Univerzita Hradec Králové
Fakulta informatiky a managementu

POSUDEK OPONENTA DIPLOMOVÉ PRÁCE

Jméno studenta: bc. Marek Laušman

Název práce: Automatizované stažení a analýza webových stránek

Autor posudku: Martina Husáková

Cíl práce: Cílem práce je charakterizovat značkovací zvyklosti vzhledem k značkovacím doporučením s využitím automatizovaného stahování webových stránek.

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

Antiplagiátorská kontrola vykazuje 1% podobnosti v systému Odevzdej.cz.

Dílčí připomínky a náměty:

Diplomant ve své práci vytvořil program (tzv. scraper) pro automatizované stahování webových stránek ze serveru Archive.org a extrakci vybraných dat z tohoto zdroje. Použil k tomu jazyk Python a framework Scrapy. Výsledná data ukládal do databáze PostgreSQL. Za velmi zajímavý aspekt práce považuji cíl, ke kterému směřovalo použití těchto technologií. Diplomantovi se podařilo vhodně zkombinovat výše uvedené technologie pro zjištění, jak se v průběhu času měnily webové dokumenty z pohledu značkovacích zvyklostí, přístupnosti i případně reprezentované sémantiky. Vybral klíčové prvky web. stránek, které byly extrahovány a podrobeny analýze, čímž si čtenář udělá dobrou představu, jak se tento typ dokumentů v čase vyvíjel. V práci jsem nenalezla informaci, na jaké časové období diplomant u webových stránek cílil, tj. kdy byly vytvořeny. Dle obr. 10 nebo 11 odhaduji, že se jednalo o období 1996 – 2021. Po stránce teorie bych měla jednu

faktickou připomínku k tvrzení: „Čím nižší číslo v HTML elementu <h...>, tím důležitější a větší nadpis.“ Určitě lze s tímto souhlasit, nicméně až od HTML5 se hovoří o důležitosti nadpisů. Ve verzích dřívějších byly nadpisy spíše o té velikosti než o významu. Velmi pozitivně hodnotím snahu diplomanta nalézt cestu pro zrychlení scrapovacího procesu. V závěru práce (kap. 9) je uveden scraping pomocí jazyka Go a frameworku Colly, který diplomant shledal jako efektivnější alternativu ke scrapingu z pohledu rychlosti. V této kapitole (str. 51) je ale velmi matoucí uvedení faktu, že demonstrační příklad pro webový scraping byl vytvořen v knihovně Beautiful Soup (BS), nikoliv v dříve deklarovaném Scrapy (str. 1, 2, 28). Pokud bylo vytvořeno více scraperů (demo-příkladů), jasně bych je od sebe odlišila. Co se týká formální a jazykové stránky práce, pak přikládám několik dodatečných poznámek:

- Na str. 10 je uvedena definice sémantického webu. Pokud se jedná o překlad anglické definice, bylo by vhodné uvést, že se o překlad jedná, příp. uvést znění definice v původním jazyce.
- Str.2: „Předmětem této práce není práce s datovými sady...“ (Myslel jste datovými sklady?)
- Str. 8: <meta name="description» (Pozor na formátování)
- Str. 21: „Musí se ale vysloveně jednat o data, na které vlastník webu vlastní autorská práva, jelikož...“, „To samé platí pro nápady lze vlastnit pouze konkrétní...“
- Str. 22: „nevyhmatatelná“
- Str. 23: [odkaz na kapitolu] (Zřejmě chybí číslo zdroje)
- Str. 36: [ParseHub a Octoparse dokumentace] (Zřejmě chybí číslo zdroje)
- Str. 27: „Z porovnání webových stránek by mělo být vidět, se drží nějakých
- standardů...“

Celkové posouzení práce a zdůvodnění výsledné známky:

Diplomant měl před sebou velmi zajímavý cíl, kterého se podařilo úspěšně dosáhnout. S určitou nadsázkou lze hovořit o tom, že se student zabývá oborem, který bychom mohli nazvat jako „websites archeology and analysis“. Výklad teoretického pozadí je čtivý a srozumitelný. Teoretická i praktická část práce jsou dobře strukturované a obsahově vyvážené. V práci lze najít drobné jazykové nesrovnalosti, které by bylo vhodné ještě odstranit a text tak vyladit. Diplomovou práci hodnotím jako velmi zdařilou a obsahově poutavou. Práce splňuje podmínky dané metodickými pokyny pro tento typ prací.

Otázky k obhajobě:

Otázka navazuje na poznámku uvedenou výše, tj. na knihovnu BS (viz str. 51). Nejedná se o faktickou chybu, když je v kap. 9 (Shrnutí a doporučení) zmíněno, že demonstrační příklad byl vytvořen v knihovně BS? V úvodním textu (str. 1, 2, 28) je uvedeno, že je použit framework Scrapy. Pokud se o chybu nejedná, z jakého důvodu byl vytvořen další scraper? Čtenář by pak očekával, že i data extrahovaná pomocí scraperu vytvořeného v BS budou analyzována a třeba porovnána s programem vytvořeným ve Scrapy.

Práci doporučuji k obhajobě.

Navržená výsledná známka: A

V Hradci Králové**, dne 13. května 2022**

podpis