Posudek oponenta diplomové práce

Student: Hubl Lukáš, Bc.
Téma: Sémantická analýza webového obsahu (id 22669)
Oponent: Rychlý Marek, RNDr., Ph.D., UIFS FIT VUT

1. Náročnost zadání **průměrně obtížné zadání**

Jedná se o průměrně obtížné zadání, kde student pro řešení využívá dostupné a dobře dokumentované technologie DBPedia a DBPedia Spotlight.

2. Splnění požadavků zadání **zadání splněno**

Zadání je splněno bez výhrad.

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

Rozsahem je technická zpráva v obvyklém rozmezí, od úvodu po závěr obsahuje 59 vysázených stran. Všechny kapitoly zprávy jsou informačně bohaté a pro téma nezbytné.

4. Prezentační úroveň předložené práce **75 b. (C)**

Technická zpráva má logickou strukturu, která odpovídá procesu vývoje výsledného programového řešení. Některé části z kap. 5 "Implementace segmentační metody" (např. kap. 5.2 či 5.3.4) by bylo vhodnější zařadit dříve, již do kap. Kapitola 4 "Návrh extrakce témat za účelem segmentace". Blokové schéma návaznosti základních kroků aplikace na obr. 5.1 na str. 28 by bylo vhodnější zapsat v notaci UML, např. jako diagram aktivity, neboť v nynějším zápisu se nepřehledně míchají data s akcemi a stavy systému. Vzorce v sekci "Výpočet procentuálního podílu tématického okruhu" na str. 37 by si zasloužily podrobnější vysvětlení, přestože jsou převzaty z odkazovaného zdroje.

5. Formální úprava technické zprávy **65 b. (D)**

Z hlediska formální úpravy obsahuje technická zpráva drobné, ale časté jazykové i typografické chyby, které působí dosti rušivě. Např. chybějící mezery před jednotkami ("99%" či "16GB" na str. 29), chybný zápis verze ("Ubuntu 18.4." na str. 29), překlepy ("Spolight" na str. 29, "vedoucíc" na str. 38, "kontola" na str. 44, "sématntticky" na str. 48, aj.), použití zdvojených apostrofů místo uvozovek v HTML ve výpisech 5.5 až 5.7 na str. 43, užití spojovníku místo pomlčky (kap. 7.1.1), chybné znaky pro uvozovky (kap. 7.1.1), aj.

6. Práce s literaturou **85 b. (B)**

Seznam literatury obsahuje 13 položek, z nichž jsou většina odborné publikace k tématu práce. Zdroje jsou použity především v první části technické zprávy, kde lze velmi dobře odlišit převzaté části od vlastních myšlenek studenta. Bylo by ale vhodné nalézt a použít relevantní zdroje také v druhé části zprávy, např. při vyhodnocení experimentů v kap. 7.1.

7. Realizační výstup **75 b. (C)**

Výsledným programovým řešením je aplikace v jazyce Python pro stažení a sémantickou segmentaci v DOM stromu webových stránek s využitím anotací získaných z DBPedia pomocí nástroje Spotlight. Výsledky jsou doplněny do vstupních HTML souborů jako nové atributy původních elementů, které však nejsou ve vlastním jmenném prostoru, a tak je výsledkem bohužel nevalidní HTML. Dle kap. 5.1, 5.3.2 a 6.2.5 se pro stahování vstupní HTML stránky používá nástroj Selenium nad prohlížečem Firefox (volání akce uživatelského rozhraní "Save as", tj. Ctrl+S), což považuji za zbytečně složité řešení a domnívám se, že by bylo vhodnější použít např. "wget --mirror" či jiný nástroj. V textu je zmíněn problém s výkonem (rychlostí) komponenty Spotlight, avšak výkon vlastní segmentace bohužel nebyl v experimentech měřen ani později diskutován, a výsledné řešení není pravděpodobně optimalizováno pro výkon (např. pomocí škálování).

8. Využitelnost výsledků

Výsledky jsou prakticky využitelné, avšak je škoda, že nebyly porovnány s jinými nástroji pro segmentaci webových stránek, přestože to zadání explicitně nevyžadovalo.

9. Otázky k obhajobě

- V kap. 6.2 popisujete různé testování celé aplikace. Jak probíhalo testování jejích komponent během vývoje (jednotkové testy)?
- Zkoušel jste také jiné možnosti stáhnutí kompletní webové stránky (např. "wget --mirror"), než Vámi implementované řešení s využitím Selenium nad prohlížečem Firefox? Proč jste se rozhodl zrovna pro složité řešení s využitím nástroje Selenium?
- V kap. 7.1.1 "Referenční výsledek - Ground truth" popisujete ruční segmentaci pro získání referenčních dat. Jaký je Váš a jaký je obvyklý postup dle literatury či obdobných nástrojů (v kap. není žádná citace)?

10. Souhrnné hodnocení

75 b. dobře (C)

Technická zpráva je sice podrobná, ale s nedostatky (prezentační a formální úroveň). Programové řešení je funkční, splňuje zadání, avšak některé části aplikace mohly být navrženy lépe a také vyhodnocení mohlo být důslednější. Celkově považuji výsledek za průměrný a navrhuji hodnotit práci stupněm **dobře (C)**.

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

V Brně dne: 25. června 2020

Rychlý Marek, RNDr., Ph.D.
oponent