BIS 2017

Projekt 2

Autor: Ondřej Valeš
Login: xvales03

Datum vytvoření: 15. 12. 2017

Návrh programu

Implementovaný antispam patří do skupiny antispamů pracujících na základě pravidel. K ohodnocení spamu používá tříúrovňový seznam zakázaných slov (`en_blacklist_low`, `en_blacklist_med` a `en_blacklist_high`), přičemž slova patřící do seznamu na vyšší úrovni mají větší bodové ohodnocení.

Seznamy obsahují slova považovaná za „spam trigger words“ v anglickém jazyce. Aby program dokázal detekovat i česky psaný spam byly do seznamů `en_blacklist_med` a `en_blacklist_high` přidány české překlady nejběžnějších slov vyskytujících se ve spamu.

Program prohledává předmět a tělo emailu a počítá výskyty zakázaných slov. Bodové ohodnocení nalezeného slova je násobeno konstantou, a to pokud je celé slovo napsané velkým písmem nebo pokud se nachází v předmětu zprávy. Toto ohodnocení tvoří hlavní rozhodovací mechanizmus programu.

Dále jsou kontrolovány pomocné příznaky. Do této skupiny patří: prázdná nebo neplatná adresa odesilatele, prázdná nebo neplatná adresa příjemce, prázdný předmět, prázdné tělo, velký počet odkazů v textu, odkazy vedoucí na stránky s podezřelou doménou a velký poměr HTML kódu vůči velikosti zprávy. Každý nalezený příznak je penalizován konstantním ohodnocením.

Součet skóre a penalizací je potom porovnám s délkou emailu a konstantní hodnotou, pokud je celkové skoré větší než jedna z těchto hodnot, je email označen za spam.

Implementace

Program je implementován v jazyce Python. Pro načtení emailu ze souboru používá knihovnu `eml_parser` a pro odstraňování HTML kódů `BeautifulSoup`. Knihovna `BeautifulSoup` není dostupná na serveru `merlin` a je doinstalována při zadání příkazu `make`.

Implementované součásti jsou popsány v kapitole Návrh programu a využívají následující ohodnocení:

word_low	1 b
word_med	15 b
word_high	70 b
header_multiplier	× 5
upper_case_multiplier	× 5
no_body	2 b
no_head	2 b
no_sender_reciever	2 b

Kontrola podezřelé domény se ukázala jako kontraproduktivní a při implementaci byla vypuštěna. Stejně tak poměr HTML kódů vůči textu, protože mnoho ham ho obsahovalo velké množství.

Celkové porovnání je potom provedeno následujícím vzorcem:

$$score + penalizations \geq size * \ percentage \ OR \ score + penalizations \geq \ threshold$$

kde `size` je počet slov v emailu, `percentage` bylo určeno jako 0,08 a `threshold` jako 90 bodů.

Testování

Program byl testován na sadách emailů programu SpamAssassin (spam + spam_2 + easy_ham_2), sadách Enron (enron1, enron2, enron3, enron4 a enron5) a na sadě českých emailů dodané v zadání projektu. Všechny sady obsahovaly převážně emaily v anglickém jazyce, sada obsahující i česky psané spamy se mi nepodařila nalézt.

Odkazy na testovací sady:
SpamAssassin – http://spamassassin.apache.org/old/publiccorpus/
Enron – https://labs-repos.iit.demokritos.gr/skel/i-config/downloads/enron-spam/preprocessed/

Výsledky testování:

Sada	Počet spam/ham	Výsledky spam/ham (správně zařazené)	Bodové hodnocení dle zadání
K projektu	21 / 7	67% / 100%	8
Spamassassin	1896 / 2501	61% / 94%	7
Enron1	1500 / 3672	37% / 99%	6
Enron2	1496 / 4361	43% / 98%	7
Enron3	1500 / 4012	39% / 94%	6
Enron4	4500 / 1500	37% / 99%	6
Enron5	3675 / 1500	43% / 99%	7
Celkem	14588 / 17549	42% / 97%	7

Závěr

Program vykazuje stabilní výsledky na testovacích sadách Enron a testovací sadě dodané k projektu, kde detekuje 37% až 43% veškerého spamu a správně označuje více jak 94% validních emailů. Jediný výkyv ve výkonnosti nastal u testovací sady SpamAssassin, kde množství detekovaného spamu vzrostlo na 60%.

Tuto nekonzistenci ve výsledcích mezi testovacími sadami Enron a SpamAssassin se nepodařilo zcela odstranit a je značně ovlivněna nastavením porovnávacího prahu. Při zvýšení porovnávacího prahu dává výborné výsledky sada SpamAssassin, naopak při snížení Enron.

Ve výsledné implementaci je hodnota porovnávací prahu kompromisem zajišťujícím pro všechny testovací sady minimum detekovaného spamu 37% a správně detekovaného hamu 93%.

Při testování na velké sadě emailů (celkem 30 000) tyto výkyvy ve výkonosti zmizí a program detekuje přes 40% veškerého spamu a správně propouští 97% validních emailů.