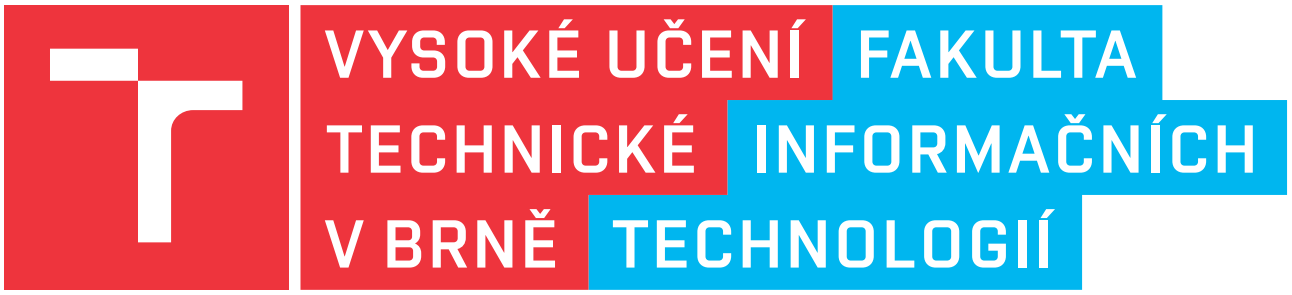

Dokumentace k projektu č. 2

Bezpečnost informačních systémů

Nikola Valešová (xvales02)

17. prosince 2017

1 Úvod

Cílem projektu je vytvořit program *antispam*, který bude vstupní e-maily klasifikovat do dvou tříd – ham a spam. Činnost programu lze logicky rozčlenit na několik fází, kterými jsou otevření e-mailu, načtení jeho obsahu a zpracování pomocí knihovny *eml_parser*, klasifikace e-mailu na základě vlastností hodnot jeho atributů a následně výpis výsledku klasifikace.

2 Použité knihovny

eml_parser Pro zpracování vstupního e-mailu je využita funkce `decode_email_b` ze knihovny *eml_parser*. Pokud se toto zpracování nezdaří úspěšně, e-mail je automaticky klasifikován jako spam.

BeautifulSoup4 Ze knihovny BeautifulSoup4 jsou využity funkce pro zpracování textu obsahujícího HTML. Tato knihovna je použita k extrakci HTML značek z textu za účelem získání délky e-mailu bez HTML kódu a dále pro určení počtu fontů definovaných v rámci e-mailu.

3 Způsob klasifikace

Klasifikace probíhá na základě výpočtu skóre daného e-mailu. Za každý negativní rys jsou klasifikovanému e-mailu přičteny záporné body, jejichž součet je na závěr porovnán s hodnotou prahu. Pokud skóre přesáhne tuto hodnotu, je e-mail klasifikován jako spam, jinak je považován za ham. V této sekci budou dále vypsány všechny kontrolované atributy a způsoby jejich penalizace.

3.1 Nevalidní odesílatel a příjemce

Pokud je pole odesílatele či příjemce prázdné, připočte se určitá hodnota k celkovému skóre. U odesílatele je dále kontrolován výskyt zavináče.

3.2 Nevhodná slova

Předmět e-mailu i obsahy všech částí těla e-mailu jsou podrobeny kontrole výskytu slov, která se často vyskytují ve spamech, jinak také označovaná jako *spam trigger words*. Kontrola je rozdělena na dvě části, nejprve se vyhledají slova nevhodná, poté slova zakázaná, která mají několikanásobně vyšší penalizaci. Získané skóre je poté poděleno délkou obsahu dané části e-mailu bez kódu HTML a výsledek přičten k celkovému skóre daného e-mailu.

3.3 Počet slov psaných velkými písmeny

Pro předmět i obsah těla e-mailu je vypočten poměr slov psaných velkými písmeny ku počtu slov celé části opět bez HTML.

3.4 Nestandardní znaky

Dále je spočten počet nestandardních znaků, tedy znaků, které nejsou součástí anglické ani české abecedy, ani se nejedná o nějaký jiný symbol běžného textu. Jejich počet je opět podělen délkou celého textu bez HTML.

3.5 Jednoznaková slova

Je spočten poměr slov, která mají pouze jeden znak, kvůli časté technice spamů, kdy jsou znaky zakázaných slov rozděleny mezerami.

3.6 HTML kód

V obsahu je spočten počet fontů definovaných v HTML kódu. Dále jsou připočteny záporné body, pokud HTML kód obsahuje značku `bgsound` a pokud obsahuje nějaké neuzavřené HTML značky.

3.7 Webové odkazy

Je připočteno skóre za počet odkazů, které jsou vnitřně rozděleny mezerami, dále je spočten celkový počet odkazů a připočten ke skóre a také jsou připočteny záporné body, pokud odkaz odkazuje na stránku na serverech „goo.gl“ a „bit.ly“, nebo stránku s doménou „ru“.

3.8 Podezřelé znaky

Ke skóre je dále přičten poměr výskytů podezřelých znaků, jako jsou četné vykřičníky, otazníky, symboly dolaru, apod.

3.9 Prázdné řádky

Tělo e-mailu je podrobeno kontrole počtu prázdných řádků, které obsahují pouze bílé znaky, a je vypočten postih za velký poměr prázdných řádků vůči celé délce e-mailu.

3.10 Prázdné tělo

Pokud je tělo prázdné, nebo zcela chybí, je opět připočten adekvátní postih.

3.11 Message-id

V poli message-id je kontrolován počet znaků zavináč a také, zda není celý unikátní identifikátor složen z velkých písmen.

3.12 Příloha

Za každou přílohu je připočteno drobné skóre, stejně tak pokud se v názvu přílohy nevyskytuje tečka.

3.13 Výstup aplikace

Pokud se e-mail nezdaří otevřít, ať už z důvodu, že není nalezen na disku, nebo kvůli nedostatečným přístupovým právům, vypíše aplikace zprávu „FAIL“. Aplikace vypíše „OK“, pokud nebylo dosaženo skóre potřebné ke klasifikaci e-mailu jako spam. V opačném případě je výstupem „SPAM“, za nímž následuje seznam atributů, díky kterým byl e-mail do této kategorie zařazen. U konkrétních slov se jedná o slovní spojení „CONTAINS nalezene_slovo“, v případě ostatních klasifikačních atributů je vypsáno pro přehlednost pouze číslo příslušného atributu. O jaký postih se jedná lze nalézt v souboru `spam_signs.py`.

4 Výsledky aplikace

Aplikace byla testována na několika sadách e-mailů, které jsou volně dostupné na internetu. Patří sem sady `enron1` až `enron6`[1], sady `20021010` a `20030228`[2] (s výjimkou složek s názvy obsahujícími slovní spojení „hard ham“) a také na sadě poskytnuté v zadání tohoto projektu. Přehled výsledků, jichž aplikace *antispam* dosahuje na testovaných sadách, je uveden v následující tabulce.

Název testovací sady	Úspěšnost spamů	Úspěšnost hamů	Počet bodů
enron1	580 (38,57 %)	3484 (94,88 %)	6
enron2	607 (40,57 %)	4083 (93,63 %)	7
enron3	549 (36,60 %)	3760 (93,72 %)	6
enron4	1711 (38,02 %)	1455 (97 %)	6
enron5	1494 (40,65 %)	1357 (90,47 %)	5
enron6	1505 (33,44 %)	1376 (91,73 %)	4
20021010	322 (64,27 %)	2486 (97,45 %)	8
20030228	1277 (67,28 %)	3781 (96,92 %)	8
Sada ze zadání	15 (71,43 %)	3 (100 %)	8

5 Závěr

Výsledná aplikace *antispam* dosahuje na testovacích sadách relativně dobrých výsledků. Chyby v klasifikaci jsou často způsobeny výskytem zcela nevhodných vulgárních slov v e-mailech, jenž by měly být klasifikovány jako ham. V opačném případě nastává problém se správnou klasifikací spamů často v případě, že e-mail obsahuje velké množství nesmyslných slov, čímž se výrazně sníží penalizace za nevhodná slova nalezená ve validní části e-mailu. Nabízí se snaha tato nevalidní slova detekovat a odstranit, avšak z důvodu nutnosti akceptace i jiných jazyků, než je angličtina, by toto rozšíření mohlo přinést více škody než užitku.

[1] Dostupná na adrese: https://labs-repos.iit.demokritos.gr/skel/i-config/downloads/enron-spam/preprocessed

[2] Dostupná zde: http://spamassassin.apache.org/old/publiccorpus