

AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

## Przetwarzanie Języka Naturalnego Lab 1

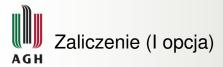
Zbigniew Kaleta zkaleta@agh.edu.pl

Wydział IEiT Katedra Informatyki

8.03.2019

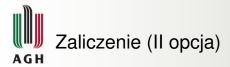


- tryb laboratoryjny
- 1.5 h tygodniowo do końca semestru
- ★ obecność obowiązkowa
- ★ konsultacje: czwartek 10:30 11:30 lub po umówieniu mailowo w 4.60



- ★ 10 zestawów zadań domowych, za każdy do zdobycia maks. 3 pkt.
- 🖈 kolokwium za 10 pkt. na koniec semestru
- w przypadku usprawiedliwionej nieobecności zadania należy oddać na pierwszych zajęciach, z których się nie ma zwolnienia
- zadania z danego laboratorium należy oddać na następnych zajęciach
- każdy tydzień spóźnienia oznacza utratę 1 pkt. z danego zadania (bez punktów ujemnych)

Z. Kaleta (KI AGH) PJN 1 2018 3 / 10



- ★ 10 zestawów zadań domowych, za każdy do zdobycia maks. 3 pkt.
- 🖈 kolokwium za 10 pkt. na koniec semestru
- w przypadku usprawiedliwionej nieobecności zadania należy oddać na pierwszych zajęciach, z których się nie ma zwolnienia
- zadania z danego laboratorium należy oddać na jednych z dwóch następnych zajęć
- 🖈 na koniec semestru można poprawić jedno dowolne zadanie

Z. Kaleta (KI AGH) PJN 1 2018 4



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

Ala me, kota, n = 2

Z. Kaleta (KI AGH) PJN 1 2018 5/10



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



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

sklean Count Vectorizer



## Odległość między wektorami

$$x = [x_1, x_2, ..., x_n]$$
  
 $y = [y_1, y_2, ..., y_n]$ 

$$\bigstar$$
 euklidesowa:  $d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$ 

$$\bigstar$$
 taksówkowa:  $d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|$ 

**Maksimum**: 
$$d(x,y) = max(|x_1 - y_1|, |x_2 - y_2|, ..., |x_n - y_n|)$$

**$$\bigstar$$** cosinusowa:  $d(x, y) = 1 - \frac{x_1 * y_1 + x_2 * y_2 + ... + x_n * y_n}{len(x) * len(y)}$ 

Normalizacja?

Normy

ma

bordro

mosing



## Miary poprawności klasyfikacji (binarnej)

Precision (precyzja): jak duży procent obiektów zaklasyfikowanych do A został poprawnie zaklasyfikowany

$$precision = \frac{|\text{true positives}|}{|\text{true positives} \cup \text{false positives}|}$$

Recall (pełność): jak duży jest procent poprawnie zaklasyfikowanych obiektów względem wszystkich obiektów w zbiorze wzorcowym

$$\mathit{recall} = \frac{|\mathsf{true}| \mathsf{positives}|}{|\mathsf{true}| \mathsf{positives} \cup \mathsf{false}| \mathsf{negatives}|}$$

F1: średnia harmoniczna miar precision i recall

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Accuracy (skuteczność):

| true positives∪true negatives | | true positives∪false positives∪true negatives | | true positives∪false positives∪false negatives |

2018



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

## Korpusy:

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

Z. Kaleta (KI AGH) PJN 1 2018 10 / 10