Univerzita Hradec Králové
Fakulta informatiky a managementu

POSUDEK VEDOUCÍHO DIPLOMOVÉ PRÁCE

Jméno studenta: Bc. Jan Thér

Název práce: Extrakce dat z webu pomocí web scrapingu

Autor posudku: dr. Martina Husáková

Cíl práce: Návrh a implementace vlastního webscrapingového nástroje spolu s jeho otestováním a vyhodnocením

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

Práce vykazuje 0% podobnosti s jinými pracemi v systému Odevzdej.cz.

Dílčí připomínky a náměty:

Pro samotné tzv. in-text citace by bylo vhodnější zvolit jejich jiné umístění, např. na str. 5, kde se aktuálně jednotlivé odkazy na zdroje váží spíše na text následující než na text, který je přímo u odrážky (např. u odrážky „oddělení klienta od serveru: text text text. [1]" se zdroj [1] váže spíše k textu následujícímu). V textu by bylo vhodné se na obrázky přímo odkazovat, viz obrázek 4 a 8, u kterých jsem odkazy nenalezla. Text, zejména anotaci, by bylo vhodné ještě min. 1x pročíst. Místy se v něm (v ní) vyskytují vyjadřovací a interpunkční chybky.

Celkové posouzení práce a zdůvodnění výsledné známky:

Teoretická část práce je pečlivě zpracována. Diplomant problematiku charakterizuje výstižně a srozumitelně. Velmi oceňuji vlastní testování a vyhodnocení existujících webscraperů, viz str. 19 – 25. Na základě výsledků analýzy pak diplomant mohl dobře zhodnotit klady, zápory současných řešení a navrhnout řešení vlastní. Diplomant navrhl, úspěšně implementoval a otestoval vlastní webscrapingový nástroj (klient-server single-page aplikaci), který se opírá o webový framework Flask, knihovnu BeautifulSoup a JS knihovnu React. Pro výběr prvků k extrakci z webových stránek

navrhl vlastní způsob zápisu (tzv. extrakčních vzorce). Velmi podrobně a přesně vysvětlil jednotlivé části aplikace a jejich funkce. Lze deklarovat, že i tato část je zpracována velmi důkladně a pečlivě. Diplomant se ještě pustil do vývoje rozšíření, které extrahuje data z dynamických webů. I když je tato část v rozpracovaném stavu, tak je malinko škoda, že se diplomant více o tomto rozšíření nerozepsal. Teoretická i praktická část dokládá zvládnutí komplexní problematiky diplomantem. Vytvořil nástroj, který lze prakticky použít mj. i díky jeho veřejné dostupnosti. Aktivitu diplomanta v průběhu tvorby práce hodnotím jako příkladnou. Diplomant komunikoval velmi intenzivně a byla více než patrná silná motivace pro vývoj nástroje.

Otázky k obhajobě:

1. Máte v plánu dále pracovat na rozšíření extrahující data z dynamických webů? Pokud ano, pak jaké kroky by byly ještě třeba podniknout ve vývoji?
2. V práci na straně 56 zmiňujete, že se v problematice datové analýzy pohybujete několik let. Můžete toto tvrzení více upřesnit?

Práci doporučuji k obhajobě.

Navržená výsledná známka: A

V Hradci Králové**, dne 25. srpna 2022**

podpis