Posudek vedoucího diplomové práce

Ladislav Hlom: Automatická klasifikace vícejazyčných dokumentů

Cílem práce je návrh a implementace systému pro automatickou klasifikaci vícejazyčných textových dokumentů. Výsledky práce budou využity v rámci výzkumné činnosti katedry (NTIS-P2). O výsledky má dále zájem Česká tisková kancelář (ČTK) s cílem automatické analýzy zahraničních zpravodajství.

Práce autora začíná popisem současného stavu poznání v oblasti řešené problematiky. Zde je popsána úloha strojového překladu, metoda LDA (Latentní Dirichletova alokace), které jsou pro vyřešení problému nezbytné, dále pak úloha klasifikace dokumentů. V této části jsou dále popsány použité nástroje a evaluační metriky pro vyhodnocení výsledků práce. Následuje popis použitých datových kolekcí.

Dále se diplomant zabývá vlastním řešením, kde je popsáno (názorně i na několika schématech), jak autor postupoval. Další kapitola se zabývá provedenými experimenty, kterých byla realizována velká řada. Autor zde srovnává dva přístupy k vícejazyčné klasifikaci: 1) se strojovým překladem (systém Mooses a Google Translate); 2) s metodou reprezentace dokumentů bez učitele pomocí LDA. Autor použil a srovnal dvě klasifikační metody: metodu maximální entropie a metodu podpůrných vektorů. Experimenty jsou provedeny na celkem čtyřech datových kolekcích. Výsledná F-míra nejlepší metody (strojový překlad + ME klasifikace) kolem 70 % je velmi dobrá.

Průvodní dokument (52 stran + přílohy) je vytvořen v systému LaTeX. Má logickou přehlednou strukturu, názvy kapitol jsou vhodně voleny. Dokument je na dobré jazykové úrovni, neobsahuje pravopisné chyby ani překlepy. Práce obsahuje několik drobných nepřesností, které odpovídají znalostem studenta magisterského studia.

Přiložené DVD má logickou strukturu. Jednotlivé adresáře obsahují readme soubory s potřebnými informacemi. DVD také obsahuje spustitelný *.jar* soubor pro provedení klasifikace. Jediné, co bych ještě uvítal, je soubor build.xml pro vytvoření *.jar* souboru pomocí nástroje ant.

Předložená diplomová práce splňuje zadání v plném rozsahu a má velký výzkumný potenciál. Dle dostupných informací neexistuje žádná podobná studie pro český a anglický jazyk. Předpokládám proto, že výsledky práce budou po rozšíření publikovány na mezinárodní konferenci z oblasti zpracování přirozeného jazyka. Je třeba dále zdůraznit, že téma práce je velmi složité a vyžadovalo nastudování řady informací z oblasti umělé inteligence. Autor zde prokázal dobré znalosti nejen z informatiky, ale i z matematiky a statistiky. Přesvědčivě ukázal, že dokáže samostatně analyzovat a řešit složité problémy. Práci doporučuji k obhajobě a hodnotím klasifikačním stupněm

„výborně“

doc. Ing. Pavel Král, Ph.D.
vedoucí DP

V Plzni 31. května 2016

SOUHLASÍ
S ORIGINÁLEM

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky