Posudek bakalářské práce

předložené na Matematicko-fyzikální fakultě
Univerzity Karlovy v Praze

posudek oponenta

Autor/ka:	Martin Majliš
Název práce:	Sumarizace textu
Studijní program a obor:	Informatika, programování
Rok odevzdání:	2008

Jméno a tituly oponenta:	Mgr. Pavel Schlesinger
Pracoviště:	ÚFAL 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:
Rozsah nastudované problematiky i samotné práce.

Nejzávažnější nedostatky:
Kvalita celé práce je výrazně snížena jazykovou úrovní předloženého textu. Pominu-li gramatické chyby (např. záměna i/y, chybná délka samohlásek, chybějící čárky při uvození vedlejších vět), pak se v textu ve velkém (a pro čtenáře nepříjemném) množství vyskytují zásadní chyby (absence slova ve větné konstrukci, nelogičnost ve formulaci a spojení vět do souvětí, problém s logickou návazností při rozvinutí myšlenky do více vět daného odstavce), které by vůbec v bakalářské práci být neměly a které mnohdy znesnadňují pochopení samého obsahu textu. Vypadá to, jakoby sám autor ani na závěr psaní textu své vlastní dílo nečetl znovu celé. Přitom by asi stačilo dát před definitivním odevzdáním text přečíst další osobě a kvalita práce by šla výrazně nahoru a díky svému obsahu a délce by byla nad průměrem. Pro výše zmíněný nedostatek jsem nucen navrhnout známku NEPROSPĚL. Zároveň navrhuji zvážit řádnou obhajobu s prezentací, kde autor dostane možnost obhájit části textu (ne všechny mohu níže uvést), které jsou chybami znehodnoceny.

Příklady chyb (po jisté době jsem upustil od opravování celého textu, zde uvádím jen některé z nalezených chyb):
str. 14 … Na webové stránce je nutné nejdřív nutné identifikovat část, která obsahuje text článku a z této části ještě navíc odstranit formátovací značky, reklamy, odkazy na další články atd. Data, pro tuto práci byla získávána z webových stránek, a proto je bylo nutné před použitím pročistit.
str. 16 … V této fázi jsou z původního dokumentu vybráno požadované množství vět na základě celkového skóre, kterého je součtem dílčích skóre.
… Výsledný extrakt je nutné prezentovat uživateli v odpovídající podobě. Pokud uživatel vložit vstupní text prostřednictvím webového formuláře, pak jako webovou stránku.
str. 17 … Ze vzorce je zřejmá nevýhoda této míry. Pokud dokument obsahoval věty 1-10, a referenční extrakt obsahoval věty 1, 2 a 5. Tak pokud systém označil za důležitou pouze větu 1, tak dosáhl skóre 1, ale výsledek není nikterak kvalitní, proto je dobré nutné tuto míru kombinovat s metodou recall nebo f-measure.
… Recall (česky: úplnost) je míra, jak velký podíl z vybraných vět odpovídá referenčnímu textu.
str. 19 … Systém by měl do začátku implementovat nejzákladnější komponenty pro jednotlivé kroku sumarizace.
…
str. 62 … Zatímco v odkazovaných článcích byly vybíráno n nejdůležitejších vět, tak...
str. 64 … Není ale jasné, zda-li by měly i srovnatelný přísnost pro uživatele.
… Tvorbu konfigurací pro sumarizaci by usnadnila podpora XML entit možnost používat v názvech souborů regulární výrazy.

Další poznámky:

Kapitola 2.2.5 "Tvorba extraktu", která je dle mého názoru zásadní pro sumarizaci samotnou, je příliš stručná a navíc vůbec neobsahuje kritérium výběru VĚT ani podmínku, kdy je výběr ukončen. Navíc je zde exlipicitně zmíněn pouze výběr vět, přesto se na str. 62 lze dočíst i o výběru na základě jednotlivých znaků.

Při definování důležitých pojmů na str. 16 a dál se pracuje s neznámým pojmem "term".

Obr. 2.1, str 13, vysvětluje proces sumarizace. Skládá se tento proces opravdu z několika na sobě NEZÁVISLÝCH kroků?

	výborně	velmi dobře	dobře	neprospěl/a
Návrh známky				X

Datum: 29. 8. 2008

Podpis: