



AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Przetwarzanie Języka Naturalnego

Lab 1

mgr inż. Zbigniew Kaleta
zkaleta@agh.edu.pl

Wydział IEiT
Katedra Informatyki

11.03.2015

- ✚ tryb laboratoryjny
- ✚ 1.5 h tygodniowo do końca semestru
- ✚ obecność obowiązkowa

- ✚ **głównie** na podstawie zadań domowych
- ✚ za każde laboratorium do zdobycia 3 pkt.
- ✚ zadania z danego laboratorium należy oddać na następnych zajęciach
- ✚ możliwe kolokwium

- ✦ n-gramem nazywamy **każdą** sekwencję n kolejnych składowych
- ✦ sekwencje mogą się zazębiać
- ✦ w przypadku analizy języka składowymi mogą być litery, sylaby lub słowa

- ✚ pozwala przedstawić korpus tekstowy w postaci wektora
- ✚ prosty
- ✚ skalowalny (ze względu na wielkość korpusu czy n ?)

Słowo: przetwarzanie

digramy: pr, rz, ze, et, tw, wa, ar, rz, za, an, ni, ie

trigramy: prz, rze, zet, etw, twa, war, arz, rza, zan, ani, nie

Zdanie: Mężny bądź, chroń pułk twój i sześć flag.

digramy: Mężny bądź, bądź chroń, chroń pułk, pułk twój, twój i, i sześć, sześć flag

Odległość między wektorami

$$x = [x_1, x_2, \dots, x_n]$$

$$y = [y_1, y_2, \dots, y_n]$$

- ✚ euklidesowa: $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$
- ✚ taksówkowa: $d(x, y) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n|$
- ✚ maksimum: $d(x, y) = \max(|x_1 - y_1|, |x_2 - y_2|, \dots, |x_n - y_n|)$
- ✚ cosinusowa: $d(x, y) = \frac{x_1*y_1 + x_2*y_2 + \dots + x_n*y_n}{\text{len}(x)*\text{len}(y)}$

Normalizacja?

- 1 Napisać program budujący statystykę n-gramów dla różnych języków (1 pkt.)
- 2 Napisać program odgadujący język zdania wprowadzonego przez użytkownika (1 pkt.)
- 3 Przeanalizować wyniki odgadywania w zależności od n (1 pkt.)

Korpusy:

<http://home.agh.edu.pl/~zkaleta/pjn/lab1.tar.gz>