Univerzita Hradec Králové
Fakulta informatiky a managementu

POSUDEK OPONENTA BAKALÁŘSKÉ PRÁCE

Jméno studenta: Filip Roškot

Název práce: Web crawlery

Autor posudku: Martina Husáková

Cíl práce: navrhnout specializovaného webového robota (web crawlera), který analyzuje webové sídlo za účelem vytvoření sitemapy; představit metodu web crawling a web scraping, metody vytvoření sitemapy a navrhnout její další využití

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

Vyjádření k výsledku anti-plagiátorské kontroly

Anti-plagiátorská kontrola vykazuje 0% podobnosti.

Dílčí připomínky a náměty:

Bakalant zvolil (si vybral) téma, které je v dnešní době velmi aktuální, hojně zmiňované, a kterému není zatím věnována dostatečná pozornost ve výuce webových technologií na UHK-FIM. Bakalant se zabývá problematikou získávání dat z webových stránek pomocí techniky web scraping a web crawling, viz teoretická část práce str. 1 - 26. Hlavní pozornost bakalant věnuje python. knihovnám, které lze k tomuto účelu využít, tj. uvádí převzaté i vlastní příklady získávání dat z webových stránek pomocí knihovny Request, Urllib, BeautifulSoup a Scrapy. Poslední zmiňovanou knihovnu bakalant použil pro vývoj vlastního webového robota (crawlera), který získává webové adresy z webových stránek vybraného webového sídla, viz praktická část práce str. 27 – 38. Obě části práce jsou přehledné, problematika je vyložena jasně bez výrazného množství gramatických chyb nebo překlepů. Níže uvádím několik poznámek, připomínek a námětů na zlepšení práce:

- Na str. 6 je uvedeno následující: *„Kupříkladu níže zobrazený příklad [8]: …"* Myšlenka v tomto odstavci s odrážkami není dokončena.
- Na str. 22/bod 2 je zmíněna zkratka lxml. Bylo by vhodné ji v textu blíže charakterizovat.

- V textu chybí odkaz na obrázek 1 a 2.
- Na str. 7/podkapitola 2.5 je uvedeno: *„Data mining je definováno jako "dolování informací z webových sídel" [16]* U zdroje 16 není přímo uvedeno, z jaké stránky knihy citace pochází, a tak je obtížné určit, zda toto tvrzení publikace opravdu zmiňuje. Spíše se domnívám, že došlo k ne zcela správné interpretaci/překladu pojmu data mining z anglického do českého jazyka. Data mining se totiž neomezuje na dolování informací z webových sídel. Jeho působnost je mnohem širší.
- Na str. 12 jsou uvedeny možnosti crawlera, ale přitom není zcela jasné, zda těmito schopnostmi nedisponuje i scraper.

Bakalant opomněl u knih Learning Scrapy, Python Web Scraping a Website scraping with python ... uvést jejich autory. Bakalant k práci přikládá několik příloh (celkem 11 dalších souborů), které nejsou v samotném textu uvedeny, resp. na str. 42/9. Přílohy. CD příloha není k práci přiložena, nicméně v systému eVSKP nahrána je.

Celkové posouzení práce a zdůvodnění výsledné známky:

Bakalářskou práci Filipa Roškota považuji za užitečnou a přínosnou, zejména po stránce praktické, kde vytvořil funkčního webového robota. Pozitivně také hodnotím využití programu Graphviz pro vizualizaci výstupu webového robota. Za negativum práce považuji skutečnost, že práce je svým stránkovým rozsahem na hraně. Práce by mohla být např. obohacena o hlubší vzájemné srovnání knihoven využitelných pro web crawling/scraping nebo o již existující práce na toto téma. I přes některá výše uvedená negativa práci k obhajobě doporučuji.

Otázky k obhajobě:

1. Objasněte rozdíl mezi web scrapingem a web crawlingem.
2. Co označujeme zkratkou lxml?
3. Jak byste práci dále rozšířil?

Práci doporučuji k obhajobě.

Navržená výsledná známka: C

V Hradci Králové**, dne 27. srpna 2019**

_____

podpis