3. ledna 2023

SPAM filter report

autoři: Jan Sadílek a Kateřina Rejchrtová

Náš SPAM filter splňuje pouze základní požadavky pro splnění zadání úlohy a nevyužívá strojové učení. Zvažovali jsme různé metody fungování algoritmu, nakonec jsme se spokojili s jednoduchou variantou.

Hlavní soubor *filter.py* používá třídy souborů *basefilter.py*, *corpus.py*, *spam_words.py* a *utils.py*. Využívá list SPAM_WORDS plný slov typických pro SPAM ze souboru *spam_words.py* k rozlišení, zda je email SPAM anebo není.

Kód přečte email a rozdělí ho do listu slov. Poté zjistí počet jednotlivých spamových slov ve SPAM_WORDS, která se objeví v emailu. Pokud je číslo nalezených spamových slov větší než 0, email je klasifikován jako SPAM. Pokud je číslo výskytu spamových slov rovno 0, email je klasifikován jako OK a program ho nevyřadí. Spamová slova byla vybrána procházením jednotlivých spamových mailů z pomocných souborů. Další slova a slovní spojení jsme doplnili pomocí internetu.

Třída *BaseFilter*, kterou hlavní soubor používá, je rodičovskou třídou pro všechny filtry a tvoří tak základní kámen projektu. Má dvě metody *train* a *test*. Metoda *train* není v našem spam filtru používána a metoda *test* slouží ke čtení obsahu mailů a jejich rozdělení do jednotlivých slov.

Třída *Corpus* (v modulu *corpus.py*), má metodu *emails()*, která je generátorem. Tato metoda si je vědoma toho, že v adresáři s emaily mohou být i soubory s metainformacemi. Názvy těchto souborů začínají vždy znakem ¨!¨, proto všechny soubory začínající vykřičníkem jsou ignorovány.

Týmová práce fungovala výhradně online při domluvě přes službu *Messenger* a sdílení programů přes *GitHub*. Největší kus práce v našem projektu odvedl Jan Sadílek, který napsal nejdůležitější část projektu *basefilter.py*. Veškerá další spolupráce proběhla v pořádku a pokaždé jsme nalezli shodu v našich názorech a postupech.

Jsme si vědomi nedokonalostí v projektu ovšem označujeme jeho splnění za dostačující. Tento úkol nám umožnil si vyzkoušet práci s více soubory a zdroji spojené se schopností dělat kompromisy i umět prosadit svůj názor.

Seznam použité literatury:

- https://cw.fel.cvut.cz/wiki/courses/b4b33rph/cviceni/spam/start
- testíky z hodin Řešení problémů a hry
- https://en.wikipedia.org/wiki/Spamming