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

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
- 648 článkov
- 20 autorov
- Celkovo 1 846 885 znakov po normalizácii
- V priemere 2 850 znakov na článok

Trigramová pravdepodobnosť

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

Trigramová pravdepodobnosť

Skúmaný autor id	Skutočný autor id	Článok id	Pravdepodobnosť
0	0	0	2006282599e-24077
0	0	1	2718316079e-61393
0	0	2	566827109e-265403
0	0	3	2783932241e-270250
0	1	0	3476529945e-75571
0	1	1	712000501e-109271
0	1	2	837546449e-44767
0	1	3	1788845615e-39972
0	2	0	1473147699e-10385
0	2	1	2745748147e-6643
0	2	2	3663149549e-114831
0	2	3	2148053289e-7469

Trigramová pravdepodobnosť

Autor id	Precision	Recall	F1	Počet dokumentov
0	0	0	0	11
1	0	0	0	11
2	0	0	0	11
3	0.101852	1	0.184874	11
4	0	0	0	11
5	0	0	0	11
6	0	0	0	11
7	0	0	0	11
8	0	0	0	11
9	1	0.090909	0.166667	11
10	0	0	0	11

Trigramová pravdepodobnosť

	Makropriemer	Mikropriemer
Precision	0.100168	0.100168
Recall	0.099174	0.099174
F1	0.031958	0.031958

Postup práce

- ✓Prehľad existujúcich riešení
- Získanie dát (zoznamu diel)
  - Slovenský národný korpus
  - Prepisy z Národnej rady
  - ✓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.