



AKADEMIA GÓRNICZO-HUTNICZA  
IM. STANISŁAWA STASZICA W KRAKOWIE

# Przetwarzanie Języka Naturalnego

## Lab 4 – Spellchecker Bayesa

Zbigniew Kaleta  
`zkaleta@agh.edu.pl`

Wydział IEiT  
Katedra Informatyki

20.03.2018

Prawdopodobieństwo zajścia zdarzenia A pod warunkiem zajścia zdarzenia B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$A, B \subset \Omega, P(B) > 0$$

w skom  
rest  
gotowec  
zimplenaty

$B_1, B_2, \dots, B_n$  wykluczają się parami,  
 $A \subset \sum_{i=1}^n B_i$

$$P(B_i|A) = \frac{P(A|B_i) * P(B_i)}{\sum_{j=1}^n P(A|B_j) * P(B_j)}$$

$C$  – zbiór form

$C \ni c$  – poprawka

$w$  – wprowadzona forma

$$P(c|w) = \frac{P(w|c) * P(c)}{P(w)}$$

$c_i$  jest najlepszą poprawką  $\Leftrightarrow P(c_i|w) = \max_{c \in C} P(c|w)$

$P(w)$  – prawdopodobieństwo wystąpienia danego napisu (błédnego).

Jest stałe dla każdego  $c$ , więc nie jest potrzebne

$P(c)$  – prawdopodobieństwo wystąpienia poprawki – jest proporcjonalne do częstotliwości występowania  $c$  w języku

$P(w|c)$  – prawdopodobieństwo wystąpienia błędu  $w$ , pod warunkiem że poprawnym wyrazem było  $c$ , może być oszacowane na podstawie metryki określającej odległość pomiędzy napisami, np. metryki Levenshteina

## Wykładanie Laplace'a (*additive smoothing*)

$N_c$  – liczba wystąpień  $c$  w korpusie

$N$  – liczba wszystkich wystąpień w korpusie ( $\sum_c N_c$ )

$$N_c = 0 \Rightarrow P(c) = \frac{N_c}{N} = 0$$

Żeby tego uniknąć należy użyć wykładania (ang. *smoothing*).

Jednym z najprostszych jest wykładanie Laplace'a:

$$P(c) = \frac{N_c + \alpha}{N + \alpha * M}$$

gdzie  $M$  jest liczbą wszystkich dopuszczalnych form (rozmiar słownika).

prechwytywanie bzdur  
rest parametrów  
w gulem

- 1 Napisać funkcję obliczającą prawdopodobieństwo błędu  $P(w|c)$  (1 pkt)
- 2 Wprowadzić modyfikację do metryki Levenshteina, uwzględniającą jeden powszechny błąd (1 pkt)
- 3 Korzystając z naiwnego klasyfikatora Bayesa zaproponować najlepszą poprawkę dla wpisanego słowa (1 pkt)

Formy:

`http://home.agh.edu.pl/~zkaleta/pjn/lab4.tar.gz`

- ✚ „Speech and Language Processing, 3rd edition”, Punkt 4.4  
<https://web.stanford.edu/~jurafsky/slp3/4.pdf>