Posudek bakalářské práce

předložené na Matematicko-fyzikální fakultě
Univerzity Karlovy v Praze

posudek vedoucího X posudek oponenta

Autor/ka: Josef Čech

Název práce: Oborová klasifikace textu

Studijní program a obor: Informatika, programování

Rok odevzdání: 2010

Jméno a tituly vedoucího/oponenta: RNDr. Miroslav Spousta

Pracoviště: UFAL MFF UK

	excelentní	odpovídající	slabší	nevyhovující
Náročnost zadaného tématu		X		
Míra splnění zadání			X	
Rozsah práce		X		
Struktura textové části práce		X		
Analýza			X	
Vývojová dokumentace				X
Uživatelská dokumentace			X	
Jazyková a typografická úroveň			X	
Návrh a design implementace		X		
Kvalita zpracování softwarové části			X	
Stabilita aplikace				X

Nejvýznamnější klady:

- zajímavé téma s možností vyzkoušení mnoha algoritmů a jejich vylepšení
- autor vyvinul pro účely práce vlastní lemmatizaci pro češtinu
- multiplatformní aplikace

Nejzávažnější nedostatky:

Autor při řešení práce postupoval svérázně, bohužel se soustředil na část (lemmatizace), která není pro výsledné řešení příliš podstatná: pokud se během práce ukázalo, že neexistuje funkční morfologická analýza češtiny pro MS Windows, bylo jistě možné pro potřeby analýzy textu použít linuxovou verzi a teprve analyzované soubory používat v aplikaci na MS Windows. Stejně tak není potřeba implementovat různé klasifikační algoritmy, ale stačilo by využít dostupné knihovny.

Očekával jsem, že těžiště práce bude v porovnání výsledků různých metod, případně alespoň v porovnání nastavení parametrů jedné vybrané metody. Autor popisuje několik možných metod a implementuje pozměněnou verzi (využívající morfologické tagy a lemmatizaci) jedné z nich – vektorového modelu. Bohužel chybí srovnání se standardní implementací, takže není možné posoudit, zda je autorův postup lepší, či horší.

Autor používá termín F-measure, ale upravil jeho výpočet tak, že neodpovídá původní definici a významu; navíc hodnoty, které používá pro výpočet, jsou špatně spočítané (Tabulka č. 4).

Algoritmus samotný je špatně popsaný, autor zmiňuje, že při implementaci nastavoval několik různých parametrů, ale není uvedeno, které všechny parametry nastavoval, jakým způsobem jednotlivé parametry ovlivňují výsledek a jakým způsobem se snažil toto nastavení optimalizovat (na str. 28 např. zmiňuje o nastavení mezí: „Jediná možnost, jak můžeme tyto hranice pro každý obor nastavit, je metodou pokusů a omylů").

Implementace je stručně popsaná, ale bohužel samotné zdrojové soubory jsou komentované málo, případně nepříliš srozumitelně. Uživatelská příručka je také velmi stručná, není dobře zdokumentován postup, jak program spustit např. pro trénování modelu. Používání programu je nestandardní, např. jako parametr je možné zadat absolutní či relativní cestu, ale relativní cesta se nevztahuje k aktuálnímu adresáři, ale k adresáři se slovníkem. V grafické aplikaci se pomocí jedné položky v menu vybere adresář se slovníkem a jiná položka je potřeba pro samotné načtení slovníku.

Aplikace na autorem vyžadované konfiguraci (Ubuntu 9.10) nefungovala (pravděpodobně z důvodu jiné verze knihoven), na jiné instalaci Linuxu ano. Pokud použité knihovny nejsou v podobných verzích kompatibilní, buď je potřeba přesně specifikovat jejich verze, nebo program linkovat staticky. Překlad ze zdrojového kódu se nepovedl, musel jsem ručně upravit zdrojové soubory.

Autor měřil rychlost programu, ale neuvedl v jakém prostředí (OS, rychlost CPU, velikost paměti), takže tato informace pozbývá smyslu.

Práce obsahuje mnoho stylistických nedostatků a nejasných formulací, např.: str. 6: „V současné době neexistuje systém, který by byl schopen se 100% úspěšností analyzovat text jakýmkoliv způsobem.", str. 11: „jako Čitatel", str. 22: spojená slova, str. 23: „úhel, který svírajíc", str. 36: „v před připravené". Autor v programu také míchá angličtinu a češtinu způsobem až matoucím (seznam souborů, folder, file, ...)

Další poznámky:

Vzhledem k uvedeným nedostatkům jak v textu práce, tak v programu samotném, nepovažuji předloženou práci za obhajitelnou.

	výborně	velmi dobře	dobře	neprospěl/a
Návrh známky				X

Datum: 17.6.2010

Podpis:

Poučení k formuláři pro hodnocení informatických bakalářských prací

Tento formulář je určen pro hodnocení vedoucího i oponenta bakalářské práce, která má formu softwarového projektu. Bakalářské projekty jiných typů (teoretická práce, srovnávací studie apod.) budou hodnoceny pomocí standardních textových posudků.

Jednotlivá políčka vyplňte nejlépe elektronicky (lze případně i ručně), je možné zaškrtnout i dvě sousední políčka (např. pro hodnocení typu 'něco mezi odpovídající a slabší'), a to i u návrhu výsledné známky. Pokud některá položka nemá vzhledem k práci smysl (např. stabilita aplikace u práce bez vlastní implementace), položku nevyplňujte. Výsledná navrhovaná známka nemusí být žádným 'průměrem' hodnocení jednotlivých kritérií. Pokud některé položky hodnotíte jako slabší nebo nevyhovující, v sekci Nejzávažnější nedostatky popište důvody vašeho hodnocení a zjištěné nedostatky.

Výklad stupňů hodnocení:

- excelentní znatelně lepší/rozsáhlejší/dokonalejší než je pro Bc práci požadováno
- odpovídající přiměřené Bc práci, student splnil to, co měl
- slabší výhrady ke kvalitě, rozsahu, hloubce nebo zpracování
- nevyhovující neodpovídá požadavkům na Bc práci, práce nemá být obhájena

Vyplněné a ručně podepsané (i v případě elektronického vyplňování) hodnocení odevzdejte na sekretariát KSI, elektronickou verzi pošlete na sekretariat@ksi.ms.mff.cuni.cz. Pokud máte emailový kontakt na autora práce, pošlete posudek i jemu.