

Review of Bachelor's Thesis

Student: Kočalka Jakub
Title: Recognition of Repeating SMS Patterns (id 24169)
Reviewer: Češka Milan, RNDr., Ph.D., DITS FIT BUT

- 1. Assignment complexity** **average assignment**
Zadání považují za průměrně složité - jedná se o adaptaci a optimalizaci algoritmů pro zarovnávání textů za účelem klasifikace SMS spamů (zejména pro detekci tzv. spamovacích kampaní).
- 2. Completeness of assignment requirements** **assignment almost fulfilled**
Na jedné straně bylo zadání rozšířeno, jelikož se ukázalo, že požadovaný Smith-Watermann algoritmus pro lokální zarovnávání nedosahuje požadované výsledky. Student tudíž rovněž adaptoval a optimalizoval Needleman-Wunsch algoritmus pro globální zarovnávání. Na druhé straně požadované experimentální vyhodnocení považují za nedostatečné (viz další body hodnocení).
- 3. Length of technical report** **in usual extent**
Rozsah technické zprávy je v obvyklém rozmezí. Jednotlivé části jsou informačně přínosné. V úvodu postrádám shrnutí dosažených výsledků. Část věnovaná implementaci je minimalistická a popis experimentálního vyhodnocení je nedostatečný (viz dále).
- 4. Presentation level of technical report** **65 p. (D)**
K prezentace technické zprávy má několik zásadních výhrad. Kromě formálních nedostatků (viz níže), je struktura a logická návaznost některých kapitol velmi nevhodná. Například samostatná kapitola 2.2 nedává smysl a technický popis zarovnávacích algoritmů je nedostatečný. Z pohledu pochopení textu jsou nejvíce problematické kapitoly 5 a 6. Ačkoliv tyto kapitoly obsahují původní a zajímavé myšlenky autora, jejich prezentace je velice chaotická a těžce pochopitelná, jelikož míchá dohromady definice pojmů, experimentální pozorování a navržené optimalizace. Má hlavní výhrada souvisí s již zmiňovanou experimentální částí. Zde bych očekával detailnější vyhodnocení zaměřené na korektnost/kvalitu získané klasifikace (zejména zda je výsledné slučování SMS spamů do jednotlivých skupin/kampaní korektní) a na vyhodnocení dopadu jednotlivých parametrů a optimalizací na kvalitu klasifikace. Ačkoliv takové vyhodnocení vyžaduje ruční anotaci vstupních dat, je pro tuto práci naprosto nezbytné. Student se omezil pouze na demonstraci a zdůvodnění klasifikace na pár vybraných příkladech, což je přínosné ale nedostatečné.
- 5. Formal aspects of technical report** **65 p. (D)**
Formální úprava práce je velice slabá (například kapitola 2 či sazba algoritmů). Hlavním problémem je obrovské množství chyb a překlepů, které svědčí o tom, že text nebyl dostatečně zkontrolován. Ačkoliv je jazyková úroveň angličtiny dostatečná, velké množství chyb kazí čitelnost a srozumitelnost textu.
- 6. Literature usage** **80 p. (B)**
Ačkoliv student demonstroval dostatečné pochopení problematiky a orientaci v existující literatuře, práce s literaturou včetně formy citací má horší úroveň. Například bych očekával lepší pokrytí relevantní literatury v úvodu a citace [1-3] nejsou vhodné/úplné.
- 7. Implementation results** **70 p. (C)**
Na základě demonstrace vytvořeného nástroje mohou konstatovat, že nástroj je funkční. Během demonstrace student prokázal orientaci a pochopení hlavních částí kódu. Kód by ale měl být lépe komentovaný. Hlavní problémem vidím ve validaci a testování, která se omezuje pouze na ruční inspekci a kontrolu několika výstupů. S tím rovněž souvisí již kritizované experimentální vyhodnocení, které považují za nedostatečné.
- 8. Utilizability of results**
BP byla realizována ve spolupráci s firmou Mavenir s.r.o., která se zabývá detekcí a klasifikací SMS spamů. Na základě prezentovaných výsledků a diskuze v rámci demonstrace se zdá, že realizační výstup poskytuje lepší výsledky než aktuální řešení ve firmě Mavenir s.r.o založené na specializovaných hashovacích funkcích. Pokud se toto prokáže i v rámci důkladnějšího experimentálního vyhodnocení, navržený a implementovaný přístup může vést k vylepšení detekce a klasifikace SMS spamů.
- 9. Questions for defence**
 1. Jak jste testoval kvalitu získané klasifikace SMS spamů (tj. detekci kampaní a získané přiřazení spamů)? Lze toto testování nějak automatizovat?
 2. Můžete rigorózně ukázat, že navržený přístup je lepší než aktuální řešení použité ve firmě Mavenir s.r.o založené na specializovaných hashovacích funkcích?

3. Jaký je vliv jednotlivých "hyper-parametrů" navrženého přístupu na výslednou kvalitu klasifikace SMS spamů.

10. Total assessment

65 p. satisfactory (D)

Z práce mám velice smíšené dojmy. Ačkoliv se zdá, že student navrhl a implementoval efektivní přístup pro netriviální problém klasifikace SMS spamů (zejména pro detekci tzv. spamovacích kampaní), textová část BP je velice slabá. Hlavní nedostatek vidím v nedostatečném experimentálním vyhodnocení získané klasifikace SMS spamů. Celkově navrhuji práci hodnotit známkou D. V případně kvalitní prezentace a zodpovězení dotazů navrhuji práci hodnotit známkou C.

In Brno 29 July 2021

Češka Milan, RNDr., Ph.D.
reviewer