

AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

## Przetwarzanie Języka Naturalnego Lab 3

Wojciech Korczyński wojciech.korczynski@agh.edu.pl

Wydział IEiT Katedra Informatyki

25.03.2015

W. Korczyński (KI AGH) PJN 3

2015



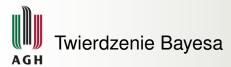
## Prawdopodobieństwo warunkowe

Prawdopodobieństwo zajścia zdarzenia A pod warunkiem zajścia zdarzenia B:

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

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

W. Korczyński (KI AGH)



 $B_1, B_2, \ldots, B_n$  wykluczają się parami

$$\Rightarrow \bigvee_{A \subset B_1 \cup B_2 \cup ... \cup B_n} P(B_k|A) = \frac{P(A|B_k) * P(B_k)}{P(A)}$$



## Twierdzenie Bayesa a sprawdzanie pisowni

C – zbiór form

 $C \ni c$  – poprawka

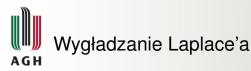
w − wprowadzona forma

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

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



P(w) – prawdopodobieństwo wystąpienia danego napisu (błędnego). 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) jest prawdopodobieństwem błędu o odległości Levenshteina równej odl. pomiędzy w a c



 $N_c$  – ilość wystąpień c w korpusie N – ilość wszystkich wystąpień w korpusie  $(\sum_c N_c)$   $N_c = 0 \Rightarrow P(c) = \frac{N_c}{N} = 0$  Żeby tego uniknąć należy użyć wygładzenia Laplace'a:

$$P(c) = \frac{N_c + 1}{N + M}$$

gdzie M jest liczbą wszystkich dopuszczalnych form



- Napisać funkcję obliczającą prawdopodobieństwo błędu P(w|c) (1 pkt)
- Zebrać statystyki występowania form w korpusie (1 pkt)
- Korzystając z naiwnego klasyfikatora Bayesa zaproponować najlepszą poprawkę dla wpisanego słowa (1 pkt)

## Materialy:

http://home.agh.edu.pl/~wojtek/pjn2015/lab3.tar.gz

7/7