Posudek oponenta bakalářské práce

Student: Perina Lukáš
Téma: Metody extrakce dat z webových stránek (id 23941)
Oponent: Křivka Zbyněk, Ing., Ph.D., UIFS FIT VUT

1. Náročnost zadání **obtížnější zadání**

Téma je průzkumně implementační a vyžaduje také vyhodnocení ve formě experimentů. Téma extrakce informací z webu je stále aktuální výzkumné téma a pořád se objevují nové přístupy a ty jsou implementovány a vyhodnocovány a v této práci se jeden studentem navržený přístup zkouší.

2. Splnění požadavků zadání **zadání splněno**

3. Rozsah technické zprávy **je v obvyklém rozmezí**

Text má rozsah kolem 65 normostran včetně autorských obrázků, což je i vzhledem k obsahové kvalitě přiměřené.

4. Prezentační úroveň předložené práce **70 b. (C)**

Text má dobrou celkovou strukturu. Nicméně obsahuje řadu menších nedostatků, které jako celek snižují kvalitu celé práce. Například, aby si čtenář ujasnil používané pojmy (např. okruh, skupina, prvek, primární prvek), tak musí přečíst většinu práce a pak se teprve vrátit k obtížnějším pasážím. Nedefinování pojmů nebo nedůsledné používání zadefinovaných pojmů či vhodné terminologie snižuje pochopitelnost textu. Např. na jednom místě popisujete "hlavní pravidla" (nadpis 4.2.1), ale později píšete o "základních pravidlech" apod. Dobré logické provázanosti také neprospívá neodkazování některých obrázků v textu (např. výpisy 5.1 a 5.2). Pro lepší orientaci by bylo vhodné se odkazovat v kapitole o implementaci též na některé význačné funkce ve zdrojovém kódu.

5. Formální úprava technické zprávy **85 b. (B)**

Práce je psaná ve slovenštině, ale zdá se být pravopisně v pořádku až na anglikanismy (např. web scrapery, setu vysledkov, level integracie). Z typografického hlediska obsahuje řadu drobných problémů jako chybějící mezery mezi číslem a znakem procenta, odkazování jen číselným odkazem bez uvedení, zda je to sekce či obrázek (např. str. 6 nebo 10).

6. Práce s literaturou **72 b. (C)**

Prameny jsou vhodně zaměřené, jen v případě již zavedených technologií (JavaScript, NodeJS) bylo jistě možné využít i knižní publikace. Podobně bych očekával i stručnou rešerši technik extrakce informací z webových stránek založenou na odbornější literatuře než jen několika webových statistikách.

7. Realizační výstup **83 b. (B)**

Program je zcela funkční, obsahuje řadu parametrů pro další experimentování a lze jej zprovoznit dle návodu v příloze B. Médium obsahuje také dvě testovací sady studenta a dva další od kolegy s podobným tématem. Zdrojové kódy jsou dostatečně komentované a rozlišují práci studenta od využitých cizích knihoven.

8. Využitelnost výsledků

Aplikace je nyní použitelná pouze pro odbornější uživatele, ale výsledky zatím vypadají slibně. Pro použití řadovými uživateli je třeba předdefinovat větší množství různých formátů informací (např. regulární výrazy pro různé formáty času) a vytvořit libivé GUI, které pomůže především s výběrem vhodného formátu informací a jejich poskládání do bloku, což by mohlo být lépe diskutováno v závěru.

9. Otázky k obhajobě

- U knihovny Puppeteer jste zmiňoval především výhody, můžete se zamyslet i nad nevýhodami použití této knihovny?
- Chápu, že v současné době není k dispozici žádná srovnávací testovací sada pro alternativní nástroje, ale dle zmínky v textu byl vytvořen jiným studentem alternativní nástroj s jiným přístupem k extrakci, ale využívající stejné testovací sady. Můžete provést alespoň srovnání s kolegou z hlediska přesnosti a časové náročnosti pro jednotlivé sady?

10. Souhrnné hodnocení **81 b. velmi dobře (B)**

Text je spíše průměrné kvality především kvůli problematické prezentaci náročnějších pasáží. Velmi však oceňuji studentův vlastní inovativní návrh automatické extrakce dat kombinující četnostní analýzu stylových tříd a popis formátu dat pomocí regulárních výrazů. Z tohoto pohledu práci pro lepší hodnocení chybí porovnání s dalšími existujícími nástroji, takže navrhuji **velmi dobře (B)**.

Prohlášení: Uděluji VUT v Brně souhlas ke zveřejnění tohoto posudku v listinné i elektronické formě.

V Brně dne: 3. června 2021

Křivka Zbyněk, Ing., Ph.D.
oponent