Autor: Bc. Radovan Fuska
Vedúci práce: prof. RNDr. Stanislav Krajči, PhD.

Určovanie autorstva neznámeho slovenského textu

Problém

- Máme text v slovenskom jazyku.
- Nepoznáme jeho autora.
- Chceme identifikovať jeho autora.

Riešenie

- Použijeme kolekciu textov od známych autorov.
- Predpokladáme, že texty majú črty špecifické pre autora.
  - Konkrétna podoba jazyka jedného človeka – idiolekt.
- Vytvoríme metódu na analýzu týchto čŕt a vytvoríme akýsi lingvistický odtlačok pre každého autora.
- Porovnáme odtlačok neznámeho textu s odtlačkami známych autorov.
  a) Autor je jedným z kandidátov
  b) Autor sa medzi kandidátmi nenachádza
  c) Binárne prisudzovanie autorstva

Znaky, resp. vlastnosti slovenského textu

- Lexikálne
  - Štatistiky konkrétnych slov a slovných spojení resp. slovných n-gramov
- Znakové
  - Frekvencie n-gramov písmen (najčastejšie n = 3)
- Syntaktické
  - Dĺžky slov, viet, riadkov, odsekov a iných jednotiek
  - Počet prázdnych riadkov
- Morfologické
  - Štatistiky slovných druhov, pádov a iných gramatických kategórií, interpunkcie
- Chyby
  - „metaznak" o iných znakoch

Metódy

- Klasifikačný problém
  - Metóda podporných vektorov (SVM)
  - Naivný Bayes
  - K-najbližších susedov (k-nearest neighbour)
  - Logistická regresia
  - Rozhodovacie stromy
  - Bayesovská regresia
  - Winnowov algoritmus

Trigramová pravdepodobnosť

- Pravdepodobnosť, že autor napísal daný text
- Na úrovni „tokenov“
- Witten-Bell smoothing
- $p\left(w_i \middle| w_{i-n+1}^{i-1}\right) = \lambda \cdot p\left(w_i \middle| w_{i-n+1}^{i-1}\right) + (1-\lambda) \cdot p(w_i | w_{i-n+2}^{i-1})$

Miery

- Zaužívané z oboru získavania informácií (information retrieval)
- Podľa autora A
  - Precision: $P_A = \frac{|spravne\ priradene(A)|}{|vsetky\ priradene(A)|}$
  - Recall: $R_A = \frac{|spravne\ priradene(A)|}{|dokumenty\ od\ autora(A)|}$
  - Harmonický priemer $F_1 = \frac{2P_A R_A}{P_A + R_A}$
- Priemer z autorov $\{A_i\}$ podľa miery M
  - $makropriemer_M(\{A_i\}) = \frac{1}{n}\sum_i M_{A_i}$, kde n je počet autorov
  - $mikropriemer_M(\{A_i\}) = \frac{1}{k}\sum_i k_i M_{A_i}$, kde k je počet dokumentov a k_i je počet dokumentov od A_i

Dáta

- Články z novín
- Stiahnuté zo stránky
- Normalizované
  - Vyfiltrovanie pomenovaných entít (vlastné mená, názvy, ...)
- 1822 článkov
- 20 autorov

Dáta

- 20 autorov: 26 až 143 článkov na autora
- Počet článkov: 1822
- Počet odsekov: 26854
- Počet viet: 61010
- Počet slov: 782333
- Počet písmen: 4379687

Dáta

- Priemerný počet odsekov v článku: 14,739
- Priemerný počet viet v článku: 33,485
- Priemerný počet viet v odseku: 2,272
- Priemerný počet slov v článku: 429,381
- Priemerný počet slov v odseku: 29,133
- Priemerný počet slov vo vete: 12,823
- Priemerný počet písmen v článku: 2 403,780
- Priemerný počet písmen v odseku: 163,093
- Priemerný počet písmen vo vete: 71,786
- Priemerný počet písmen v slove: 5,598

Odskúšané znaky

- STA
  - Štatistiky počtu a dĺžok odsekov, viet, slov, písmen
- CW
  - Štatistiky počtov unigramov, bigramov a trigramov najčastejších slov z celého datasetu
- CNG
  - Štatistiky počtov bigramov a trigramov najčastejších písmen z celého datasetu

Odskúšané metódy

- Pomocou prostredia WEKA
- J4.8
  - Rozhodovací strom
- NB
  - Naivný Bayes
- NBM
  - Naivný Bayes viactredový
- SL
  - Logistická regresia
- SMO
  - Metóda podporných vektorov

Výsledky

Precision	J4.8	NB	NBM	SL	SMO
STA	0.728	0.356	0.296	0.338	0.353
CW	0.779	0.365	0.424	0.572	0.611
CNG	0.822	0.279	0.382	0.536	0.464
CW+CNG	0.831	0.366	0.539	0.726	0.634
STA+CW+CNG	0.851	0.413	0.591	0.742	0.699

Výsledky

F1	J4.8	NB	NBM	SL	SMO
STA	0.723	0.238	0.249	0.323	0.336
CW	0.771	0.287	0.418	0.551	0.383
CNG	0.816	0.158	0.381	0.515	0.266
CW+CNG	0.829	0.246	0.533	0.713	0.456
STA+CW+CNG	0.847	0.259	0.556	0.732	0.543

Postup práce

- Prehľad existujúcich riešení
- Získanie dát (zoznamu diel)
  - Články z novín
- Porovnanie existujúcich metód
- Návrh novej metódy

Zdroje

- ARGAMON, Shlomo; JUOLA, Patrick. Overview of the international authorship identification competition at PAN-2011. In: *CLEF (Notebook Papers/Labs/Workshop)*. 2011.
- KOPPEL, Moshe; SCHLER, Jonathan; ARGAMON, Shlomo. Computational methods in authorship attribution. Journal of the American Society for information Science and Technology, 2009, 60.1: 9-26.