



AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Przetwarzanie Języka Naturalnego

Lab 4

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

Wydział IEiT
Katedra Informatyki

01.04.2015

- ✚ Levenshteina (edycyjna)
- ✚ N-gramowa
- ✚ Longest Common Substring

✚ x, y - napisy

✚ $DICE(x, y) = 1 - \frac{2 \times |Ngrams(x) \cap Ngrams(y)|}{|Ngrams(x)| + |Ngrams(y)|}$
($Ngrams(x)$ – zbiór wszystkich n-gramów występujących w x)

✚ $COSINE(x, y) = 1 - \frac{|Ngrams(x) \cap Ngrams(y)|}{|Ngrams(x)| \cdot |Ngrams(y)|}$
($Ngrams(x)$ – statystyka n-gramów w postaci wektora)

Metryka LCS (Longest Common Substring)

- ✚ x, y - napisy
- ✚ $f(x, y)$ - najdłuższy wspólny podciąg napisów x i y
- ✚ $LCS(x, y) = 1 - \frac{|f(x, y)|}{\max(|x|, |y|)}$

- ✚ SOUNDEX (1918)
- ✚ Metaphone (1990)
- ✚ Double Metaphone (2000)
- ✚ są to algorytmy stratne

Miary poprawności klasyfikacji

- ✚ Precision (precyzja): jak duży procent obiektów zaklasyfikowanych do danego zbioru został poprawnie zaklasyfikowany

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

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

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

- ✚ F1: średnia harmoniczna miar precision i recall

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

- ➊ Napisać program klasteryzujący nazwy firm z pliku *lines.txt*:
 - ▶ Wykonać potrzebny preprocessing (**stworzyć stoplistę**, etc.) (1 pkt)
 - ▶ Dokonać klasteryzacji przy pomocy wybranej metryki (1 pkt)
 - ▶ Przy pomocy miar precision, recall i F1 porównać otrzymany wynik z klastrami z pliku *clusters.txt* (1 pkt)

Materiały:

<http://home.agh.edu.pl/~wojtek/pjn2015/lab4.tar.gz>