AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

System automatycznego rozpoznawania wybranych osób na podstawie analizy ich aktywnego słownika słów i zwrotów

praca dyplomowa

Paweł Miczko, WEAIiE
Katedra Automatyki

promotor:
dr Adrian Horzyk

Kraków, 15 grudnia 2009

Plan prezentacji

1. Cel pracy
2. Wstęp
3. Opis rozwiązania
4. Wyniki
5. Podsumowanie

Cel pracy

- Celem pracy jest skonstruowanie słowników frekwencyjnych słów i zwrotów dla wybranych mówców lub pisarzy oraz ich wykorzystanie do rozpoznania autorstwa innych tekstów wypowiadanych przez tych mówców lub pisarzy.

- Celem pracy jest skonstruowanie systemu pozwalającego na zidentyfikowanie dzieł literackich jak również dowolnych innych tekstów napisanych przez jedną z wybranej grupy osób.

- Dodatkowo system rozpoznaje osobowość autorów na podstawie pewnych słów kluczowych

12 typów osobowości wg dr Adriana Horzyka

1. Dominujący 
2. Maksymalista 
3. Inspirujący 
4. Odkrywczy 
5. Weryfikujący 
6. Systematyczny 
7. Asekuracyjny 
8. Oszczędny 
9. Harmonijny 
10. Empatyczny 
11. Zadaniowy 
12. Równoważący 

Aplikacja internetowa

Dlaczego?

- **łatwość dostępu**
- **klient nie musi nic instalować**
- **użycie popularnych technologii webowych: HTML, CSS, PHP, JavaScript**

Opis działania systemu

1. **Stworzenie bazy słów i zwrotów dla poszczególnych programów osobowości.**

2. **Pozyskanie bazy tekstów i budowa słowników frekwencyjnych dla ich autorów wraz ze stworzeniem ich profilów osobowościowych.**

3. **Wprowadzenie do systemu rozpoznawanego tekstu.**

Pozyskiwanie danych do systemu

1. Parsowanie tekstu, usunięcie znaków interpunkcyjnych
2. Podział tekstu na atomy – słowa oraz frazy dwu- i trzywyrazowe
3. Usunięcie niektórych spójników i przyimków
4. Stworzenie słownika frekwencyjnego słów i zwrotów
5. Obliczenie intensywności programów osobowości
6. Normalizacja

Miara podobieństwa tekstów

Wymagania:

- **Dla identycznych tekstów - 100%**
- **Maleje w stronę zera im bardziej teksty się różnią**
- **Zależy od ilości powtarzających się słów i zwrotów, ich częstotliwości oraz od różnic osobowości autorów**

Obliczenie miary podobieństwa

$$\Delta = 100 - (A+B+C+D+E)$$

A ~ różnice w programach osobowości

B ~ różnice w częstotliwości występowania wspólnych słów

C ~ procent powtarzających się słów

D, E ~ procent powtarzających się zwrotów 2- i 3- wyrazowych

Wyniki

Wynik rozpoznawania wśród autorów z bazy

Autor	Programy osobowości	POWT	Słowa	Zwroty 2-wyr	Zwroty 3-wyr	Suma	% dopasowania	Porównanie	
Chłopi	27	51	48	67	93	285	71.45%	tabele	wykres
Sienkiewicz	55	70	63	83	98	369	63.09%	tabele	wykres
Quo_Vadis	50	78	64	82	97	371	62.88%	tabele	wykres
Szkice_węglem	50	104	64	82	98	398	60.22%	tabele	wykres
Prus	72	105	67	82	98	423	57.67%	tabele	wykres
Mickiewicz	68	104	66	89	99	426	57.41%	tabele	wykres
DołęgaMostowicz	67	115	64	82	98	426	57.41%	tabele	wykres
Konopnicka	72	115	65	80	98	430	56.96%	tabele	wykres
Hłasko	79	118	63	80	98	437	56.33%	tabele	wykres
Kraszewski	93	119	60	80	98	449	55.08%	tabele	wykres
Orzeszkowa	90	159	67	86	99	501	49.87%	tabele	wykres
Krasicki	100	155	74	87	99	515	48.55%	tabele	wykres
Przedwiośnie	102	167	70	87	99	524	47.59%	tabele	wykres
Schulz	120	151	80	94	100	545	45.51%	tabele	wykres

Wyniki

Najczęściej używane słowa i zwroty

Pojedyncze wyrazy

rozpoznawany		Chłopi	
słowo	**ilość**	**słowo**	**ilość**
nie	137	nie	133
że	110	że	103
już	52	już	46
ale	38	tak	38
tak	34	jeno	38
jeszcze	29	kiej	34
aż	28	ale	33
antek	28	jej	28
zaś	26	było	25
ino	25	zaś	22
jeno	24	jak	19
kiej	23	te	18
ze	21	jeszcze	17
go	21	ten	16
było	20	nawet	16
mu	20	jakby	16
jak	19	był	16
też	19	ją	16
był	18	ano	15
przy	18	jako	15

Zwroty 2-wyrazowe

rozpoznawany		Chłopi	
zwrot	**ilość**	**zwrot**	**ilość**
i nie	12	to i	16
i w	12	się jej	13
to i	11	i nie	11
się do	11	i tak	10
że to	11	że i	8
aż się	10	kiej te	8
się z	10	się już	8
się nie	9	się do	8
juści że	9	się w	8
się w	8	że już	7

Zwroty 3-wyrazowe

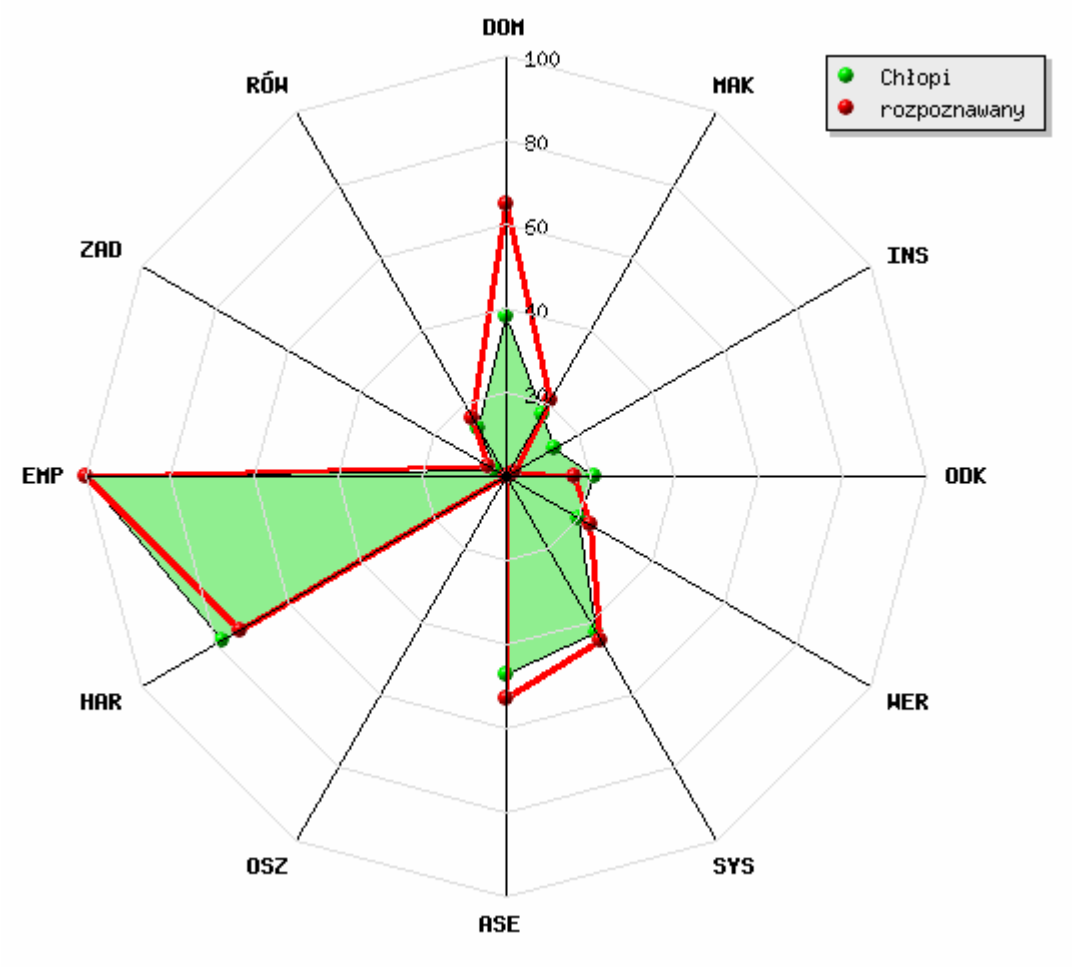
rozpoznawany		Chłopi	
zwrot	**ilość**	**zwrot**	**ilość**
z całej mocy	3	że i nie	4
w ten mig	3	i bez to	4
bartek z tartaku	3	się w sobie	3
się do niej	3	raz po raz	3
że i nie	3	to juści że	3
jest tylko jedna	2	a i bez	3
tylko jedna rada	2	nie dziwota że	3
i nie wypowiedzieć	2	i nie dziwota	3
i tany szły	2	że się już	3
a za nim	2	jakby z musu	2

Wyniki

Porównanie osobowości autorów

rozpoznawany	typ	Chłopi
65%	Dominujący	35%
21%	Maksymalista	16%
4%	Inspirujący	18%
16%	Odkrywczy	20%
25%	Weryfikujący	19%
59%	Systematyczny	54%
53%	Asekuracyjny	44%
0%	Oszczędny	0%
73%	Harmonijny	73%
100%	Empatyczny	100%
5%	Zadaniowy	3%
16%	Równoważący	12%

Wyniki


DOM
MAK
INS
ODK
WER
SYS
ASE
OSZ
HAR
EMP
ZAD
RÓW
100
80
60
40
20
Chłopi
rozpoznawany

Wnioski

System może znaleźć zastosowanie w:

- **rozpoznawaniu autorstwa tekstów**
- **detekcji plagiatów**
- **tworzeniu portretów psychologicznych autorów**

Pytania?

Dziękuję za uwagę!