Strukturovaný posudek bakalářské práce

Veronika Černá

Automatická klasifikace textových dokumentů

1. Informace k zadání

Zadání této práce vzniklo pro potřeby České Tiskové Kanceláře (ČTK). Cílem práce je návrh a realizace systému pro automatickou klasifikaci textových dokumentů. Vytvořená aplikace by měla být použita ke klasifikaci dokumentů z produkce ČTK. Výsledky práce budou dále použity v rámci výzkumné činnosti katedry.

2. Aktivita během řešení, konzultace, komunikace – 15 bodů (max. 15 bodů)

Studentka byla aktivní během celého řešení. Velkou část teoretických znalostí nastudovala samostatně s použitím dodané i vyhledané literatury, která byla zejména v anglickém jazyce. Také se samostatně seznámila s nástrojem MinoTrird, který je použit pro implementaci klasifikačních metod.

3. Splnění požadavků zadání – 25 bodů (max. 25 bodů)

Požadavky zadavatele byly splněny ve všech bodech.

4. Hodnocení formální stránky předložené práce – 24 bodů (max. 25 bodů)

První část práce se zabývá teoretickými poznatky z oblasti řešené problematiky. Je zde popsána klasifikační úloha obecně, dále je podrobněji rozebrána úloha klasifikace textových dokumentů. Následuje popis samotných klasifikačních algoritmů. Druhá část popisuje vlastní řešení. Je zde popsána a zdůvodněna volba klasifikačního nástroje. Dále je popsán korpus textových dokumentů spolu s provedenými úpravami a předzpracováním. Následuje návrh a popis provedených experimentů. Práce je přehledně členěna, neobsahuje překlepy ani pravopisné chyby, jen několik drobných nepřesností, které jsou ale plně v souladu s úrovní znalostí studentky v dané oblasti. Dokument je vytvořen pomocí systému LaTeX a je kvalitní.

5. Hodnocení realizačního výstupu – 30 bodů (max. 35 bodů)

V rámci práce se podařilo navrhnout a vytvořit systém pro automatickou klasifikaci textových dokumentů. Jako předzpracování dokumentů je provedena tokenizace, lemmatizace a POS-tagging. Dle slovních druhů je provedena filtrace slov použitých ke klasifikaci. Jako parametrizační metoda byla použita dokumentová frekvence (angl. document frequency). Studentka zvolila na základě prostudované literatury tři klasifikační algoritmy, které jsou použity v této oblasti: Naivní Bayesův klasifikátor, Support Vector Machines a model Maximální Entropie. Funkčnost systému byla ověřena na korpusu dodaném ČTK. Dosažená přesnost klasifikace, 87%, je velmi dobrá. Za drobné nedostatky práce považuji poměrně malý počet provedených experimentů a dále pak, že bylo vytvářeno „několik" trénovacích /testovacích sad na místě, kde bych očekával pouze jednu sadu dokumentů. Vstupy klasifikátoru by bylo lepší řešit přímo pomocí parametrizačního modulu. Na závěr bych chtěl uvést, že vzhledem ke kvalitě, práci i přes uvedené drobné nedostatky doporučuji do soutěže CCA.

6. Otázky k obhajobě

V této práci klasifikujete dokumenty do jedné kategorie. Jak byste postupovala, pokud byste uvažovala, že jeden dokument může patřit do více kategorií?

7. Závěrečné shrnutí – celkem dosaženo 94 bodů (max. 100 bodů)

Celkem bylo dosaženo 94bodů (max. 100). Bakalářskou práci doporučuji k obhajobě.

Král

V Plzni dne 31.5.2012

Ing. Pavel Král, Ph.D.
KIV – FAV - ZČU