

# Przetwarzanie Języka Naturalnego

## Lab 8

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

**Wydział IEiT**  
**Katedra Informatyki**

**06.05.2015**

- ✚ model algebraiczny dla reprezentacji tekstu
- ✚ wektor dla i-tego tekstu:  $t_i = (w_1, w_2, \dots, w_n)$
- ✚ „bag of words”: nie uwzględniamy kolejności występowania wyrazów
- ✚ Wady:
  - ▶ wyrazy występujące tylko w jednym tekście niosą z sobą dużo informacji (vide prawo Zipfa), ale nie mają dużego wpływu na podobieństwo (vide np. metryka cosinusowa)
  - ▶ wyrazy występujące często nie niosą ze sobą informacji, a mają mocny wpływ na pozorne podobieństwo
  - ▶ zgodność tekstów sprowadza się do zgodności wyrazów

- ✦ usuwamy *hapax legomena*
- ✦ usuwamy wyrazy, które występują w więcej niż 70% tekstów
- ✦ w macierzy term-document wartość w danej komórce zawiera wagę danego wyrazu w danym tekście

TF-IDF (Term Frequency - Inverted Document Frequency)

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

- ✦ jest to również sposób wyznaczania słów kluczowych
- ✦ mierzona jest jednocześnie częstość wyrazu w dokumencie i jego rzadkość korpusowa
- ✦ wada: porównywana jest **dokładna zgodność** wyrazów

# Latent Semantic Analysis (LSA)

- ✦ czasem nazywane Latent Semantic Indexing (LSI)
- ✦ metoda analizy podobieństwa między dokumentami i wyrazami oparta na tworzeniu zbioru *pojęć* (*concepts*)
- ✦ założenie: słowa bliskoznaczne pojawiają się w podobnych fragmentach tekstu
- ✦ rozkład macierzy term-document przy pomocy dekompozycji głównych składowych (ang. *singular value decomposition* - SVD)
- ✦ zmniejszona zostaje liczba wierszy (wyrazów) przy zachowaniu podobieństwa między kolumnami (dokumentami)

$A$  – macierz term-document o wymiarach  $n \times m$

$$A = U\Sigma V^T$$

$U$  – macierz pojęć o wymiarach  $n \times l$  (wektory własne)

$\Sigma$  – przekątniowa macierz wartości własnych o wymiarach  $l \times l$

$V$  – macierz dokumentów o wymiarach  $m \times l$  (wektory własne)

Wymiary nowego układu współrzędnych wyznaczonego przez wektory własne to *pojęcia* lub *tematy* (*concepts*, *topics*).

Wybierając  $k$  największych wartości własnych dokonujemy redukcji wymiarów:

$$A' = U' \Sigma' V'^T$$

Wymiary macierzy  $U'$ ,  $\Sigma'$ ,  $V'$  to teraz kolejno:  $n \times k$ ,  $k \times k$ ,  $m \times k$   
Możemy teraz porównywać wyrazy i dokumenty w przestrzeni o mniejszej ilości wymiarów.

Zalety:

- ✚ oszczędność reprezentacji ( $k$  jest często rzędu setek)
- ✚ zwiększona skuteczność (usuwany jest szum)

## Wady:

- ✚ pojęcia, jako wektory, często nie mają zrozumiałej dla człowieka postaci (niejasne komponenty powstałe przy redukcji wymiarów, ujemne wartości wag, etc.), na przykład:
  - ▶  $[(auto), (motor), (kwiat)] \rightarrow [(1.34 * auto + 0.28 * motor), (kwiat)]$
  - ▶  $[(auto), (butelka), (kwiat)] \rightarrow [(1.34 * auto + 0.28 * butelka), (kwiat)]$
- ✚ ponieważ waga dla każdego słowa to pewien punkt w przestrzeni, LSA jest nieczułe na polisemię
- ✚ konsekwencje wynikające z użycia modelu „bag of words”



- ✦ dokument to „mieszanka” tematów, gdzie każdy temat to zbiór słów z przypisanymi im prawdopodobieństwami
- ✦ wektor tematów dla każdego dokumentu jest bardziej zrozumiały niż w przypadku LSA
- ✦ dzięki bardziej zrozumiałej reprezentacji, model jest modelem generatywnym
- ✦ prezentowany jest rozkład tematów w dokumencie i rozkład słów w temacie (zgodnie z rozkładem Dirichleta)

Przykładowy algorytm:

- ❶ dla każdego dokumentu każdemu słowu losowo przypisz jeden z  $K$  tematów
- ❷ dla każdego słowa  $w$  w każdym dokumencie  $d$ :
  - ❶ dla każdego tematu  $t$  oblicz:  $P(t|d)$  i  $P(w|t)$
  - ❷ przypisz słowu  $w$  nowy temat z prawdopodobieństwem  $P(t|d) * P(w|t)$

Po powtórzeniu tych kroków wiele razy otrzymujemy całkiem dobry rozkład tematów w dokumentach.

- 1 zbudować modele LSA i LDA przy użyciu np. biblioteki gensim (<http://radimrehurek.com/gensim/tutorial.html>). Proszę pamiętać o sprowadzeniu wyrazów do formy podstawowej (1.5 pkt.)
- 2 napisać program, który dla danej notatki z korpusu PAP przedstawi N najbardziej zbliżonych tematycznie notatek i wypisze M najistotniejszych tematów tych notatek w modelach LSA i LDA (1.5 pkt.)

Materiały:

<http://home.agh.edu.pl/~wojtek/pjn2015/lab8.tar.gz>