Dokumentácia k projektu do predmetu Bezpečnosť informačných systémov: Detekcia spamu

Autor: Filip Gulán (*xgulan00*)

1 Úvod

Úlohou tohoto projektu bolo vytvoriť program v jazyku *Python* pre detekciu spamu v emailovej komunikácií. Program mal brať na svoj vstup cesty k emailovým súborom, o ktorých mal rozhodovať, či sa jedná o nevyžiadanú poštu.

2 Problematika

Termín spam označuje nevyžiadanú a hromadne rozosielanú správu, ktorá sa šíri internetom (e-mail, diskusné fóra, komentáre, instant messaging). Emailový spam je posielanie nevyžiadaných e-mailových správ, často s komerčným obsahom, veľkému množstvu príjemcov. Používa sa taktiež skratka *UBE/UCE* (*Unsolicited Bulk/Commercial Email*). Nevyžiadaná pošta začala byť problémom v 90. rokoch, kedy bol internet otvorený verejnosti. Od tej doby množstvo nevyžiadanej pošty exponenciálne rastie a momentálne tvorí asi 80-85% všetkých e-mailov. Spam je v súčastnosti odosielaný zväčša cez zombie siete. Detekcia spamu je veľmi náročná a nie je stopercentná. Existuje niekoľko prístupov k detekcií spamu, ktoré sa spolu kombinujú. Detekcia a následná filtrácia môže prebiehať podľa spôsobu dopravy mailu (Blacklisting, Greylisting), podľa obsahu (filtre založené na pravidlách, filtre založené na učení) a na strane príjemcu (*Sender Policy Framework*, *DomainKeys Identified Mail*). Opak spamu, teda pošta, ktorá je považovaná za žiadúcu sa označuje ako ham. [1]

3 Popis testovacej databáze

Pre overenie funkčnosti programu bola zvolená databáza *CSDMC2012 SPAM*, ktorá je jednou z databáz pre súťaž v dolovaní dát *ICONIP 2010*. Databáza obsahuje dáta určené pre trénovanie a testovanie vo formáte *EML*. Trénovacie dáta sú v počte 4327 emailov, z toho je 2949 ham správ a 1378 spam správ. Testovacie dáta nemajú uvedenú svoju príslušnosť k hamu, alebo spamu. Pre účely projektu boli využité iba trénovacie dáta. [2]

4 Popis použitých knižníc

V rámci projektu bola použitá iba jedna neštandardná knižnica, ktorá nesie meno *eml_parser* a je dostupná z https://github.com/GOVCERT-LU/eml_parser. Knižnica slúži na analýzu súborov typu *EML* a je vďaka nej možné získať rôzne nájdené, vypočítané, či odvodené informácie z mailu.

5 Popis alternatív

V rámci tohoto projektu, bola najprv skúšaná metóda detekcie spamu na základe strojového učenia. Konkrétne sa využíval *Bayesovský naivný klasifikátor*, ktorý sa najprv natrénoval na ⅔

databáze popísanej v kapitole 3 a následne sa testovala jeho funkčnosť na zostávajúcej ⅓ databáze. Výsledky na danej databáze boli pomerne uspokojivé. Klasifikácia dosahovala úspešnosť okolo *90%* pri detekcií spamu a *98%* pri detekcií hamu (lemizácia nebola vykonávaná, stop slová boli detekované a vymazávané). Avšak po naučení a spustení klasifikátoru nad dátami, ktoré boli ponúknuté v rámci zadania, klasifikátor vyhodnotil ⅓ hamu ako spam (čo môže byť zapríčinené ne anglickým jazykom e-mailu) a spam správne označil iba v *20%.* Z čoho bolo usúdené, že na tento projekt metódy založené na učení nie sú úplne vhodné, keďže český/slovenský otvorený dataset mailov nebolo možné na internete zohnať.

6 Popis riešenia

Projekt bol nakoniec riešený metódou detekcie zakázaných slov, ktoré sa nachádzajú v predmete mailu. Zakázané slová boli získané z rôznych internetových zdrojov a boli následne spracované. Všetky internetové zdroje, z ktorých boli slová získané sú odkazované priamo v súbore *black_words.txt*. Spracovanie týchto slov prebiehalo zväčša ručne, kedy sa prechádzalo slovo po slove a vyhadzovali sa potenciálne nechcené (hlavne mená pornoherečiek). Vzniknutý súbor so zakázanými slovami obsahuje viac ako 1175 slov a slovných spojení. Ide zväčša o slová z porno odvetvia. Či už priamo slová a slovné spojenia, alebo rôzne úmyselne schmolené verzie, aby boli ťažko detekovateľné antispamom (napr. slovo *b00bs*), alebo rôzne iné explicitne ladené slová. Súbor obsahuje iba naozaj slová, vysoko pravdepodobné pre spam poštu. Ne predpokladá sa, že užívatelia si medzi sebou posielajú správy s ponukami sexu a podobne.

Program rozhoduje o e-mailoch, ktoré sú zadané ako argumenty. Tieto e-maily sa najprv otvoria a následne rozparsujú knižnicu *eml_parser*. Získa sa konkrétne predmet mailu vo svojej originálnej forme a následne vo forme poľa slov. Tento predmet sa následne porovnáva so slovami, ktoré sú uvedené v spomínanom súbore *black_words.txt*. Ak sa nejaké slovo z predmetu nachádza v tomto súbore, tak je e-mail označený ako spam a je na výstup zobrazené slovo, ktoré toto zaradenie zapríčinilo.

7 Záver

Program bol riadne testovaný na operačnom systéme *Zorin OS 12* a na školskom systéme *CentOS*. Behom testovania sa nevyskytli žiadne chyby a všetky navrhnuté testy dopadli úspešne. Vyhodnocovanie úspešnosti detekcie spamu prebiehalo pomocou dátovej sady popísanej v kapitole 3. Program na tejto dátovej sade správne označil spam v *40%*, a ham v *90%*. Program na dátovej sade, ktorá bola dostupná v rámci zadania, správne detekoval *43%* spamu a *100%* hamu.

Referencie

[1] Spam. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2017-12-03]. Dostupné z: https://cs.wikipedia.org/wiki/Spam

[2] Spam email datasets. *Csmining Group* [online]. New Zealand [cit. 2017-12-03]. Dostupné z: http://csmining.org/index.php/spam-email-datasets-.html