Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Petr Fejfar		
Název práce	Interactive web crawling and data extraction		
Rok odevzdání	2018		
Studijní program	Informatika	**Studijní obor**	Softwarové systémy

Autor posudku	Mgr. Pavel Ježek, Ph.D.	**Role**	Vedoucí
Pracoviště	UK MFF KDSS		

Text posudku:

Autor v práci naimplementoval nástroj pro interaktivní extrakci dat z RIA webových stránek (kde jsou data typicky "generována"/"dostahována" dynamicky JavaScriptem do načtené stránky), na kterých selhávají běžně dostupné konkurenční nástroje zaměřené na statický web. Výsledná aplikace je plně funkční a efektivně kombinuje různé technologie - velmi oceňuji, že přestože je aplikace prototyp takového autorova řešení, tak se zaměřuje na budoucí reálné použití, a je navržena jako kvalitní framework s možností budoucího snadného rozšíření o další možnosti.

Dle mého názoru je zásadní, že aplikace je navržena s head-less infrastrukturou pro crawlování webů na pozadí (a tedy do budoucna umožňuje snadnou škálovatelnost), plus to, že pro "simulaci" chování webu používá přes web-driver API reálný head-less prohlížeč, a tedy že i v budoucnu s dalším rozvojem webových technoligií bude možné aplikaci snadno udržovat aktuální. Jelikož se většina existujících nástrojů i aktuálního výzkumu crawlování webů stálě zaměřuje hlavně na statický web, a tedy jsou z pohledu aktuálního vývoje webových aplikací velmi pozadu, tak je výsledek autorovy diplomové práce velice unikátní.

Na textové části práce bych nejvíce ocenil, že autor během celého vývoje prováděl velmi rozsáhlou a komplexní rešerži aktuálního stavu crawlování webů (prostudoval a v práci se odkazuje na velké množství vědeckých článků k tématu, a otestoval možství běžně používaných nástrojů pro crawlování webu, které v práci též srovnává). Dále že autor provedl rozsáhlou rešerži různých RIA webů, a reprezentativní množinu prezentoval v práci spolu s vlastnotmi ovlivňujícími návrh crawlovací aplikace. V neposlední řadě je přínosné, že se autor snažil zaměřit na potřeby reálných zákazníků, a reálné case-studies využití crawlovací aplikace v praxi pro dolování dat z webů (dle reálných požadavků zákazníků české pobočky firmy IBM, pro kterou nástroj autor vytvářel).

Co práci trochu sráží, je fakt, že byla dokončována dost na poslední chvíli, a tak některé části textu neprošly dostatečným množstvím revizí, a úroveň anglického jazyka a stylistiky je zde slabší.

Nicméně se celkově domnívám, že se jedná o velmi kvalitní diplomovou práci.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 24.8.2018 **Podpis**