Posudek bakalářské práce

předložené na Matematicko-fyzikální fakultě
Univerzity Karlovy v Praze

❑ posudek oponenta

Autor/ka: Evelina Gabašová

Název práce: Text clustering and classification (Klastrování a klasifikace textů)

Studijní program a obor: Informatika, obecná informatika

Rok odevzdání: 2007

Jméno a tituly vedoucího/oponenta: RNDr. Jan Hric

Pracoviště: KTIML MFF UK

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

Nejvýznamnější klady:
Text: dobrý popis (prevzatých) techník, popis práce so systémom ako running example v prílohe B
Primeraný výber relevantných metód
Integrácia C++, Csharp a cudzích knižníc (Fortran77)
Primerané možnosti nastavovania parametrov (prahov) jednotlivých fázi spracovania, návrh programu umožňuje vďaka integrovanému prostrediu expertovi zasahovať medzi fázami.

Nejzávažnější nedostatky:
Pomerne pomalé a teda neškálovateľné (Klustrovanie na 100 dokumentoch rádovo minúty, 15', pravdepodobne použitá externá knižnica), dlhé výpočty sa (mne) nedarí prerušiť.
Chýbajú použité experimentálne datá na CD (a ich konkrétne rozdelenie na trénovacie a testovacie)
Malo byť umožnené a popísané spúšťanie dávok pre jednotlivé fázy „z príkazovej riadky" s nastavením parametrov (vlastnú prácu vykonávajú 3 dávkové filtry).
Text: niektoré časti sú dlhé, nie sú zdoraznené doležité časti a nie vždy je jasné, čo bolo nakoniec implementované.

Další poznámky:
Z hladiska užívateľa uzavretý systém – nejdú pridať metódy. Ale výber metód je primeraný.
Zlepšiť podporu konceptov a ich zobrazovania expertovi - na to bola práca o.i. zameraná
Dialog klasifikácie po chybnom zadaní mena suboru oznámi chybu a zavrie sa so stratou všetkých (predtým nastavených) dát. (a iné drobnosti mi nevyhovovali v user interface, resp. musel by som si zvyknúť)
Použitý prístup, že sa klastruje/klasifikuje vždy celý adresár je vhodný pre dávkové použitie, menej vhodný pre experimentovanie (napr. hľadanie vhodných parametrov) na čiastočných kolekciách dokumentov. Mať možnosť použiť zoznam súborov.
Vhodný by bol záznam ručných zmien experta a Undo.
Formát dát v niektorých súboroch (napr. ClassFrequencies.xml) mi pripadal neprirodzený a nevhodný na prípadné ďalšie spracovanie. Oddelil by som slovo a jeho frekvenciu (atribút alebo samostatný tag).
Pretože sa pracuje s mnohými druhmi informácií (v XML a TXT), mala sa použiť pre ich rozlíšenie druhá prípona.
Keď už je použité užívateľské prostredie, je možné z výsledných dát poskytnúť užívateľovi viac informácií - napr. podobnosť dokumentu k iným, rozsah hodnot al. histogram pre lepšiu diskretizáciu ...
Priebeh hierarch. klastrovania končí podľa počtu dokumentov. Mohlo končiť (aj) podľa iných mier (uvedených v práci).

	výborně	velmi dobře	dobře	neprospěl/a
Návrh známky	XX	x		

Datum: 6,9,2007

Podpis: