

AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Przetwarzanie Języka Naturalnego Lab 9 – Grafowa reprezentacja tekstu

Zbigniew Kaleta zkaleta@agh.edu.pl

Wydział IEiT Katedra Informatyki

20.05.2015



Wektorowy model tekstu:

- 🖈 nie uwzględniamy kolejności występowania wyrazów
- * łatwe i efektywne obliczeniowo



Graf odległości (distance graph)

C – korpus

D – dokument ($\in \mathcal{C}$)

k – rząd ($\in \mathbb{N}$)

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

Z. Kaleta (KI AGH) PJN 8 2015 3 / 10



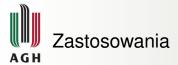
 $N(\mathcal{C})$ – zbiór wierzchołków – każdemu unikatowemu słowu z \mathcal{C} odpowiada jeden wierzchołek

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.

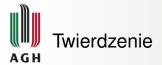


- zaaplikowanie stoplisty
- utworzenie wierzchołka dla każdego słowa występującego w tekście
- \odot utworzenie okna przesuwnego zawierającego k+1 słów
 - dla każdego słowa w oknie dodanie krawędzi między pierwszym słowem a danym
 - przesunięcie okna o 1 i powtórzenie operacji

5/10

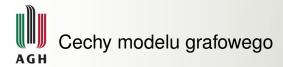


- klasteryzacja
- klasyfikacja (Bayesa, kNN, centroid, regułowa)
- ★ indeksowanie i wyszukiwanie
- wykrywanie plagiatów



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

Z. Kaleta (KI AGH) PJN 8 2015 7 / 10



Zalety:

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

Wady:

★ większy koszt pamięciowy i obliczeniowy niż SVM



Konwersja na SVM (ale nie "bag of words")

$$w_1, w_2, ..., w_n$$
 – słowa

Bag of words: $x = [f(w_1), f(w_2), ..., f(w_n)]$ f(w) – liczba wystąpień w lub jej funkcja

SVM z reprezentacji grafowej:

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

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

Z. Kaleta (KI AGH) PJN 8 2015 9 / 10



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

Materialy:

http://home.agh.edu.pl/~zkaleta/pjn/lab6.tar.gz

Z. Kaleta (KI AGH) PJN 8 2015 10 / 10