

Posudek oponenta diplomové práce

Student: Piják Marek, Bc.
Téma: Klasifikace emailové komunikace (id 20270)
Oponent: Herout Adam, prof. Ing., Ph.D., UPGM FIT VUT

1. **Náročnost zadání** **průměrně obtížné zadání**
Řešitel vytvořil klasifikátor pro zpracování zadané datové sady (e-mailů konkrétní obchodní společnosti).
2. **Splnění požadavků zadání** **zadání splněno s drobnými výhradami**
Podle bodu 3. měl řešitel klasifikátor zdokonalovat. To se stalo pouze dílčím způsobem a v prezentovaných výsledcích nelze vypořádat nějaký trend vylepšování pomocí různých technik.
3. **Rozsah technické zprávy** **splňuje pouze minimální požadavky**
Technická zpráva patří mezi kratší a méně výživné.
4. **Prezentační úroveň předložené práce** **65 b. (D)**
Textová zpráva obsahuje prakticky vše podstatné. Nicméně, bylo by na místě více diskutovat předzpracování dat - hledat lepší řešení stemmatizace, vyřazení stopslov atd. Naměřené výsledky ukazují, že různé použité klasifikátory všechny dosahují hodně podobných výsledků. To může naznačovat, že limitem celého řešení je právě raná fáze zpracování, pořizování příznaků ze zpráv - možná v takto krátkých a monotematických zprávách použití TF/IDF není to pravé a třeba použití n-gramů či podobných příznaků by bylo správnější. Jak již bylo naznačeno v bodě 2. výše, řešitel vytvořené řešení příliš nezdokonaloval, ale navrhl jednu extrakci slov, jedno řešení převodu slov do příznakového vektoru a vyzkoušel pár klasifikátorů nad tímto příznakovým vektorem.
5. **Formální úprava technické zprávy** **75 b. (C)**
Po formální stránce je textová zpráva průměrná - je vysázená přijatelně a srozumitelně, ale obsahuje ne úplně vzácné chyby ve formátování a jazykové.
6. **Práce s literaturou** **65 b. (D)**
Seznam zdrojů není nijak impozantní. Jak bylo naznačeno výše, řešitel si s jednotlivými částmi celého řešení příliš nehrál a nehlídal alternativní přístupy ke "kanonickému" řetězci pro klasifikaci přirozeného jazyka.
7. **Realizační výstup** **75 b. (C)**
Řešitel pořídil/anotoval dataset rozumných rozměrů, vytvořil řešení pro extrakci příznakových vektorů a klasifikaci jednotlivých položek (e-mailů) a vytvořené řešení vyhodnotil. Vyhodnocení není příliš podrobné, takže není zřejmé, co je limitujícím faktorem - data samotná? zvolené třídy? extrakce slov? extrakce příznakového vektoru?
8. **Využitelnost výsledků**
Práce byla dělana pro praxi, ale dosažené výsledky zřejmě nejsou tak přesvědčivé, aby řešení (bez dalšího výrazného rozvinutí) bylo nasazené.
9. **Otázky k obhajobě**
 - Kdybyste měl vylepšit jedinou část celého řetězce, která by přinesla největší zlepšení a proč?
 - Bylo by možné - spíše než do zvolených deseti tříd - hodnotit strojem nějakou jinou kvalitu textu tak, aby to bylo pro "objednatele" aplikačně zajímavé a přitom by stroj dosahoval lepších výsledků? Hodnotit důležitost/urgentnost zprávy? Hodnotit relevanci pro danou pracovní pozici? ...
10. **Souhrnné hodnocení** **65 b. uspokojivě (D)**
Téma je zajímavé a netriviální a řešitel vytvořil základ řešení. Vyhodnocení konstatuje hrst naměřených čísel, ale nevrhá skutečné světlo na možnosti a omezení zvoleného přístupu. Odevzdaný produkt je prvním řešením zadaného problému - řešitel příliš neiteroval jednotlivé jeho části, aby došel k optimu. Textová zpráva je svým rozsahem, obsahem i po formální stránce průměrná či lehce podprůměrná.

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

V Brně dne: 5. června 2018

.....
podpis

