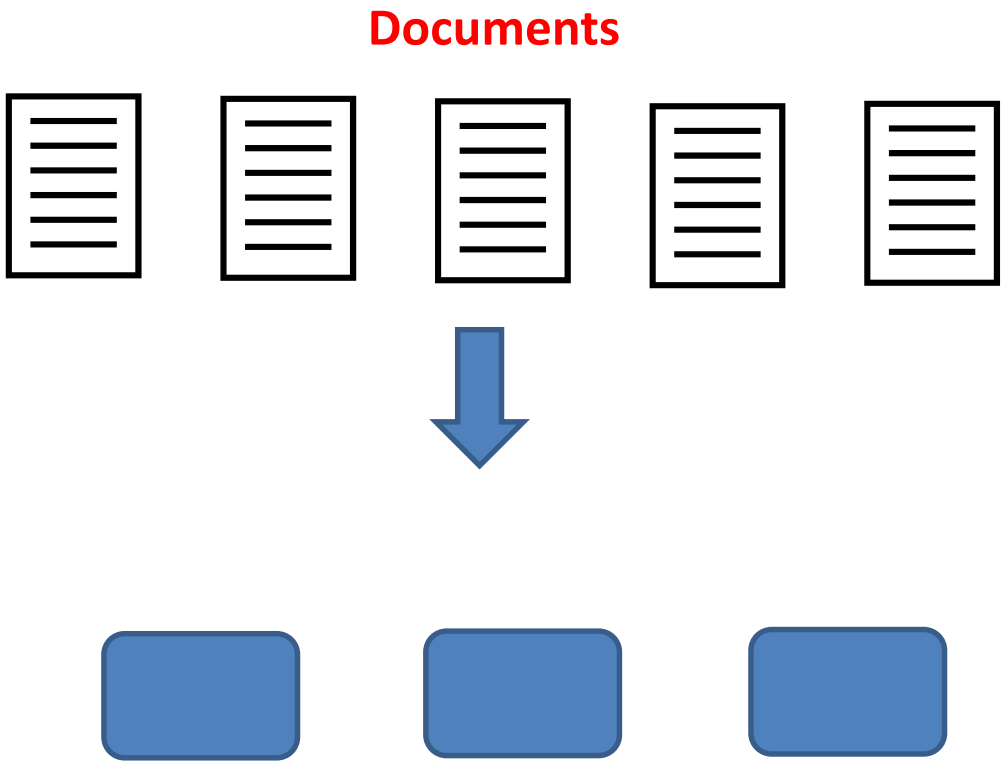
Latent Dirichlet Allocation Models and their Evaluation

IT for Practice 2016

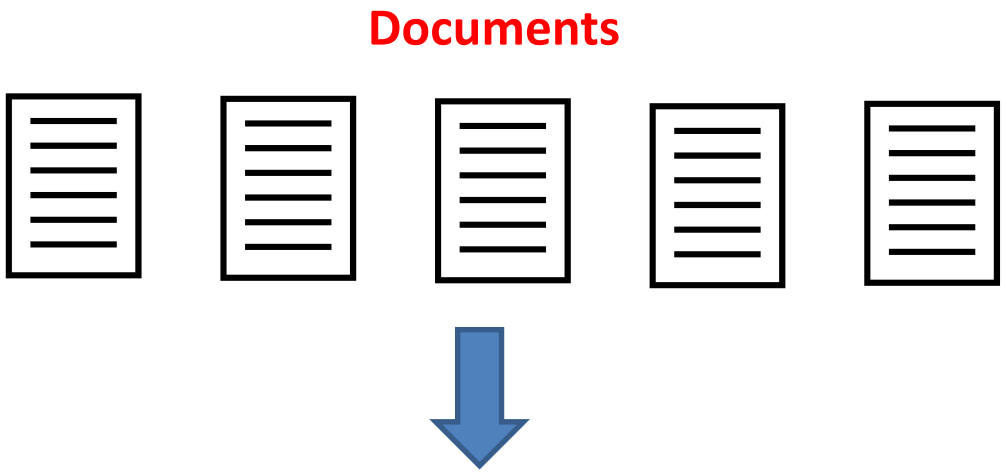
Paweł Lula
Cracow University of Economics , Poland
pawel.lula@uek.krakow.pl

Latent Dirichlet Allocation (LDA)



Latent Dirichlet Allocation *– completely* ***unsupervised*** *method of topics identification.*

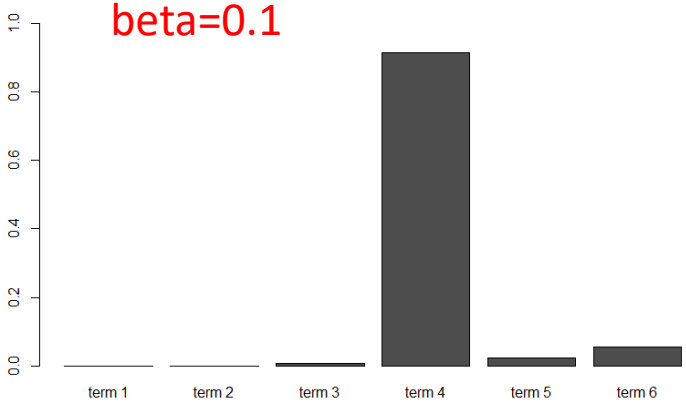
Latent Dirichlet Allocation (LDA)

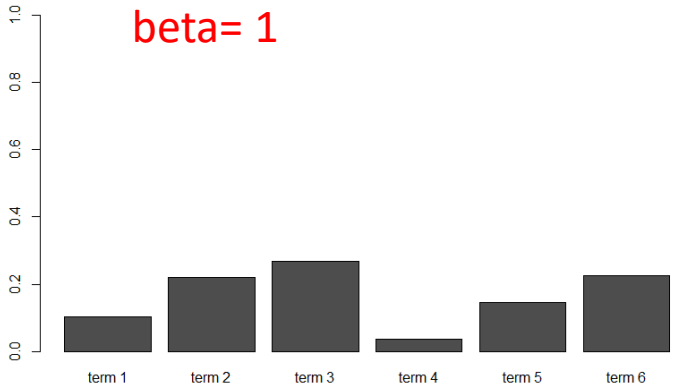


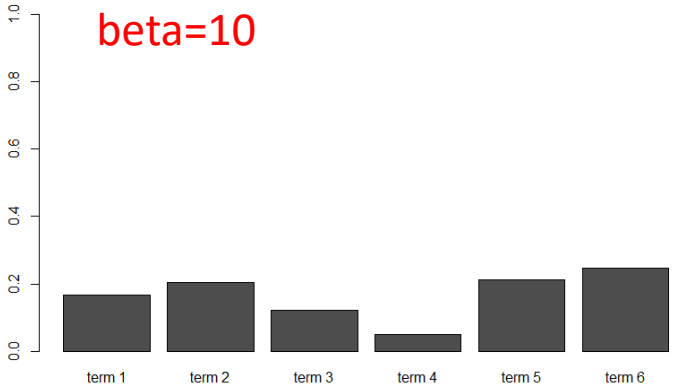
Latent Dirichlet Allocation *– completely* ***unsupervised*** *method of topics identification.*

Topics are described in terms of discrete probabilities over words.

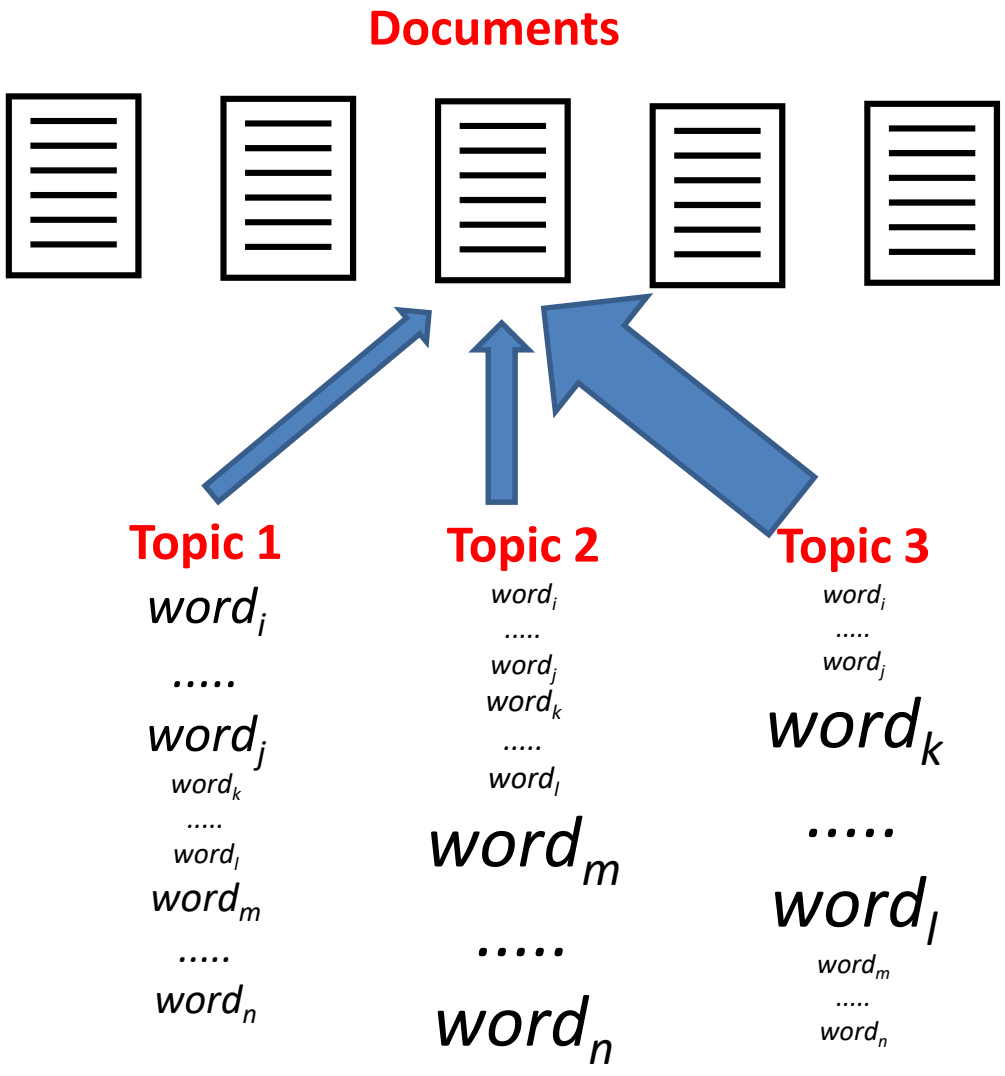
Description of topics: Dir(beta)







Latent Dirichlet Allocation (LDA)



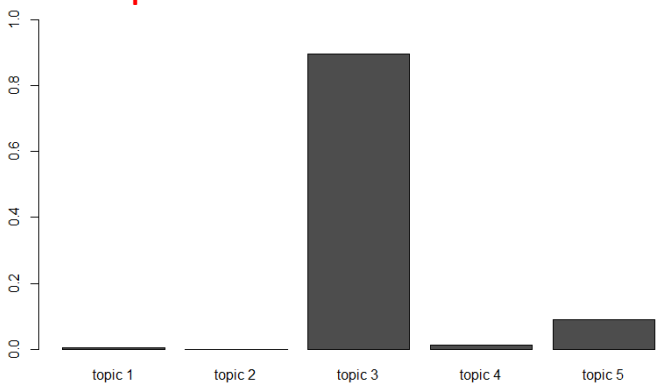
Latent Dirichlet Allocation *– completely* ***unsupervised*** *method of topics identification.*

Topics are described in terms of discrete probabilities over words.

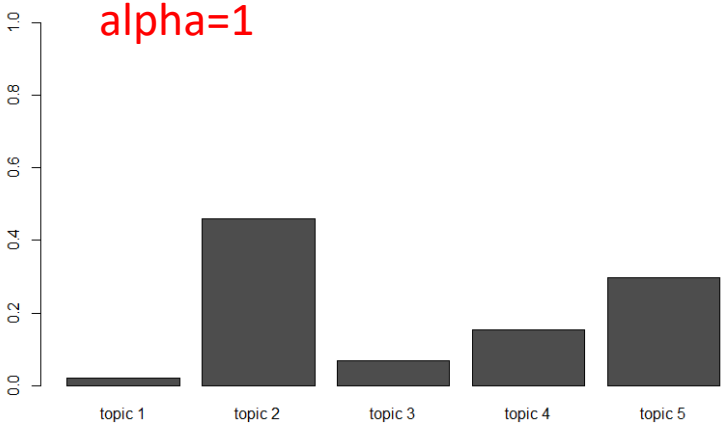
Each document can be modeled as a mixture of topics. Documents are describes in terms of discrete probabilities over topics.

Description of documents: Dir(alpha)

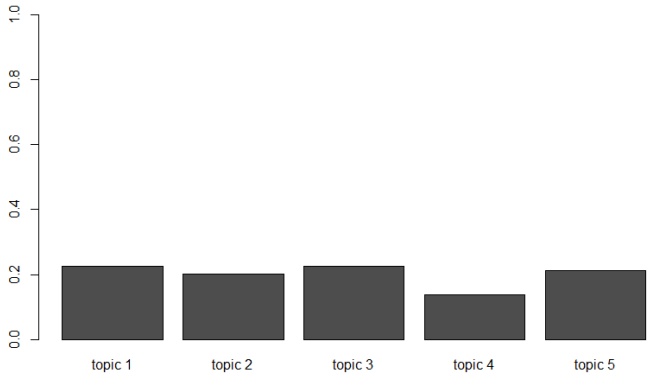
alpha=0.1



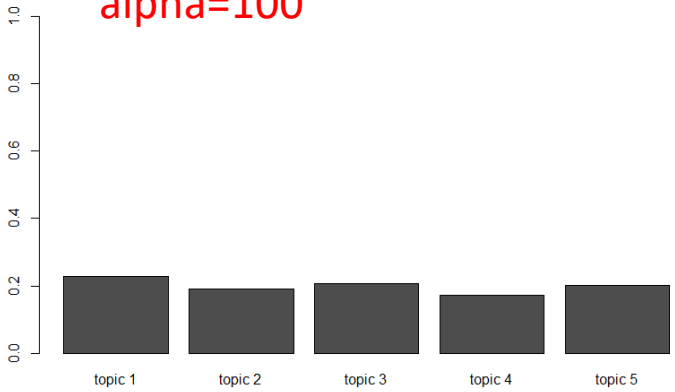
alpha=1



alpha=10

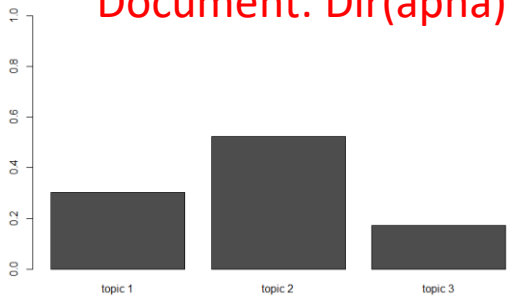


alpha=100



Generating a new document

Document: Dir(apha)



Topic 1: Dir(beta)

Topic 2: Dir(beta)

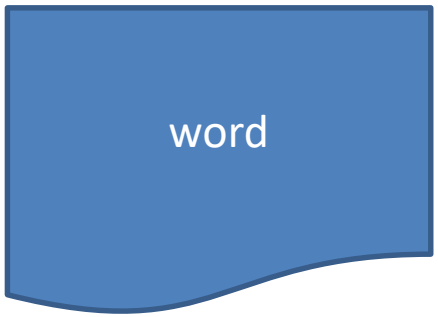
Topic 2: Dir(beta)

for i= 1: docLength {

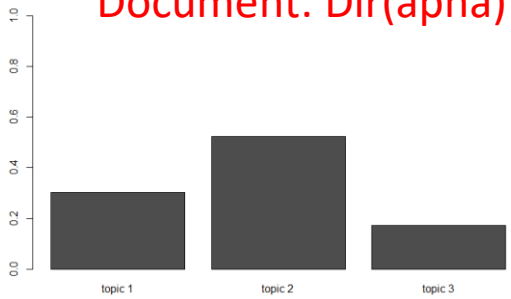
- *choose topic*
- *choose word from chosen topic*

}

Assigning the most probable topic to words



Document: Dir(apha)



Topic 1: Dir(beta)

Topic 2: Dir(beta)

Topic 2: Dir(beta)

p(word|Topic 1) * p(Topic 1)

p(word|Topic 2) * p(Topic 2)

p(word|Topic 3) * p(Topic 3)

choose max value

Evaluation of LDA models

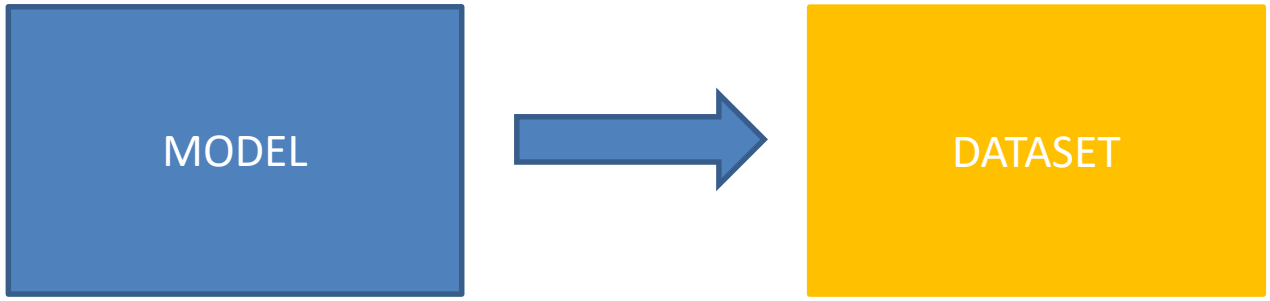
- There are two main forms of *LDA* models evaluations:
  - human assessment of model's results,
  - evaluation based on measures calculated during simulation experiments.

	Human evaluation	Automatic evaluation
Advantages	• multifaceted, • based on external expert knowledge.	• objective, based on well-defined indicators, • repeatable.
Disadvantages	• subjective, • prone to errors, • based on rules difficult for identification and explicit expression.	• difficult for conducting at the semantic level.

Measures of quality for LDA models

- measures of model's ability to dataset reconstruction
  - likelihood function
  - perplexity
- measures of topic's diversity
  - average Kullback-Leibler divergence
  - Bhattacharyya distance
- measures of topic's coherence
  - extrinsic coherence (Newman et al., 2010)
  - intrinsic coherence (Mimno et al. 2011)

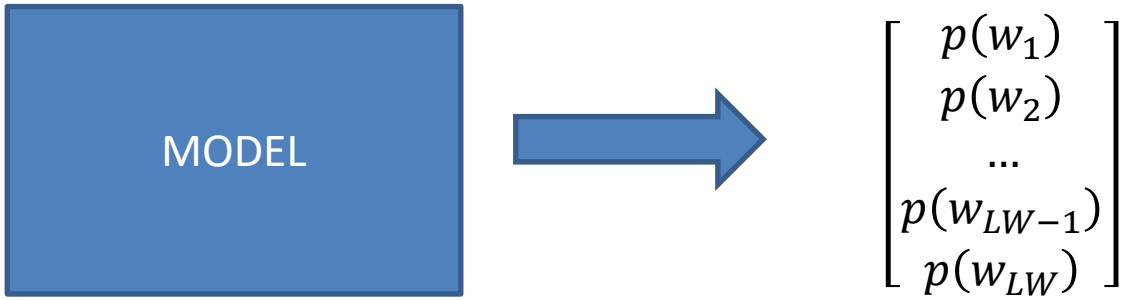
Likelihood



likelihood(MODEL|DATASET) = *probability*(DATASET|MODEL)

The greater value of likelihood calculated, the better quality of the model

Perplexity



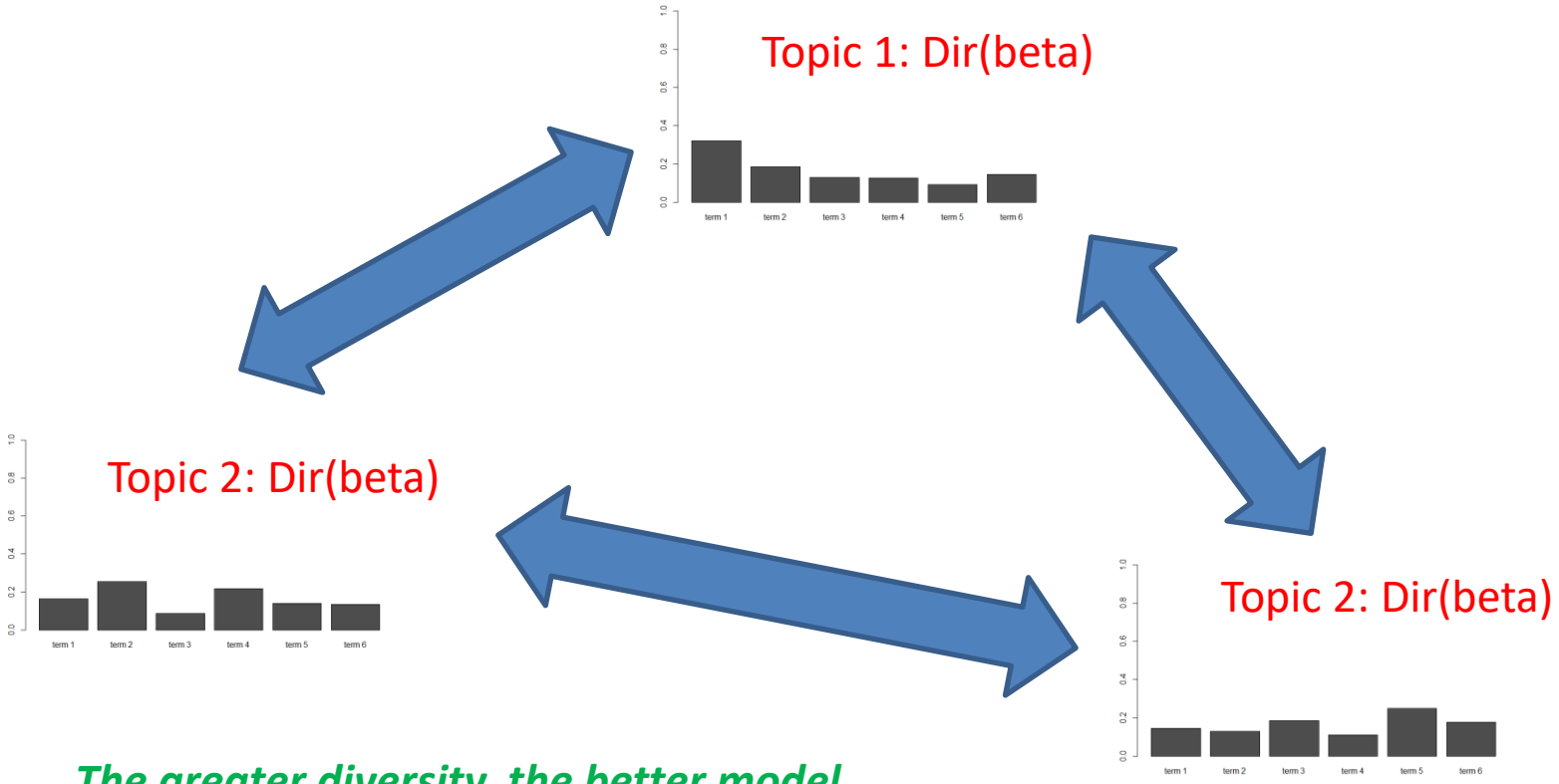
$$perplexity(\boldsymbol{D}) = \sqrt[LW]{\frac{1}{\prod_{i=1}^{LW} p(w_i)}}$$

The lower perplexity, the better quality of the model.

Topic's diversity

Measures for topic's diversity

- average Kullback-Leibler divergence
- Bhattacharyya distance



The greater diversity, the better model.

Topic' coherence

Topic:

word.1
word.2
word.3
word.4
word.5
word.6
word.7
word.8
word.9
word.10
word.11
word.12
.....

$$coherence(t) = \sum_{i=2}^{n} \sum_{j=1}^{n-2} association\left(w_i^{(t)}, w_j^{(t)}\right)$$

Association between two words depends on the number of times they appear together in the same document

The greater coherence, the better model.

Multi-criteria analysis of LDA quality indicators

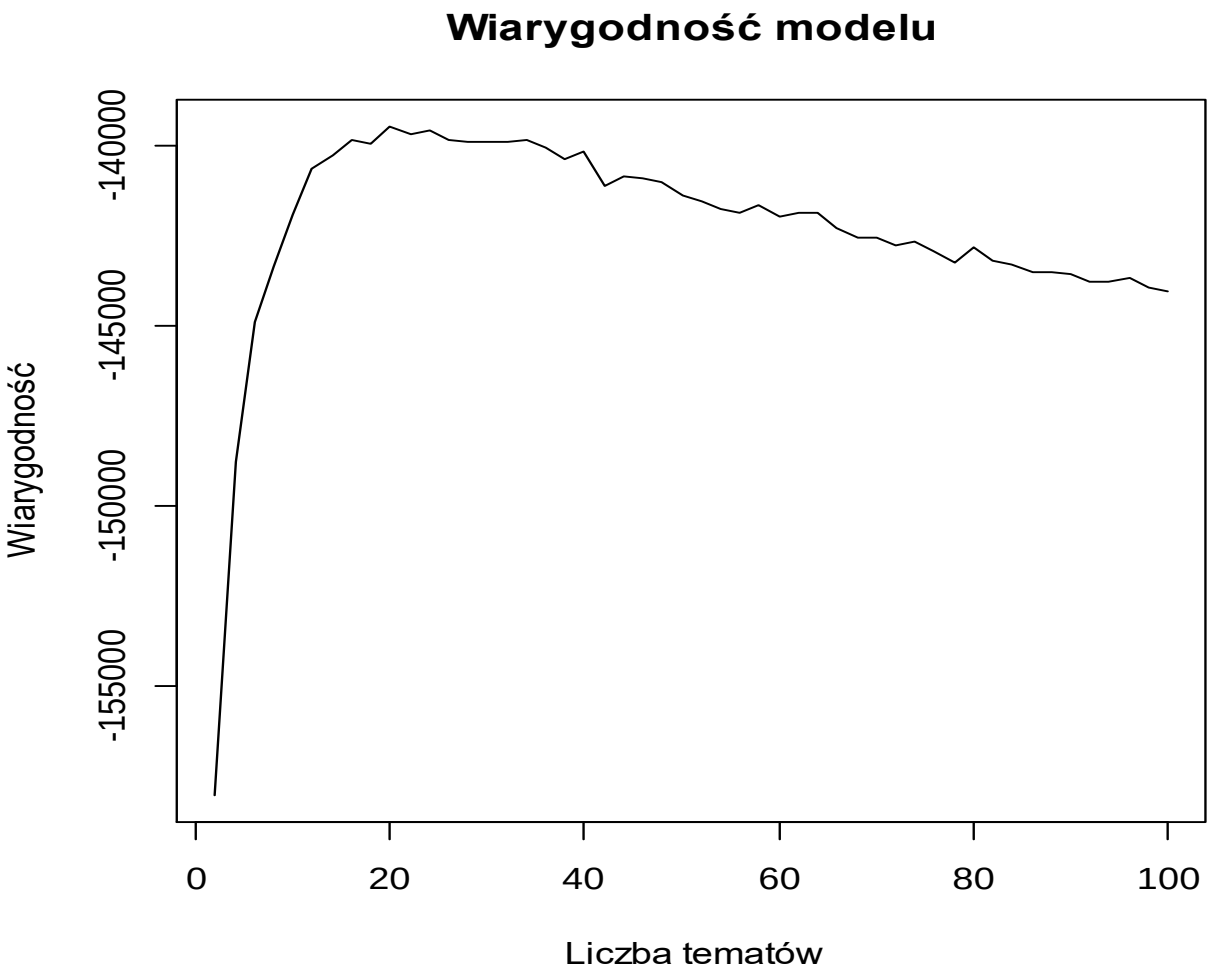
Indicators of *LDA* quality		Q_1	Q_2	Q_3	...	Q_P
Weights		w_1	w_1	w_1	...	w_1
Models	M_1	$q_{1,1}$	$q_{1,1}$	$q_{1,1}$	...	$q_{1,1}$
	M_2	$q_{1,1}$	$q_{1,1}$	$q_{1,1}$	...	$q_{1,1}$
	...	...	...	...	...	...
	M_T	$q_{1,1}$	$q_{1,1}$	$q_{1,1}$	...	$q_{1,1}$

Hellwig development pattern method (Hellwig 1968)

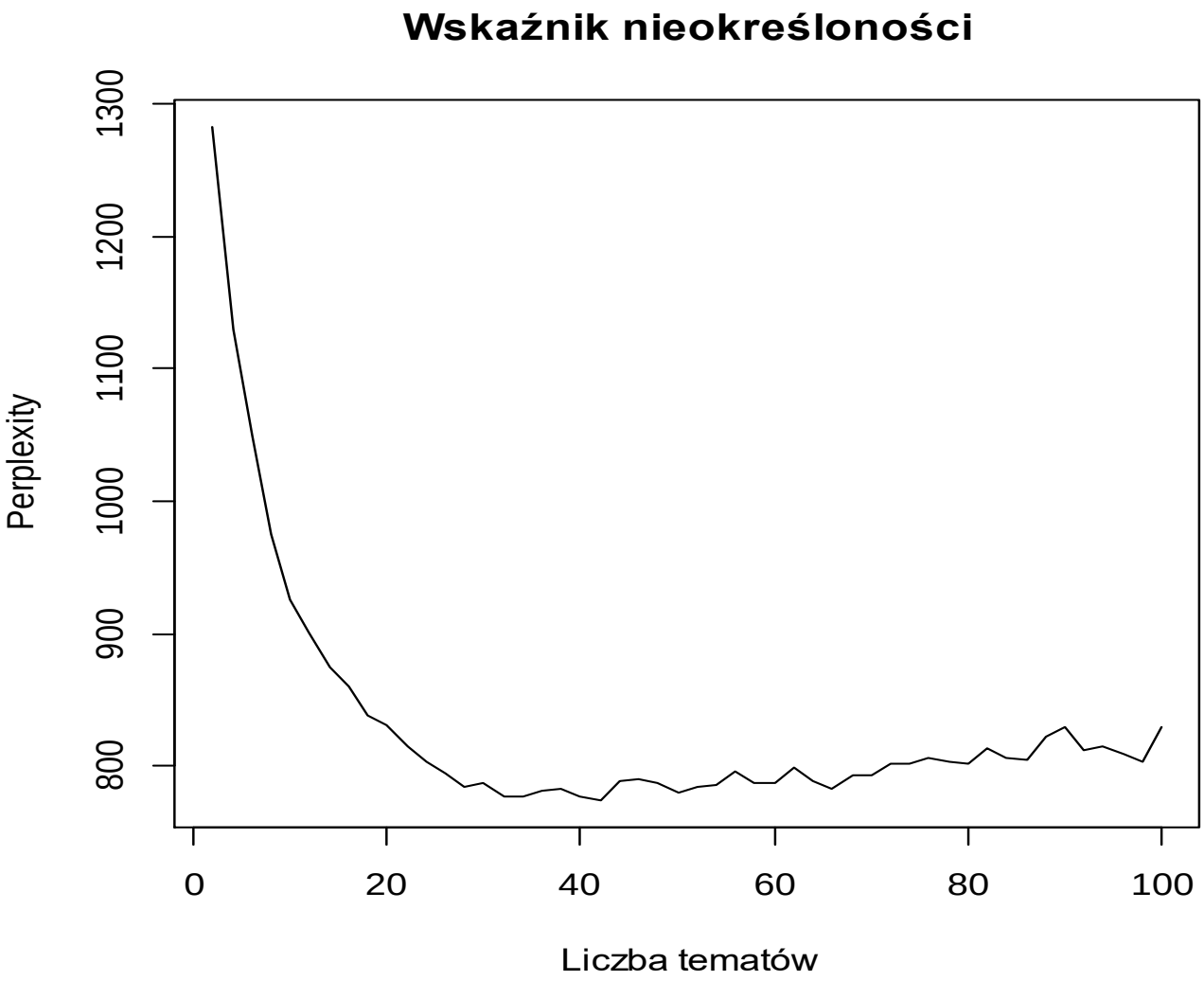
Simulation experiment

- The approach presented here was used for evaluation of abstracts of PhD thesis prepared in Polish language at the Cracow University of Economics in the period 2010-2015.
- The corpus was composed of 159 documents.
- Stemming process was performed with the help of *Morfologik* system .
- Several *LDA* models were prepared using *topicmodels* packet for *R* system.
- For every model the analysis of likelihood, perplexity, topic diversity and topic coherence was performed.
- Next an aggregated quality measure was calculated with Hellwig development pattern method (Hellwig 1968).
- As a result the LDA model with ***six*** topics was chosen.

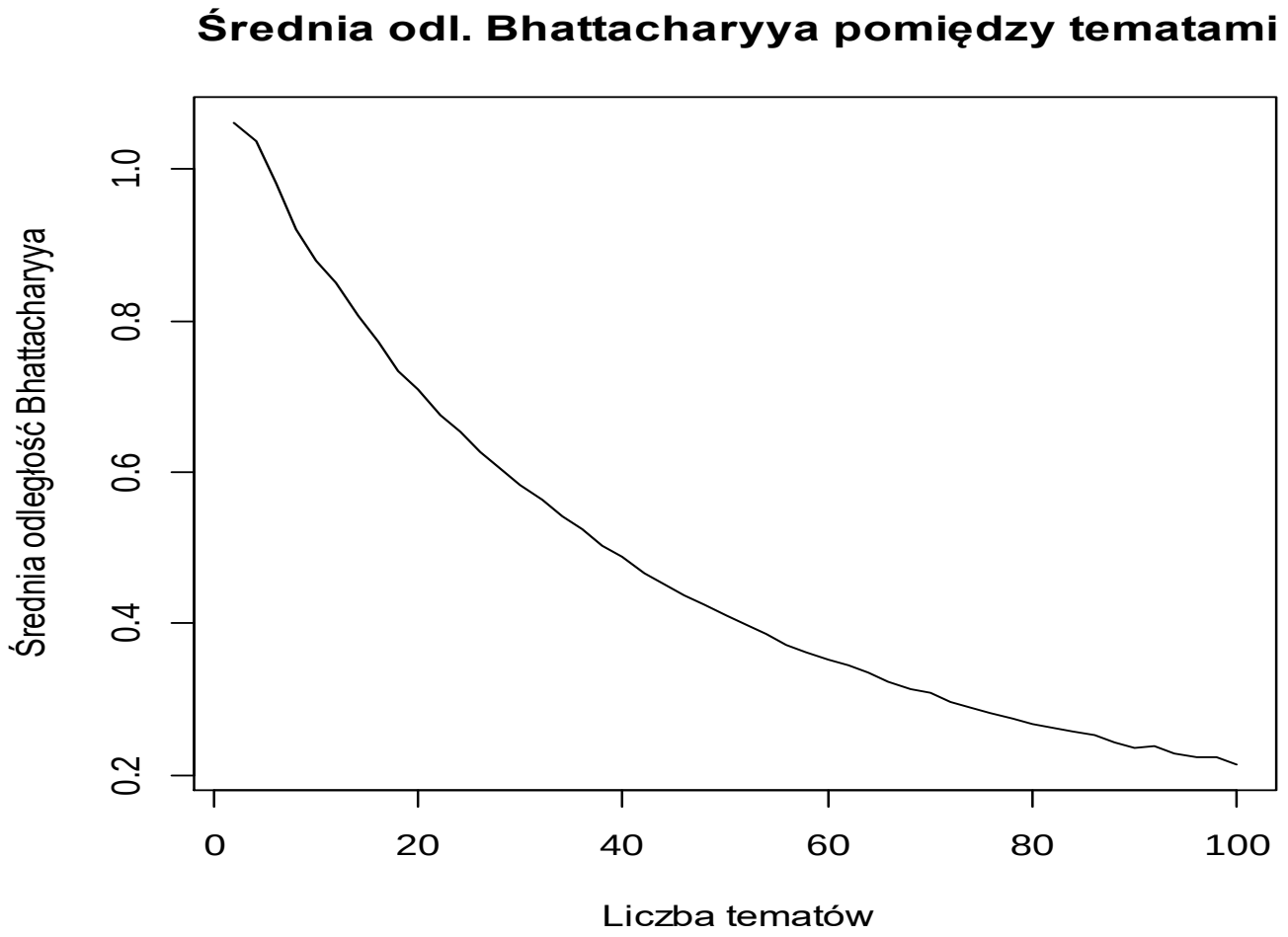
Likelihood function



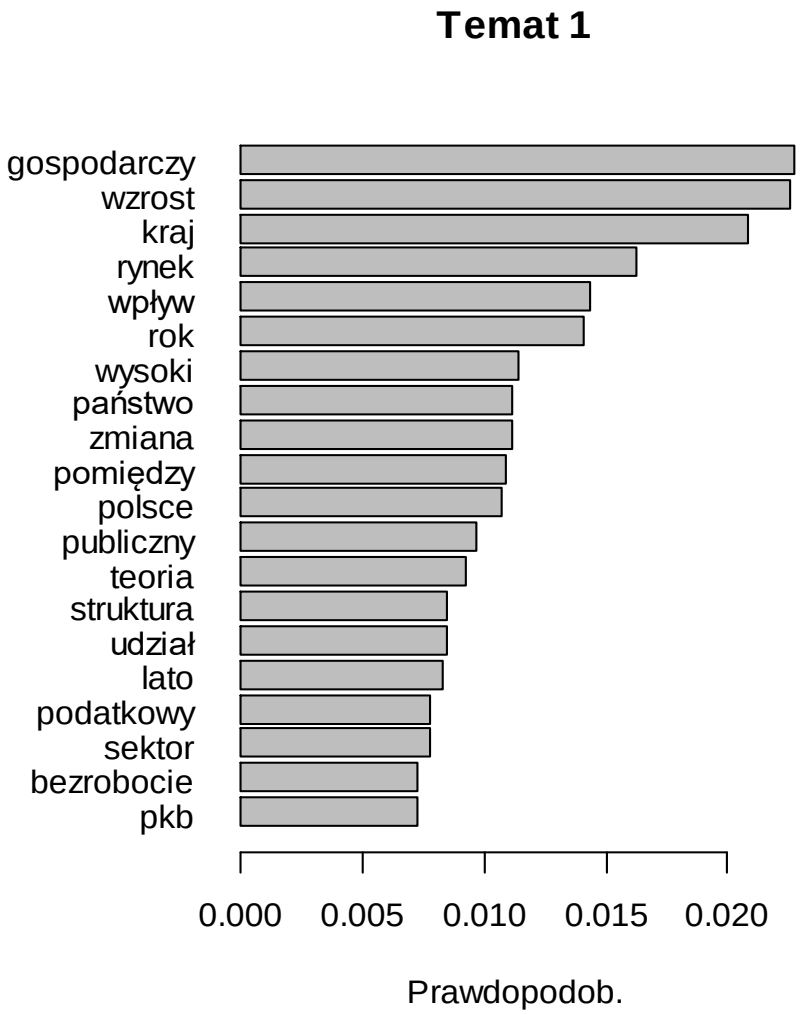
Perplexity


Wskaźnik nieokreśloności
Perplexity
800
900
1000
1100
1200
1300
0
20
40
60
80
100
Liczba tematów

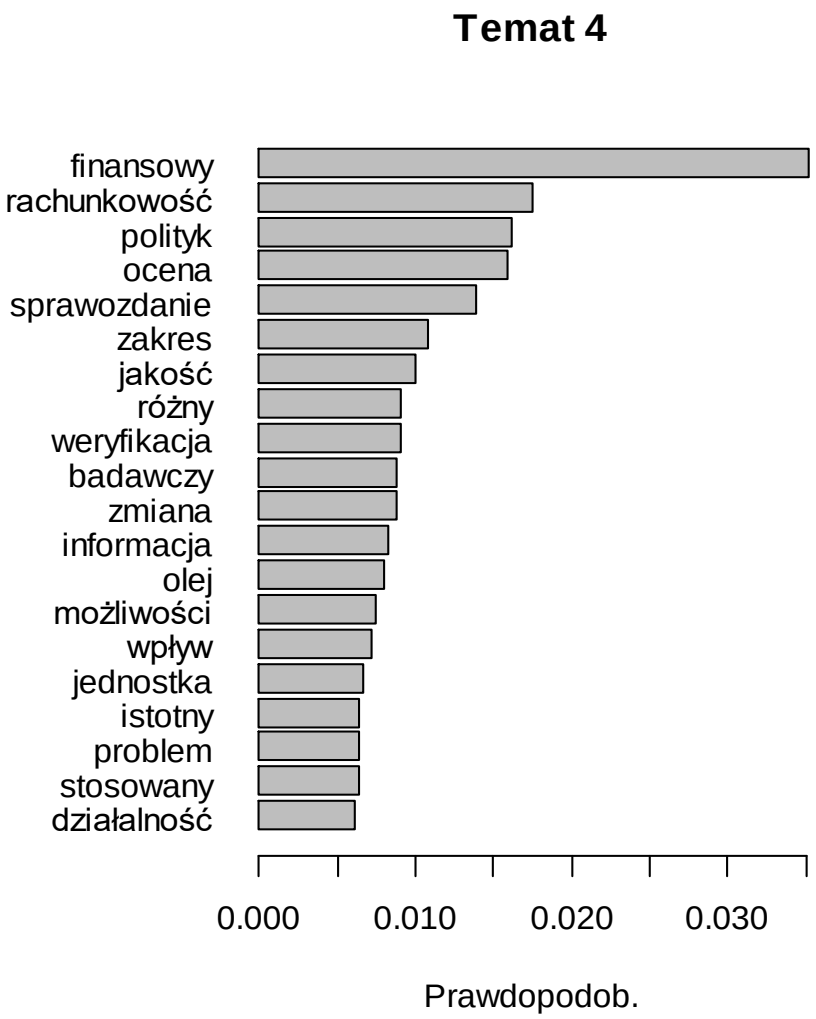
Average distance between topics (topics' diversity)



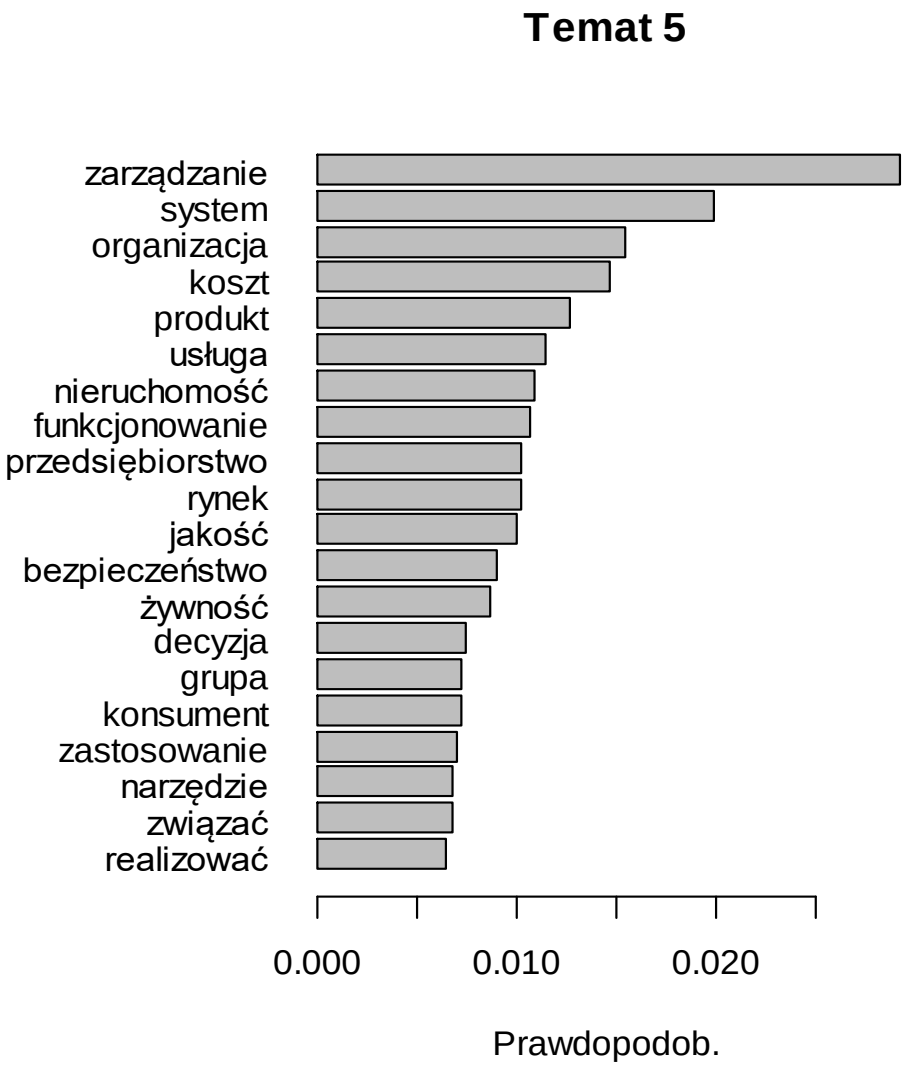
Topic 1


Temat 1
gospodarczy
wzrost
kraj
rynek
wpływ
rok
wysoki
państwo
zmiana
pomiędzy
polsce
publiczny
teoria
struktura
udział
lato
podatkowy
sektor
bezrobocie
pkb
0.000
0.005
0.010
0.015
0.020
Prawdopodob.

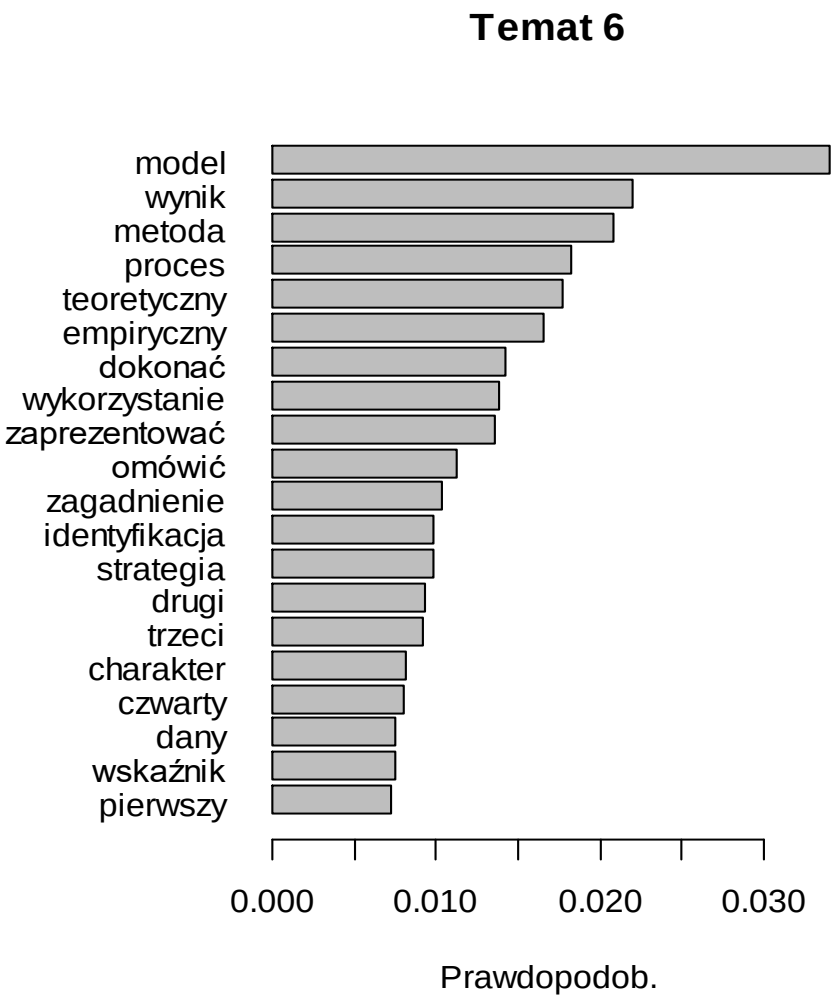
Topic 4


Temat 4
finansowy
rachunkowość
polityk
ocena
sprawozdanie
zakres
jakość
różny
weryfikacja
badawczy
zmiana
informacja
olej
możliwości
wpływ
jednostka
istotny
problem
stosowany
działalność
0.000
0.010
0.020
0.030
Prawdopodob.

Temat 5


Temat 5
zarządzanie
system
organizacja
koszt
produkt
usługa
nieruchomość
funkcjonowanie
przedsiębiorstwo
rynek
jakość
bezpieczeństwo
żywność
decyzja
grupa
konsument
zastosowanie
narzędzie
związać
realizować
0.000
0.010
0.020
Prawdopodob.

Topic 6

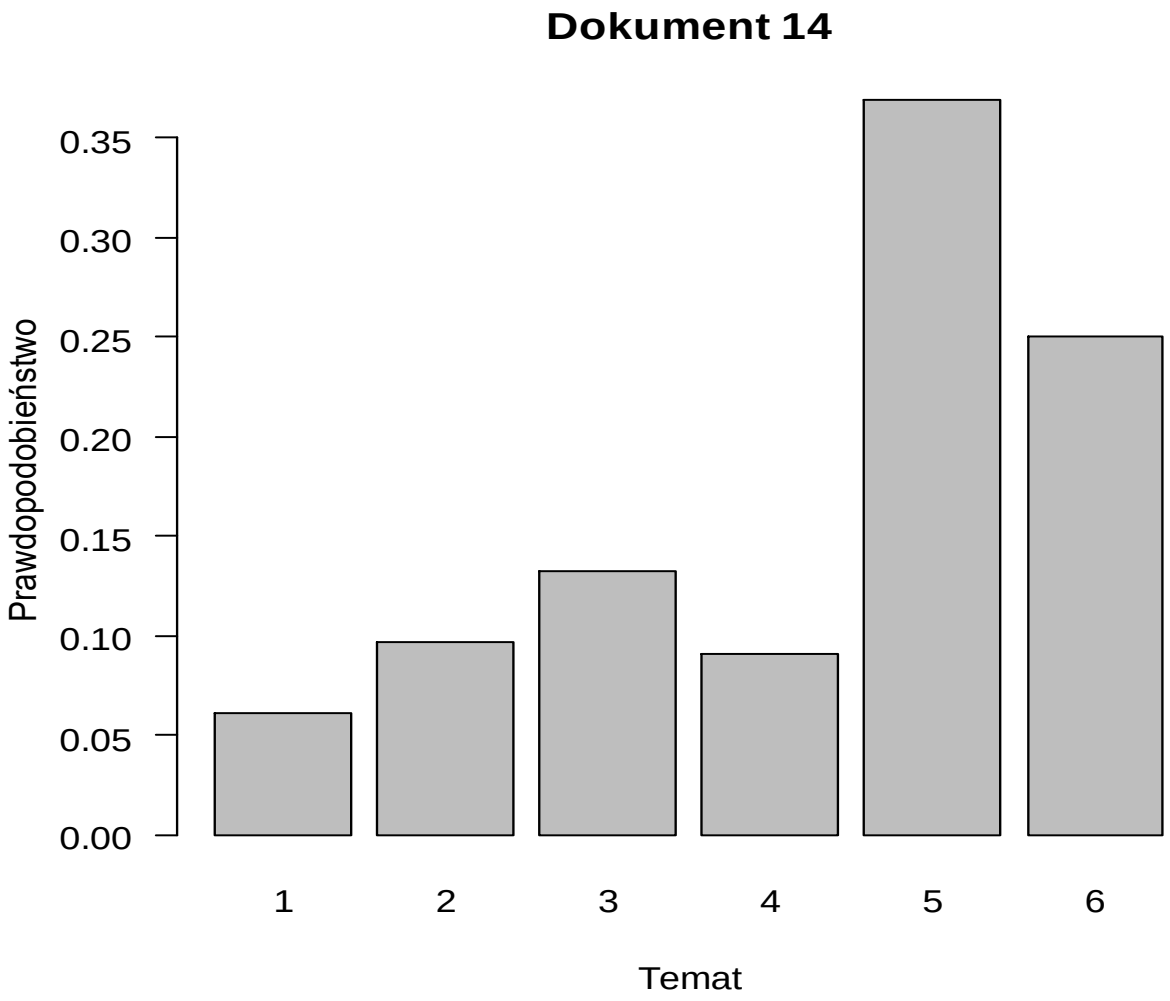

Temat 6
model
wynik
metoda
proces
teoretyczny
empiryczny
dokonać
wykorzystanie
zaprezentować
omówić
zagadnienie
identyfikacja
strategia
drugi
trzeci
charakter
czwarty
dany
wskaźnik
pierwszy
0.000
0.010
0.020
0.030
Prawdopodob.

An exemplary abstract of PhD thesis

Modelowanie rachunku kosztów i dobór metod budżetowania w szpitalu

Rozprawę poświęcono zagadnieniom modelowania rachunku kosztów w szpitalu z uwzględnieniem przydatności opracowanych rozwiązań w procesie budżetowania. W rozdziale pierwszym omówiono prawne aspekty funkcjonowania szpitala jako zakładu opieki zdrowotnej. Rozdział drugi poświęcony jest zagadnieniom rachunku kosztów ze szczególnym uwzględnieniem jego specyfiki w szpitalu. Poddano analizie definicje rachunku kosztów zawarte zarówno w literaturze krajowej, jak i zagranicznej, omówiono przekroje ewidencyjne kosztów stosowane w szpitalach identyfikując ośrodki i nośniki kosztów, zwrócono także uwagę na przeobrażenia i rozwój rachunku kosztów w ostatnich latach oraz omówiono kalkulację kosztów procedur medycznych, jako podstawowych nośników kosztów w szpitalu. W rozdziale trzecim poruszone zostały zagadnienia związane z metodą budżetową i jej rolą w procesie zarządzania szpitalem. W rozdziale tym omówiono metody budżetowania i dokonano oceny ich przydatności w sporządzaniu budżetu dla szpitala. Przedstawiono również poszczególne etapy procesu budżetowania w szpitalu zwracając uwagę na zagadnienia problematyczne. Rozdział czwarty zawiera badania empiryczne, których przedmiotem są koszty szpitala specjalistycznego. Koszty te poddane zostały szczegółowej analizie w celu zidentyfikowania czynników wpływających na ich poziom. Rezultatem badań jest wielowymiarowa funkcja opisująca koszty operacyjne szpitala, którą może mieć zastosowanie w procesie budżetowania kosztów.

Analysis of the exemplary abstract



Conclusions

- Multifaceted approach to *LDA* model evaluation allows to take into account many different aspects of models
- Experiments show that results obtain with the help of this approach can be useful for analysis of real sets of documents.
- Further research will be focus on methods of introducing domain knowledge into the process of model's evaluation.

Thank you for your attention!