Posudek vedoucího diplomové práce

Michal Hrala: Automatická klasifikace dokumentů s podobným obsahem

Diplomová práce vznikla na základě požadavku České tiskové kanceláře (ČTK). Cílem práce je návrh a realizace systému pro automatickou klasifikaci textových dokumentů dle tématu. Výsledky práce budou dále použity v rámci výzkumné činnosti katedry.

Práce autora začíná popisem současného stavu poznání v oblasti řešené problematiky. Je zde popsána klasifikační úloha spolu s obvyklými použitými metodami předzpracování dokumentů (tokenizace, filtrace, lemmatizace a určení slovních druhů). Dále jsou podrobně popsány použité metody pro výběr příznaků (Dokumentová Frekvence (DF), Information Gain (IG), χ^2 test, Mutual Information (MI) a GSS koeficient) a použité klasifikační metody (Naivní Bayesův (NB) klasifikátor, Support Vector Machines (SVM) a metoda Maximální Entropie (ME)). Volba metod je dobře zdůvodněna zejména odkazy na prostudovanou literaturu. V další kapitole autor popisuje dostupné nástroje pro klasifikaci textů. Na závěr kapitoly diplomant zvolil systém MinorThird. Volba nástroje je dobře zdůvodněna v návaznosti na výslednou aplikaci. Práce pokračuje popisem struktury databází ČTK a použitých evaluačních metrik.

Dále se diplomant zabývá vlastním řešením. Popisuje návrh aplikace a samotnou implementaci. Systém je vhodně navržen – modulární architektura je výhodná z mnoha důvodů: možnost nahrazení/vypuštění modulu, samostatné spouštění jednotlivých modulů, apod. Defaultní spouštění systému je z příkazové řádky s předem nastavenými parametry, nicméně autor implementoval i grafické uživatelské rozhraní. Výstupy systému jsou ve dvou různých formátech: 1) textový soubor obsahující úspěšnost klasifikace; 2) seznam rozpoznaných kategorií ve formátu *xml*. Autor provedl celou řadu experimentů (viz kapitola 5), které na sebe navazují a jsou dobře rozmyšleny. Experimenty jsou navrženy s cílem nalezení nejlepší množiny parametrů a optimálního nastavení klasifikátoru. Nejlepší výsledky (9,44 % *Hamming loss*) dosahuje autor při použití klasifikátoru SVM a příznaků vybraných pomocí metody MI.

Průvodní dokument (47 stran + přílohy) je vytvořen v systému LaTeX. Má logickou přehlednou strukturu, názvy kapitol jsou vhodně voleny. Dokument je na kvalitní jazykové úrovni, neobsahuje nepřesnosti, pravopisné chyby ani překlepy. Kladně hodnotím vytvořený přehled termínů a zkratek, který čtenáři usnadní orientaci v práci samotné.

Předložená diplomová práce splňuje zadání v plném rozsahu. ČTK je spokojena s kvalitou práce. Bylo dohodnuto nasazení a rozšíření této práce v rámci další spolupráce, která bude formou doktorského studia. Výsledky diplomové práce budou publikovány na mezinárodní konferenci SOFSEM 2013. Autor zde prokázal kvalitní znalosti nejen z informatiky, ale i z matematiky a statistiky. Přesvědčivě prokázal, že dokáže samostatně efektivně analyzovat a řešit zadané problémy. Práci doporučuji k obhajobě a hodnotím klasifikačním stupněm

„výborně“

Ing. Pavel Král, Ph.D.
vedoucí DP

V Plzni 5. června 2012