IB047

Četnosti a kolokace

Pavel Rychlý

pary@fi.muni.cz

10. dubna 2017

Zipfův zákon

Jaké je rozložení slov v korpusu?

- $f * r = C$
  součin četnosti a pořadí v seznamu četností je zhruba konstantní
- slova, slovní spojení
- vlastní jména, velikosti měst
- nejfrekventovanější jevy pokrývají většinu jazyka

Redukovaná četnost

- pravděpodobnost výskytu slova vs. četnost slova v korpusu
- některá slova jsou pouze v jenom dokumentu, ale mnohokrát
- redukované četnosti normalizují výskyty
- $RF <= F$
- $RF >= 1$
- dokumentová četnost, ARF

Kolokace

Jaká slova se vyskytují v kontextech daného výrazu?

- četnosti
- relativní četnosti
- skóre – asociační míry

Asociační míry

Počítáme na základě kontingenční tabulky.

	$V = v$	$V \neq v$
$U = u$	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$U \neq u$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

expected frequencies

	$V = v$	$V \neq v$	
$U = u$	O_{11}	O_{12}	$= R_1$
$U \neq u$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

observed frequencies

O_{ij} – pozorované hodnoty (observed) E_{ij} – očekávané hodnoty (expected)

Asociační míry

- T-score: $T = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}} = \frac{f_{xy} - \frac{f_x f_y}{N}}{\sqrt{f_{xy}}}$
- MI-score: $MI = \log_2 \frac{O_{11}}{E_{11}} = \log_2 \frac{f_{xy} N}{f_x f_y}$
- Log-likelihood:
  $LL = -\log_2 \frac{L(O_{11}, C_1, r) L(O_{12}, C_2, r)}{L(O_{11}, C_1, r_1) L(O_{12}, C_2, r_2)}$
  $L(k, n, r) = r^k (1 - r)^{n-k}$
  $r = \frac{R_1}{N}; r_1 = \frac{O_{11}}{C_1}; r_2 = \frac{O_{12}}{C_2}$

Asociační míry

- Minimum sensitivity: $MS = \min\{\frac{O_{11}}{C_1}, \frac{O_{11}}{R_1}\} = \min\{\frac{f_{xy}}{f_x}, \frac{f_{xy}}{f_y}\}$
  – minimum z relativních četností
- Dice: $D = \frac{2O_{11}}{R_1+C_1} = \frac{2f_{xy}}{f_x+f_y}$
- logDice: $lD = 14 + \log_2 D = 15 + log_2 f_{xy} - log_2(f_x + f_y)$

Filtrování

- vybíráme jen ty kolokace, které splňují podmínku na značkách
- ADJ NN
- NN NN
- word sketches – jednostránkový souhrn chování slov

Word Sketches

Jak jej lze vytvořit

- Velký vyvážený korpus
- Vyhledáme závislé prvky (subjects, objects, heads, modifiers etc)
- Seznam kolokací pro každou gramatickou relaci
- Statistika pro třídění každého seznamu