Ukryta alokacja Dirichleta w ujęciu praktycznym

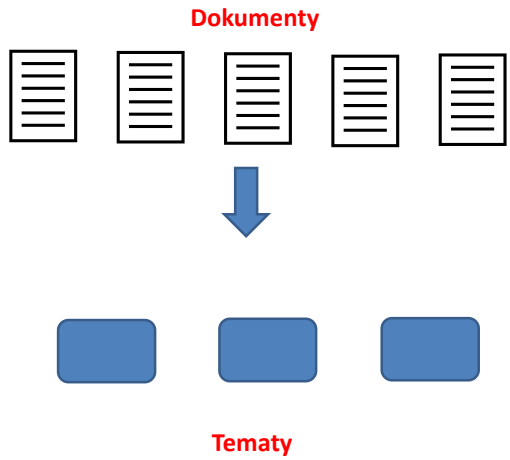
Przetwarzanie języka naturalnego
Ćwiczenia 3.

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).

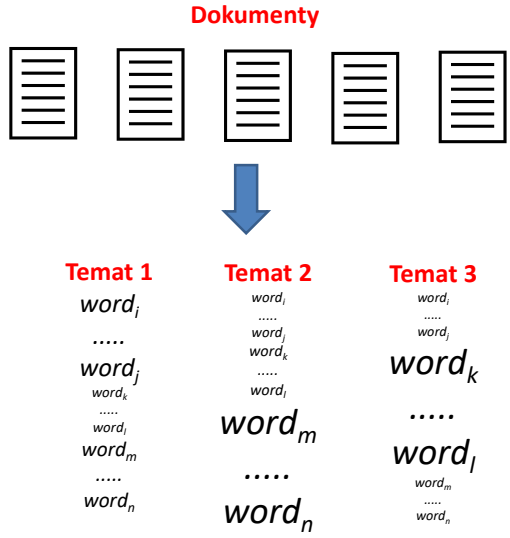
Latent Dirichlet Allocation (LDA)



Latent Dirichlet Allocation *– metoda identyfikacji tematów; realizowana w trybie bez nauczyciela.*

Tematy są opisywane poprzez podanie prawdopodobieństwa wystąpienia poszczególnych wyrazów.

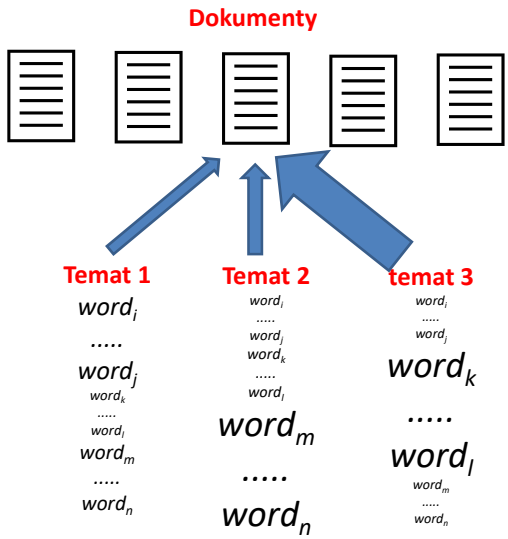
Latent Dirichlet Allocation (LDA)



Latent Dirichlet Allocation *– metoda identyfikacji tematów; realizowana w trybie bez nauczyciela.*

Tematy są opisywane poprzez podanie prawdopodobieństwa wystąpienia poszczególnych wyrazów.

Latent Dirichlet Allocation (LDA)



Labeled Latent Dirichlet Allocation *– metoda identyfikacji tematów realizowana w trybie uczenia z nauczycielem.*

Tematy reprezentowane są przez etykiety przypisane do dokumentów (liczba tematów = liczba różnych etykiet).

Tematy są opisane poprzez prawdopodobieństwa wystąpienia w nich różnych słów.

Model LDA

Dostępny jest słownik $\boldsymbol{V}$ złożony z LV terminów:

$$\boldsymbol{V} = \begin{bmatrix} v_1 \\ \dots \\ v_{LV} \end{bmatrix}$$

Przetwarzany korpus $\boldsymbol{D}$ składa się z LD dokumentów:

$$\boldsymbol{D} = \begin{bmatrix} D_1 \\ \dots \\ D_{LD} \end{bmatrix}$$

Treść dokumentów jest mieszanką różnych tematów. Niech $\boldsymbol{T}$ będzie zbiorem LT tematów:

$$\boldsymbol{T} = \begin{bmatrix} t_1 \\ \dots \\ t_{LT} \end{bmatrix}$$

Model LDA

Udział poszczególnych tematów w rozpatrywanych dokumentach opisany jest za pomocą macierzy $\boldsymbol{\Theta}$:

$$\boldsymbol{\Theta} = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,LT} \\ \cdots & \cdots & \cdots \\ \theta_{LD,1} & \cdots & \theta_{LD,LT} \end{bmatrix}$$

której element $\theta_{i,j}$ może być interpretowany jako prawdopodobieństwo wystąpienia j-tego tematu w i-tym dokumencie.

Model LDA

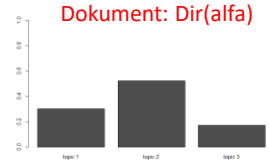
Każdy z tematów definiowany są poprzez informację o rozkładzie występujących w nim słów. Definicje tematów ujęte są w postaci macierzy $\boldsymbol{\Phi}$:

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_{1,1} & \cdots & \phi_{1,LV} \\ \cdots & \cdots & \cdots \\ \phi_{LT,1} & \cdots & \phi_{LT,LV} \end{bmatrix}$$

Element $\phi_{i,j}$ może być interpretowany jako prawdopodobieństwo wystąpienia j-tego słowa w i-tym temacie.

LDA jako model generatywny

Dokument: Dir(alfa)



Temat 1: Dir(beta)

Temat 2: Dir(beta)

Temat 3: Dir(beta)

for i= 1: docLength {

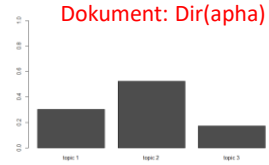
- *wybierz temat*
- *wybierz słowo z wybranego tematu*

}

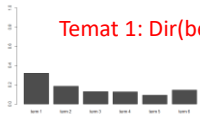
Określanie przynależności słów do tematów

słowo

Dokument: Dir(apha)



Temat 1: Dir(beta)



Temat 2: Dir(beta)

Temat 3: Dir(beta)

p(słowo|Temat 1) * p(Temat 1)

p(słowo|Temat 2) * p(Temat 2)

p(słowo|Temat 3) * p(Temat 3)

wybierz wartość maksymalną

Tworzenie modelu LDA

```
library(tm)
library(topicmodels)

katalog<-"C:/Users/pawel_000/Documents/Ksiazka/Ksiazka 2015 - obliczenia/Literatura polska - stem/"
korpus <- VCorpus(DirSource(katalog,encoding="UTF-8"), readerControl = list(reader=readPlain))
korpus<-tm_map(korpus,removeNumbers)
stoplista <-readLines("C:/Users/pawel_000/Documents/Ksiazka/Ksiazka 2015 - obliczenia/R/AnalizaSkupien/stoplista_PL.txt",encoding="UTF-8")
korpus<-tm_map(korpus,removeWords,stoplista)

dtm<-DocumentTermMatrix(korpus)
```


Tworzenie modelu LDA

```
n.words <- ncol(dtm)

n_group <- 6

lda.model6<-LDA(dtm,k=n_group, method = "Gibbs",control = list(burnin = 2000,thin = 100, iter = 3000))

perp6 <- perplexity(lda.model6,dtm)

res6<-posterior(lda.model6)
```

Udział tematów w dokumentach

```
> res6$topics
                                      1           2          3           4
AM_Dziady_III.txt             0.008542110 0.585339228 0.02090497 0.020741226
AM_Pan_Tadeusz.txt            0.008267569 0.072828229 0.03479588 0.016711532
BP_Katarynka.txt              0.103282003 0.102348732 0.05101882 0.025353865
EO_Nad_Niemnem_t_I.txt        0.020224584 0.022206209 0.02350327 0.006965712
GZ_Moralnosc_Pani_Dulskiej.txt 0.733515613 0.036772864 0.02096212 0.041409098
HS_Janko_Muzykant.txt         0.334574336 0.081409779 0.07172996 0.059071730
HS_Krzyzacy_t_I.txt           0.022248414 0.010967054 0.81558063 0.005107482
JS_Kordian.txt                0.011545919 0.628547506 0.03388860 0.025559669
MK_Nasza_Szkapa.txt           0.635259259 0.011970370 0.04432593 0.016059259
SW_Noc_LIstopadowa.txt        0.011304709 0.843367559 0.02802842 0.038918274
SW_Wesele.txt                 0.051815133 0.067870179 0.02429220 0.747509125
WR_Chlopi_t_I.txt             0.837017280 0.002332726 0.02433077 0.017121328
                                      5           6
AM_Dziady_III.txt             0.204055455 0.16041701
AM_Pan_Tadeusz.txt            0.651504345 0.21589244
BP_Katarynka.txt              0.084149946 0.63384663
EO_Nad_Niemnem_t_I.txt        0.037807002 0.88929322
GZ_Moralnosc_Pani_Dulskiej.txt 0.008123316 0.15921699
HS_Janko_Muzykant.txt         0.045668900 0.40754530
HS_Krzyzacy_t_I.txt           0.018762979 0.12733345
JS_Kordian.txt                0.125506787 0.17495152
MK_Nasza_Szkapa.txt           0.039881481 0.25250370
SW_Noc_LIstopadowa.txt        0.019160962 0.05922008
SW_Wesele.txt                 0.023034428 0.08547894
WR_Chlopi_t_I.txt             0.012050950 0.10714694
>
```

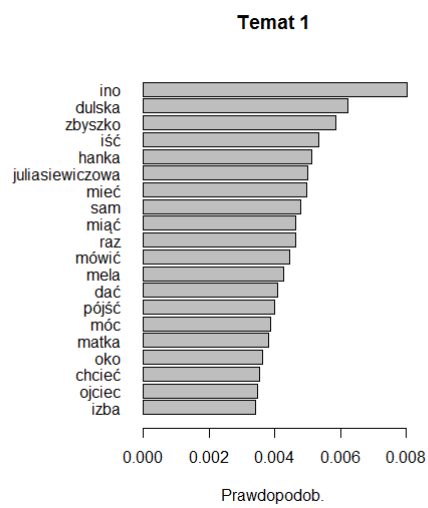

Udział tematów w słowach

```
> res6$terms[,c("ojczyzna","dom","miłość","śmierć")]
      ojczyzna          dom       miłość       śmierć
1 2.085258e-06 2.525247e-03 2.085258e-06 4.608420e-04
2 5.118147e-04 3.178973e-06 1.719824e-03 1.529086e-03
3 1.788215e-06 1.788215e-06 6.097812e-04 2.183410e-03
4 6.226379e-06 6.226379e-06 1.687349e-03 6.226379e-06
5 7.124698e-04 1.632634e-03 2.629040e-06 2.918234e-04
6 1.720676e-06 4.940060e-03 2.770288e-04 1.720676e-06
>
```

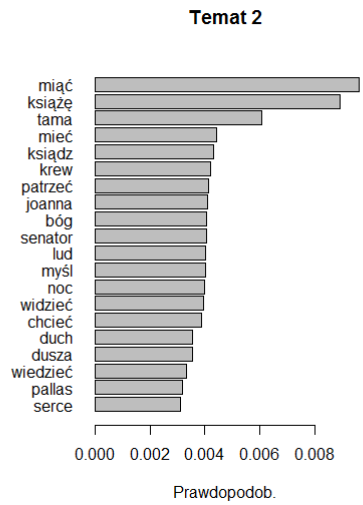
Prezentacja tematu

```
#Prezentacja tematu 1

par(mai=c(1,2,1,1)) #wielkość marginesów

t1=head(sort(res6$terms[1,],decreasing=TRUE),20)

barplot(rev(t1),horiz=TRUE,las=1,main="Temat
1",xlab="Prawdopodob.")
```

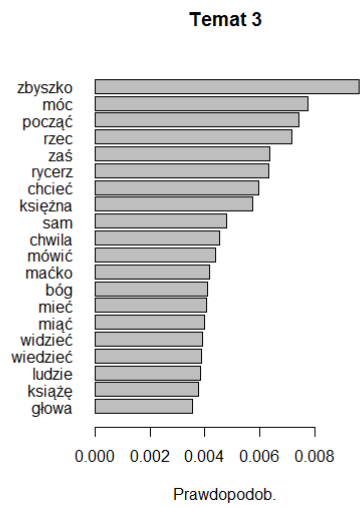

Temat 1



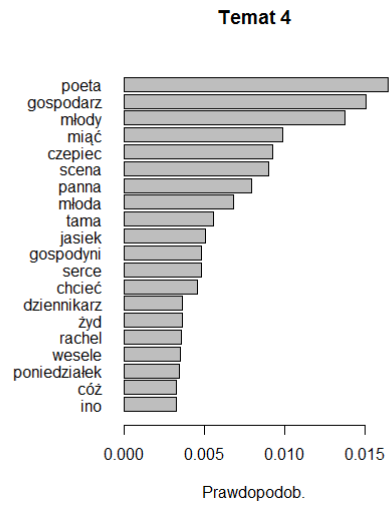
Temat 2



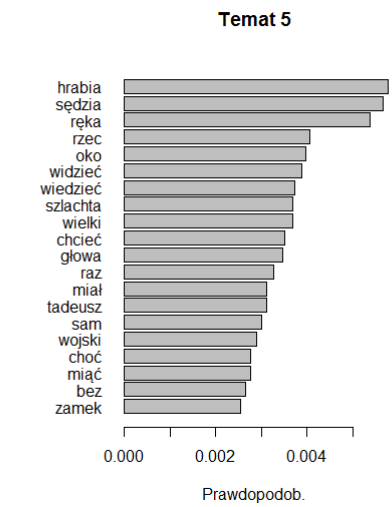
Temat 3



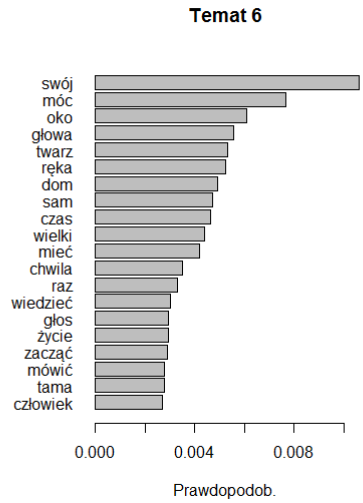
Temat 4



Temat 5



Temat 6



Funkcja wiarygodności

- **Wiarygodność modelu** – prawdopodobieństwo odtworzenia przez model posiadanego zbioru danych
- **Funkcja wiarygodności** – prawdopodobieństwo wygenerowania przez model posiadanego zbioru danych
- $\boldsymbol{doc} = \{w_1, w_2, w_3, ..., w_N\}$
- $L(\boldsymbol{doc}) = p(w_1|model)*p(w_2|model) * ... * p(w_N|model)$
- $\log\text{-}L(\boldsymbol{doc}) = \log(p(w_1|model)) + ... + \log(p(w_N|model))$

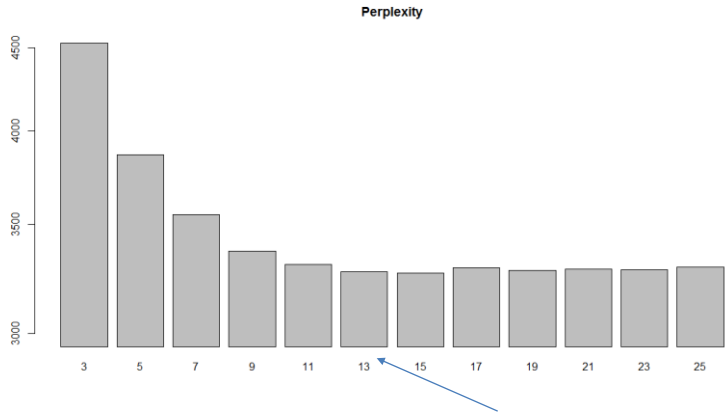
Nieokreśloność modelu (perplexity)

- log-L(***doc***) = log(p(w_1|*model*)) + ... + log(p(w_N|*model*))
- Perplexity (nieokreśloność, niezdecydowanie)
- *perplexity*(***doc***) = exp(- *log-L*(***doc***)) = exp (1 / *log-L*(***doc***))

Określenie optymalnej struktury modelu

- Struktura modelu jest zdeterminowana przez liczbę klas ukrytych
- Dobór liczby klas:
  - na podstawie wartości funkcji wiarygodności (maksymalizacja funkcji) lub wartości funkcji *perplexity* (minimalizacja funkcji) – wybierany jest model, który najlepiej potrafi odtworzyć posiadany korpus
  - inne kryteria (wiedza merytoryczna dotycząca badanego zjawiska).

Wybór optymalnej liczby klas w modelu LDA



Optymalna liczba klas

Wybór optymalnej liczby klas w modelu LDA

Select number of topics for LDA model

Murzintcev Nikita

2016-10-24

Package can be installed from CRAN

```
install.packages("ldatuning")
```

or downloaded from the GitHub repository (developer version).

```
install.packages("devtools")
```

Pakiet ldatuning