Univerzita Hradec Králové
Fakulta informatiky a managementu

POSUDEK OPONENTA DIPLOMOVÉ PRÁCE

Jméno studenta: Bc. Jan Thér
Název práce: Extrakce dat z webu pomocí web scrapingu
Autor posudku: doc. Ing. Pavel Čech, Ph.D.

Cíl práce: Seznámit čtenáře s problematikou získávání dat z internetu, způsoby, jak k získávání dat přistupovat, jejich úskalími a možnými řešeními těchto úskalí.

Povinná kritéria hodnocení práce	**Stupeň hodnocení** (známka)					
	A	**B**	**C**	**D**	**E**	**F**
Práce svým zaměřením odpovídá studovanému oboru	☒	☐	☐	☐	☐	☐
Vymezení cíle a jeho naplnění	☐	☒	☐	☐	☐	☐
Zpracování teoretických aspektů tématu	☐	☐	☒	☐	☐	☐
Zpracování praktických aspektů tématu	☐	☒	☐	☐	☐	☐
Adekvátnost použitých metod, způsob jejich použití	☐	☒	☐	☐	☐	☐
Hloubka a správnost provedené analýzy	☐	☐	☒	☐	☐	☐
Práce s literaturou	☐	☐	☒	☐	☐	☐
Logická stavba a členění práce	☐	☐	☒	☐	☐	☐
Jazyková a terminologická úroveň	☐	☐	☒	☐	☐	☐
Formální úprava a náležitosti práce	☒	☐	☐	☐	☐	☐
Vlastní přínos studenta	☐	☐	☒	☐	☐	☐
Využitelnost výsledků práce v teorii (v praxi)	☐	☒	☐	☐	☐	☐

Vyjádření k výsledku anti-plagiátorské kontroly

Systém anti-plagiátorské kontroly uvádí celkovou podobnost 0 %.

Dílčí připomínky a náměty:

Práce je psána velmi čtivě a z textu je patrný nepochybný vhled autora do vybrané problematiky. Nicméně, ani tak se autor nevyhnul některým drobným terminologickým nepřesnostem. Například označení API jako technologie není zcela přesné. Stejně tak se v obecné rovině nejedná o řadu pravidel. Podobně označení REST, SOAP, RPC jako architektur či srovnání API a web scrapingu není úplně šťastné. Rozdělení technik web scrapingu v kap. 3.2 nemá zcela jasnou koncepci, protože techniky jsou spíše komplementární, což však z textu není patrné. V některých částech by autor měl více odkazovat na konkrétní zdroje, ze kterých čerpal (viz například kap. 3.1.). Výsledné porovnání aplikací v kap. 3.4.4 by mohlo být rozšířeno o další parametry. Předložená aplikace je funkční a dle přiloženého zdrojového kódu vhodně strukturována. Aplikace však mohla být podrobněji srovnána s již stávajícími aplikacemi např. z hlediska míry

úspěšnosti vyhnutí se detekci. Implementace mohla být také obohacena o některé sémantické funkce, které by umožnily efektivnější extrakci dat.

Je otázkou, zda pro česky psaný text nebylo vhodnější použít český ekvivalent jako např. termín extrakce webových dat. V textu je také několik drobných stylistických nedostatků. Například větu: „*Ačkoliv tato úprava dat již není součástí samotného procesu práce s API, jedná se o velmi důležitou součást procesu jak práce s API, tak web scrapingu.*“ (str. 4), by bylo vhodné srozumitelněji přeformulovat. Slovo základní není vhodné stupňovat (viz např. str. 22).

Celkové posouzení práce a zdůvodnění výsledné známky:

Předložená práce se zabývá zajímavou oblastí extrakce dat. Autor prokazuje dostatečnou orientaci ve vybrané oblasti, i přesto, že má práce určité terminologické nejasnosti. Součástí práce je také aplikace pro extrakci webových dat, kterou autor reaguje na některé nedostatky stávajících aplikací. Implementace aplikace je v textu pečlivě popsána a prokazuje autorovu schopnost vytvořit nástroj pro extrakci webových dat za použití odpovídajících programovacích technik a technologií. Autor se však mohl pokusit implementovat i některé pokročilejší sémantické funkce, které by napomohly efektivnější extrakci dat.

Otázka k obhajobě.

- Jaké sémantické funkce se při extrakci webových dat uplatňují?
- Jaké nevýhody má uplatnění session pro uchování dat aplikace?

Práci doporučuji k obhajobě.

Navržená výsledná známka: C

V Hradci Králové, dne 2. září 2022

podpis