



Personalizovaná sumarizácia textu

Róbert Móro
prof. Mária Bieliková

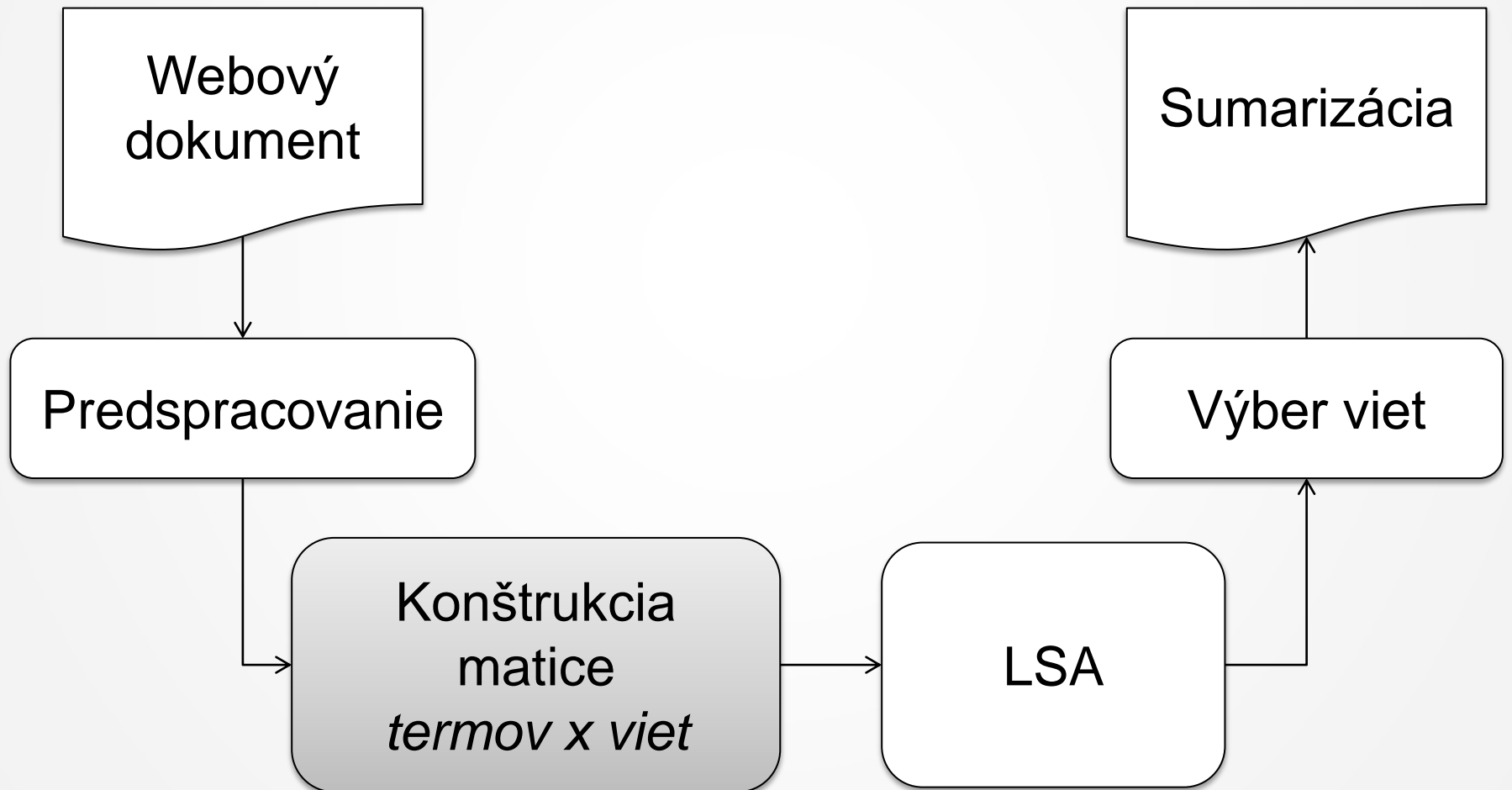
Motivácia

- Problém zahltenia informáciami
 - Automatická sumarizácia textu
 - Personalizácia
- Charakteristiky používateľa
- Metadáta
- Doména výučby
 - Sumarizácia pre opakovanie

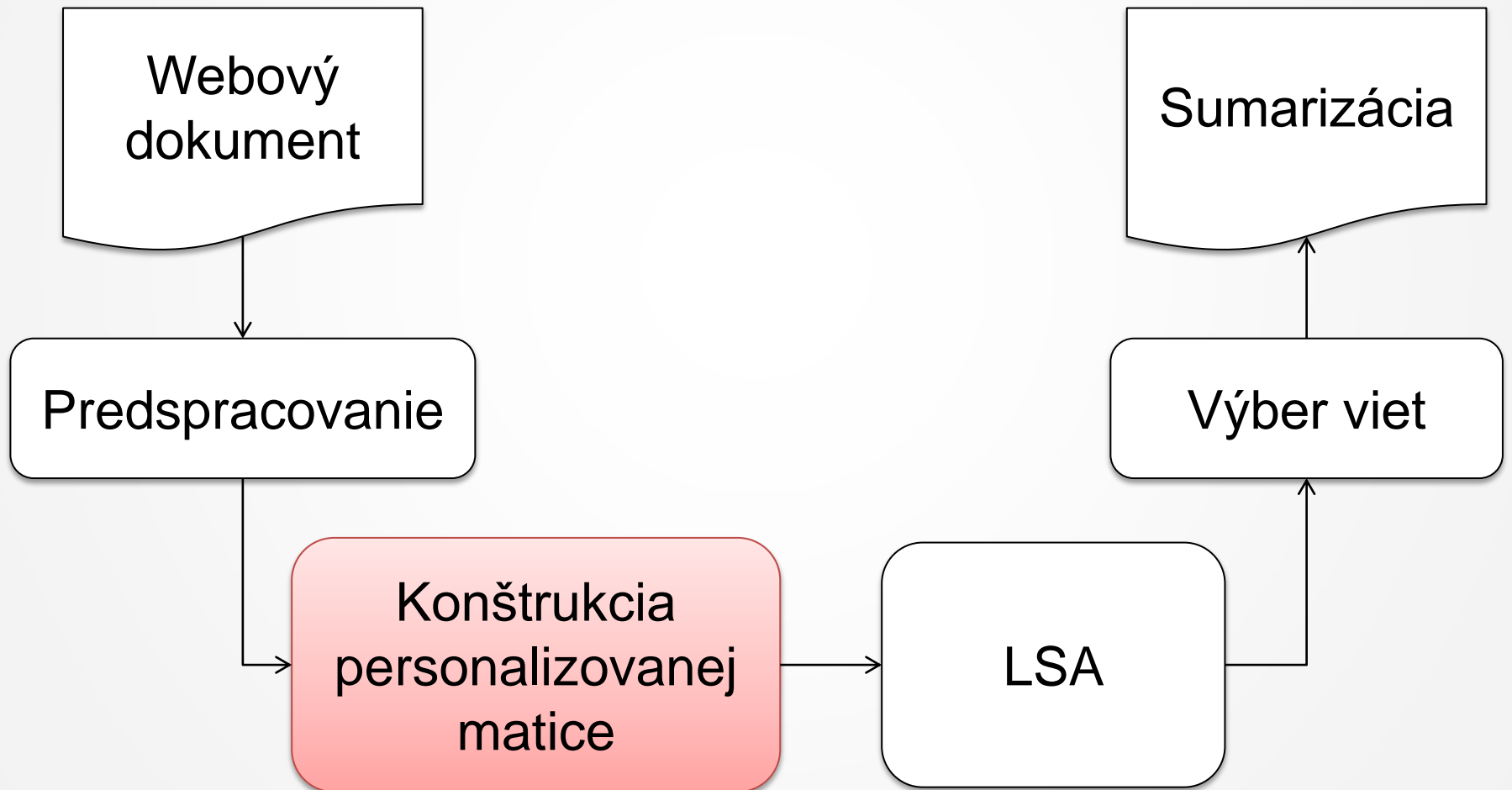
Automatická sumarizácia textu

- Konvenčné metódy
 - Frekvencia termov, strojové učenie, grafové (PageRank), algebrické
- Metódy uvažujúce ďalšie informácie
 - Kontextová, sociálna sumarizácia
- Metódy personalizovanej sumarizácie
 - **Čo uvažovať pri personalizácii**
 - **Ako navzájom kombinovať rôzne zdroje personalizácie**

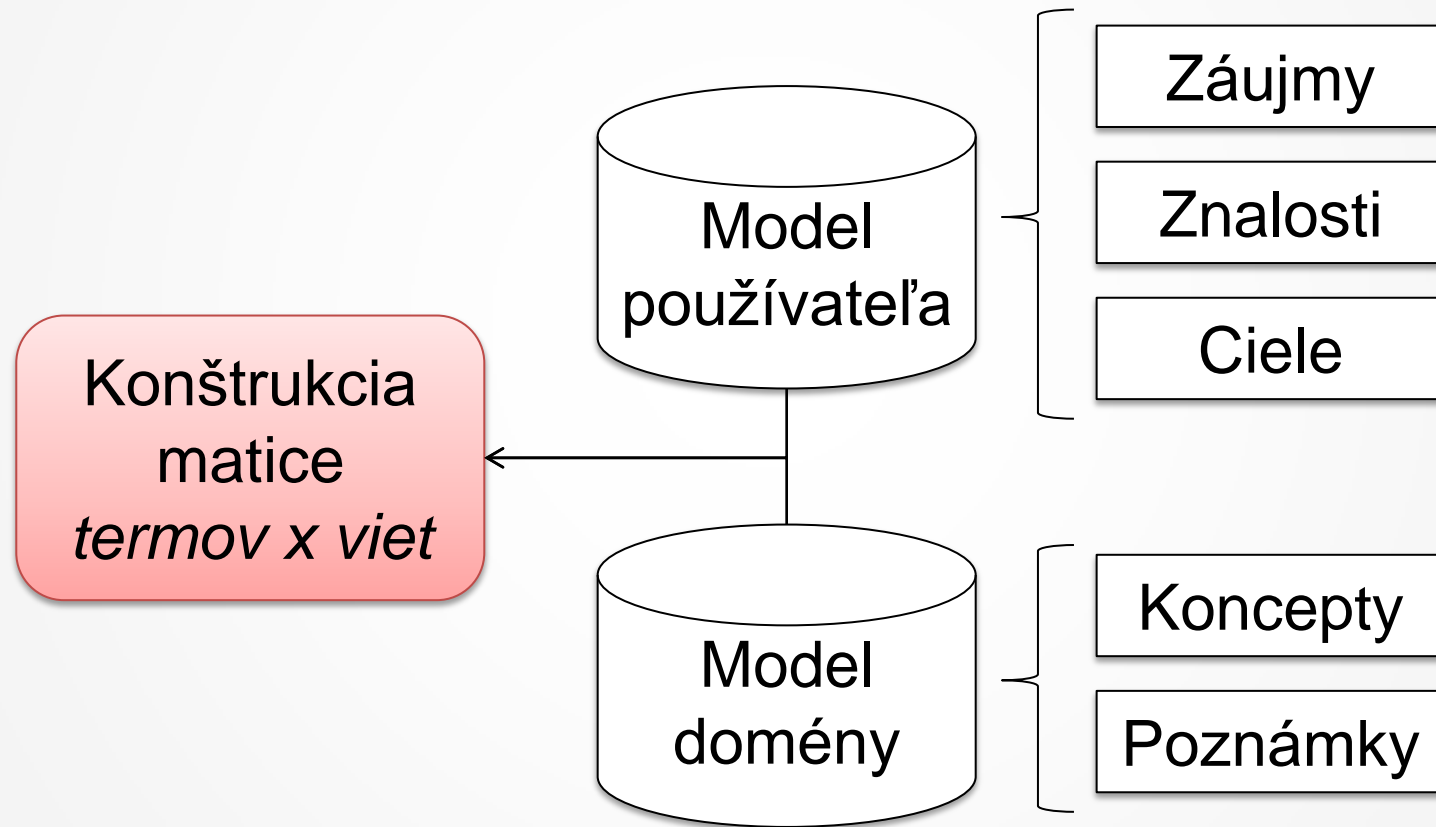
Metóda personalizovanej sumarizácie



Metóda personalizovanej sumarizácie



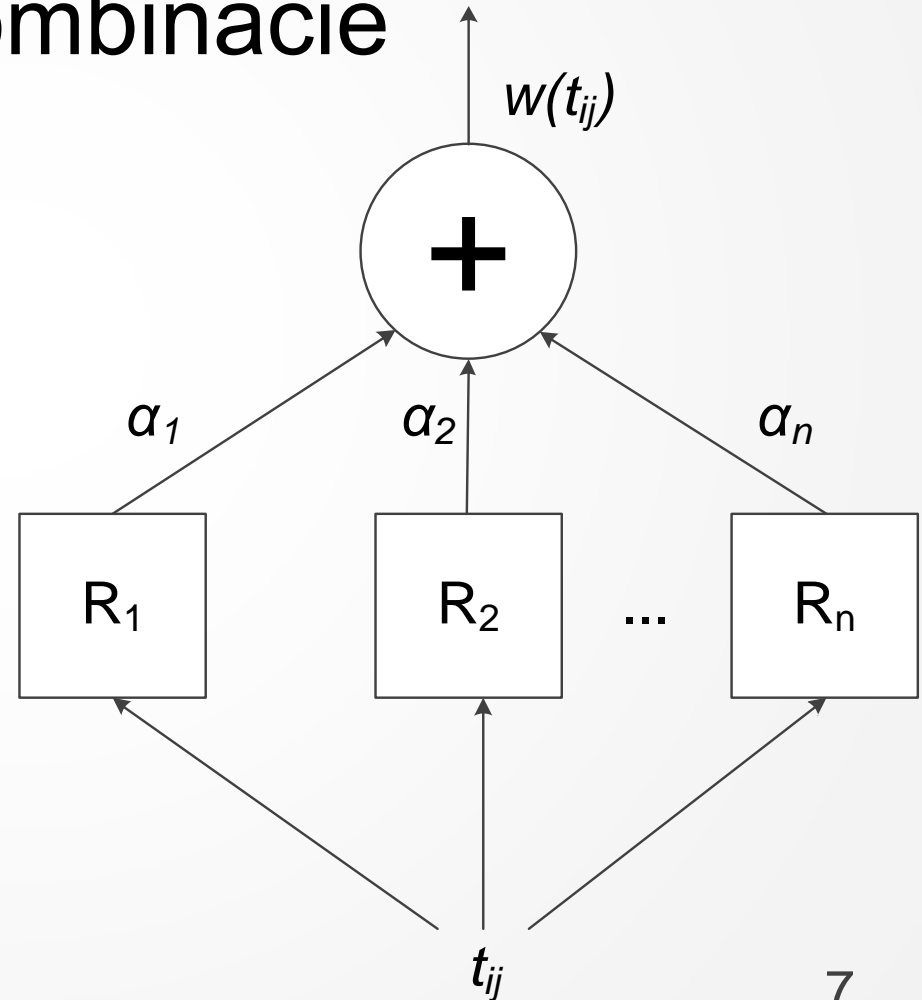
Konštrukcia personalizovanej matice



Konštrukcia personalizovanej matice

- Váhy pomocou kombinácie hodnotičov
 - generické
 - personalizované

$$w(t_{ij}) = \sum_k \alpha_k R_k(t_{ij})$$



Konštrukcia personalizovanej matice

- Generické
 - Hodnotič frekvencie výskytu termov
 - Hodnotič polohy termov
 - Hodnotič relevantných doménových pojmov

$$w(t_{ij}) = w_i \quad \text{ak } t_i \in S_j \cap C_d$$

$$w(t_{ij}) = 0 \quad \text{inak}$$

Konštrukcia personalizovanej matice

- Generické
 - Hodnotič frekvencie výskytu termov
 - Hodnotič polohy termov
 - Hodnotič relevantných doménových pojmov
- Personalizované
 - Hodnotič vedomostí
 - Hodnotič poznámok
 - Hodnotič cieľov
 - Hodnotič záujmu

Zhodnotenie navrhnutej metódy

- Nezávislá od zvolenej domény
- Nezávislá od jazyka sumarizovaného textu
- Umožňuje uvažovať rôzny kontext sumarizácie
 - Rôzne scenáre použitia
- Obmedzenia
 - Štruktúrovanosť textu
 - Dĺžka textu

Možnosti použitia metódy personalizovanej sumarizácie

- Vhodné domény
 - Novinové portály
 - Systémy digitálnych knižníc
 - Výučbové systémy
- Scenáre použitia
 - Sumarizácia výsledkov vyhľadávania
 - Jednoduchá rešerš
 - **Opakovanie vo výučbovom systéme**

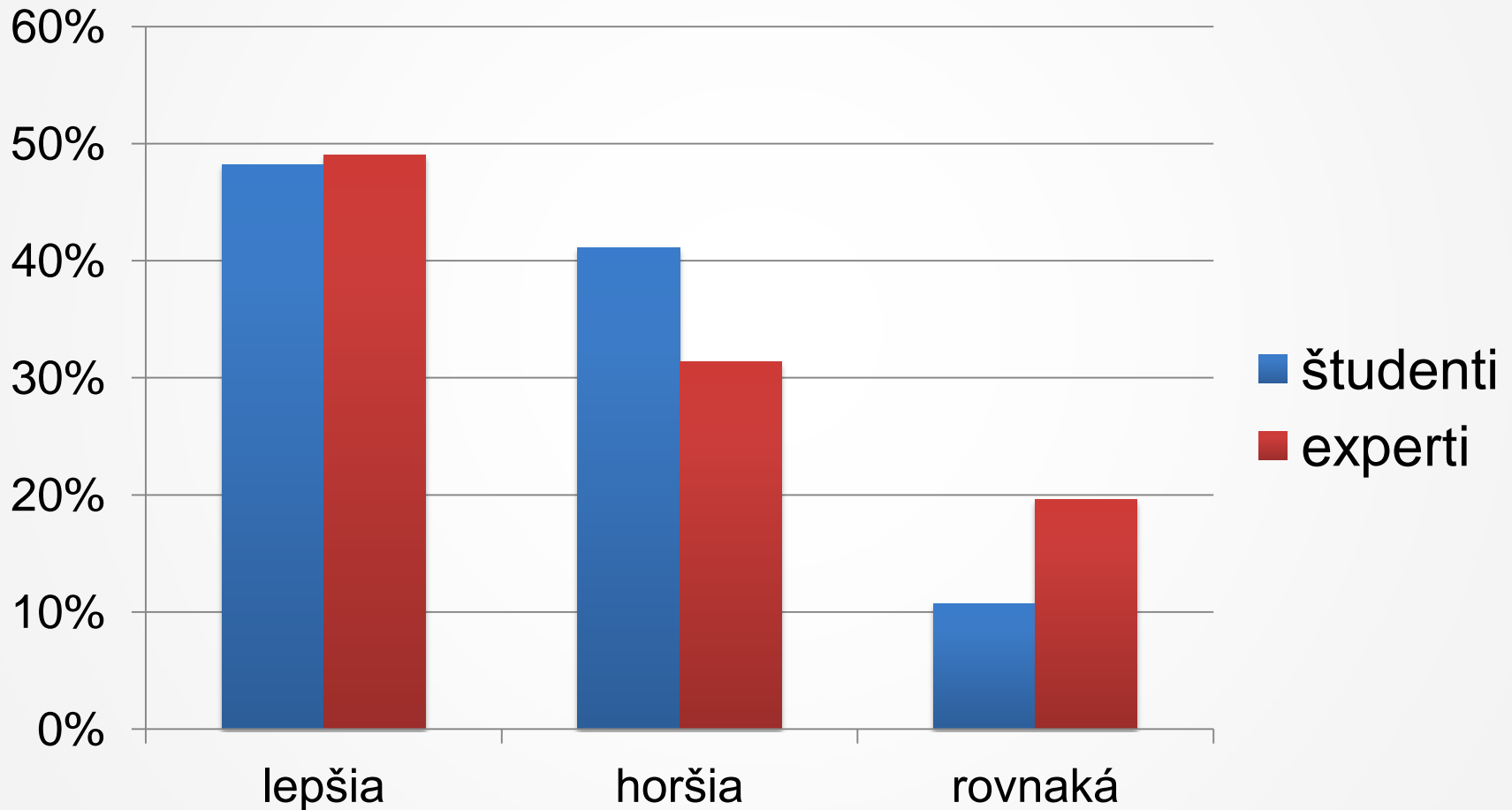
Realizácia metódy sumarizácie

- Vlastný sumarizátor ako REST webová služba
 - Využitie externých služieb predspracovania textu
 - Nízka previazanosť, znovupoužitelnosť
- Komunikačný JSON protokol
 - Konfigurácia sumarizátora
 - Sumarizácia textu
- Integrácia so systémom ALEF
 - Používanie
 - Experimentovanie

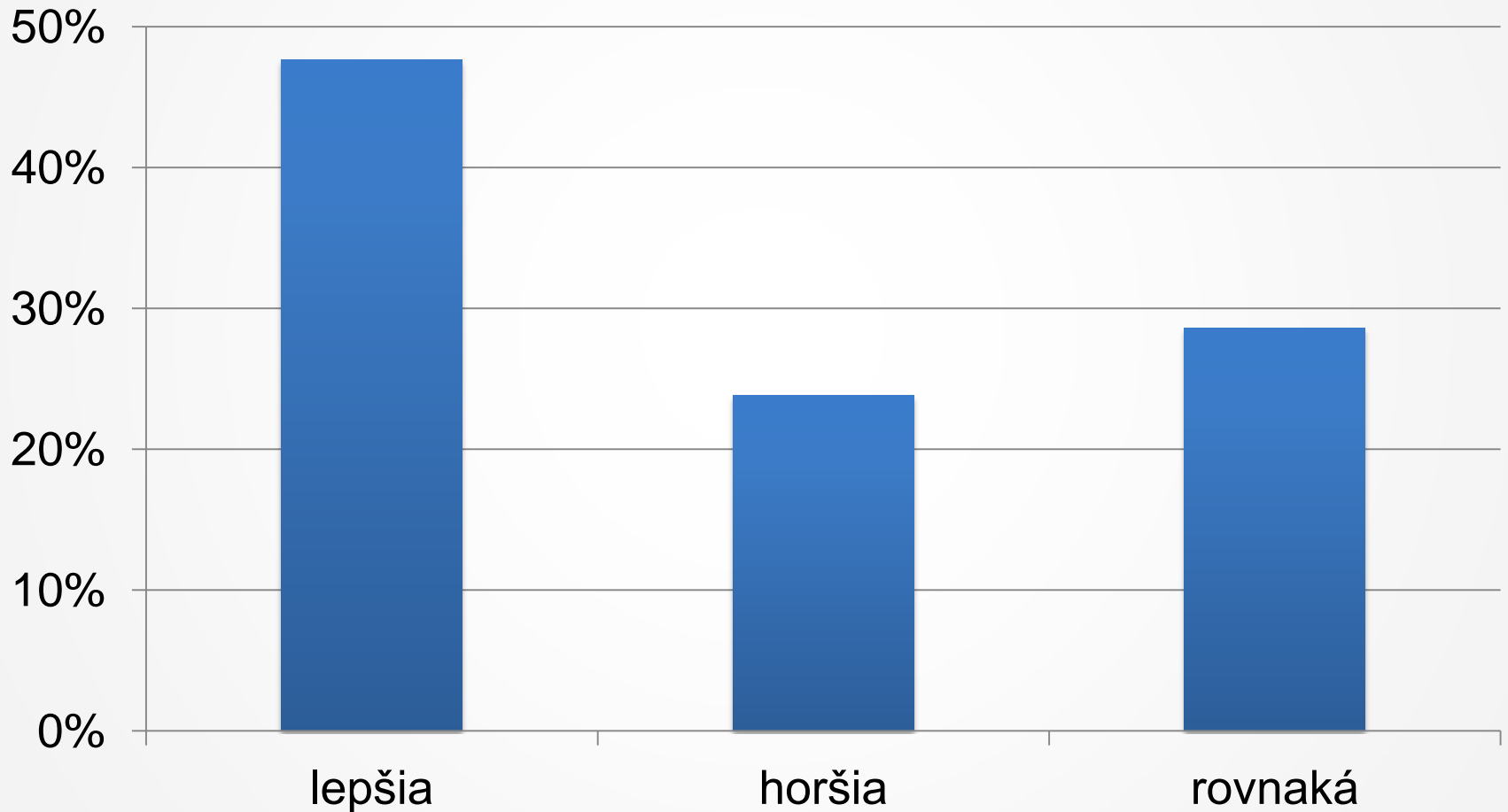
Overenie metódy sumarizácie

- 2 experimenty
 - Na predmetoch FLP a PSI
 - Sumarizácia zohľadňujúca relevantné doménové pojmy vs. generická
 - Sumarizácia zohľadňujúca poznámky vs. generická
- 75 študentov, 5 študentov – expertov, 303 vzdelávacích textov
- 2242 hodnotení, 479 odpovedí na doplňujúce otázky od študentov
- 385 porovnaní od študentov – expertov

1. Sumarizácia zohľadňujúca relevantné doménové pojmy



2. Sumarizácia zohľadňujúca poznámky



Zhodnotenie výsledkov experimentov

- Zohľadnenie relevantných doménových pojmov aj poznámok prináša zlepšenie oproti generickému variantu
- Sumarizáciu možno použiť pri
 - rozhodovaní o relevancii dokumentu
 - opakovaní
- Zistené aj niektoré problémy
 - Segmentácia textu, anafory (koreferencie)

Vlastný príspevok

- Návrh spôsobu kombinácie hodnotičov
 - umožňuje uvažovať rôzne parametre a kontext sumarizácie
- Návrh konkrétnych hodnotičov zohľadňujúcich
 - termy relevantné v danej doméne
 - znalosti a ciele používateľa
 - poznámky v podobe zvýraznení a tagov

Ďalšie možnosti rozvoja

- Automatické (dynamické) nastavovanie koeficientov kombinácie hodnotičov
- Využitie tagov namiesto expertom identifikovaných relevantných pojmov
- Použitie pokročilejších metód spracovania prirodzeného jazyka
- Rozvinutie navrhnutej metódy výberu dokumentov na opakovanie

Otázka 1: Váhovacia schéma

- Koeficienty α_k
 - z intervalu $\langle 0;1 \rangle$, pričom $\sum_k \alpha_k = 1$
 - vo všeobecnosti však ľubovoľné z \mathcal{R}
- Iné schémy
 - násobenie, max, min, priemer
 - rozšírenie o dynamické nastavenie váh v závislosti od kontextu
 - kombinácia na nižšej úrovni \rightarrow súhrny (vety)

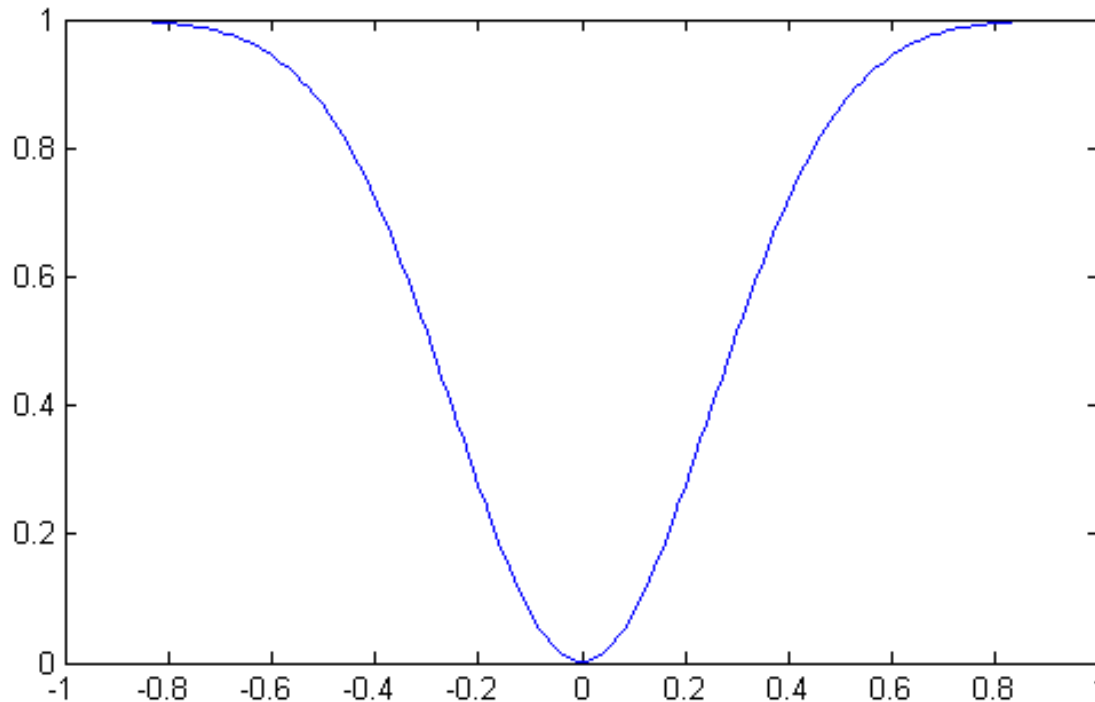
Otázka 2: Možnosti zapojenia strojového učenia

- Interpretácia váhy ako dôvery v daný hodnotič (zdroj)
 - Dynamicky v závislosti od kontextu
- Určenie váh hodnotičov
 - rozhodovacie stromy
 - Bayesove siete
 - v závislosti od typu textu → klasifikátor

Personalizovaná sumarizácia pre opakovanie

- Čas opakovania
- Výber dokumentov na opakovanie
 - Prečítané dokumenty
 - Vzťahy medzi konceptmi (prerekvizity)
 - Ciele používateľa
 - **Zmena znalostí konceptov**

Zmena znalostí konceptov



- Vektor zmien znalostí konceptov
- Kosínusová podobnosť