

Využití Gibbsova algoritmu ve shlukovacích úlohách komputační lingvistiky

Pavel Schlesinger
MFF UK

Ústav formální a aplikované lingvistiky



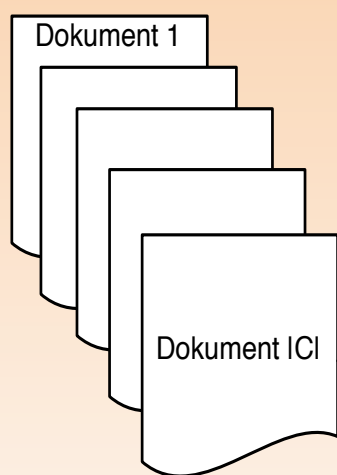
XVI ROBUST
Králíky 31. 1. – 5. 2. 2010

- *určení žánru*
- *autorství*
- *eliminace duplicitních textů*
- *nalezení odvozených dokumentů, např. plagiátů*

*příklad: určení autorství sporného dokumentu:
Karel Čapek, nebo Vladislav Vančura?*

aplikace shlukové analýzy
v počítačové lingvistice

Kolekce C



Dokument d

Shluková analýza **je** často používaná metoda při řešení **úloh** počítačové lingvistiky, ať už se jedná například o zkoumání podobnosti dokumentů nebo jednotlivých slov. Specifickou vlastností této shlukovací **úlohy je** ...

Dokument d – lemmata

Shlukový analýza **být** často používaný metoda při řešení **úloha** počítačové lingvistiky, ať už se jedná například o zkoumání podobnosti dokument nebo jednotlivý slovo. Specifický vlastnost tento shlukovací **úloha být** ...

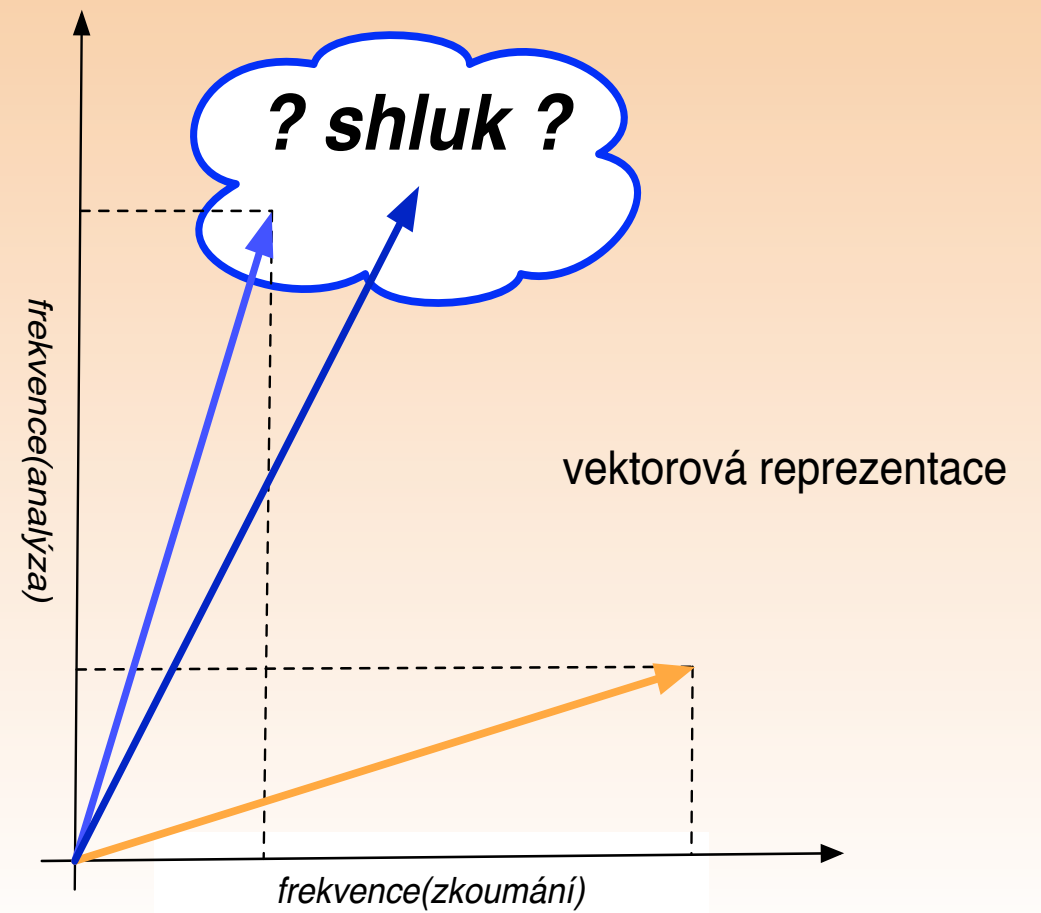
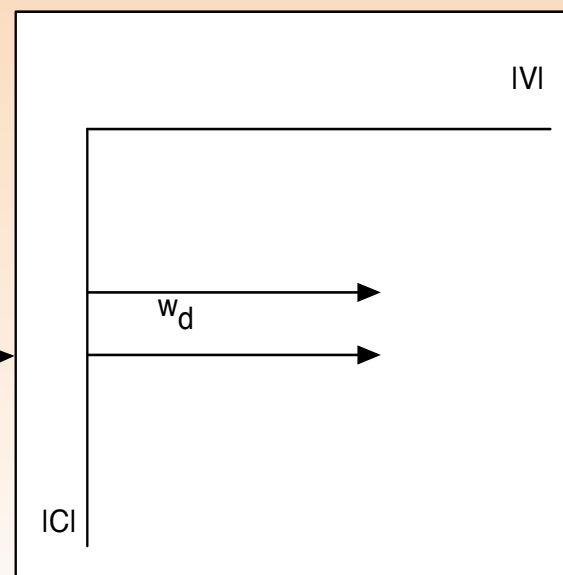
Slovník V

abonent	0
...	
analýza	1
být	2
...	
úloha	2
...	
zkoumání	1
...	
želva	0

kolekce dokumentů

$$\mathbf{w}_d = (w_1, \dots, w_{|V|}), d = 1, \dots, |C|$$

Matice W



Zařazení dokumentu do shluku:

$$z_d = k, k = 1, \dots, K$$

Zařazení celé kolekce:

$$\mathbf{z} = (z_1, \dots, z_d, \dots, z_{|C|})$$

K shluků ... K distribucí nad slovníkem V

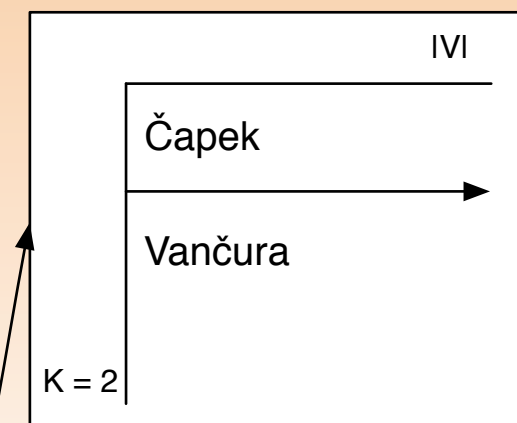
Multinomické rozdělení:

$$\mathbf{w}_d | z_d \sim \text{Mult}(\sum \mathbf{w}_{di}, \boldsymbol{\Theta}^k)$$

$$\boldsymbol{\Theta}^k = (\theta_1^k, \dots, \theta_{|V|}^k), k = 1, \dots, K$$

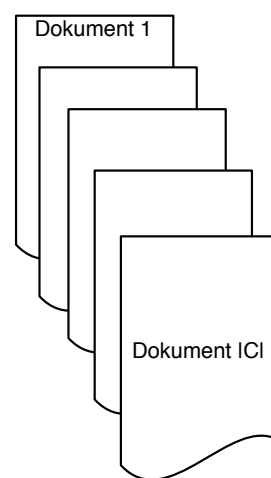
příklad shluků = {Čapek, Vančura}, K = 2

Agregovaná matice F

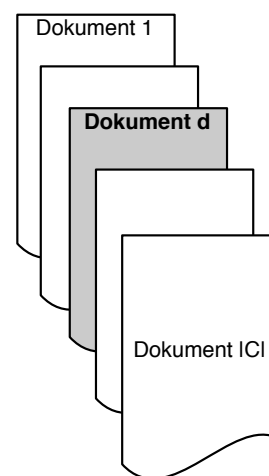


pravděpodobností
model

Kolekce C



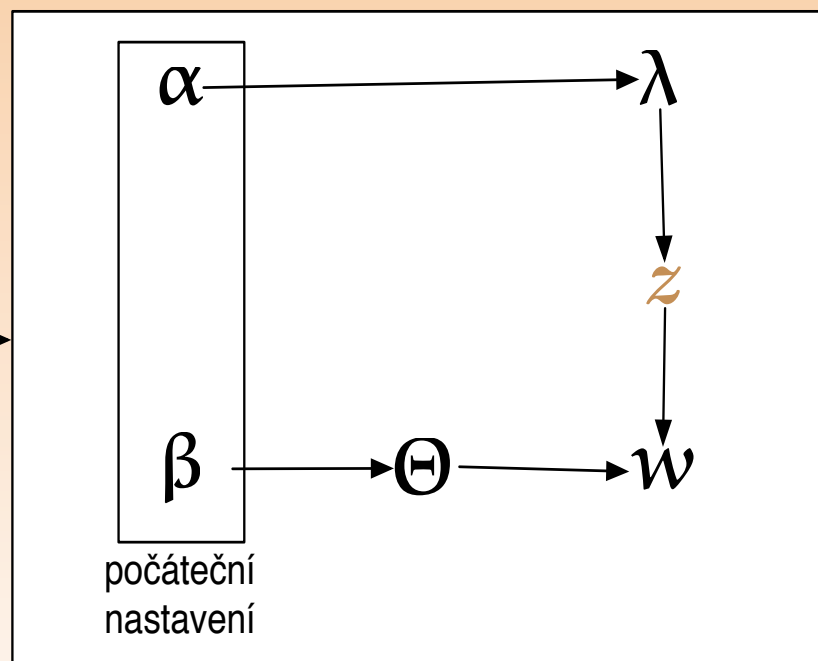
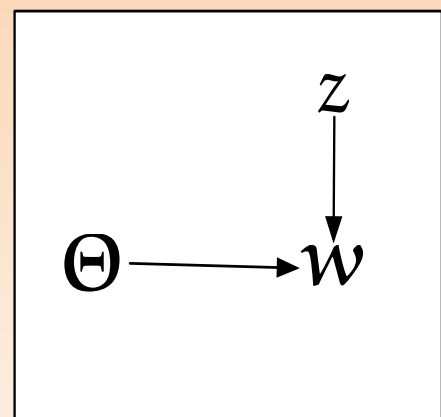
Iterace přes jeden dokument d



Přeřazení ze shluku starého do nového

Update matice F i parametrů

iterativní změna
(dvou) shluků



$$z_d \sim Mult(1; \lambda)$$

$$\lambda \sim Dirichlet(\alpha)$$

$$\Theta^k \sim Dirichlet(\beta)$$

Proč Dirichletovo rozdělení?

- Konjugovaný pár k multinomickému rozdělení
- Snadný update parametrů - součet stávajících hodnoty + napozorované četnosti

Iterativní (kolapsovaný) sampling:
pouze z_d na základě přepočtu $p(z_d = j)$

bayesovský přístup,
(kolapsovaný)
Gibbsův algoritmus

$$p(z, \mathbf{W}, \Theta, \lambda | \alpha, \beta) \Rightarrow$$

$$p(z_d = j | \mathbf{W}, z_{-d}) = \frac{p(z, \mathbf{W})}{p(z_{-d}, \mathbf{W})} = \frac{p(z_d=j, z_{-d}, \mathbf{W})}{p(z_{-d}, \mathbf{W})}$$

$$p(z_d = j | \mathbf{W}, z_{-d}) \propto (\alpha_j + n_j - 1) \frac{B(\beta_{j*})}{B(\beta_{-dj*})}$$

$$\beta_k^* = (\beta_1 + f_{k1}, \dots, \beta_1 + f_{k|V|})$$

$$\beta_{-dj*} = (\beta_1 + f_{j1} - w_{d1}, \dots, \beta_{|V|} + f_{j|V|} - w_{d|V|}).$$

update matice F i parametrů

Děkuji za pozornost.

Dobrou chut'!