Separating Named Entities

Marek Grác, Barbora Ulipová

Motivation

Extracting information about people ->

finding people's names?

Capitalized

Separating cluster of capitalized words such as:

PSP Miroslava Němcová

LogDice and MI-score - bigrams

Count LogDice and Miscore for each possible division

PSP – Miroslava Němcová

Bigram: PSP Miroslava

PSP Miroslava – Němcová

Bigram: Miroslava Němcová

LogDice and MI-score - bigrams

	LogDice	MI-score		LogDice	MI-score
PSP-Miroslava	**8.77**	**-14.26**	D. - Cerekve	**-0.49**	-18.18
Miroslava - Němcová	10.21	-13.71	**Cerekve – Zdeněk**	1.08	-18.19
			Zdeněk - Jirsa	2.69	**-19.44**

LogDice and MI-score - bigrams

Results (precision):

LogDice: 29.5 %

MI-score: 11.8 %

LogDice and MI-score - ngrams

	MI-score	logDice
PSP – Miroslava Němcová	10.07	-11.23
PSP Miroslava – Němcová	**10.89**	**-9.86**
D. – Cerekve Zdeněk Jirsa	-1.49	-13.83
D. Cerekve – Zdeněk Jirsa	**10.64**	**-4.30**
D. Cerekve Zdeněk - Jirsa	4.44	-8.01

LogDice and MI-score - ngrams

Results (precision):

LogDice: 41.2 %

MI-score: 29.5 %

Negative n-grams

PSP not followed by Miroslava

CQL :

[word="PSP"][word!="Miroslava"]

Negative n-grams

N-grams:

freq(PSP) – freq(PSP Miroslava Němcová)

Problem:

PSP Miroslava not Němcová

freq(PSP Miroslava) – freq(PSP Miroslava Němcová) = 0

Diff2

Sums negative n-grams:

For division PSP – Miroslava Němcová:

freq(PSP Miroslava) + freq(PSP Miroslava Němcová)

Diff4

Positive + positive – negative – negative

for PSP – Miroslava Němcová:

freq(PSP) + freq(Miroslava Němcová) – freq(PSP Miroslava) – freq(PSP Miroslava Němcová)

Diff Results

Results (precision):

Diff2: 94.1 %

Diff4: 35.2 %

Thank you for attention