Aplikace `the_next_word_prediction`

4IZ470 Dolování znalostí z webu

Lubomír Štěpánek

Katedra biomedicínské informatiky
Fakulta biomedicínského inženýrství
České vysoké učení technické v Praze
—
Centrum podpory multimediálních forem výuky
Oddělení výpočetní techniky
1. lékařská fakulta
Univerzita Karlova v Praze

24. dubna 2017

Pipeline projektu

1. získání textového korpusu
   - včetně jeho obohacení vlastním webscrapovaným textem
2. processing textových dat korpusu
3. n-gramming nad korpusem pro $n \in \{2, 3, 4\}$
   - včetně Kneserova-Neyova smoothingu
4. implementace koncové webové aplikace predikující i-té slovo, které nejpravděpodobněji následuje uživatelem zadané $(i - 1)$-členné slovní spojení, kde $i \in \{1, 2, 3\}$

Získání textového korpusu

- použita část známých HC korpusů (*Helsinki corpora*) různorodých anglických textů
- je dostupná online na

  http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/

- pracovní korpus byl konkrétně sestaven
  - ze zpravodajských příspěvků
  - z tweetů
  - z blogerských textů
- dohromady $>$ 3 miliony anglických vět
- v plánu obohacení vlastním webscrapingem částí anglicky psaného webu
  - twitteru prostřednictvím balíčku `twitteR` jazyka R
  - Wikipedie, protože nabízí statické HTML

Processing textových dat korpusu

- odstranění větné interpunkce
- odtranění stop slov
  - existují slovníky anglických stop slov
- odstranění vulgárních, nevhodných slov
  - rovněž pomocí existujících slovníků

n-gramming nad korpusem

- jde o vytvoření "slovníku" n-členných slovních spojení pro $n \in \{2, 3, 4\}$
- např. {i like}, {how are you}, {what about your own} apod.
- smyslem n-grammingu je nakonec predikce i-tého slova, které nejpravděpodobněji následuje uživatelem zadané $(i-1)$-členné slovní spojení, kde $i \in \{1, 2, 3\}$
- v plánu Kneserovo-Neyovo vyhlazování, jeho principem je provážení pravděpodobností n-gramů pro nízká a vysoká n; momentálně implementován MAP (Maximum-Aposteriori-Probability) odhad, tedy slovo w_i^* následující frázi $w_{i-1} \dots w_1$ takové, že

$$w_i^* = \arg\max_{\forall w_i} \{\hat{p}(w_i^* w_{i-1} \dots w_1 \mid w_{i-1} \dots w_1)\}$$

Koncová webová aplikace

- implementována v R, uložena na R-serveru 1. lékařské fakulty UK
- beta verze dostupná online na

http://shiny.statest.cz:3838/the_next_word_prediction/

Děkuji za pozornost!

lubomir.stepanek@fbmi.cvut.cz
lubomir.stepanek@lf1.cuni.cz