Wydział Matematyki i Nauk Informacyjnych PW

ZGŁOSZENIE TEMATU PRACY DYPLOMOWEJ MAGISTERSKIEJ
Na rok akademicki 2014/2015

Imię, nazwisko, tytuł, stopień naukowy:	dr Marek Gągolewski
Zakład, telefon, e-mail:	ZPSiMF, gagolews@mini.pw.edu.pl
Tytuł zgłaszanej pracy:	Automatyczna kategoryzacja tematyczna tekstów przy użyciu metryk w przestrzeni ciągów znaków Text clustering basing on string metrics
Kierunek:	MATEMATYKA
Imię i nazwisko dyplomanta/ki:	Natalia Potocka

Tematyka zgłaszanej pracy:

W literaturze znanych jest wiele metryk określonych w przestrzeni napisów (ciągów o dowolnej długości nad pewnym zbiorem skończonym, zwanym alfabetem), np. metryki Hamminga-Navarro, Levenshteina, Damerau-Levenshteina, Soundex, por. [3]. Kluczowym etapem w klasycznym podejściu do wykrywania skupień w danych tekstowych (np. celem automatycznej kategoryzacji tematycznej zbioru napisów) jest wyznaczanie liczby unikalnych słów w tekście (po jego odpowiedniej normalizacji). Dalej odpowiednie metody maszynowego uczenia się bez nadzoru (ang. *unsupervised machine learning*) wyznaczane są na podstawie danych typu $\{(słowo_i, liczność_i), i=1,2,...,m\}$. Łatwo zauważyć, że proces ten nie jest odporny na zniekształcenia tekstu (błędy transmisji danych, błędy ortograficzne itd.), a także własności gramatyki przetwarzanego języka naturalnego, m.in. fleksję. Można spodziewać się, że zastosowanie metryk w przestrzeni ciągów znaków w ww. metodach poprawi szeroko pojętą jakość kategoryzacji tematycznej tekstów.

Celem niniejszej pracy dyplomowej jest więc zbadanie wpływu doboru tychże metryk na różnych zbiorach *benchmarkowych*, m.in. tekstach z polskiej Wikipedii oraz plikach pomocy pakietów R. Dokonywana kategoryzacja ma mieć charakter wielopoziomowy, tzn. teksty zostaną pogrupowane na tematy, jak i ich podtematy, przy użyciu metod hierarchicznych. Co więcej, często zdarzać się będzie, że dany tekst należy do wielu kategorii, zatem w analizowanym problemie wykorzystane zostaną metody rozmytej analizy skupień, czyli takie, które mogą przydzielać jeden element (tekst) do więcej niż jednej kategorii (tematu). Dodatkowym elementem w rozpatrywanym modelu będzie wykorzystanie naturalnych powiązań między tekstami w postaci grafu odniesień (hiperlinków), w oparciu o założenie, że częściej to dokumenty pokrewne sobie tematycznie są ze sobą związane takimi odniesieniami.

Literatura pomocnicza:

[1] M. van der Loo, The stringdist package for approximate string matching, *The R Journal*, **6**, 2014, s. 111-122
[2] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady*, **10**, 1955, s. 707-710
[3] G. Navarro, A guided tour to approximate string matching, *ACM Computing surveys*, **33**, 2001, s. 31-88
[4] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, C. Watkins, Text classification using string kernels, *Journal of Machine Learning Research*, **2**, 2002, s. 419-444

.....
data i podpis