Povzetek

Pričujoče diplomsko delo obravnava v rudarjenje večjezičnih besedil. Na začetku predstavi osnove rudarjenja teksta s poudarkom na predstavitvi dokumentov, algoritem za iskanje po bazi dokumentov ter na kratko tudi algoritem za avtomatsko kategorizacijo dokumentov. Za predstavitev uporabi vrečo besed, ter preuči njene prednosti in slabosti. Iskanja dokumentov se loti z metodo najbljižjih sosedov, kategorizacije pa z metodo podpornih vektorjev.

V nadaljevanju je podrobno predstavljena kanonična korelacijska analiza (KKA). To je metoda, ki za par slučajnih vektorjev poišče smeri, vzdolž katerih sta slučajna vektorja visoko korelirana. Podana je klasična definicija KKA ter njena posplošitev na metodo jeder. Ker ima klasična KKA težave pri delu z visoko dimenzionalnimi podatki se vpelje regularizacijo. Predstavljen je tudi numerični postopek za reševanje KKA.

Za konec je predstavljena še uporaba KKA pri rudarjenju večjezičnih besedil. Prikazani so rezultati opravljenih poskusov na večjezičnih bazah in prototip iskalnika po večjezičnih dokumentih. Le-ta se nahaja tudi na priloženi zgoščenki.

Math. Subj. Class. (MSC 2000): 62H20, 62H30, 62P99, 65F15, 68U15

Ključne besede:
rudarjenje teksta, večjezično iskanje, vreča besed, kanonična korelacijska analiza, metoda podpornih vektorjev, metode jeder

Keywords:
text mining, cross-lingual information retrieval, bag of words, canonical correlation analysis, support vector machine, kernel methods

Literatura

[1] John Shawe-Taylor, Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004

[2] William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery. *Numerical Recipes in C – The Art of Scientific Computing, Second Edition*. Cambridge University Press, 1997

[3] Bai, J. Demmel, J. Dongarra, A. Ruhe and H. van der Vorst, editors. *Templates for the solution of Algebraic Eigenvalue Problems: A Practical Guide*. SIAM, Philadelphia, 2000