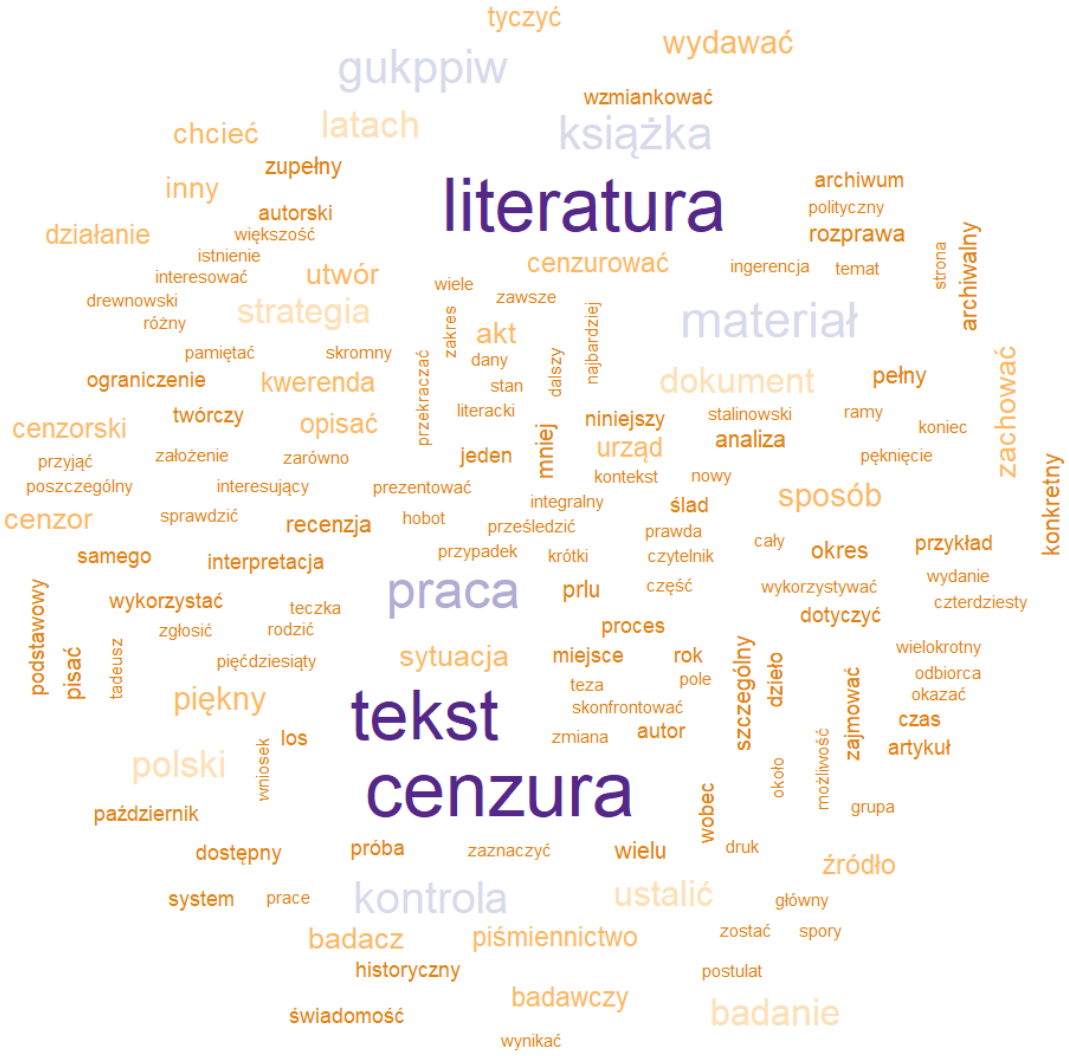


- Zdrowie ("Religia a zdrowie - czy religia może sprzyjać trosce o zdrowie"

- COVID-19 (“Koronawirus 2019-nCoV – transmisja zakażenia, objawy i leczenie”



- Literatura (“Literatura i pisarze wobec cenzury PRL - wprowadzenie”)



7. Bibliografia/netografia

A. Wawrzyniak, K. Kuczborska, A. Lipińska-Opałka i inni, *Koronawirus 2019-nCoV - transmisja zakażenia, objawy i leczenie*, Pediatr Med Rodz, 2020

Podręcznik prewencji i leczenia COVID-19, The First Affiliated Hospital, Zhejiang University School of Medicine, LIANG Tingbo, 2019

Epidemia koronawirusa SARS-CoV-2: Informacje i tymczasowe wytyczne dla farmaceutów i pracowników aptek, Międzynarodowa Federacja Farmaceutyczna, 2020

J. Duszyński, A. Afelt, A. Ochab-Marcinek i inni, *Zrozumieć COVID-19*, Polska Akademia Nauk, 2020

R. Nycz, *Literatura jako trop rzeczywistości*, Universitas, 2001

R. Sulima, *Folklor i literatura*, Ludowa Spółdzielnia Wydawnicza, 1985

R. Nycz, *Literatura nowoczesna: cztery dyskursy (tezy)*, FNP, 2002

K. Budrowska, *Literatura i pisarze wobec cenzury PRL 1948-1958*, Wydawnictwo Uniwersytetu w Białymstoku, 2009

H. Sęk, *Jakość życia a zdrowie*, Wydział Prawa i Administracji UAM, 1993

D. Ponczek, I. Olszowy, *Styl życia młodzieży i jego wpływ na zdrowie*, Probl Hig Epidemiol, 2012

Ł. Kaczmarek, H. Sęk, M. Ziarko, *Sprężystość psychiczna i zmienne pośredniczące w jej wpływie na zdrowie*, 2011

J. Pawlikowski, K. Marczewski, *Religia a zdrowie - czy religia może sprzyjać trosce o zdrowie?*, Kardiologia po dyplomie, Tom 7 Nr 10, Listopad/Grudzień 2008

P. Niemcewicz, *Ekonomia Behawioralna - Hybryda Teorii i Eksperymentu*, Wydział Nauk Ekonomicznych i Zarządzania Uniwersytetu Szczecińskiego, 2018

J. Hausner, *Ekonomia i społeczne imaginarium*, Biuletyn PTE, 2019

I. Gosk, et al., *Ekonomia Ekonomia społeczna jako aktor rynku pracy*, Ekonomia Społeczna - Teksty, 2006

M. Ratajczak, *Ekonomia w dobie finansyzacji gospodarki*, Ruch prawniczy, ekonomiczny i socjologiczny, 2014

R. Orzechowski. *Efektywne zastosowanie IT w przedsiębiorstwie*, Od redakcji, 2007

F. Liebert, *Zarządzanie projektami w przedsiębiorstwach branży IT–studium literaturowe*, Zeszyty Naukowe. Organizacja i Zarządzanie/Politechnika Śląska, 2017

I. Chomiak-Orsa; A. Kołtonowska, *Metody i narzędzia rozwiązywania problemów komunikacji w relacji IT–biznes w projektach informatycznych*, Informatyka Ekonomiczna, 2016

R. Orzechowski; A. Tarasiewicz, *Kreowanie wartości poprzez efektywne zarządzanie usługami IT*, e-mentor, 2008

Materiały z wykładów oraz ćwiczeń