PA153 Počítačové zpracování přirozeného jazyka

12 – Strojové účení a ZPJ

Jiří Materna

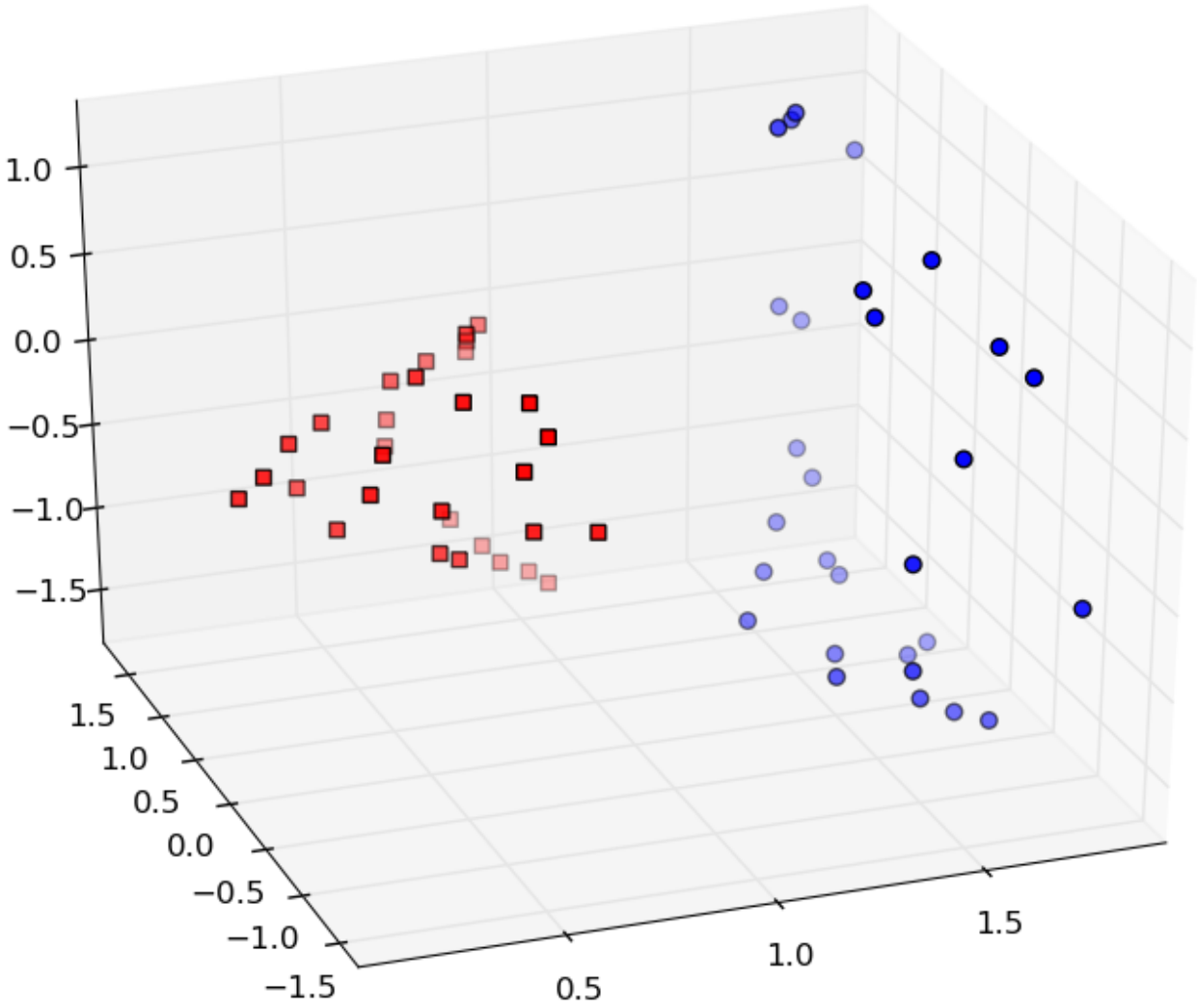
Centrum ZPJ, FI MU, Brno

9. prosince 2013

Strojové učení

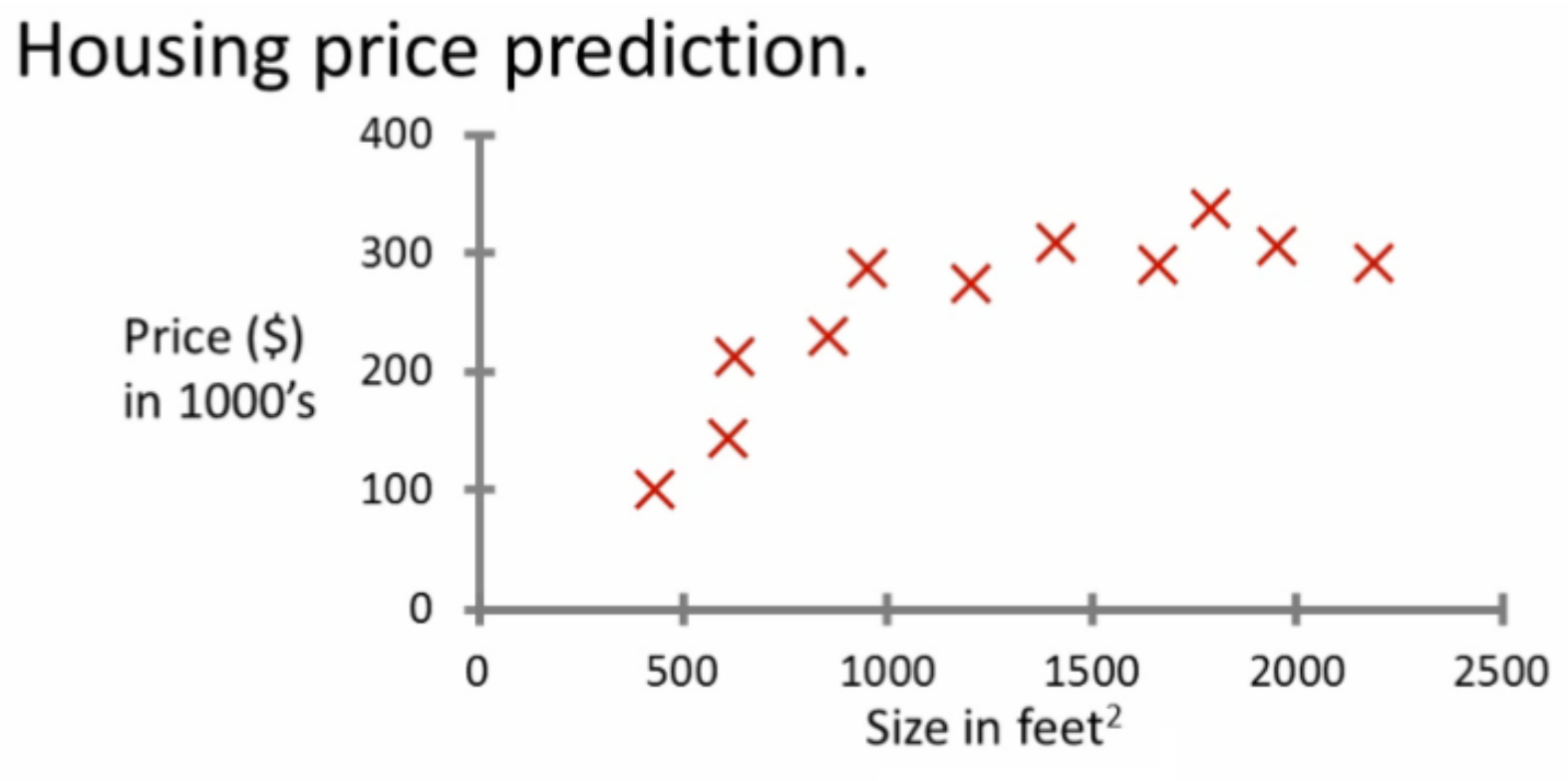
- učení s učitelem (supervised)
- učení bez učitele (unsupervised)
- kombinace předchozího (semi-supervised)
- zpětnovazební učení (reinforcement learning)
- optimalizační úloha

Klasifikační úloha



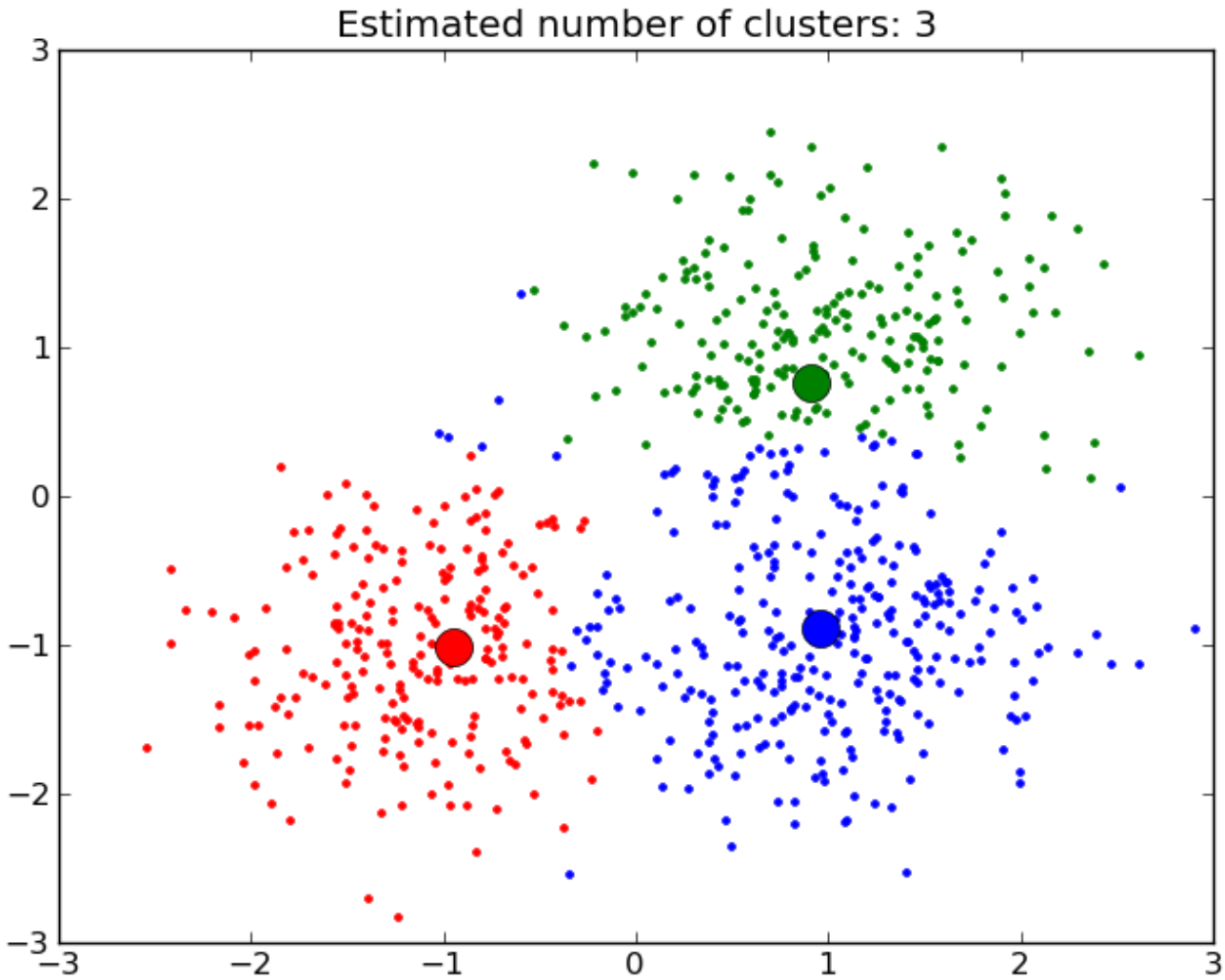
Zdroj: `http://gromgull.net/blog/category/machine-learning/`

Regresní úloha



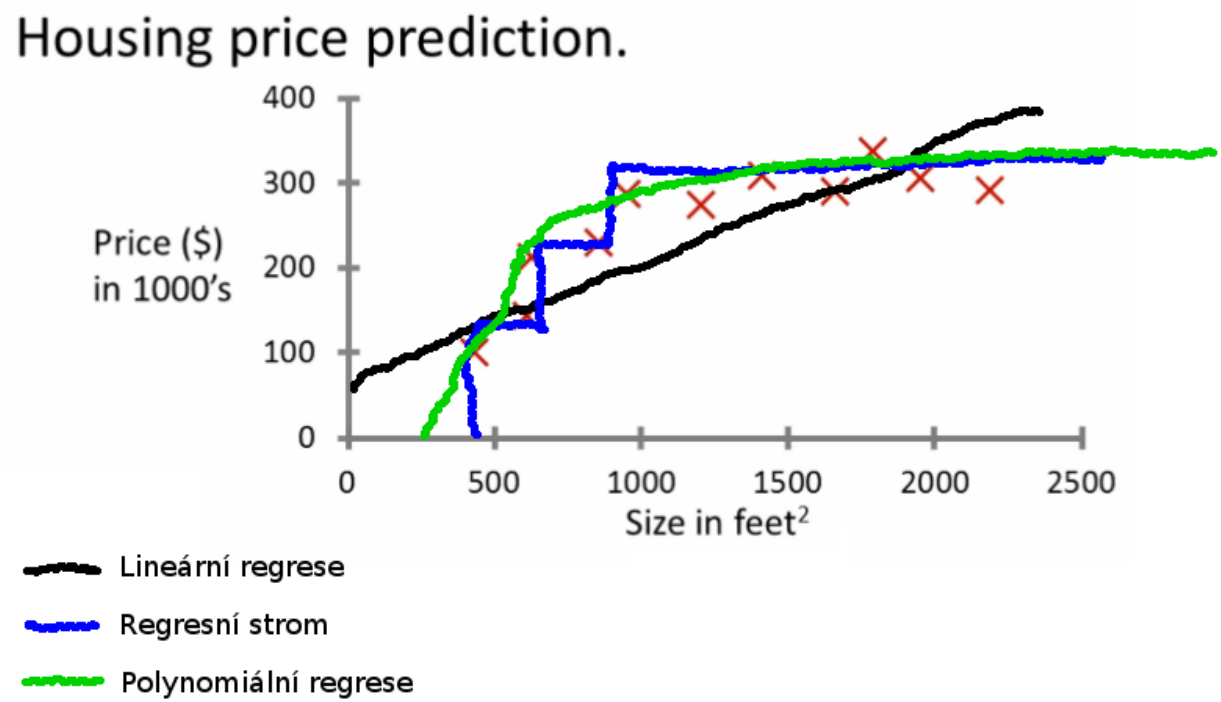
Zdroj: `https://class.coursera.org/ml/class`

Shlukování



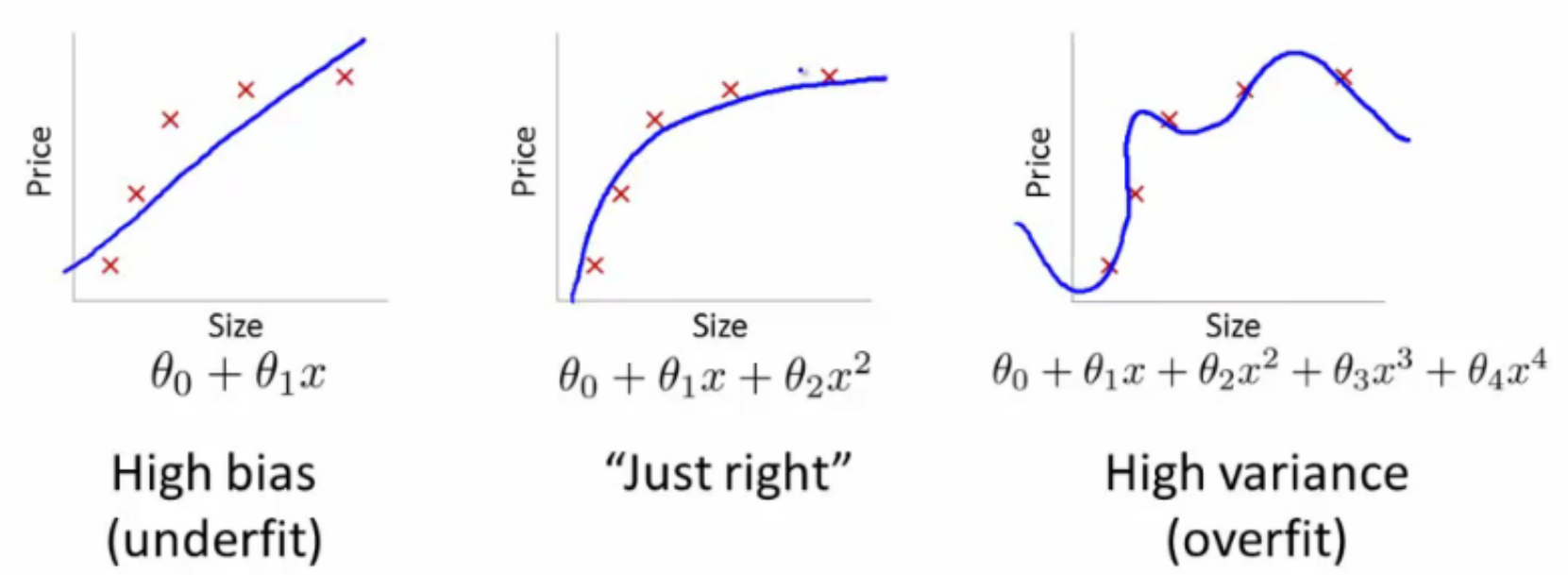
Zdroj: `http://scikit-learn.org/0.5/modules/clustering.html`

Nedostatečná expresivita



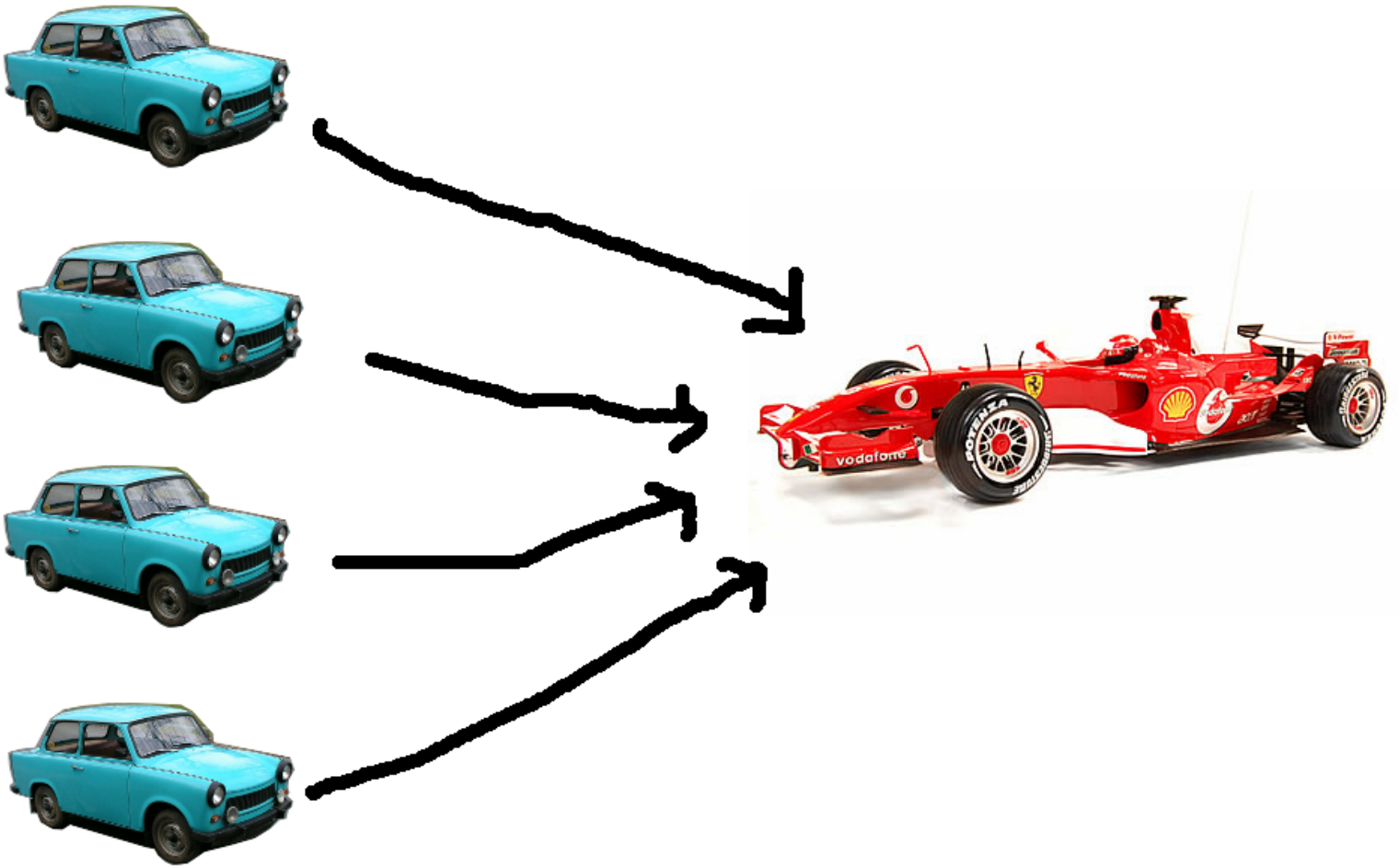
Zdroj: `https://class.coursera.org/ml/class`

Problém s přeučováním



Zdroj: `https://class.coursera.org/ml/class`

Bagging & Boosting



Klasifikace dokumentů

Bag-of-words reprezentace dokumentů

1. the man walked the dog
2. the man took the dog to the park
3. the dog went to the park

[dog, man, park, the, to, took, walked, went]

1. [1, 1, 0, 1, 0, 0, 1, 0]
2. [1, 1, 1, 1, 1, 1, 0, 0]
3. [1, 0, 1, 1, 1, 0, 0, 1]

TF-IDF

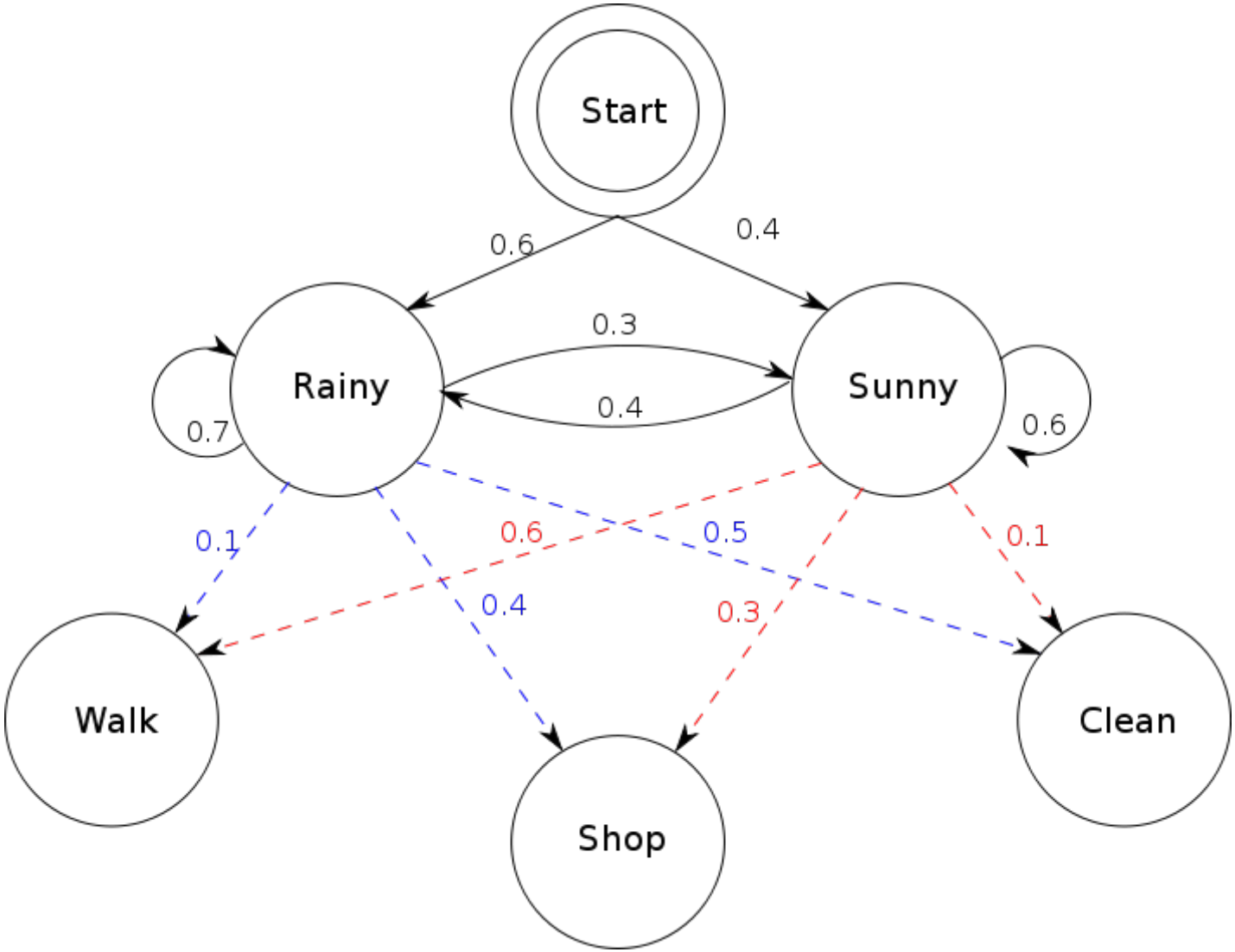
$$\mathrm{TF}(t, d) = \frac{n_{t,d}}{\sum_k n_{k,d}} \tag{1}$$

$$\mathrm{IDF}(t) = \log \frac{|D|}{|j : t_i \in d_j|} \tag{2}$$

$$\text{TF-IDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t) \tag{3}$$

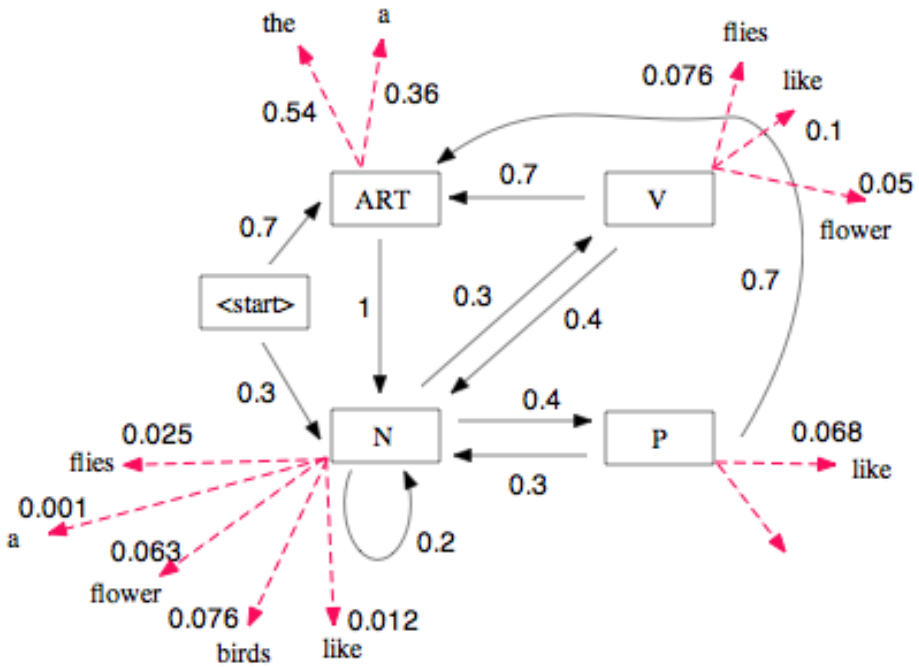
1. [0, 0.18, 0, 0, 0, 0, 0.48, 0]
2. [0, 0.18, 0.18, 0, 0.18, 0.48, 0, 0]
3. [0, 0, 0.18, 0, 0.18, 0, 0, 0.48]

Skryté Markovovy modely



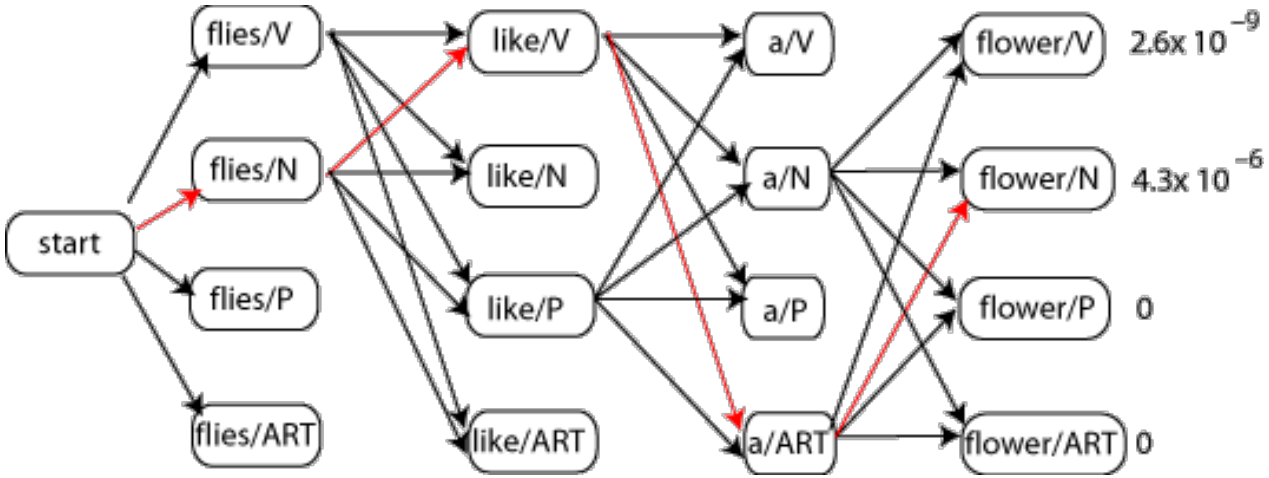
Zdroj: `http://en.wikipedia.org/wiki/Hidden_Markov_model`

Morfologické značkování



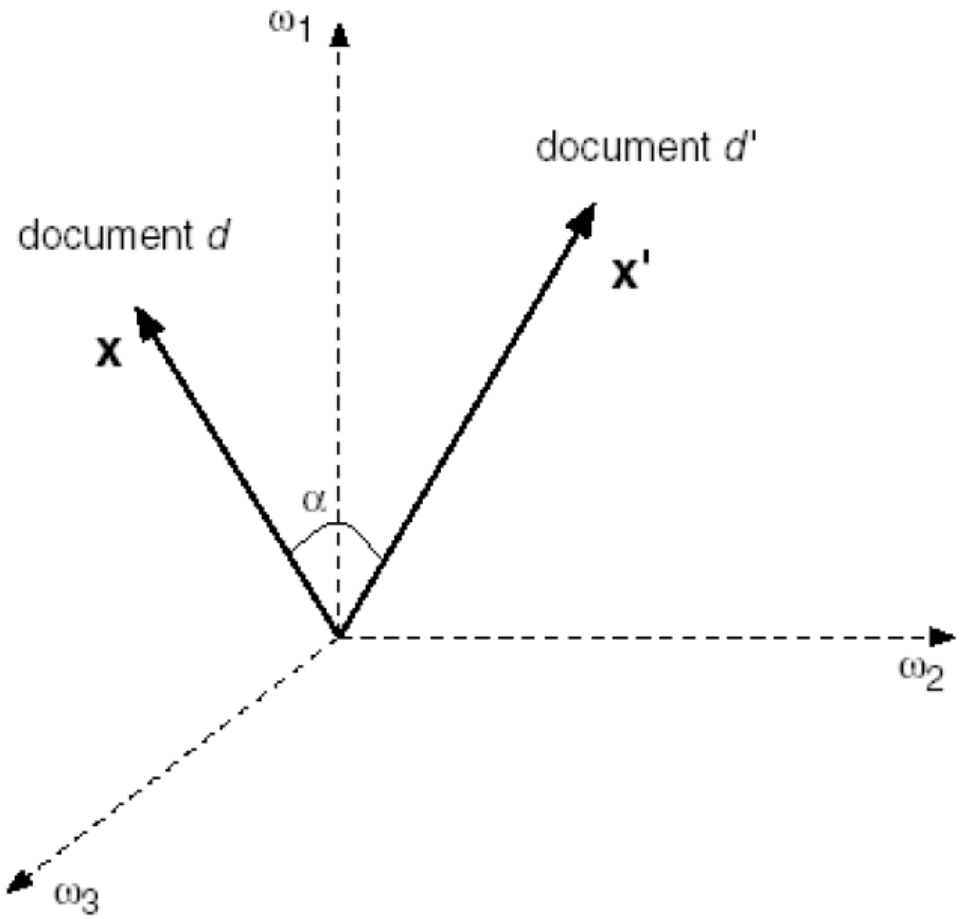
Zdroj: `http://www.cse.unsw.edu.au/~billw/`

Viterbiho algoritmus



Zdroj: `http://www.cse.unsw.edu.au/~billw/`

Modelování témat dokumentů

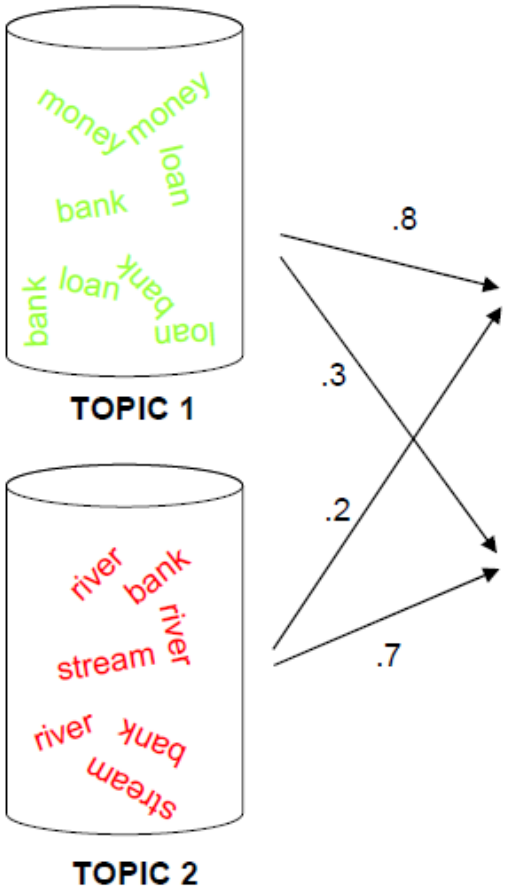


Latentní sémantická analýza

$$
\begin{array}{ccccccc}
 & & X & & U & \Sigma & V^T \\
 & & (\mathbf{d}_j) & & & & (\hat{\mathbf{d}}_j) \\
 & & \downarrow & & & & \downarrow \\
(\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} & = & (\hat{\mathbf{t}}_i^T) \rightarrow & \begin{bmatrix} \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} & \dots & \begin{bmatrix} \\ \mathbf{u}_l \\ \\ \end{bmatrix} \end{bmatrix} & \cdot \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{bmatrix} & \cdot \begin{bmatrix} [& \mathbf{v}_1 &] \\ & \vdots & \\ [& \mathbf{v}_l &] \end{bmatrix}
\end{array}
$$

Zdroj: http://en.wikipedia.org/wiki/Latent_semantic_analysis

Latentní Dirichletovská alokace



DOCUMENT 1: money[1] bank[1] bank[1] loan[1] river[2] stream[2] bank[1] money[1] river[2] bank[1] money[1] bank[1] loan[1] money[1] stream[2] bank[1] money[1] bank[1] bank[1] loan[1] river[2] stream[2] bank[1] money[1] river[2] bank[1] money[1] bank[1] loan[1] bank[1] money[1] stream[2]

DOCUMENT 2: river[2] stream[2] bank[2] stream[2] bank[2] money[1] loan[1] river[2] stream[2] loan[1] bank[2] river[2] bank[2] bank[1] stream[2] river[2] loan[1] bank[2] stream[2] bank[2] money[1] loan[1] river[2] stream[2] bank[2] stream[2] bank[2] money[1] river[2] stream[2] loan[1] bank[2] river[2] bank[2] money[1] bank[1] stream[2] river[2] bank[2] stream[2] bank[2] money[1]

Zdroj: Probabilistic Topic Models, Mark Steyvers and Tom Griffiths, 2007