Využití jazyka a profilování autora: Určení pohlaví a věku

Use of Language and Author Profiling: Identification of Gender and Age

Úvod

Francisco Rangel, Paolo Rosso

- Natural Language Engineering Lab, ELiRF, Universitat Politecnica de Valencia, Spain
- NLPCS 2013

- Oblíbené téma
- PAN 2013

Využití profilování autora

- Forenzní lingvistika: jazyk jako důkaz
- Zabezpečení: profilování možných delikventů
- Marketing: segmentace uživatelů

Předchozí práce: Používání jazyka na internetu

- Jaké gramatické kategorie lidé používají v různých kanálech
  - Wikipedie, informační letáky, blogy, diskuzní fóra, Twitter a Facebook
- Identifikace základních emocí
  - Radost, překvapení, smutek, znechucení, hněv, strach
  - Sada stylistických vlastností
  - Analýza: jak se jazyk liší podle pohlaví, tématu a emocí

Současná práce

- Kognitivní rysy: jak se lišíme na základě pohlaví a věku
- Sada stylistických rysů: modelace rozdílů
- Metoda SVM (Support vector machines)
- Datová sada: PAN-AP-13
  - Velké množství anonymních textů

Neurologie: teoretický rámec

- Pohled mluvčího: CO se říká a JAK se to říká
- Pohled posluchače: CO se říká a KDO to říká

- Profilování autora: JAK/KDO

Automatická identifikace pohlaví a věku na základě stylistických rysů

- Pennebaker:
  - Angličtina
  - Profilování autora na základě stylistických rysů
- Získání frekvence použitých gramatických kategorií

Výzkum 1/3

Table 1. Distribution of grammatical categories per channel

POS	WIKI	NEWS	BLOGS	FORUMS	TW	FB
ADJ	13.57	12.50	13.67	9.27	6.62	12.06
ADV	2.78	3.46	3.87	**4.74**	**6.30**	3.49
CONJ	1.52	2.10	1.80	4.18	**7.00**	2.64
Q	3.34	4.47	4.15	5.34	5.53	4.29
DET	2.88	3.48	2.78	4.18	6.40	4.02
INTJ	0.35	0.04	0.06	0.42	0.38	0.07
MD	0.01	0.03	0.02	0.00	0.00	0.00
PREP	4.00	5.49	5.07	**8.94**	**13.81**	6.15
PRON	0.65	0.92	1.12	**2.22**	**3.32**	1.39
NOM	**50.33**	47.05	46.59	42.63	34.08	47.07
VERB	20.55	20.47	**20.88**	18.08	16.56	18.83

Výzkum 2/3

Table 2. Frequency of person and number in pronouns and verbs

POS	PER	NUM	WIKI	NEWS	BLOG	FOR	TW	FB
PRON	1	SIN	13.61	14.58	18.85	**54.47**	**65.81**	22.3
		PLU	0.00	0.00	0.00	0.00	0.00	0.00
	2	SIN	4.58	1.18	2.23	1.54	3.53	3.95
		PLU	1.92	1.75	5.31	4.61	5.62	3.49
	3	SIN	**55.06**	**50.75**	39.26	24.08	12.70	34.68
		PLU	13.42	18.22	16.93	8.91	3.35	17.14
	OTHER		11.41	13.52	17.42	6.39	8.99	18.44
VERB	1	SIN	19.95	17.41	17.50	28.94	24.00	16.61
		PLU	2.10	2.42	4.19	2.68	4.68	4.89
	2	SIN	6.02	1.55	3.58	3.55	6.77	2.95
		PLU	0.46	0.42	0.69	0.98	1.65	0.76
	3	SIN	31.40	34.00	29.92	28.80	31.21	31.21
		PLU	40.07	44.20	45.11	35.05	31.69	43.59

Výzkum 3/3

Table 3. Distribution of grammatical categories by gender

POS	ALL	MALE	FEMALE
ADJ	6.49	6.53	6.45
ADV	3.93	3.94	3.91
CONJ	9.51	9.55	9.46
Q	5.46	5.76	5.12
DET	7.25	6.81	**7.74**
INTJ	0.23	0.18	**0.30**
MD	0.00	0.00	0.00
PREP	6.06	**6.25**	5.85
PRON	2.45	2.24	**2.67**
NOM	31.89	32.21	31.53
VERB	15.38	15.44	15.32

Stylistické rysy

- Frekvence: poměr počtu unikátních slov a celkového počtu slov, slov začínajících velkým písmenem, slov psaných velkými písmeny, délka slov, počet zdvojených hlásek (např. Heeeelloooo)
- Interpunkční znaménka : frekvence interpunkčních znamének
- Frekvence využívání jednotlivých gramatických kategorií, vlastních jmen, slov mimo slovníky apod.
- Emoticony: Poměr mezi počtem smajlíků a celkovým počtem slov, počet různých druhů smajlíků
- Spanish Emotion Lexicon (SEL)
- ! Žádné rysy závislé na obsahu či kontextu.

Metoda

- Soubor dat PAN-AP-13
- 3 věkové skupiny:
  - 10s (13-17)
  - 20s (23-27)
  - 30s (33 až 47).

Table 4. Distribution of number of authors by age

AGE	NUM. OF AUTHORS	
	TRAIN	TEST
10s	2,500	240
20s	42,600	3,840
30s	30,800	2,720

Metoda

- Strojové učení:
  - SVM metoda
  - Gaussovo jádro
- Poměr mezi počtem správně určených autorů a celkovým počtem autorů.

Výsledky

Table 5. PAN ranking for Author Profiling by Gender and by Age (Spanish)

POS	TEAM	GENDER		POS	TEAM	AGE
1	Santosh	0.6473		1	Pastor	0.6558
2	Pastor	0.6299		2	Santosh	0.6430
3	Haro	0.6165		**3**	**(*Rangel*)**	**0.6350**
4	Ladra	0.6138		4	Haro	0.6219
5	Flekova	0.6103		5	Flekova	0.5966
6	Jankowska	0.5846		6	Ladra	0.5727
7	**(*Rangel*)**	**0.5713**		7	Yong	0.5705
8	Kern	0,5706		8	Ramirez	0.5651
9	Jimenez	0.5627		9	Aditya	0.5643
10	Ayala	0.5526		10	Jimenez	0.5429
11	Cagnina	0.5516		11	Gillam	0.5377
12	Yong	0.5468		12	Kern	0.5375
13	Mechti	0.5455		13	Moreau	0.5049
14	Weren	0.5362		14	Meina	0.4930
15	Meina	0.5287		15	Weren	0.4615
16	Ramirez	0.5116		16	Jankowska	0.4276
17	*Baseline*	*0.5000*		*17*	Cagnina	0.4148
18	Aditya	0.5000		18	Hidalgo	0.4000
19	Hidalgo	0.5000		19	Farias	0.3554
20	Farias	0.4982		*20*	*Baseline*	*0.3333*
21	Moreau	0.4967		21	Ayala	0.2915
22	Gillam	0.4784		22	Mechti	0.0512