Przetwarzanie Języka Naturalnego
Lab 4 – Spellchecker Bayesa

dr inż. Aleksander Smywiński-Pohl
apohllo@agh.edu.pl

Wydział IEiT
Katedra Informatyki

21.03.2017

Prawdopodobieństwo warunkowe

Prawdopodobieństwo zajścia zdarzenia A pod warunkiem zajścia zdarzenia B:

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

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

Twierdzenie Bayesa

$B_1, B_2, \ldots, B_n$ wykluczają się parami, $A \subset \sum_i^n B_i$

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

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

Założenia

$P(w)$ – 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

Wygładzanie Laplace'a

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

Zadania

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

Formy: `http://apohllo.pl/text/lab4.tar.gz`