Spam filtr (report)

Na projektu se podíleli:

-Matvej Safrankov (saframa9)

-Nadzeya Shchahlova (shchanad)

Úkolem bylo vytvořit filtr nevyžádané pošty, který by rozlišoval mezi normálními zprávami a spamem.

Filtr je založen na vyhledání klíčových slov v textu. Algoritmus je maximálně primitivní: [V složce otevři postupně každý mail, je-li spamové slovo v mailu, označ ho jako spam, jinak označ jako ham. Zapiš výsledky do slovníku]
Klíčová slova se vybrala ručně. Hlavním parametrem na výběr slova, byl náš subjektivní názor na to, jak moc vnímáme takové slovo jako klišé v spamových emailech.

Filtr neaktivuje funkci samoučení. Není k tomu žádná potřeba. Jediné, co se nabídlo v průběhu vývoje filtru, bylo přidáním nových slov na základě trénovacích dat. Princip by byl jednoduchý, rozbití těla mailu na tokeny, jež by se ukládali do slovníku. Z tohoto slovníku by se vymazávali slova s nejnižším výskytem. Následně tokeny vyskytující se nejvíce krát, by se přidali do seznamu předem vybraných slov (popsaného v popisu algoritmu).
Avšak to bylo zavrženo, neboť trénovací data nebyla dostatečně obsáhlá, tudíž nová slova měli tendenci zhoršovat výsledky filtru.

Test byl proveden na datech nabízených v příloze úlohy. Byla vyzkoušena různá nastavení filtru související s výběrem slov. Výsledkem pečlivého výzkumu byl optimální seznam slov, který dosáhl nejvyšší kvality filtru (kvalita se u předložených testovacích dat zvýšila z 0,37 na 0,66).
Ovšem nemůžeme očekávat stejně vysokých výsledků na jiném setu dat. Neboť se tam mohou vyskytovat úplně jiné myšlenky.

Práce na projektu probíhaly soudržně a všechny problémy byly po diskusích společně vyřešeny. Je obtížné určit jasné rozdělení odpovědností. Komunikace mezi členy týmu probíhala pomocí webové služby GitHub.

Níže je uvedena tabulka, která ukazuje silně přibližné procento splnění jednotlivých částí úkolu jednotlivými členy týmu.

Část úlohy	Matvej Safrankov	Nadzeya Shchahlova
Obecná struktura algoritmu	50	50
Hledání slovních tokenů (seznam spamových slov)	30	70
Strukturování programu, rozdělení kódu na samostatné, čitelné části.	80	20
Testování pro nalezení optimálního seznamu tokenů.	32	68
Zpráva a prezentace projektu	70	30