Spam Filter

Autoři:
Ondřej Chvátal (chvatond@fel.cvut.cz)
Lukáš Cafourek (cafoulu1@fel.cvut.cz)

Úvod:
Naší třetí a poslední úlohou v RPH bylo vytvořit Spam Filter, který do výsledného souboru vypíše předpovědi, které emaily jsou SPAM a které jsou OK na základě nějakých dat. Spam Filter mohl být jak primitivní bez schopnosti učení, tak pokročilý s využitím učení.

Popis našeho Spam Filtru:
Náš Spam Filter je mírně pokročilejší a využívá i funkce train(). Využívá i pravidel, jako je délka těla e-mailu, přílišné využití CapsLock nebo určité výrazy psané v těle e-mailu. První verzí filtru byl jen Spam Filter, který se rozhodoval náhodně a měl prázdnou funkci train(). Na něm jsme si zkusili odevzdání, jestli prochází testy v BRUTE.

Způsob trénování filtru:
Náš Spam Filter ve funkci train() nejdříve otevře soubor !truth.txt v adresáři s trénovacími daty, podle kterého načítá slova použitá ve SPAMových zprávách a v OK zprávách do seznamů spam_words a ok_words respektive. Poté odečetl průnik těchto seznamů ze seznamu spam_words, čímž by měl vzniknout seznam čistě “spamových” slov, který je následně použit ve funkci test().

Výsledky (míra kvalit filtrů)**:**

	Testing corpus	Training Corpus
Random Filter	0.22	N/A
Keyword Filter	0.47	N/A
Advanced Filter	0.63	0.75

Rozdělení práce v týmu:
Cafourek - vypracování Random Filtru a zkouška implementace na testovací sadě dat (míra kvality filtru).
Chvátal - vypracování Keyword Filtru (s použitím určitých pravidel) a Advanced Filtru (se schopností učení) a zkouška implementací na testovacích i trénovacích sadách dat a porovnání výsledků (míra kvalit filtrů).

Organizace práce v týmu:
Soubory jsme si sdíleli přes GitHub nebo Discord a komunikace probíhala převážně přes Discord.

Zhodnocení a závěr:
Úloha Spam Filtru byla zajímavá a určitě nám přispěla ke zlepšení znalostí Pythonu. Zajímavé bylo pozorovat výsledky kvalit jednotlivých filtrů, kde se nám podařilo dosáhnout hodnot 0.63 až 0.75 lokálně, tyto výsledky ale nekorespondují hodnotám v BRUTE, kde byla nejvyšší dosažená kvalita 0.46. To může být způsobeno “přetrénovaným” filtrem, nebo pouze nevhodnou implementací našeho řešení. Na filtru je spousta prostoru pro zlepšení, například přidání dalších principů pro rozlišení mezi spamy a ok emaily, nebo optimalizace již implementovaného kódu.

Seznam použitých zdrojů:
Prezentace a kódy z přednášek, kódy z cvičení, pomocné tipy na stránkách předmětu