



AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Przetwarzanie Języka Naturalnego

Lab 9

Wojciech Korczyński
`wojciech.korczynski@agh.edu.pl`

Wydział IEiT
Katedra Informatyki

20.05.2015

Wektorowy model tekstu:

- ✚ „*bag of words*”, *space-vector model*
- ✚ nie uwzględniamy kolejności występowania wyrazów
- ✚ łatwe i efektywne

\mathcal{C} – korpus

D – dokument ($\in \mathcal{C}$)

k – rząd ($\in (N)$)

$$G(\mathcal{C}, D, k) = (N(\mathcal{C}), A(D, k))$$

$N(\mathcal{C})$ – zbiór wierzchołków – każdemu unikatowemu słowu z \mathcal{C} odpowiada jeden wierzchołek

$A(D, k)$ – zbiór krawędzi – wierzchołki (słowa) x i y łączy krawędź skierowana $:\Leftrightarrow$ w dokumencie D x poprzedza y o co najwyżej k słów. Może zawierać krawędzie wielokrotne, a także każdy wierzchołek zawiera krawędź do siebie samego.

- ❶ zaaplikowanie stoplisty
- ❷ utworzenie wierzchołka dla każdego słowa występującego w tekście
- ❸ utworzenie okna przesuwnego zawierającego $k + 1$ słów
 - ❶ dla każdego słowa w oknie dodanie krawędzi między pierwszym słowem a danym
 - ❷ przesunięcie okna o 1 i powtórzenie operacji

- ✦ klasteryzacja
- ✦ klasyfikacja (Bayesa, kNN, centroid, regułowa)
- ✦ indeksowanie i wyszukiwanie
- ✦ wykrywanie plagiatów

Zalety:

- ✦ zachowuje informacje o następstwie słów
- ✦ konwertowalny na SVM, więc wszystkie algorytmy są nadal stosowalne
- ✦ czytelny dla człowieka

Wady:

- ✦ większy koszt pamięciowy i obliczeniowy niż SVM

Konwersja na SVM (ale nie „bag of words”)

w_1, w_2, \dots, w_n – słowa

Bag of words: $x = [f(w_1), f(w_2), \dots, f(w_n)]$

$f(w)$ – ilość wystąpień w lub jej funkcja

SVM z reprezentacji grafowej:

$x = [F(w_1, w_1), F(w_1, w_2), \dots, F(w_1, w_n), F(w_2, w_1), \dots, F(w_n, w_n)]$

$F(w_a, w_b)$ – liczność krawędzi od w_a do w_b lub jej funkcja

Jeśli tekst B zawiera się w tekście A , to graf utworzony z tekstu B jest podgrafem grafu utworzonego z tekstu A .

- 1 Stworzyć stoplistę (0.5 pkt.)
- 2 Przekonwertować korpus PAP na model grafowy (1 pkt)
- 3 Dla wybranych 10 notatek znaleźć po 10 najbliższych notatek dla k od 0 do 4 (1.5 pkt.)

Materiały:

<http://home.agh.edu.pl/~wojtek/pjn2015/lab9.tar.gz>