Hackovani

Jan Macošek

Cíl

- Program, který přidá diakritiku do (slovenského) textu
- K dispozici trénovací data (z Wikipedie)
- Vstup: Text bez diakritiky
- Výstup: Text s přidanou diakritikou

Trénování

- Program zpracovává trénovací data slovo po slovu:
  - Každé slovo „odháčkuje“
  - K odháčkované variantě si zapamatuje jeho variantu s diakritikou
  - Pro každou možnou variantu s diakritikou si zaznamenává počet výskytů
- Na závěr vypíše na výstup celou získanou statistiku
  - Odhackovane: Diakritika1=Četnost1; ...; DiakritikaN=ČetnostN

Háčkovač

- Program nejprve načte ze souboru předem získanou statistiku
- Poté načítá ze vstupu jednotlivá slova, přidá jim diakritiku a vytiskne je na výstup
  - Přidání diakritiky: Zvolí se nejpravděpodobnější varianta, tedy oháčkování s nejvyšší četností
- Např.: Doplnit -> Doplniť ->OUT
- Úspěšnost: 84%
  - Počet shodných slov/celkový počet slov
  - Měřeno na trénovacích datech
  - Po rozdělení na trénovací a testovací data 82%

Statistické poznatky

- Slovo s nejvíce nejednoznačným háčkováním:
  - stat:štát=51;stať=44;stát=14;stat=9;stáť=5;
- Slova neumožňující přidání diakritiky: 51,6%
- Slova vyžadující přidání diakritiky právě jedním způsobem: 12,6%
- Slova s více variantami oháčkování: 35,8%
- Průměrná míra nejednoznačnosti na jeden slovní výskyt: 1,46
  - (tj. průměrný počet možných odpovědí)

Šum

- Data z Wikipedie obsahují šum
- Je možné ho statisticky odstranit
  - Odebereme ze „slovníku“ všechna oháčkování, jejichž počet výskytů nepřesahuje 10% počtu výskytů slova ze stejné kategorie
- S tímto slovníkem opět zpracujeme předchozí statictické poznatky

Statistické poznatky – bez šumu

- Slova s nejvíce nejednoznačným háčkováním:
  - stat:štát=51;stať=44;stát=14;stat=9;stáť=5;
  - zavazne:záväzné=6;závažné=4;zavazne=2;závazné=1;
  - zastava:zástava=3;zastáva=3;zastává=1;zastava=1;
  - wikipediach:wikipediach=5;wikipédiách=2;wikipediách=2;wikipédiach=1;
- Slova neumožňující přidání diakritiky: 58,4%
- Slova vyžadující přidání diakritiky právě jedním způsobem: 18,9%
- Slova s více variantami oháčkování: 22,7%
- Průměrná míra nejednoznačnosti: 1,24