Porovnávání kontextů

A dalších podmnožin textu

Jiří Milička
ÚBVA

A boy

He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish. In the first forty days a boy had been with him.

But after forty days without a fish the boy’s parents had told him that the old man was now definitely and finally salao, which is the worst form of unlucky, and the boy had gone at their orders in another boat which caught three good fish the first week. It made the boy sad to see the old man come in each day with his skiff empty …

A boy

He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish. In the first forty days a boy had been with him.

But after forty days without a fish the boy's parents had told him that the old man was now definitely and finally salao, which is the worst form of unlucky, and the boy had gone at their orders in another boat which caught three good fish the first week. It made the boy sad to see the old man come in each day with his skiff empty…

A boy
He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish. In the first forty days a boy had been with him.
But after forty days without a fish the boy’s parents had told him that the old man was now definitely and finally salao, which is the worst form of unlucky, and the boy had gone at their orders in another boat which caught three good fish the first week. It made the boy sad to see the old man come in each day with his skiff empty…

A / an

Boy
Old
Skiff
Fish
Boy
Fish

The

Gulf
First
Boy
Old
Worst
Boy
First
Boy
Old

A / an

Boy
Old
Skiff
Fish
Boy
Fish

The

Gulf
First
Boy
Old
Worst
Boy
First
Boy
Old

A / an

Boy

Old

Skiff

Fish

Boy

Fish

The

Gulf

First

Boy

Old

Worst

Boy

First

Boy

Old

A / an

Boy
Old
Skiff
Fish
Boy
Fish

The

Gulf
First
Boy
Old
Worst
Boy
First
Boy
Old

Tři tokeny „mají bratříčka“ , je to moc, nebo málo?

Jak tyto množiny porovnávat?
Spousta metrik, ztížená interpretace.

Kdyby množina A byla z textu náhodně vybraná
a množina B náhodně vybraná z toho,co zbude.
Kolik průměrně tokenů bude „mít bratříčka“?

Typ *Day* (celkem 4x)

Například:

A (2x)	x	**B** (1x)	x	Zbytek textu (1x)
000011		000000001		00000… …00000001
000101		000000100		00000… …00000010
000110		000001000		00000… …00000100
001001		000010000		00000… …00001000
001010		000100000		00000… …00010000
001100		001000000		00000… …00100000
010001		010000000		…
010010		100000000		10000… …00000000
010100				
011000				
100001				
…				
110000				

Jaká je pravděpodobnost,
že bude například v prvním výběru dvakrát
a v druhém výběru jednou?

Day (celkem 4x)

A (2x)	x	B (1x)	x	Zbytek textu (1x)
000011		000000001		00000… …00000001
000101		000000100		00000… …00000010
000110		000001000		00000… …00000100
001001		000010000		00000… …00001000
001010		000100000		00000… …00010000
001100		001000000		00000… …00100000
010001		010000000		…
010010		100000000		10000… …00000000
010100				
011000				
100001				
…				
110000				

$$\frac{|A|!}{m!(|A|-m)!} \qquad \frac{|B|!}{n!(|A|-n)!} \qquad \frac{[d-(|A|+|B|)]!}{[f_i-(m+n)]!\{d-(|A|+|B|)-[f_i-(m+n)]\}!}$$

Day (celkem 4x)

Kolik permutací může být v celém textu?

```
00000…  …00000000001111
00000…  …00000000010111
00000…  …00000000011011
00000…  …00000000011101
00000…  …00000000011110
00000…  …00000001000111
00000…  …00000001001011
00000…  …00000001001101
00000…  …00000001001110
00000…  …00000001010011
00000…  …00000001010101
…
11101…  …00000000000000
11110…  …00000000000000
```

$$\frac{d!}{f_i!(d-f_i)!}$$

$$r_{m;n} = \frac{\frac{|A|!}{m!(|A|-m)!}\,\frac{|B|!}{n!(|A|-n)!}\,\frac{[d-(|A|+|B|)]!}{[f_i-(m+n)]!\{d-(|A|+|B|)-[f_i-(m+n)]\}!}}{\frac{d!}{f_i!(d-f_i)!}}$$

$$g_{A;B} = \sum_{i=1}^{M}(\sum_{m=1}^{|A|}(m\sum_{n=m}^{|B|} r_{m;n}) + \sum_{n=1}^{|B|}(n\sum_{m=n+1}^{|A|} r_{m;n}))$$

M … Počet typů

f_i … frekvence i-tého typu

A … První zkoumaná podmnožina

B … Druhá zkoumaná podmnožina

$$g_{A;B} = \sum_{i=2}^{M} \sum_{m=1}^{|A|} \sum_{|B|}^{n=1} min(m,n) \frac{\binom{|A|}{m}\binom{|B|}{n}\binom{d-|A|-|B|}{f_i-m-n}}{\binom{d}{f_i}}$$

Pokud A sjednoceno B dávají celý text pak:

$$g_{T;A} = \sum_{i=2}^{M} \sum_{m=1}^{|A|} min(m, f_i - m) \frac{\binom{|A|}{m} \binom{d-|A|}{f_i - m}}{\binom{d}{f_i}}$$

Kolem milionu náhodných výběrů
z The Last of the Mohicans

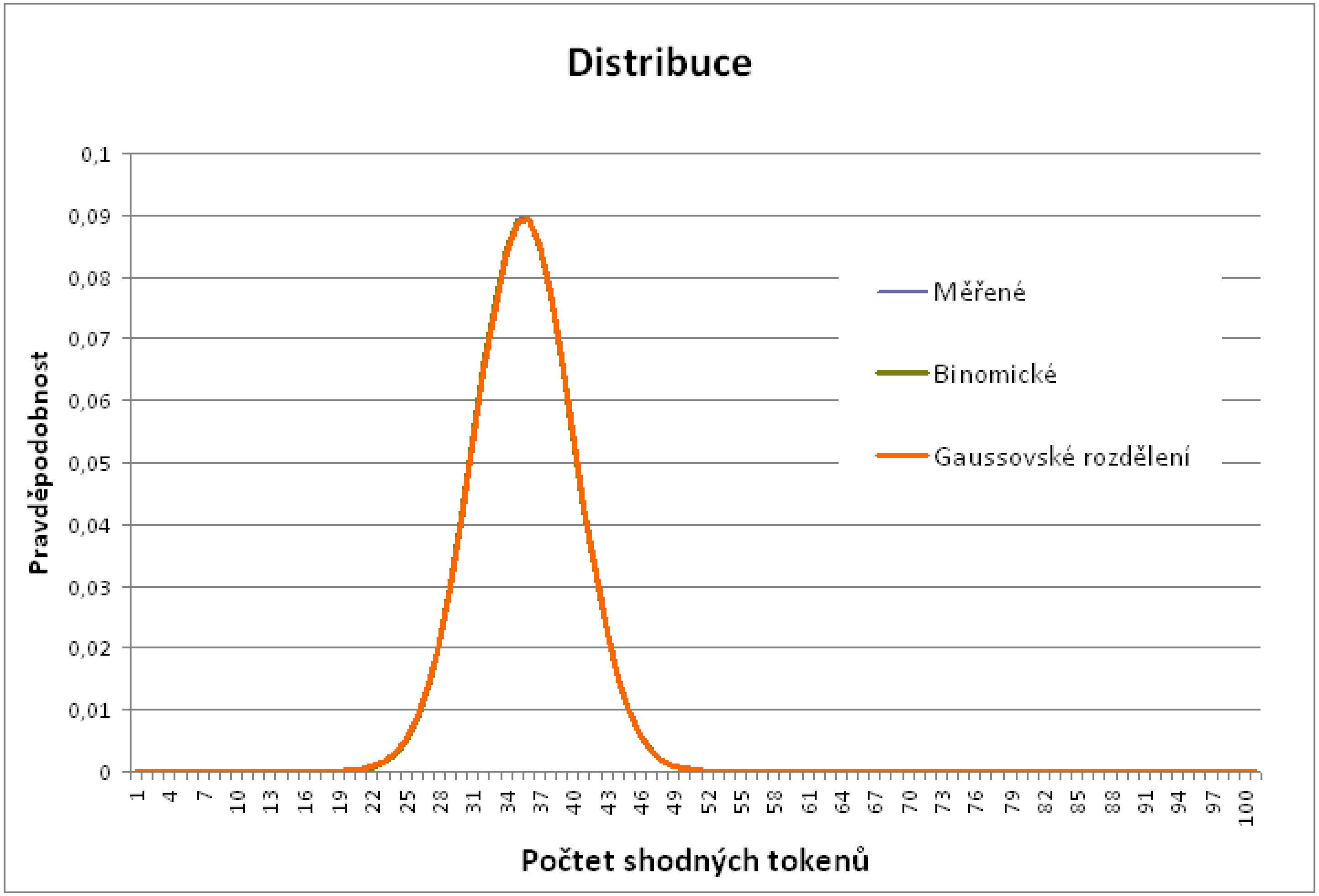
|A|	|B|	Měřeno	Model
1000	1000	505,598	505,586
150	100	34,6007	34,6000
50	25	6,05841	6,05592
20	5	0,81816	0,81827
2	2	0,05180	0,05202
10 000	10 000	7333,08	7333,07
1000	5	3,03604	3,03390

Příklad užití (bezprostřední pravý kontext):

Slovo	Frekvence
say	567
says	64
said	991

Dvojice	Model	Změřeno	Poměr
say – says	35,02	47	1,34
say – said	324,44	292	0,90

Jaká je pravděpodobnost, že při náhodném výběru podmnožin bude počet shodných tokenů stejná a vyšší?



Využití

- Testování lingvistických hypotéz
- Určování autorství / plagiátorství / tématu
- Desambiguace
- Hnízda v různých stromech

Děkuji za pozornost

milicka@centrum.cz