Bayes spam filter
by Timushev/Masmalieva

1) Úvod

Naivní Bayesovi klasifikátoři fungují tak, že korelují použití tokenů (obvykle slov nebo někdy jiných věcí) se spamem a nevyžádanými emaily a poté pomocí Bayesovy věty vypočítají pravděpodobnost, že e-mail je nebo není spam.

Filtrování spamu Naive Bayes je základní technikou pro řešení spamu, která se může přizpůsobit e-mailovým potřebám jednotlivých uživatelů a poskytnout nízkou míru detekce nevyžádané pošty, která je pro uživatele obecně přijatelná. Jedná se o jeden z nejstarších způsobů filtrování nevyžádané pošty s kořeny v 90. letech.

2) Popis principu použitého spam filtru

Podstatou metody statistické filtrace je použití Bayesovské matematické věty na příchozí písmena. Tato věta vám umožňuje vypočítat pravděpodobnost úspěšného dokončení nějaké události na základě statistik této události v minulosti. Pokud jde o filtrování spamu: pokud je 9 z 10 e-mailů obsahujících notoricky známé slovo "kráva" spam, a pouze jeden je "dobrý", vypočítá Bayesova věta pravděpodobnost, že další e-mail obsahující toto slovo bude spam.

Bayesova metoda zahrnuje použití statistické základny - dvou sad ("korpusů") písmen, z nichž jedna je složena ze spamu a druhé - "dobrých" písmen. Při vytváření této základny se počítá počet výskytů každého jednotlivého slova (tokenu) v každém korpusu a na základě toho se pro každý token vypočítá skóre nebo "spam".

"Spam" se měří na stupnici od 0..1. Hodnota "0" znamená žádný spam, "1" znamená úplnou jistotu, že jde o spam. Neutrální hodnota - 0,5 - vyjadřuje absenci jakéhokoliv jednoznačného posouzení. Tokeny, jejichž "spam" je téměř neutrální, jsou pro vyhodnocení zprávy málo zajímavé. Naopak ti, jejichž hodnoty se velmi liší od 0,5, jsou nápadnými ukazateli psaní.

Nechte dopis obsahovat n tokenů s odhady S1 ... Sn. Celkové skóre písmene S pak lze snadno vypočítat pomocí následujícího vzorce:

a = S1 * S2 * ... * Sn;
b = (1-S1) * (1-S2) * ... * (1-Sn);
S = a / (a + b).

Výsledným odhadem bude hodnota "spamu" pro určité písmeno na základě stávající základy statistického odhadu.

3) Popis způsobu trénování filtru

"Vytváření inteligence" filtru bylo prováděno pomocí naivního testování založeného na příchozích informacích ve velkém objemu. To znamená použití příkladů založených na psaní prvních kroků filtru (n-tý počet souborů obsahujících obsah známý pouze testerům (vývojový tým)). Postupem testování se měnilo také množství "krájení".

4) Výsledky dosažené spam filtrem

K naší velké lítosti jeden z členů vývojového týmu (Timushev Fedor) po mezilehlém testování, jak se říká, kvůli nepředvídaným okolnostem (nikoli zákeřným záměrem), zcela naformátoval fyzické paměťové médium (pevný disk), které obsahovalo datovou sadu pro testování v plně.

V té chvíli však ukazatele filtrace uspokojily zbytek vývojového týmu, takže bylo rozhodnuto dokončit práci na základě výsledků získaných při průběžném testování a rozhodnutí o propuštění účastníků incidentu (Fedor Timushev) bylo zrušeno.

5) Stručný popis rozdělení práce v týmu

Timushev F. pracoval nad "test", Masmalieva D. pracovala nad funkcí "train". Po skončení hlavní práce proběhlo nezávislé testování, vzájemné ověření kódu všemi členy týmu, revize a oprava logických chyb. Posledním krokem byl proces přidávání komentářů a také "coding style". Algoritmus vytvořen společně.

6) Stručný popis organizace práce v týmu

V souvislosti se situací epidemie koronavirů bylo rozhodnuto vyvinout proces vývoje „vzdáleného programu“. Jako komunikační prostředky byly vybrány následující platformy (software): Discord [1], Telegram [2].

Vývojové prostředí - PyCharm (2020.2.3) [3], python 3.8 [4].

Médiem pro vedení záznamů a dokumentace je Google Docs [5].

Operační systémy: Windows 10 (2004) [6], Ubuntu 20.04 [7];

[1] https://discord.com/
[2] https://web.telegram.org/
[3] https://www.jetbrains.com/ru-ru/pycharm/
[4] https://docs.python.org/3/whatsnew/3.8.html
[5] https://docs.google.com/
[6] https://docs.microsoft.com/en-us/windows/whats-new/whats-new-windows-10-version-2004
[7] https://releases.ubuntu.com/20.04/

7) Zhodnocení, závěr

Závěrem stojí za zmínku, že zkušenosti získané během kurzu b4b33rph v dostatečné míře pomohly při vytváření jednoduchých (naivních) řešení v této práci. V konečném důsledku vývojový tým výše uvedeného softwaru doufá v adekvátní hodnocení své práce jak od vedení (učitele), tak od testovací platformy a také očekává, že v práci plně upozorní na nedostatky a rady od zkušenějších programátorů.

8) Seznam použité literatury, online zdrojů

[1] - Naive Bayes spam filtering: https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering

[2] - Spam filter - krok 6: https://cw.fel.cvut.cz/wiki/courses/b4b33rph/cviceni/spam/krok6

[3] - How to build and apply Naive Bayes classification for spam filtering: https://towardsdatascience.com/how-to-build-and-apply-naive-bayes-classification-for-spam-filtering-2b8d3308501