

Spam Filter

Poludina Sofia
poludsof

Pimenova Olga
pimenol1

Leden 2023

1 Úvod

Filtrování spamu je typ softwaru určený k identifikaci a klasifikaci nežadoucích e-mailů, známých také jako spam. Jednou z běžných metod pro vytváření filtru je metoda Naive Bayes, která je založena na myšlence použití pravděpodobností k určení celkové pravděpodobnosti, že daný e-mail je spam/ne spam. V této práci jsme implementovali filtr spamu pomocí Bayesovy metody v jazyce Python.

2 Algoritmus

Metoda Naive Bayes zahrnuje trénování filtru spamu na velkém datasetu označených e-mailů, což umožňuje filtru spamu naučit se charakteristiky spamu a ne-spamu. Jednou trénovaný, může filtr spamu použít tyto informace k třídění nově příchozích e-mailů jako spam nebo ne-spam s určitou úrovní přesnosti.

K provedení této metody jsme použili archiv zip se soubory získaný z [webové stránky](#) předmětu.

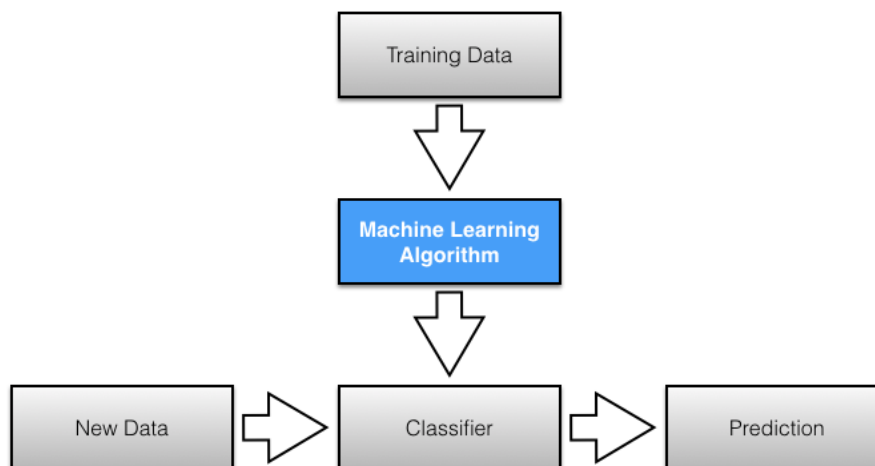


Fig. 1: Algoritmus filtru

Příprava dat

Nejprve z dokumentu odstraníme slova a znaky, které nemusí odpovídat informacím, které chceme zpracovat. E-maily mohou obsahovat mnoho nežádoucích znaků, které mohou rušit vyhledávání nevyžádané pošty. Za tímto účelem jsme odstranili interpunkční znaménka, číslovky, některá nejpoužívanější slova, zájmena a předložky.

```
our database here: <a href="http://63.205.154.38/RemoveList/RemoveMe.aspx?WHOAMI=myemail@email.com">REMOVE ME<span class="GramE">&nbsp;</span></a><span class="GramE"></span> Please allow 24hours for removal.</span></font></P><P><i style="mso-bidi-font-style: normal"><font face="Arial" size="1"><span style="FONT-SIZE: 8pt; FONT-STYLE: italic; FONT-FAMILY: Arial; mso-bidi-font-style: normal">This e-mail is sent in compliance with the Information Exchange Promotion and Privacy
```

Text 1: e-mail před úpravami

```
database allow hours removal arial size italic arial sent compliance information exchange promotion privacy
```

Text 2: e-mail po úpravách

Trénování spam filtru

Pro trénování filtru musí program načíst soubory e-mailů a jejich klasifikaci, zda je nový e-mail spam, nebo není.

Pro všechna slova v každém tréninkovém e-mailu filtr vybere pravděpodobnosti výskytu jednotlivých slov ve spamovém nebo nespamovém e-mailu ve získané databázi a vytvoří dva slovníky, do kterých se budou ukládat vygenerované pravděpodobnosti.

Vytváříme také seznam odesílatelů spamu. Použijeme jej k určení, zda je zpráva spam nebo není.

Funkčnost

Při testu první věc, kterou zkontrolujeme - zda je odesílatel v tomto seznamu odesílatelů spamu. Pokud ano, okamžitě označíme zprávu jako spam.

Jinak používáme pravděpodobnost slov vypočítanou v tréninku k výpočtu pravděpodobnosti, že e-mail s určitou sadou slov patří do jedné z kategorií. Každé slovo v e-mailu přispívá k pravděpodobnosti spamu v e-mailu. Tento příspěvek (posterior probability) se vypočítá pomocí Bayesovy věty:

$$P(\text{spam}|\text{words}) = \frac{P(\text{words}|\text{spam}) * P(\text{spam})}{P(\text{words})}$$

- $P(\text{spam}|\text{words})$ - pravděpodobnost, že zpráva je spam, za předpokladu že máme přesně tento soubor slov
- $P(\text{words}|\text{spam})$ - pravděpodobnost, že se tento soubor slov vyskytuje pouze ve spamových zprávách.
- $P(\text{words})$ - pravděpodobnost výskytu této sady slov v jakékoli zprávě bez ohledu na to, zda se jedná o spam nebo ne.
- $P(\text{spam})$ - pravděpodobnost, že zpráva je spam bez použití dalších informací o ní. (V našem příkladu je to poměr počtu přijatých spamových e-mailů k celkovému počtu e-mailů).

Vypočítáme údaje pro dvě kategorie.

Pokud je celková pravděpodobnost spamu vyšší než celková pravděpodobnost, že to spam není, filtr e-mail označí jako spam.

3 Kvalita spam filtru

Vyzkoušeli jsme napsat několik filtrů, například pomocí určitého seznamu slov pro klasifikaci spamu, náhodného filtru a dalších jednoduchých metod. A zjistili jsme, že kvalita Naive Bayes klasifikátoru je lepší než u jiných jednoduchých algoritmů.

Výsledky

V tréninkových datech jsme získali dvě složky s e-maily.

Pokud byl filtr natrénován na jednom z nich a testován na druhém, průměrná kvalita byla:

$$q = 0.77176781$$

Když byl filtr natrénován a otestován na jedné skupině e-mailů, kvalita dosáhla hodnoty:

$$q = 0.87050359$$

Po odevzdání programu do systému BRUTE jsme dostali tyto výsledky:

Výsledky kvality filtru na skupinách e-mailů:

- **dataset 1:** $q = 0.773087$
- **dataset 2:** $q = 0.759628$
- **dataset 3:** $q = 0.841875$

4 Rozdělení práce v týmu

Rozdělení odpovědnosti za psaní kódu bylo rozděleno mezi soubory:


<code>utils.py</code>	<code>filter.py</code>	<code>training.py</code>	<code>corpus.py</code>
pimenol1	poludsof	pimenol1	poludsof

- `utils.py` - načtení klasifikace mailů z textového souboru
- `training.py` - trénování dat potřebných ke klasifikaci e-mailů na základě získaných dat.
- `filter.py` - základní implementace filtru, tréninkové a testovací funkce.
- `corpus.py` - zpracování příchozích e-mailů


Na psaní reportu jsme pracovali společně, diskutovali jsme o jednotlivých tématech a upravovali text.

5 Organizace práce

Ke spolupráci na projektu jsme použili:

git [Git](#) a [Github](#)  - Sloužili ke sledování změn ve zdrojovém kódu a umožňovali nám oběma pracovat na projektu společně.

[Overleaf](#)  - pro paralelní společnou editaci textového dokumentu.

[NTK](#)  - Knihovna byla využívána ke komunikaci a diskusi o průběhu projektu offline.

6 Závěr

Pomocí jazyka Python jsme implementovali spamový filtr s použitím Bayesova algoritmu a využitím poskytnutých trénovacích dat. Celkově lze konstatovat, že spamový filtr vytvořený pomocí Bayesovské klasifikace je dostatečně účinným prostředkem pro zpracování nevyžádané elektronické pošty.

7 Zdroje

- [Wikipedia - Naive Bayes spam filtering](#)
- [Youtube Video - "Naive Bayes, Clearly Explained!!!"](#)
- [IEEE - Classification of Spam Mail using different machine learning algorithms](#)
- [GeeksForGeeks - Naive Bayes Classifiers](#)
- [Springboard - Email Spam Filtering](#)
- [DuoCircle - Popular Spam Filtering Techniques](#)
- Brain