Posudek oponenta závěrečné práce

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

Student: Bc. Jiří Šmolík
Oponent práce: Ing. Petr Špaček, Ph.D.
Název práce: Crawler zaměřený na sběr Web API dokumentace
Obor: Webové a softwarové inženýrství (magisterský)

Datum vytvoření: 30. 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:
Cílem práce bylo vytvořit webový crawler pro sběr Web API dokumetací na webu, který má eliminovat problém stárnutí (ne-aktuálnosti) veřejných registrů Web API (jako je např. v práci často zmiňovaný web ProgrammableWeb.com). Obtížnost spadá do kategorie středně obtížných zadání. Autor se nejříve musel seznámit s metodami crawlingu a s metodami strojového rozpoznávání (učení) dokumentů. Dále musel autor nastudovat způsob práce s open-source crawlerem Apache Nutch, který byl doporučen v zadání. Po získání teoretických základů autor hledal způsob jak nejlépe najít pro crawler množinu výchozích URL. Zde autor zvolil nástroj Google CSE, který po patřičné konfiguraci vracel adresy potenciálních míst pro crawling. Dále autor nástroj Apache Nuth propojil s databází Apach Sorl (k uložení otisků crawlovaných dokumentů), kterou následně používa v procesu klasifikace. Pro klasifikaci autor použil algoritmy strojového učení z knihovny Weka.

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 ve stanoveném rozsahu.

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:
Počet stran práce bez příloh činí 75, což je v souladu s pravidly. Rozložení objemu stran odpovídá povaze zadání, tj. pomineme-li výpisy konfigurační kódů v kapitole "Realizace", je nejrozsáhlejší částí kapitola "Analýza".

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

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 je po faktické stránce i logické stránce dobře vystavěna. V určitých pasážích je až zbytečně detailní, např. popis výhod jazyka Java, str. 35,36. Naopak v jiných pasážích by zasloužila lepší argumentaci pro zvolené nástroje a technologie, viz.: "volím Weku, protože s ním mám již nějaké zkušenosti" (str. 46 dole) Jako velmi zdařilou hodnotím část "Analýza", které je dobrým startem pro všechny zájemce o oblast crawlingu a strojového učení.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
5. Formální úroveň práce	*75 (C)*

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:
Práce je napsána čtivě, avšak dobrá čtivost je místy přerušena syntaktickými, gramatickými či stylistickými chybami jako např.:
"vhledavací" odst. 1. str. 4.; "výhodou je usnadnění vývojáři nalezení API" 5. odst. str 4.; "narozdíl od webových služeb, definovaných konzorciem W3C, aby k nim mohlo být přistupováno" odst 8. str. 10.; "kbybys nikdo nevěděl" odst. 1. str. 13.; "stažené dokumenty byla poté" poslední odst. str. 17, apod.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Práce se zdroji	*90 (A)*

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:
Krom webových odkazů na dokumentace použitých komponent, autor v rámci rešerše správně cituje i odbornou literaturu.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
7. Hodnocení výsledků, publikační výstupy a ocenění	*80 (B)*

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:
Práce pozitivně přispívá k problému vyhledávání Webových API, způsobem, který lze považovat za inovativní, jelikož překonává limity v dostupnosti sesbíraných dokumentů u svého nejbližšího konkurenta, viz literatura [16]. Kód komponent použitých při realizaci cíle této práce je užit v souladu s jejich licencí.

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:
Výstup práce hodnotím jako reálně použitelný, jelikož dokáže správně rozpoznat cca 3/4 zpracovaných dokumentací, což je i podle výstupů z kapitoly "Vyhodnocení" porovnatelné s výsledky u konkurenčních řešení.

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:
Jaké výhody/nevýhody má software Weka, použitý pro strojové učení, oproti alternativám jako RapidMiner (rovněž zmiňovaný ve vaší práci)?
Proč jste vybral jako jádro vaší implementace open-source crawler Apache Nutch namísto open-source crawleru Heritrix, který je rovněž Java-based a má údajně lepší dokumentaci [http://blog.blikk.co/comparison-of-open-source-web-crawlers/]?

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
10. Celkové hodnocení	*80 (B)*

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:
Ačkoli je práce po své implementační stránce menšího rozsahu (počet zdrojových souborů: 28, počet řádků celkem: 1731), lze tento faktor pominout z důvodu kvality rešeršní, analytické a návrhové práce. Autor nastudoval všechny potřebné informace precizně a své znalosti poté uplatnil v implementační části. Výsledek práce je tedy kvalitní a reálně použitelnou kompozicí poměrně složitých komponent (Apache Nutch, Apach Soir, Google CSE, Weka.)

Podpis oponenta práce: