

Posudek oponenta bakalářské práce

Student: Stejskal Jakub
Téma: Kontrola konzistence informací extrahovaných z textu (id 18808)
Oponent: Otrusina Lubomír, Ing., UPGM FIT VUT

- 1. Náročnost zadání** průměrně obtížné zadání
- 2. Splnění požadavků zadání** zadání nesplněno

Cílem práce bylo vytvořit systém pro kontrolu konzistence informací extrahovaných z Wikipedie a dalších zdrojů. Student vytvořil velmi jednoduchý systém pro extrakci informací, nicméně tento systém neobsahuje žádnou kontrolu konzistence extrahovaných dat na základě definovaných omezení, což bylo v zadání. Systém byl vyhodnocen na velmi malé datové sadě a ne na datových sadách získaných z rozsáhlých projektů, jak tomu mělo být dle zadání.
- 3. Rozsah technické zprávy** je v obvyklém rozmezí

Rozsah technické zprávy je v obvyklém rozmezí. Jednotlivé části jsou informačně bohaté. Kapitola s evaluací systému by ale měla být delší a měla by obsahovat podrobnější vyhodnocení systému.
- 4. Prezentací úroveň předložené práce** 60 b. (D)

Logická struktura práce je dobrá. Rozsahy jednotlivých kapitol jsou v pořádku. Většina práce je pochopitelná pro čtenáře, ale najde se v ní i několik nepodložených a nepravdivých tvrzení. Některé pasáže jsou hůře srozumitelné.
- 5. Formální úprava technické zprávy** 60 b. (D)

Po formální stránce práce obsahuje několik nedostatků. Chybí odkazy na zmiňované nástroje. Některá slova student nepřekládá (např. lower) a jiná překládá špatně (např. false negatives jako špatně negativní; escape sequence jako únikové sekvence).
- 6. Práce s literaturou** 50 b. (E)

Výběr studijních pramenů je dobrý, nicméně se tam občas vyskytují položky, které by se více hodily do poznámky pod čarou (např. odkaz na weby použitých technologií). Jako problém ale vidím fakt, že student v některých podkapitolách neuvádí odkud čerpá (např. kapitola popisující syntaktickou analýzu).
- 7. Realizační výstup** 40 b. (F)

Student vytvořil velmi jednoduchý systém pro extrakci informací. Vyhodnocení systému nesplňuje očekávání. Z textu práce je znatelné, že student nějaké operace provádí neefektivně, např. průchod dumpem Wikipedie trval desítky hodin. Systém neumožňuje kontrolu konzistence informací extrahovaných z Wikipedie a dalších zdrojů.
- 8. Využitelnost výsledků**

Jedná se o práci kompilačního charakteru. Výsledky práce pravděpodobně nejsou využitelné.
- 9. Otázky k obhajobě**
 1. V práci píšete, že jste implementoval vlastní rozpoznávač jmenných entit, protože data, která byla zpracována používaným systémem NER, byla nepřesná. Dělal jste nějaké exaktní porovnání nebo je to jen Vaše domněnka? Proč jste nepoužil nějaký standardní nástroj pro rozpoznávání jmenných entit a pustil jste se do vlastní implementace?
 2. V práci píšete, že jste vyextrahoval pouhých 435 473 vztahů pro 7 481 416 entit, přesto tvrdíte, že dosahujete pokrytí 74,46 %. Jak je to možné?
- 10. Souhrnné hodnocení** 49 b. nevyhovující (F)

Student nesplnil zadání práce. Výsledkem měl být systém pro kontrolu konzistence informací extrahovaných z Wikipedie a dalších zdrojů, student však vytvořil pouze jednoduchý systém pro extrakci informací. Vyhodnocení vytvořeného systému bylo provedeno nedostatečně. Některé pasáže neobsahují citace. Z těchto důvodů navrhuji hodnocení stupněm F.

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

V Brně dne: 1. června 2016

.....
podpis