



AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Przetwarzanie Języka Naturalnego

Lab 7

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

Wydział IEiT
Katedra Informatyki

29.04.2015

Reprezentacja wektorowa tekstu

- ✚ każda składowa wektora odpowiada jednemu słowu
- ✚ wymaga ustalenia pewnego słownika (bazy przestrzeni)
- ✚ umożliwia matematyczne traktowanie tekstu
- ✚ zachowuje informację o częstotliwości występowania słów
- ✚ traci informację o kolejności słów oraz o gramatyce
- ✚ „Ala ma kota” i „Kot ma Alę” mają taką samą reprezentację wektorową
- ✚ stosowane w przypadku kolekcji dokumentów

Reprezentacja „bag of words”

- ✦ najprostsza reprezentacja wektorowa tekstu
- ✦ wartość składowej jest równa liczbie wystąpień danego słowa w tekście
- ✦ największą wagę mają słowa występujące najczęściej ALE niosą najmniej informacji

W obliczaniu składowych wektora można posłużyć się następującymi współczynnikami:

- ✚ *tf* – *term frequency*, częstotliwość (liczba) wystąpień termu w tekście
- ✚ *df* – *document frequency*, liczba dokumentów, w których występuje term
- ✚ *cf* – *collection frequency*, liczba wystąpień termu w całym korpusie

Funkcje tłumiące (spłaszczające)

- ✦ zmniejszają wpływ bardzo częstych termów na reprezentację wektorową
- ✦ najczęściej stosowane: pierwiastek, logarytm
- ✦ może to być dowolna (monotoniczna?) funkcja o pochodnej z przedziału $]0; 1[$

Term frequency - inversed document frequency:

t – term

d – dokument

N – ilość wszystkich dokumentów

w – waga (składowa wektora)

$$w(t, d) = tf(t, d) * \log\left(\frac{N}{df(t)}\right)$$

- 1 zbudować macierz tf-idf dla korpusu PAP (0.5 pkt.)
- 2 dla każdej notatki wygenerować słowa kluczowe (0.5 pkt.)
- 3 napisać program wyszukujący notatki na podstawie słów (1 pkt)
- 4 napisać program wyszukujący notatki podobne do wybranej (1 pkt)

Materiały:

<http://home.agh.edu.pl/~wojtek/pjn2015/lab7.tar.gz>