Autor: Bc. Radovan Fuska
Vedúci práce: doc. 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

Súčasný stav

- Anglický jazyk
  - Skúmaný dlho
  - Veľa odskúšaných metód
  - Mnoho dát na analýzu
  - Jednoduchšia gramatika na spracovanie štatistickými metódami
- Slovenský jazyk
  - Málo voľných, už spracovaných textov
  - Ohýbanie slov si vyžaduje predspracovanie textu

Predchádzajúce prístupy

- Jednorozmerné ohodnotenie
  - Mendenhall (1887) – závislosť medzi dĺžkou slov a ich frekvenciou
  - Yule (1944) – dĺžka viet
- Viacrozmerné ohodnotenie
  - Mosteller, Wallace (1964) – frekvencie „function words" s Naivným Bayes
  - Burrows (1987, 1989) – analýza hlavných komponentov (PCA) na frekvenciách slov
  - Závislosť po sebe idúcich slov
- Strojové učenie

Znaky anglického textu

- Zložitostné mierky
- Function words
- Syntax, vetné členy
- Functional Lexical Taxonomies
- Content words
- Písmenové n-gramy

1. Zložitostné mierky

- Priemerná dĺžka slov (rozdelenie dĺžok slov)
  - Počtom slabík (Fucks, 1952)
  - Počtom písmen (Brinegar, 1963; Mendenhall, 1887)
- Priemerný počet slov vo vete (Morton, 1965; Yule, 1944)
- Samé nie sú veľmi efektívne
- Lepšie ako doplnok iných

2. Function words

- Nezávislé od obsahu textu
- Manipulácia nepravdepodobná
- Rozdelenie FW (Mosteller, Wallace 1964)
- V súčasnosti zoznamy stoviek slov (vrátane zámen, predložiek, modálnych slovies, spojok, členov)

3. Syntax, vetné členy

- Relatívne frekvencie vetných členov, ich postupností a ich kombinácií s konkrétnymi slovnými druhmi (Argamon-Engelson et al., 1998;)

4. Functional Lexical Taxonomies

- Reprezentujú gramatické a sémantické rozdiely medzi triedami FW v rôznych úrovniach abstrakcie (Matthiessen, 1992)
- Stromy
  - Koreň je slovný druh
  - Potomkovia sú zmysluplné podtriedy rodiča (napr. typy zámien)
  - Listy sú konkrétne slová
- Sú tvorené z uzavretej množiny, takže označovanie slov textu nie je potrebné

5. Content words

- Relatívne frekvencie synoným (Argamon et al., 2008;)
- Slová s rovnomerným rozdelením sa vyraďujú (Forman, 2003)
- Závislé na obsahu textu

6. Písmenové n-gramy

- Vedia zachytiť lexikálne, gramatické a ortografické preferencie autorov bez potreby lingvistického vzdelania
- Úspešne používané v:
  - Angličtine (Clement & Sharp, 2003;)
  - Holandčine (Hoorn, Frank, Kowalczyk, & van der Ham, 1999)
  - Ruštine (Kukushkina, et al. 2001)
  - Taliančine (Benedetto et al., 2002)
  - Gréčtine (Keselj et al., 2003;)

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

Príklad

Features/learner	NB (%)	J4.8 (%)	RMW (%)	BMR (%)	SMO (%)
FW	60.2	58.7	66.1	68.2	63.8
POS	61	59	66.1	66.3	67.1
FW + POS	65.9	61.6	68	67.8	71.7
SFL	57.2	57.2	65.6	67.2	62.7
CW	67.1	66.9	74.9	78.4	74.7
CNG	72.3	65.1	73.1	80.1	74.9
CW + CNG	73.2	68.9	74.2	83.6	78.2

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.