Posudek oponenta závěrečné práce

České vysoké učení technické v Praze **Fakulta informačních technologií**

Student: Bc. Štěpán Škorpil
Oponent práce: Ing. Jan Hořínek
Název práce: Aplikace pro analýzu textových zpráv pro potřeby vyšetřovatelů
Obor: Webové a softwarové inženýrství (magisterský)

Datum vytvoření: 31. 5. 2015

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 5:*
1. Náročnost a další komentář k zadání	*1=mimořádně náročné zadání,* ***2=náročnější zadání,*** *3=průměrně náročné zadání,* *4=lehčí, ale ještě dostatečně náročné zadání,* *5=nedostatečně náročné zadání*

Popis kritéria:
Podrobněji charakterizujte diplomovou (bakalářskou) práci a její případné návaznosti na předchozí nebo běžící projekty. Dále posuďte, čím je zadání této ZP náročné. (U obtížnější ZP lze dále tolerovat některé nedostatky, které by u ZP standardní obtížnosti tolerovány nebyly; a naopak u jednoduché ZP mohou být zjištěné nedostatky hodnoceny přísněji.)

Komentář:
Zadání bylo obecně pojaté. Student si mohl po konzultaci se zadavatelem zvolit téma, které ho nejvíc z problematiky NLP pro češtinu zaujalo. Vybral si fulltext search, clustering a similarity search. Další témata jako NER, semantic analysis apod. by byly již mimo rozsah DP.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
2. Splnění zadání	***1=zadání splněno,*** *2=zadání splněno s menšími výhradami,* *3=zadání splněno s většími výhradami,* *4=zadání nesplněno*

Popis kritéria:
Posuďte, zda předložená ZP splňuje zadání. V komentáři uveďte body zadání, které nebyly zcela splněny, případně rozšíření ZP oproti původnímu zadání. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Komentář:
Zadání bylo splněno dle požadavku zadavatele.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
3. Rozsah písemné zprávy	***1=splňuje požadavky,*** *2=splňuje požadavky s menšími výhradami,* *3=splňuje požadavky s většími výhradami,* *4=nesplňuje požadavky*

Popis kritéria:
Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části.

Komentář:
Písemná zpráva splňuje Směrnice děkana c. 9/2011, článek 3.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Věcná a logická úroveň práce	*80 (B)*

Popis kritéria:
Posuďte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti. Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře.

Komentář:
Práce obsahovala úvod, analýzu, implementaci, testování a závěr.
Ve všech částech student postupoval věcně a logicky správně.
Student správně pojmenoval a realizoval základní prvky pro fulltext:
- tokenizace
- převod na lower case
- stop words
- stemming
- indexace (insert, update, delete)
- hledání

Pouze v části 1.1.1 bod 8) student chápe pojem "jmenné entity" odlišně od standardního pojetí. Správný název je "pojmenované entity" a jde o rozdělení objektů v textu do kategorií, např. osoby, lokality, firmy apod. V problematice extrakce entit nejde tedy o určování větných členů a slovních druhů, čímž se zabývá syntaktická a morfologická analýza.
Dále mi scházel kompletní popis všech podstatných parametrů, jimiž se konfiguruje Crawler a vlastní SOLR.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
5. Formální úroveň práce	*90 (A)*

Popis kritéria:
Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 12/2014, článek 3.

Komentář:
Typografická a jazyková stránka splňuje Směrnici děkana č. 9/2011, článek 3.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Práce se zdroji	*80 (B)*

Popis kritéria:
Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení ZP. Charakterizujte výběr studijních pramenů. Posuďte, zda student využil všechny relevantní zdroje nebo zda se pokoušel řešit již vyřešené problémy. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Komentář:
Student užil standardní množství relevantních teoretických zdrojů o fulltextových technologiích.
Zdroje byly v českém i anglickém jazyce.
Nedošlo k porušení citační etiky, student postupoval v souladu se zvyklostmi a normami.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
7. Hodnocení výsledků, publikační výstupy a ocenění	*90 (A)*

Popis kritéria:
Vyjádřete se k úrovni dosažených hlavních výsledků ZP, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, apod. Případně také zhodnoťte, zda software nebo zdrojové texty, které nevytvořil sám student, byly v ZP použity v souladu s licenčními podmínkami a autorským právem. Popište případnou publikační činnost a získaná ocenění související s řešením této ZP.

Komentář:
Student již od začátku uvažoval o možnosti rozšíření systému o další moduly, čemuž přizpůsobil své řešení.
Zvolil podle mě efektivní open source technologii SOLR, jíž doplnil o vlastní prvky:
- crawler pro zpracování vstupních dat (plain text i další druhy souborů)
- GUI dle specifikace zadavatele (logické, čisté, robustní, multiplatformní)
- čestina (různé druhy stemmerů a jejich kaskádní řazení)
- clustering pomocí externích technologií (Carrot2)

Výsledný produkt funguje dle požadavků zadavatele.

Hodnotící kritérium:	*Způsob hodnocení - nehodnotí se*
8. Komentář o využitelnosti výsledků	

Popis kritéria:
Uveďte, zda hlavní výsledky ZP rozšiřují již publikované známé výsledky a/nebo přinášející zcela nové poznatky. Uveďte možnosti využití výsledků ZP v praxi.

Komentář:
Hlavní výsledky ZP rozšiřují známé výsledky. Není mi známo, že by jiná DP obsahovala problematiku fulltextu, clustering a podobnostního hledání textů pro český jazyk ve stejné šíři a v tak komplexním pojetí včetně reálného SW řešení.
V praxi lze výsledky plně použít. Pro uživatele by se hodilo doplnit systém o možnost hledání ve více kolekcích najednou, dále i uvážit práva jednotlivých uživatelů na dokumenty v indexu či práva na kolekce.
Studentem zmiňované téma extrakce pojmenovaných entit a vazeb mezi nimi by rovněž zvýšilo přínos uvedeného SW pro praxi.

Hodnotící kritérium:	*Způsob hodnocení - nehodnotí se*
9. Otázky k obhajobě	

Popis kritéria:
Uveďte případné dotazy, které by měl student zodpovědět při obhajobě ZP před komisí (body oddělte odrážkami).

Otázky:
Vybral jste si pro fulltextové řešení open source technologii SOLR. Diskutujte, proč právě ji? V čem se liší od dalších, např. Lucene, Elastic Search, Sphinx, fulltext v rámci db PosgreSQL?
Jak byste postupoval při realizaci dalších funkcionalit viz návrhy v odstavci 8. Komentář o využitelnosti výsledků?

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
10. Celkové hodnocení	*90 (A)*

Popis kritéria:
Shrňte stránky ZP studenta, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení **nemusí** být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích 1 až 9.

Text hodnocení:
Student se k zadání postavil systematicky a vyřešil ho precizně. Postupoval dle standardů pro vývoj SW: analýza zadání, návrh řešení s implementací, testování na množině různorodých uživatelů a předání zadavateli.
Podle mě zadání splnil a dodal plně funkční řešení.

Podpis oponenta práce: