Posudek oponenta bakalářské práce

Student: Cakl Jan
Téma: Automatická detekce jazyka textového dokumentu (id 18569)
Oponent: Pešán Jan, Ing., UPGM FIT VUT

1. Náročnost zadání **obtížnější zadání**

Práce je obtížnějsího charakteru, vyžaduje implementaci netriviálních algoritmů, včetně získání datové sady a vlastního návrhu testovacích dat.

2. Splnění požadavků zadání **zadání splněno**

Zadání bylo splněno se zajímavými rozšířeními (např. rozpoznání češtiny a slovenštiny s diakritikou a bez ní)

3. Rozsah technické zprávy **je v obvyklém rozmezí**

Rozsah práce je adekvátní problému. V přílohách se nachází dostatečné množství podpůrných dat a výsledků k experimentům.

4. Prezentační úroveň předložené práce **95 b. (A)**

Práce je logicky strukturována, s jasnými návaznostmi jednotlivých kapitol. Téma práce je prezentováno srozumitelným způsobem. Nejprve jsou prezentovány obecné základy klasifikace a rozpoznávání jazyka a poté již diskutována samotná práce vypracovaná studentem.

5. Formální úprava technické zprávy **89 b. (B)**

Práce je psána v češtině, bez pravopisných chyb. Občas je použit poněkud krkolomnější konstrukt, ale předpokládám že je výsledkem překládání anglických výrazů, které nemají definovaný český ekvivalent.

6. Práce s literaturou **95 b. (A)**

Práce s literaturou je bez problémů, včetně odpovídajících citací.

7. Realizační výstup **100 b. (A)**

Programové řešení je na bakalářskou práci vyjímečně kvalitní, s rozsáhlou vlastní invencí autora. Programové řešení vyžadovalo získání vlastních trénovacích dat, segmentaci na datové sady a implementaci netriviálních algoritmů.

8. Využitelnost výsledků

Práce je zčásti výzkumného charakteru, přináší užitečné srovnání několika metod pro rozpoznávání jazyka ze strojově čitelného textu. Pokud by byla dále rozšířena, nabízela by se možnost pro nasazení v reálném produktu.

9. Otázky k obhajobě

- Při získávání datových sad jste se zaměřil na paralelní korpusy, má to nějaký vliv na trénování systému? A pokud ano, jaký?
- U klasifikace češtiny a slovenštiny s diakritikou a bez (tabulka 5.3) se objevuje častá záměna Cestina za Slovenstina/Neznamy, ale tento výsledek není symetrický (Slovenstina má 98% přesnost). Proč tomu tak je?
- Při návrhu datových sad (tabulka 3.2) není stejné množství slov ve všech jazycích. Jak to může ovlivnit výkon klasifikátoru?

10. Souhrnné hodnocení **95 b. výborně (A)**

Práce byla výborně zpracována po formální, jazykové, ale především obsahové stránce. Programový výstup a vyhodnocení výkonu klasifikátoru jsou na vynikající úrovni, přesahující rozsah obyklý pro BP. Z výše uvedených důvodů navrhuji studentovi udělit hodnocení výborný (A).

Prohlášení: Uděluji VUT v Brně souhlas ke zveřejnění tohoto posudku v listinné i elektronické formě.

V Brně dne: 1. června 2016

.....
podpis