Analiza taksonomiczna dokumentów w oparciu o macierz częstości

Przetwarzanie języka naturalnego
Ćwiczenia 2.

Rok akademicki: 2016/2017

Programy do samodzielnego wykonania

- Przekształcenie wyrazów do formy podstawowej
- Identyfikacja słów i fraz kluczowych (z prezentacją w postaci wykresu słupkowego i chmury słów):
  - tf (częstości)
  - tf-idf (ważone częstości logarytmiczne)
  - LSA
  - LDA
  - RAKE
- Analiza dokumentów z wykorzystaniem ontologii (np. analiza ogłoszeń dotyczących sprzedaży nieruchomości lub samochodów).

Analiza taksonomiczna (analiza skupień)

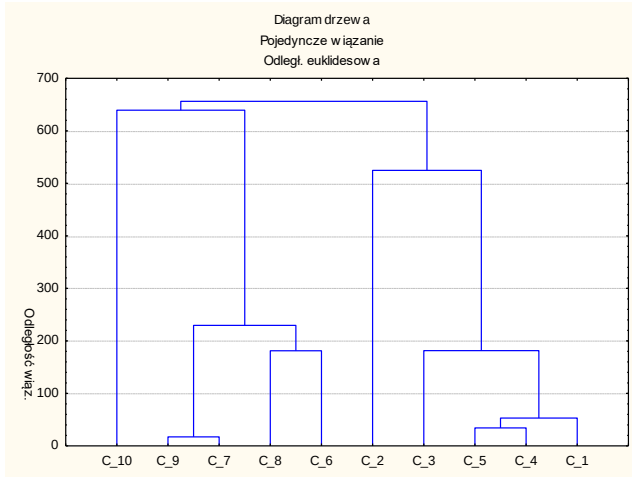
- Analiza skupień – analiza zbioru obiektów w celu określenia jego struktury (identyfikacji klas obiektów podobnych).
- W zależności od podejścia możliwe jest uzyskanie klas:
  - hierarchicznych (klasy dzielą się na podklasy),
  - wykluczających się (każdy obiekt należy do jednej klasy),
  - niewykluczających się (obiekt może należeć do kilku klas),
  - opisanych w kategoriach rozkładów prawdopodobieństwa (tworzone są modele klas).

METODY HIERARCHICZNE

Metody hierarchiczne i ich podział

- Metody hierarchiczne – metody pozwalające na odtworzenie hierarchii klas obiektów. Pokazują wszystkie stany pośrednie pomiędzy przypadkiem, w którym wszystkie obiekty tworzą jedną klasę i przypadkiem, w którym każdy z obiektów jest samodzielną klasą.
- Rodzaje metod hierarchicznych:
  - metody aglomeracyjne – w pierwszym kroku każdy z obiektów tworzy oddzielną klasę. Na każdym kolejnym dwie najbardziej podobne klasy są ze sobą łączone. Na ostatnim etapie wszystkie obiekty tworzą jedną klasę.
  - Metody podziałowe – w pierwszym kroku wszystkie obiekty tworzą jedną klasę. W trakcie każdego kolejnego kroku jedna klasa jest dzielona na dwie. W ostatnim kroku obiekty tworzą jednoelementowe klasy.

Dendrogram



Dendrogram jako wynik działania metod hierarchicznych.

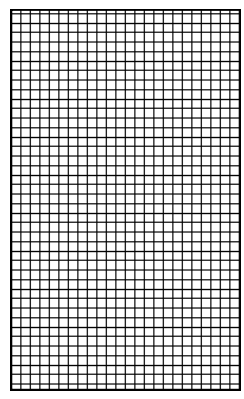
Etapy działania metod hierarchicznych

- określenie celu badań
- przygotowanie zbioru danych
- wstępne przetworzenie danych (np. standaryzacja)
- obliczenie macierzy odległości
- wykonanie obliczeń
- prezentacja wyników (drzewko połączeń)
- wybór podziału optymalnego

Struktura zbioru danych

zmienne

obiekty

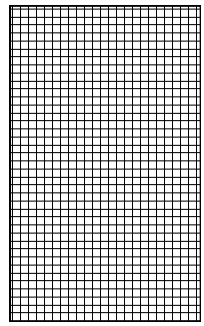


Cel badań:
• klasyfikacja obiektów,
lub
• klasyfikacja zmiennych.

Wstępne przetworzenie danych

zmienne

obiekty



$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}$$

$$z_{ij} = \frac{x_{ij}}{\max_i(x_{ij})}$$

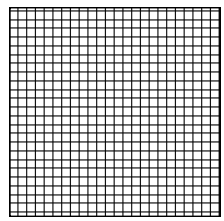
$$z_{ij} = \frac{x_{ij}}{\min_i(x_{ij})}$$

$$z_{ij} = \frac{x_{ij}}{\bar{x}_j}$$

Wyznaczenie macierzy odległości

obiekty

obiekty



Odległość miejska:

$$d_{ik} = \sum_{j=1}^{m} \left| z_{ij} - z_{kj} \right|$$

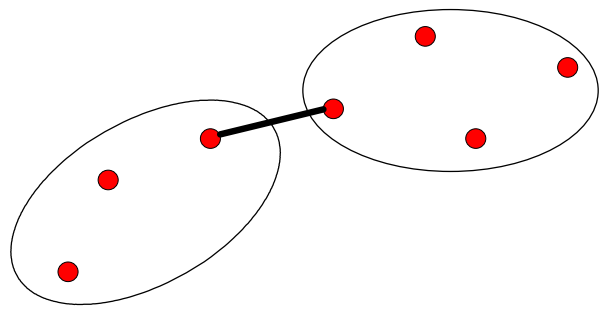
Odległość Euklidesa:

$$d_{ik} = \sqrt{\sum_{j=1}^{m} \left(z_{ij} - z_{kj} \right)^2}$$

Metody aglomeracyjne

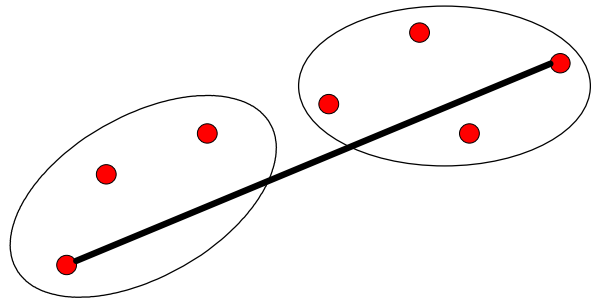
1. każdy obiekt tworzy oddzielne skupienie
2. **następuje łączenie dwóch najbliższych elementów**; połączone elementy tworzą grupę;
3. modyfikacja macierzy odległości – połączone elementy reprezentuje jeden wiersz (i jedna kolumna); aktualizacja elementów macierzy odległości;
4. jeżeli obiekty nie tworzą jednej grupy to przejście do kroku 2.

Metoda najbliższego sąsiedztwa



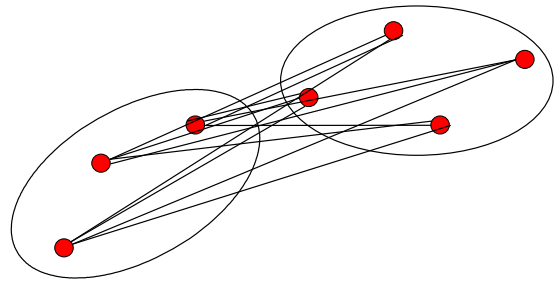
Single linkage method

Metoda najdalszego sąsiedztwa



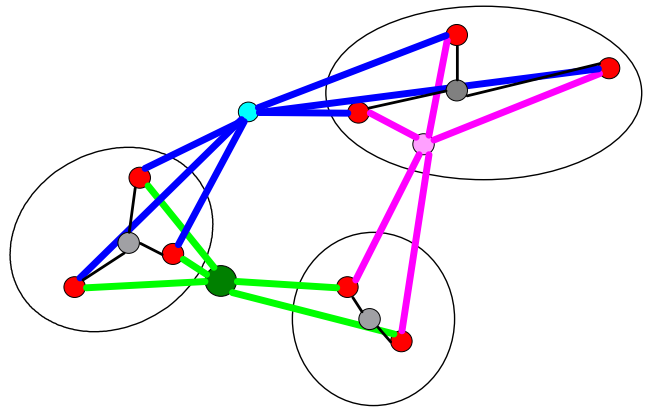
Complete linkage method

Metoda uśrednionego sąsiedztwa



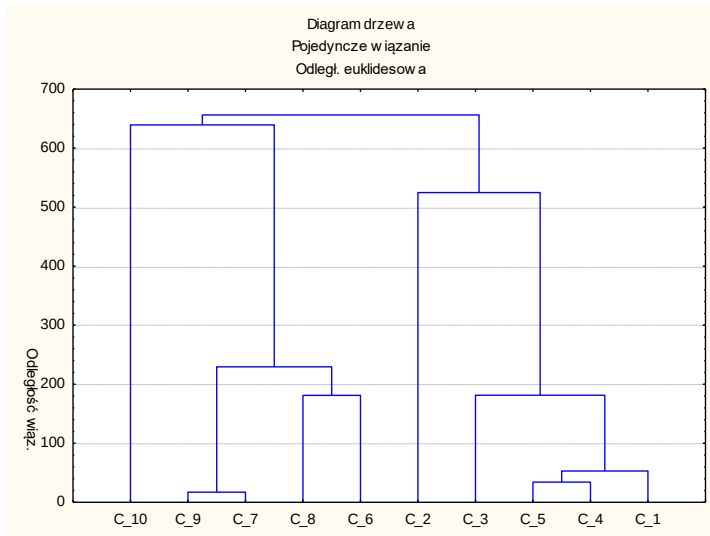
Average linkage method

Metoda Warda

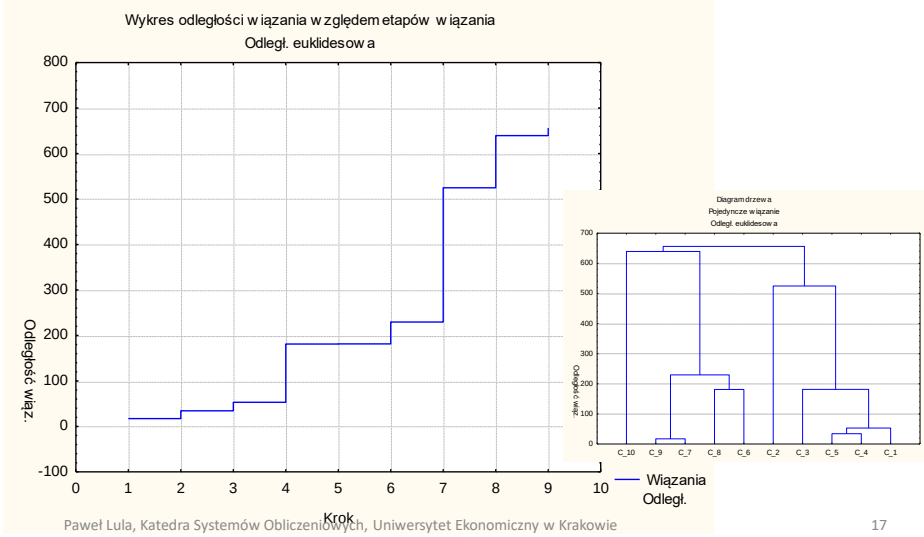


Łączenie dokonywane jest w taki sposób, aby w najmniejszym stopniu zwiększyć wariancję wewnątrzgrupową

Prezentacja wyników



Wybór podziału optymalnego



TAKSONOMICZNA ANALIZA TEKSTÓW

Zbiór analizowanych tekstów

- Adam Mickiewicz, *Dziady III*,
- Juliusz Słowacki, *Kordian*,
- Stanisław Wyspiański, *Noc Listopadowa*,
- Stanisław Wyspiański, *Wesele*,
- Bolesław Prus, *Katarynka*,
- Henryk Sienkiewicz, *Janko Muzykant*,
- Maria Konopnicka, *Nasza Szkapa*,
- Gabriela Zapolska, *Moralność Pani Dulskiej*,
- Adam Mickiewicz, *Pan Tadeusz*,
- Henryk Sienkiewicz, *Krzyżacy (t. I)*,
- Eliza Orzeszkowa, *Nad Niemnem (t. I)*,
- Władysław Reymont, *Chłopi (t. I)*.

Utworzenie macierzy częstości

teksty po transformacji do formy podstawowej

```
library(tm)

katalog<-"C:/Literatura polska – forma podstawowa/"

korpus <- VCorpus(DirSource(katalog,encoding="UTF-8"),
readerControl = list(reader=readPlain))

korpus<-tm_map(korpus,removeNumbers)

stoplista <-readLines("C:/stoplista_PL.txt",encoding="UTF-
8")

korpus<-tm_map(korpus,removeWords,stoplista)

dtm<-DocumentTermMatrix(korpus)
```

macierz częstości – bez transformacji

Charakterystyka macierz częstości

liczba terminów

```
> dtm

<<DocumentTermMatrix (documents: 12, terms: 26667)>>
Non-/sparse entries: 57289/262715
Sparsity           : 82%
Maximal term length: 20
Weighting          : term frequency (tf)

>
```

oryginalna macierz częstości - brak transformacji

Analiza taksonomiczna

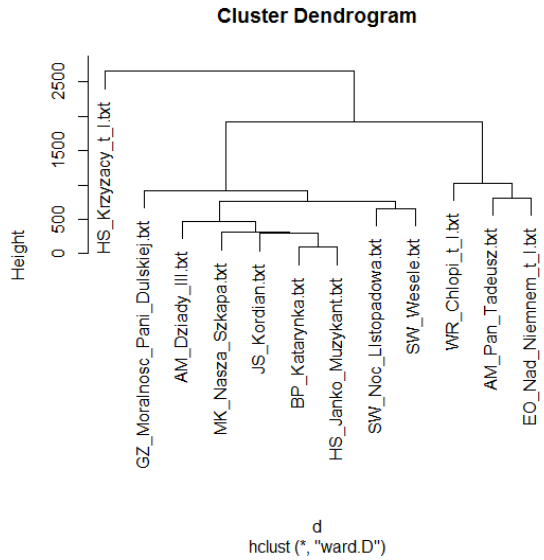
Tworzenie macierz odległości

```
d<-dist(dtm,method="euclidean")

fit1<-hclust(d=d,method="ward.D")

plot(fit1)
```

Metoda Warda

Rysowanie dendrogramu

Wyniki klasyfikacji

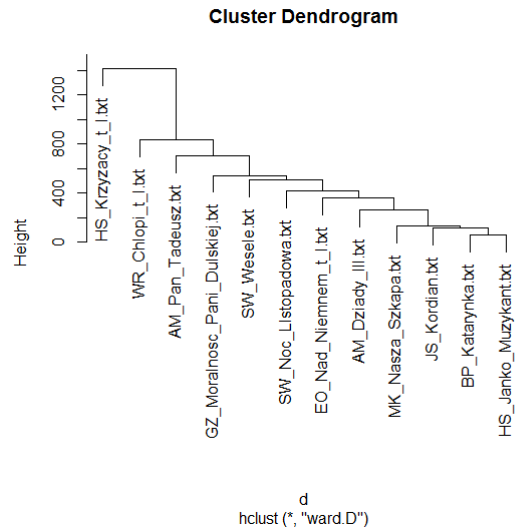


Oryginalna macierz częstości, 26667 wyrazów

Ograniczenie liczby wyrazów

```
dtm<-DocumentTermMatrix(korpus,control=list(bounds = list(global = c(2,6))))

dtm
<<DocumentTermMatrix (documents: 12, terms: 9280)>>
Non-/sparse entries: 28709/82651
Sparsity           : 74%
Maximal term length: 17
Weighting          : term frequency (tf)

d<-dist(dtm,method="euclidean")

fit2<-hclust(d=d,method="ward.D")

plot(fit2)
```

Wyniki klasyfikacji

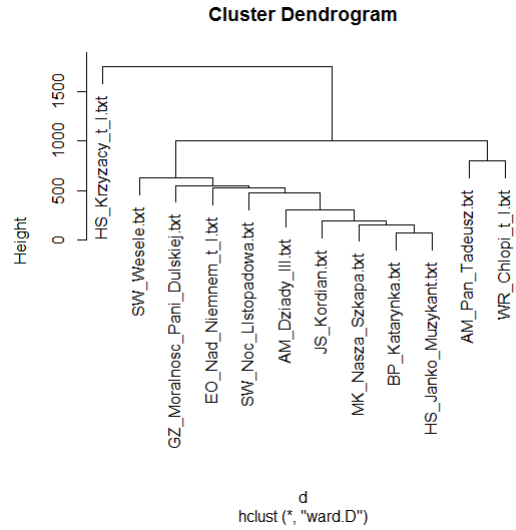


Oryginalna macierz częstości, 9280 wyrazów

Ograniczenie liczby wyrazów

```
dtm<-DocumentTermMatrix(korpus,control=list(bounds = list(global = c(2,9))))

dtm
<<DocumentTermMatrix (documents: 12, terms: 10361)>>
Non-/sparse entries: 37122/87210
Sparsity           : 70%
Maximal term length: 17
Weighting          : term frequency (tf)

d<-dist(dtm,method="euclidean")

fit3<-hclust(d=d,method="ward.D")

plot(fit3)
```

Wyniki klasyfikacji

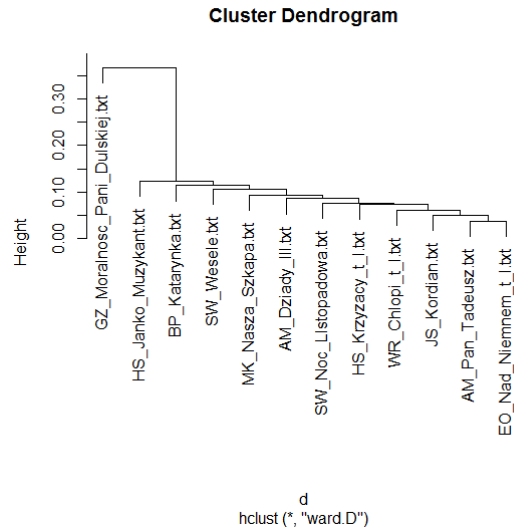


Oryginalna macierz częstości, 10361 wyrazów

Ważenie elementów macierzy częstości

```
dtm<-DocumentTermMatrix(korpus,control=list(weighting=weightTfIdf,bounds =
list(global = c(2,9))))

dtm
<<DocumentTermMatrix (documents: 12, terms: 10361)>>
Non-/sparse entries: 37122/87210
Sparsity           : 70%
Maximal term length: 17
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)


d<-dist(dtm,method="euclidean")

fit4<-hclust(d=d,method="ward.D")

plot(fit4)
```

Wyniki klasyfikacji



Ważona macierz częstości (tf-idf), 10361 wyrazów

Analiza ukrytych składowych semantycznych

przyłączenie pakietu

```
library(lsa)
```

przygotowanie macierzy częstości
- *transponowanie,*
- *zmiana formatu*

```
txt_mat<-as.textmatrix(t(as.matrix(dtm)))
```

```
lsa_model<-lsa(txt_mat)
```

analiza LSA

```
lsa_model$tk
lsa_model$dk
lsa_model$sk
```

Wyniki obliczeń:
tk – współrzędne wyrazów (terms)
dk – współrzędne dokumentów,
sk – znaczenie składowych

Współrzędne wyrazów

```
> lsa_model$tk[1:10,]
                     [,1]         [,2]          [,3]          [,4]
absolucja     4.599393e-06 0.0001364202 -0.0001629819  8.608651e-05
abych         1.589329e-04 0.0002505512 -0.0002417917  4.007406e-04
acta          1.729507e-05 0.0007573478 -0.0014356718  4.173626e-03
adam          2.017347e-05 0.0006287145 -0.0011606356  1.277217e-03
adama         2.201749e-05 0.0008879754 -0.0010308298  1.190413e-03
adieu         4.541661e-05 0.0016514440 -0.0033900206  3.086003e-03
adiutant      9.021928e-06 0.0002504629 -0.0004479297  5.991047e-04
administracja 6.464737e-06 0.0003244155 -0.0006883801  3.618225e-04
adwokat       6.392616e-05 0.0025824020 -0.0151139346 -6.569154e-03
aha           5.273037e-03 0.0005873601 -0.0012672410  2.605714e-03
>
```


Współrzędne dokumentów

```
> lsa_model$dk
                                      [,1]        [,2]        [,3]        [,4]
AM_Dziady_III.txt                0.003096394  0.06338349 -0.13142630  0.06244693
AM_Pan_Tadeusz.txt               0.003341607  0.03974654 -0.05319153  0.04364800
BP_Katarynka.txt                 0.004874450  0.15634516 -0.89562034 -0.38712411
EO_Nad_Niemnem_t_I.txt           0.004613845  0.04868442 -0.04822840  0.01431293
GZ_Moralnosc_Pani_Dulskiej.txt 0.987428365 -0.01386878  0.01589846 -0.01566446
HS_Janko_Muzykant.txt            0.002196762  0.95993023  0.23375081 -0.12336700
HS_Krzyzacy_t_I.txt              0.157119159  0.04730740 -0.04434170  0.04784093
JS_Kordian.txt                   0.002710335  0.04390221 -0.04592703  0.05745819
MK_Nasza_Szkapa.txt              0.003103212  0.08036331 -0.09893039  0.10499887
SW_Noc_LIstopadowa.txt           0.004924220  0.04905009 -0.08495365  0.11698928
SW_Wesele.txt                    0.007712236  0.15777613 -0.30598279  0.87793836
WR_Chlopi_t_I.txt                0.011289967  0.08948460 -0.07641022  0.16608217
>
```

Utworzenie macierzy S (przechowywane są jedynie wartości diagonalne)

```
lsa_model$sk
[1] 0.23659235 0.08751664 0.08119857 0.07857230

s<-matrix(rep(0,16),4,4)

diag(s)<-lsa_model$sk

s
          [,1]       [,2]       [,3]      [,4]
[1,] 0.2365923 0.00000000 0.00000000 0.0000000
[2,] 0.0000000 0.08751664 0.00000000 0.0000000
[3,] 0.0000000 0.00000000 0.08119857 0.0000000
[4,] 0.0000000 0.00000000 0.00000000 0.0785723
```

elementy leżące na głównej przekątnej macierzy S

tworzenie macierzy o elementach zerowych

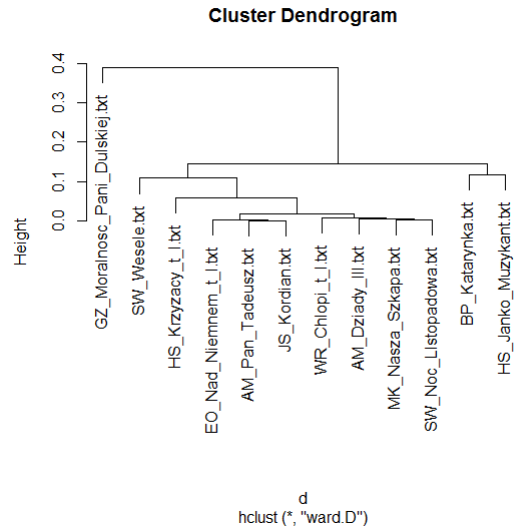
modyfikacja elementów na głównej przekątnej

Analiza taksonomiczna dokumentów w oparciu o LSA

```
d<-dist(lsa_model$dk%*%s)

fit5<-hclust(d=d,method="ward.D")

plot(fit5)
```

macierz odległości pomiędzy ważonymi współrzędnymi dokumentów (wagi = ważność składowych)

Wyniki klasyfikacji



Współrzędne LSA obliczone w oparciu o ważoną macierz częstości (tf-idf),

4 składowe

Ważność słów wyznaczana w oparciu o LSA

obliczenie mierników ważności słów

```
t_imp<-diag(lsa_model$tk%*%s%*%t(s)%*%t(lsa_model$tk))

tail(sort(t_imp),25)
       zosia      marysia     radczyni      łopucha       stójka
0.0001894955 0.0001937524 0.0002018845 0.0002464164 0.0002464709
       janku    gospodarz        scena      kredens       maryna
0.0002467565 0.0002548480 0.0002617411 0.0002835208 0.0002839436
         ino      zbyszka       matula       jasiek        janko
0.0003038068 0.0003324925 0.0003723161 0.0003743513 0.0003849106
 dziewczynka     skrzypki      czepiec        janek    katarynka
0.0005021971 0.0005903454 0.0008078759 0.0009268861 0.0013753210
      ciocia       tomasz        poeta        hanka      zbyszko
0.0015183021 0.0016668532 0.0017066843 0.0114086953 0.0407135450
>
```

25 najważniejszych słów w korpusie

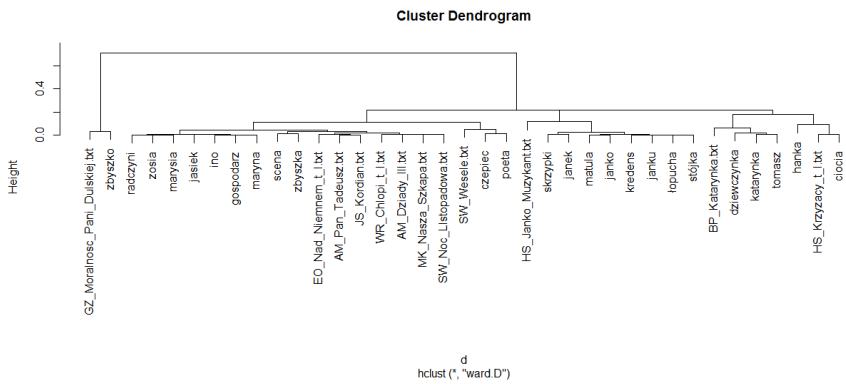
Łączna analiza dokumentów i słów

Lista najważniejszych słów

```
words<-names(tail(sort(t_imp),25))

dane<-rbind(lsa_model$dk%*%s,lsa_model$tk[words,]%*%s)

d<-dist(dane)

fit9<-hclust(d=d,method="ward.D")

plot(fit9)
```

Połączenie danych opisujących dokumenty i wyrazy (współrzędne ważone wartościami macierzy S)

Wyniki klasyfikacji dokumentów i 25 najważniejszych wyrazów



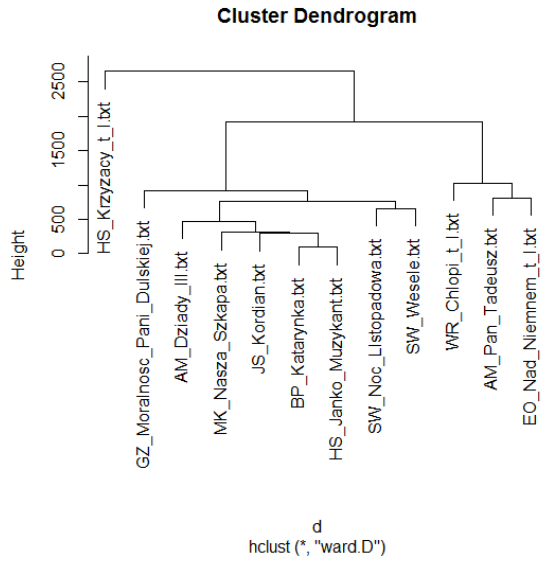
PORÓWNYWANIE WYNIKÓW KLASYFIKACJI

Porównywanie wyników klasyfikacji

- Oznaczenia:
  - K_1, K_2 – klasyfikacje,
  - TP – liczba par obiektów, które zostały zaliczone do tej samej grupy obiektów w klasyfikacji K_1 oraz w klasyfikacji K_2,
  - FP – liczba par obiektów, które zostały zaliczone do tej samej grupy obiektów w klasyfikacji K_1 oraz do różnych grup w klasyfikacji K_2,
  - FN – liczba par obiektów, które zostały zaliczone do różnych grup w klasyfikacji K_1 oraz do tej samej grupy w klasyfikacji K_2.
- Indeks Fowlkes'a – Mallows'a:

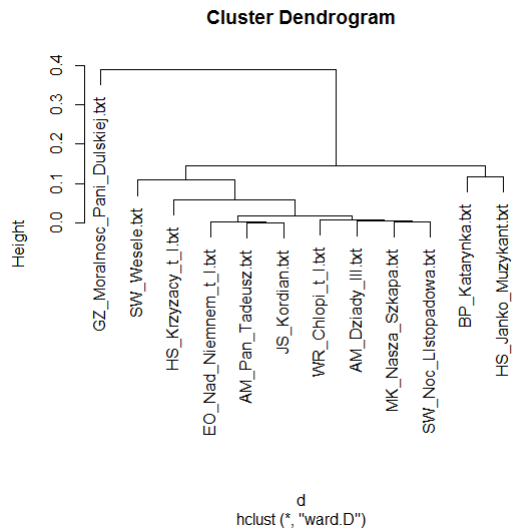
$$FM = \sqrt{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}$$

Wyniki klasyfikacji



Klasyfikacja K1: Oryginalna macierz częstości, 26667 wyrazów

Wyniki klasyfikacji



Klasyfikacja K2: Współrzędne LSA obliczone w oparciu o ważoną macierz częstości (tf-idf),

4 składowe

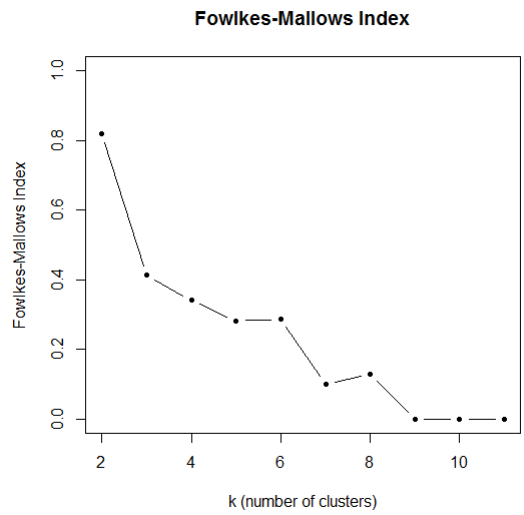
Porównywanie wyników klasyfikacji

```
library(dendextend)

dend1<-as.dendrogram(fit1)
dend2<-as.dendrogram(fit5)

Bk_plot(dend1,dend2,add_E=FALSE,rejection_line_asymptotic=FALSE,main="Fowlkes-Mallows Index",ylab="Fowlkes-Mallows Index")
```

Klasyfikacja K1

Klasyfikacja K2

Porównanie wyników klasyfikacji



Porównywane dendrogramy:

- *klasyfikacja K1*
- *klasyfikacja K2*