statistické zpracování přirozeného jazyka

OJ205

statistické zpracování přirozeného jazyka

- užití: **vyhledávání kolokací**

- **problémy:**
- - rozložení přirozeného jazyka
- - stoplist
- - nastavení vyhledávání
- - interpretace výsledků

n-gramové jazykové modely

- Jako n-gram označujeme obvykle posloupnost slov délky , která se vyskytuje v korpusu.
- Známe posloupnost − 1 slov v korpusu. Jaké slovo za nimi bude následovat?
- Formálně se n-gramový jazykový model skládá z ch pravděpodobností:

(| 1, ..., −1)

čili pravděpodobností, že se vyskytlo slovo za předpokladu, že před ním se vyskytla slova 1, ..., −1. Souhrn všech chto pravděpodobností pro všechny možné kombinace slov v korpusu se vá n-gramový jazykový model.

n-gramové jazykové modely

- 1, ..., nemusí t pouze slova – můžeme vytvořit n-gramový model znaků, fonémů, pádů, ch značek apod., případně i komplikovanější modely, kde např. bude morfologická značka a 1, ..., −1 budou slova.
- Využití - ch modelů ve zpracování jazyka: existuje jich velké množství, z nichž nejpoužívanější je pravděpodobně trigramový (= 3) jazykový model.

- (|) = 3/4
- (|) = ¼
- (|) = 2/3
- (|) = 1/3
- (|) = 1/2
- (|) = 1/2
- (|) = 1
- (|) = 1
- ($|) = 1
- (|) = 0 š *h* .

rozbor věty: Máma mele maso.