VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií

Bezpečnost informačních systémů

Program pro detekci spamu v e-mailové komunikaci

December 17, 2017

Autor: Jan Pawlus, xpawlu00@stud.fit.vutbr.cz

1 Princip funkčnosti

Můj e-mailový filtr je složen ze tří modulů - `bis.py`, kde je spuštěna logika, `ArgParser` (`argParser.py`), kde probíhá zpracovávání jednotlivých argumentů (e-mailů) a parsování, a `EmailAnalyzer` (`analyze.py`), kde probíhá analýza a aplikování regulárních výrazů na jeden e-mail. Projekt jsem implementoval v jazyce *python3*.

2 ArgParser

V tomto modulu probíhá zpracování jednotlivých e-mailů poskytnutých od uživatele jako argumenty skriptu. Pokud je argument zadán správně a soubor se podaří načíst (v opačném případě skončí analýza výstupem `FAIL`), je využita knihovna *eml-parser*, která e-mail zpracuje vstupní soubor do lepší struktury (`dictionary`). Každý e-mail ještě skript upraví tak, že substituuje diakritiku standardními znaky. Tento převod zajišťuje knihovna *unidecode*. Díky tomu je možné lépe analyzovat české či slovenské e-maily.

3 EmailAnalyzer

Analýza jednotlivých e-mailů funguje na principu aplikování několika regulárních výrazů rozdělených dle závažnosti a významu na e-mail. Každý takový regulární výraz tedy obsahuje významově podobná slova a má přiřazen určitý koeficient, neboli skóre. Pokud tedy některý z regulárních výrazů uspěje, přičte se k celkovému skóre e-mailu tento koeficient (znásobený počtem výskytu klíčových slov daného regulárního výrazu). Vyhodnocení pak funguje jako porovnání celkového skóre s určitým prahem - pokud je skóre vyšší, než práh, e-mail je označen jako spam.

3.1 Příklady aplikovaných významových regulárních výrazů

Aby bylo lépe pochopitelné, jak skript funguje, je třeba se podívat na určitý regulární výraz. Například detekce klíčových slov spojené se sexem je rozděleno do tří regulárních výrazů. V prvním, které má nejvyšší koeficient, jsou obsaženy slova jako `naked photos` nebo `suck your dick`. Pokud jsou taková slova v e-mailu nalezena, je v podstatě rovnou označen jako spam.

Ve druhém regulárním výrazu, který již při nalezení dává e-mailu nižší skóre, tedy k označení jako spam je nutné nalézt alespoň dvě taková slova, jsou obsaženy fráze jako `erotic`, `premature ejaculate` nebo `enlarging oil`. V nejméně závažném fráze jako `your husband` nebo `your wife`. Tímto je zlepšena ochrana před označením validního e-mailu jako `SPAM`, pokud se v e-mailu vyskytne méně závažná fráze.

Podobným stylem jsou ošetřeny i další typy nevyžádané pošty, jako léky, dědictví, výhry v loteriích, snadné triky k vydělání peněz a jiné. Tyto okruhy a klíčová slova jsem dal dohromady pouze pocitově z osobní analýzy různých datových setů se spamem. Většinu klíčových slov jsem se snažil také překládat do češtiny či slovenštiny, aby skript dokázal detekovat také náš spam.

3.2 Detekce klíčových slov bez významu

Tento skript nepracuje pouze s hledáním významových frází - snažil jsem se také aplikovat některá z pravidel z projektu `SpamAssassin`. Kupříkladu pokud doména odesilatele obsahuje slova jako `offer`, skóre e-mailu narůstá. Dalším příkladem může být například *HTML* odkaz, který odkazuje na určitou IP adresu (využívané například u *phishingu*), nebo pokud e-mail má pouze *HTML* část, ale ne *plaintext*. Skript také přičítá malé skóre, pokud e-mail obsahuje větší množství slov napsaného v kapitálkách.

4 Výstup a použití

Skript obsahuje soubor `Makefile`, který vytvoří symbolický odkaz. Poté je možno s programem pracovat jako se spustitelným souborem `antispam`. Skript ověří všechny e-maily zadané argumenty. Pokud soubor neexistuje, vypíše k němu skript `FAIL`, následně proběhne analýza, kdy skript k zadanému e-mailu vypíše `SPAM` či `OK`. Pokud je e-mail označen jako `SPAM`, jsou vypsána i klíčová slova, na základě kterých skript e-mail jako `SPAM` vyhodnotil (díky formě aplikování regulárního výrazu funkcí `findall` má skript přístup k těmto klíčovým slovům).