Vysoké učení technické v Brně
Fakulta informačních technologií

Bezpečnost informačních systémů

2. projekt – dokumentace

16. prosinec 2017

Ján Jusko

1 Úvod

Tento dokument popisuje riešenie druhého projektu v predmete Bezpečnosť Informačních Systémů (BIS). Jeho cieľom bolo zoznámiť sa s problematikou spamových správ a následne implementovať spam filter, ktorý úspešne klasifikuje vstupné súbory vo formáte .eml do dvoch skupín – *spam* a *ham*.

2 Použitie

Po prevedení príkazu *make* bude k dispozícii spustiteľný súbor "antispam"ktorý prijíma voliteľný počet parametrov. Každý parameter predstavuje cestu k súboru .eml s emailom. Po spustení sa na štandardný výstup terminálu vypíše na samostatný riadok:
název souboru - hodnocení - důvod hodnocení

3 Implementácia

Program antispam je implementovaný v jazyku python 2.7, testovaný na referenčnom stroji *merlin*.

3.1 Algoritmus

Na klasifikáciu emailov bola zvolená technika *Bayesovho naivného klasifikátora* ktorý je pre tento účel často používaná. Tento klasifikátor je založený na vypočítavaní pravdepodobnosti príslušnosti istého slova k triede spamu respektíve hamu na základe početnosti v natrénovanej množine spamov respektíve hamov. Z názvu "naivný"vyplýva že pravdepodobnosti jednotlivých slov (tokenov) sa neovplyvňujú.
Vzorec Bayesovho klasifikátora prispôsobený pre filtrovanie spamov je nasledovný:

$$\mathbf{Pr}(S|W) = \frac{\mathbf{Pr}(W|S)}{\mathbf{Pr}(W|S) + \mathbf{Pr}(W|H)}$$

Obrázek 1: Bayesov Naivný klasifikátor

Kde:

- Pr(SlW) - Pravdepodobnosť že email je spam, vzhľadom že obsahuje slovo W
- Pr(WlS) - Pravdepodobnosť že slovo W sa nachádza v spamových správach
- Pr(WlH) - Pravdepodobnosť že slovo W sa nachádza v hamových správach

Každý email je spracovaný a pre každé slovo v časti *body* je vypočítaná pravdepodobnosť Pr(SlW) a Pr(HlW) ktoré sú následne medzi sebou násobené. Dostávame teda dve hodnoty, a to pravdepodobnosti, že správa je spam a ham.
Porovnaním týchto dvoch hodnôt vyhodnotíme správu buď spam alebo ham.

4 Trénovanie klasifikátora

Bayesov klasifikátor si vyžaduje vopred natrénovať. V tomto prípade sa pod trénovaním rozumie spracovať vhodnú množinu spamu a hamu a vytvoriť z oboch tzv. *bag of words*, teda dvojice slovo:početnosť Ako databáza emailov vo formáte .eml bol použitý CSDMC2010 SPAM corpus [1] ktorý obsahuje 2953 správ ham a 1399 správ spamu prevažne v anglickom jayzyku. Výstupom trénovania sú 2 súbory pre každú skupinu, napríklad pre spamy:

- trained-spams.json - dvojice slovo:početnosť v JSON formáte
- total-spams.txt - mohutnosť trénovacej množiny spamov

5 Spracovanie vstupu

Pre parsovanie súborov formátu eml bol použitý package email.

6 Dodatočné opatrenia

Pre vylepšenie klasifikátora sa pred procesom klasifikácie odstránia so vstupného emailu tzv. stopwords, teda slová, ktoré nemajú významnú sémantickú hodnotu ako napríklad spojky, častice, atd. Dalším dodatočným vylepšením je zavedenie tzv. spam triggerov. Ak sa v klasifikovanom emaile objaví nejaké slovo z množiny slov v súbore *triggers.json*, správa sa ihneď klasifikuje ako spam. Jedná sa o typický spam žargon napr. (sex, viagra, ..).

7 Záver

V projekte bol vytvorený jednoduchý Bayesov klasifikátor ktorý je schopný klasifikovať emailové správy vo formáte .eml do skupín spam alebo ham. Klasifikátor bol počiatočne natrénovaný na množine spamu aj hamu v anglickom jazyku. Po implementácii klasifikátor vykazoval pri testovaní uspokojivé výsledky.

Obsah archívu

- main.py - zdrojový kód
- Makefile
- trained-spams.json - dvojice slovo:početnosť v JSON formáte trénované zo spamu
- total-spams.txt - mohutnosť trénovacej množiny spamov
- trained-emails.json - dvojice slovo:početnosť v JSON formáte trénované z hamu
- total-emails.txt - mohutnosť trénovacej množiny hamu
- triggers.json - list spam trigger slov
- stopwords-en.json - list anglických stopwords

Reference

[1] Csmining group. http://csmining.org/index.php/spam-email-datasets-.html.