SPAM FILTER

Adam Železný, David Škarda

5.1.2025

1 Úvod

V současné době probíhá velká část komunikace prostřednictvím elektrické pošty. S tímto způsobem komunikace je ovšem spojována nevyžádaná pošta, neboli spam. Cílem této práce bylo vytvořit funkční filtr v jazyce Python, který by takovou poštu rozlišil.

2 Princip

Spam filtr využívá Bayesův teorém, který je rozšířen o Laplaceovu úpravu. Samotný Bayesův teorém je naimplementován tak, že bere v potaz počet spamové pošty z testovací sady, ke kterému je připočtena četnost spamových nebo nespamových slov. Finální vyhodnocení je kromě Bayesova teorému, ovlivněno také tím, zda byla emailová adresa odesílatele již dříve označena za spamovou. V tomto případě je email automaticky označen za nevyžádaný.

3 Implementované metody

Filtr má implementovanou jednoduchou metodu $preprocess(text)$, která všechna písmena daného textu převede na malá a text rozdělí na jednotlivá slova, která pak vrátí jako seznam. Díky tomu může filtr jednoduše vypočítat četnost jednotlivých slov ve zprávě.
Mimo metodu preprocess filtr obsahuje i metodu $logarithm(x)$, která vrací přirozený logaritmus z proměnné x. Logaritmus je využíván především kvůli problému nízkých pravděpodobností, a také z důvodu zjednodušení dalších výpočtů.
Metoda $train(dir)$ slouží k učení filtru. *(viz. 4 Trénování filtru)*
Další důležitou metodou je $predict(message)$. Tato metoda slouží k předpovědění, zda je obsah zprávy spam či nikoliv. Pro určení výsledku se vypočítá skóre následujícím způsobem:

$$spam\_score = \ln(spam\_prior) + \sum_{word \in words} ln(spam\_word\_probs[word])$$

,kde *spam_score* je celkové skóre pro určení spamu, *spam_prior* je celková pravděpodobnost, že je zpráva spam bez ohledu na obsah, *words* je obsah zprávy rozdělený na slova, *word* je jedno slovo zprávy, *spam_word_probs(word)* je pravděpodobnost výskytu slova ve spamové zprávě. Obdobně metoda počítá *ham_score*. Vyhodnocení následně probíhá na základě porovnání obou skóre, při čemž je jako výsledek určeno skóre s vyšší hodnotou. Pokud tedy je $spam\_score > ham\_score$ bude zpráva vyhodnocena jako spam.
Poslední metodou je $get\_adress(file\_lines)$. Ta hledá adresu na základě předem stanoveného formátu, ve kterém se odesílatel vždy nachází na vlastním řádku, který začíná znaky *"From:"*. Řádek se rozdělí na jednotlivá slova, a pokud je v některém ze slov znak *"@"*, tak je slovo považováno za adresu odesílatele.

4 Trénování filtru

K trénování filtru je použita metoda *train(dir)* (*dir* je cesta k adresáři s daty určenými k učení). V této metodě nejdříve dojde k rozdělení trénovacích zpráv na spamové a nespamové pomocí souboru *!truth.txt*. Ve spamových zprávách najde zapomocí metody $getadress(file\_lines)$ adresy odesílatelů,

a uloží je do seznamu spamových adres. Dále dojde k vypočtení prevděpodobnosti, zda je zpráva spam či naopak bez ohledu na obsah zprávy, podle následujícího vzorce:

$$P(spam\_prior) = \frac{len(spam\_files)}{len(truth\_data)}$$

,kde *spam_prior* je celková pravděpodobnost že zpráva je spam, *spam_files* je seznam všech spamových zpráv a *truth_data* je seznam všech zpráv. Dochází i k výpočtu *ham_prior*, kdy se ve výpočtu využívá seznamu *ham_files* namísto *spam_files*. Následně metoda vytvoří dva slovníky, které obsahují četnosti všech slov, zvlášť pro spamové i nespamové zprávy. Poté se vytvoří množina všech unikátních slov, které se nachází v obou typech zpráv. Nakonec dojde k výpočtu pravděpodobností výskytu slov ve spamové zprávě s využitím Laplaceovy úpravy podle následujícího vzorce:

$$P(word) = \frac{spam\_word\_freq[word] + laplace}{len(spam\_words) + laplace * vocab\_size}$$

,kde *word* je jedno slovo z množiny všech slov, *spam_word_freq[word]* je četnost výskytu slova ve spamové zprávě, proměnná *laplace* je konstanta pro Laplaceovu úpravu (ve filtru je hodnota konstanty nastavena na 1), *spam_words* je celkový počet slov ve spamových zprávách a *vocab_size* je celkový počet unikátních slov. Podobně se pak počítá pravděpodobnost výskytu slov v nespamové zprávě.

5 Výsledky spam filtru

Výsledky filtru na třech testovacích sadách *(tyto výsledky byly získány nahráním spam filtru na web https://cw.felk.cvut.cz/brute)* jsou uvedeny v tabulce níže:

Dataset	TP*	TN**	FP***	FN****	Quality
1	438	151	2	23	0.932
2	444	140	10	6	0.846
3	584	193	7	16	0.900

*True Positive , **True Negative, ***False Positive, ****False Negative

,kde *True (False)* znamená správné (špatné) vyhodnocení a *Positive (Negative)* znamená spamová (nespamová) zpráva (FP je tedy zpráva špatně vyhodnocena jako spamová) . Lze si všimnout, že průměrná kvalita filtru je $\approx$ 0,892, což znamená, že filtr ve většině případů správně vyhodnotí, zda je obsah zprávy spam. Například na testovací sadě 1 se z celkového počtu 614 zpráv pouze 25 vyhodnotilo špatně, při čemž pouze 2 nespamové zprávy byly vyhodnoceny jako spamové. Tedy celková přesnost byla v rámci první sady $\approx$ 96%. K finálnímu výpočtu míry kvality filtru je použit následující vzorec:

$$Quality = \frac{TP + TN}{TP + TN + 10 * FP + FN}$$

, při čemž *False Positive* je vynásoben koeficientem 10, jelikož nespamová zpráva vyhodnocená jako spamová může mít velký dopad na uživatele internetové pošty.

6 Práce v týmu

6.1 Rozdělení práce

Tabulka rozdělení práce na úloze spam filter:

Kostra programu	Bayesův teorém	Spamové adresy	Finální úpravy programu
Adam Železný	Adam Železný	David Škarda	Adam Železný

pokračování tabulky:

Report	Prezentace
David Škarda	David Škarda

6.2 Organizace práce

Hlavním nástrojem pro komunikaci se stala platforma Discord, přes kterou jsme vedli téměř veškerou komunikaci. Dalším důležitým nástrojem byl web *https://pastebin.com*, díky kterému jsme byli schopni rychle sdílet upravený kód filtru.

7 Závěr

Cílem této práce bylo vytvořit spam filtr využívající Bayesův teorém s Laplaceovou úpravou, který bere v potaz adresu odesílatele. Výsledkem práce je filtr, který má průměrnou přesnost $\approx 96{,}8\%$ a dosahuje průměrné kvality $\approx 0{,}892$ *(tato data vyplívají ze 3 testovacích sad)*. To znamená, že filtr velmi spolehlivě rozlišuje spamové a nespamové zprávy. Část filtru zabývající se hledáním adresy odesílatele aktuálně závisí na předem stanoveném formátu zprávy. Míra kvality filtru je také úzce spojena s velikostí dat, na kterých se může filtr učit. S tím také souvisí délka doby učení, která se může pohybovat v řádech několika minut.

8 Zdroje

https://www.wikiskripta.eu/w/Bayesova_věta
https://cs.wikipedia.org/wiki/Bayesova_věta
https://en.wikipedia.org/wiki/Additive_smoothing