Spam filter

Na projektu pracovali:

Matvej Safrankov (saframa9)

Nadzeya Shchahlova (shchanad)

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

Filtr je založen na tokenizaci textu. Slova tokenů byla nalezena ručně na základě důkladné analýzy spamu. Filtr neaktivuje funkci samoučení. Test byl proveden na datech nabízených v podmínce úlohy. Byla vyzkoušena různá nastavení filtru související s výběrem slov token. Výsledkem pečlivého výzkumu byl nejoptimálnější seznam dříve nalezených slov, který dosáhl nejvyšší kvality filtru (kvalita se u navrhovaných testovacích dat zvýšila z 0,37 na 0,66).

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 procento splnění jednotlivých částí úkolu jednotlivými členy týmu.

|  |  |  |
| --- | --- | --- |
|  | Matvej Safrankov | Nadzeya Shchahlova |
| Obecná struktura algoritmu | 50 | 50 |
| Hledání slovních tokenů | 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 |