Linnéuniversitetet
Kalmar Växjö

Examensarbete, Nivå B

Bayesisk filtrering i syfte att motverka spam

En studie om bayesisk filtrering i olika programvaror

Författare: Andreas Bengtsson, Johan Kindstrand, Stefan Persson
Handledare: Marcus Wilhelmsson
Examinator: Jacob Lindehoff
Termin: VT13
Ämne: Datavetenskap
Nivå: B
Kurskod: 1DV41E

Linnéuniversitetet
Kalmar Växjö

Abstrakt

Ett konstant problem med e-post är mängden skräppost som skickas dagligen och bidrar till en osäkerhet bland hemanvändare samt medför stora kostnader för företag. Att kunna skydda sig och filtrera bort skräppost är av stor vikt. Vad är egentligen skräppost?
Programvaror mot skräppost använder flera metoder för att lösa problemet. Arbetet behandlar en av dessa metoder och hur effektivt den används i olika programvaror. Den metod som arbetet fokuserar på är bayesisk filtrering och programvarornas förmåga att utnyttja den. I studien kommer en analys huruvida Spamassassin och GFI MailEssentials utnyttjar bayesisk filtrering utföras. Tester kommer att genomföras med samma förutsättningar på de två programvarorna, det vill säga alla filter och skydd kommer att vara inaktiverade förrutom bayesisk filtrering. Testerna kommer att ge resultat som sedan analyseras där effektiviteten av filtret visar sig.

Nyckelord

Spam, Ham, Skräppost, E-post, MailEssentials, SpamAssasin, Bayesisk filtrering

Tack

Vi vill tacka vår handledare Marcus Wilhelmsson för alla tips och hjälp med skrivandet av arbetet.

Abstract

A constant problem with email is the amount of spam sent daily that contributes to uncertainty among home users as well as imposing significant costs on businesses. Being able to filter and protect against spam is of great importance. What is really spam?
Antispam software uses several methods to solve the problem. This thesis work addresses one of these methods and how efficiently it is used in different software. The method that this thesis work focus on is Bayesian filtering and softwares abilities to utilize it. In the study an analysis how efficiently GFI MailEssentials and SpamAssasin utilize Bayesian filtering. Tests will be conducted with the same conditions on the two software products, ie all filters and protection will be disabled except bayesian filtering. The tests will provide results for analyzes where the efficiency of the filter proves.

Innehåll

1 Introduktion

Avsnittet innehåller en kort inledning och även tidigare forskning kring ämnet. Syftet med arbetet och även avgränsningar kommer att diskuteras vidare. Arbetet behandlar skillnaderna mellan olika programvarors bayesiska filter och hur deras inlärningsförmåga påverkar klassificeringen av e-post.

1.1 Inledning

Skräppostfiltrering är ett viktigt element vid hanterande av e-post, både ur privat- och ett företagsperspektiv. Kostnaderna för all skräppost uppgår till väldigt höga summor och bara i USA rapporterades det att kostnaderna uppgick till 20 miljarder dollar år 2011. Enligt statistik från år 2010 skickades cirka 100 miljarder e-postmeddelanden varje dag, där andelen skräppost uppgick till 88 procent [1].

Bayesisk filtrering är ett sätt för e-postservrar att kunna filtrera e-post beroende på innehåll, samt att det lär sig hur skräppost är uppbyggt och vilka ord som är vanligt förekommande i skräppost [10]. Tillsammans med andra metoder som blacklist, greylist och virusskanning utgör det ett starkt skydd mot skräppost.

Ett problem som förekommer med skräppostfilter är att de kan ge ut många så kallade falska positiva meddelanden vilket inte är önskvärt när ett skräppostfilter ska implementeras. Med hjälp av bayesisk filtrering försöker programmet lösa problemet då den lär sig av vad som kommer in till e-postservern och även vad som skickas, genom att gå igenom ett antal meddelanden [10]. Metoden kan således anpassas efter vad som är skräppost och legitim e-post. En utförligare beskrivning återfinns under teknisk bakgrund.

1.2 Tidigare forskning

Tidigare forskning kring ämnet har gjorts, men jämförelser mellan olika programvarors förmåga att filtrera med hjälp av bayesisk filtrering har det ej forskats kring. I arbetet (*An Evaluation of Naive Bayesian Anti-Spam Filtering Techniques, 2007)* [7] görs en jämförelse på olika klassifikationer av bayesisk filtrering samt hur de skiljer sig från varandra. I samma dokument beskrivs även hur de går till väga för att testa sina teorier. Deras slutsats är att spam försöker utvecklas i samma takt som spamfilter utvecklas.

1.3 Problemformulering

Arbetet kommer att behandla skillnaderna mellan olika e-postservrars förmåga att filtrera skräppost med hjälp av bayesisk filtrering. Arbetet fokuserar på hur effektivt programvaror nyttjar bayesisk filtrering genom att undersöka skräppostens innehåll, reagerar programvarorna på samma sätt? Kommer resultatet skilja sig mellan programvarorna? Är det möjligt att med endast bayesisk filtrering uppnå ett fullständigt skydd?

1.4 Syfte

Arbetets syfte är att undersöka olika programvarors förmåga att hantera skräppost. Flera metoder används vid bekämpning av skräppost. Bayesisk filtrering är en av metoderna, vilket arbetet kommer att handla om. Undersökningen kommer ge en djupare förståelse kring hur bayesisk filtrering fungerar och hur det fungerar i de olika programvarorna.

Den tilltänkta målgruppen är personer eller företag som är intresserade att se bayesisk filtrerings effekt i olika programvaror. Frågeställningar som tas upp under problemformuleringen kommer att besvaras i arbetet.

1.5 Avgränsningar

Det finns många inställningar och parametrar som kan konfigureras i de olika programmen, vilket kan påverka resultaten av testerna. Därför används filtrets standardinställningar för att kunna utföra ett så likvärdigt test som möjligt. Optimering av program kan utföras efter verksamhetskrav, men arbetet behandlar inte det. Programvarornas utformning och implementation av den bayesiska algoritmen på programkodsnivå är något vi ej kommer att gå in på, då undersökningen behandlar programvarornas effektivitet gällande bayesisk filtrering.

2 Teknisk Bakgrund

Ordlista
Bayesisk filtrering - Ett skräppostfilter som använder sannorlikhetslära.
Blacklisting - En databas med IP-adresser som kan användas för att identifiera anslutningar som bör blockeras.
Body - Meddelandets body är själva meddelandet som avsändaren vill skicka och är separerat från headern med en linjebrytning
Greylistning - En metod att kunna motverka spam. MTA använder grålistning och avvisar tillfälligt meddelanden från en ny e-postserver. Om ett meddelande är legitimt kommer servern att vänta en angiven tid och sedan skicka igen.
Ham - Önskvärd e-post och anses ej vara spam.
Header - En del i ett e-postmeddelande med information om hur det ska skickas.
Header check - En analys av e-post headern där vägen meddelandet har tagit för att nå destinationen kontrolleras.
IP reputation - Kontrollerar varje anslutningsbegäran mot en databas med IP-adresser för att fastställa om en avsändare är legitim eller en känd spamavsändare.
MTA - Message Transfer Agent, är mjukvara som skickar elektroniska meddelanden från en dator till en annan genom att använda en klient-server arkitektur.
Phishing - En olaglig metod för att lura till sig elektroniska resurser eller annan känslig information.
Spam - Oönskade meddelanden via elektronisk kommunikation.
Whitelisting - Adresser som ligger i whitelist tillåts alltid, oavsett vad meddelandet har för innehåll.

2.1 Operativsystem

Kapitlet behandlar information och bakgrund till de operativsystem som används i de olika testmiljöerna.

2.1.1 Ubuntu

Linux var redan väletablerat på marknaden år 2004. Fri mjukvara var fortfarande inte en stor del av användandet, det var med det argumentet som Mark Shuttleworth satte ihop ett lag av utvecklare från Debian för att kunna skapa ett enklare operativsystem, därav skapades Ubuntu [11].

Ubuntu släpper fortfarande nya versioner var sjätte månad och med varannat år följer även ett LTS (long-term support) med. LTS är det supportsystem som följer med gratis som Ubuntu använder sig av [11].

2.1.2 Microsoft Windows Server 2008

Windows Server 2008 R2 är ett avoperativsystemen som används i arbetets testmiljön. Active Directory släpptes i en förbättrad utgåva med Windows Server 2003 och används än idag, Active Directory är en katalogtjänst och är till för att agera som en centraliserad plats att spara information om bland annat nätverksenheter, användare, datorer och grupper [2].

2.2 Programvara

Kapitlet behandlar information om de programvaror som används i undersökningen. Valet av programvara har grundats på att tester ska utföras i både Linux- och Windows-miljö. Möjligheten att kunna stänga av samtliga filter utom bayesisk filtrering var ett krav. De valda programmen är väletablerade på marknaden.

2.2.1 SpamAssassin

SpamAssassin är ett e-postfilter som är skapat för att identifiera skräppost. SpamAssassin är ett intelligent filter som använder flera olika tester för att identifiera om e-post är legitimt eller om det är identifierat som skräppost. Med hjälp av statistiska metoder utförs tester på meddelandets header och body för att klassificera det. SpamAssassin är skapat för att kunna användas på vilket e-postsystem som helst. SpamAssassin använder flera olika metoder för att bestämma vad som är spam. Några av dessa är bland annat blacklists, header tests, whitelists och bayesisk filtrering [3].

2.2.2 GFI MailEssentials

GFI MailEssentials är ett prisbelönat säkerhets- och antispamprogram för Windows Exchange Server. Programmet skyddar nätverk från e-postvirus och andra hot från skadlig kod och har enligt studier fångat in över 99% av all spam [4].

GFI MailEssentials filtrerar bort spam, phishing och virus genom att använda flera olika tekniker, varav några av dessa är fem stycken antivirusskanningsmotorer och flera anti-spam filters, varav en av dem är bayesisk filtrering. För att nämna några av filterna använder programmet IP reputation, greylisting, och flera andra [4].

GFI MailEssentials är ett modulärt e-postprogram då möjlighet finns att själv bestämma vilka metoder som ska användas för antispam, men även i vilken ordning filtreringen ska ske. För att nämna ett exempel kan whitelisting användas för att filtrera före bayesisk filtrering, då kan användaren ställa in det i programmet [5].

2.3 E-postservrar

Kapitlet behandlar information om de två e-postservrar som används i arbetet.

2.3.1 Postfix

Postfix är standard MTA (Message Transfer Agent) för Ubuntu. Den finns i huvuddatabasen för Ubuntu, vilket betyder att den ständigt får säkerhetsuppdateringar. Postfix skapades när Wietse Venema behövde en snabb och säker MTA. Istället för att använda en tidigare programvara skapade han postfix, eller vMailer som det kallades från början. Postfix lanserades år 1997 och används fortfarande idag [8].
År 2012 rapporterade E-soft, inc, att ungefär 23% av alla publika e-post-servrar på Internet använder sig av Postfix [9].

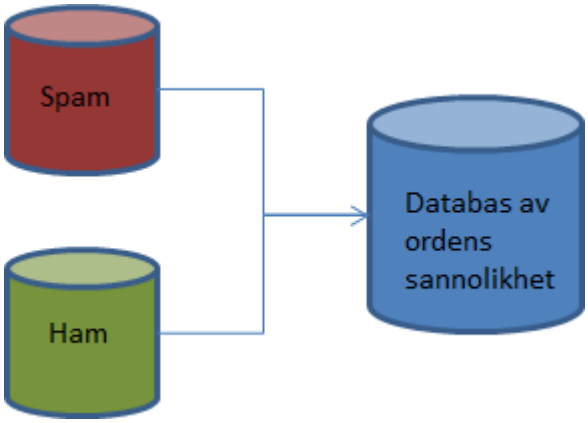
2.3.2 Microsoft Exchange Server 2010

Exchange Server 2010 är en e-postserver som körs på Windows Server 2008 och som sin föregångare Exchange Server 2007 även kan integreras med telefonsystem. Det är den sjunde versionen av produkten och fast det inte är särskilt revolutionerande har den ett par nyheter och förbättringar från sin föregångare. Skalbarheten i Exchange Server

2010 har blivit förbättrad jämfört med Exchange Server 2007:s komplexa förvaringskrav.
Exchange Server finns i två versioner, Standard Edition och Enterprise Edition [6].

2.4 Bayesisk filtrering

Bayesisk filtrering i e-postsammanhang baseras på en textbit eller ett ord ofta förekommande i skräppost men inte i legitim e-post. För att ett bayesiskt filter ska fungera måste det ha en databas med olika textsträngar eller ord från både spam (skräppost) och ham (legitim e-post). Ett ord även kallad en token får sedan ett värde av filtret och det är beroende på förekomsten och redan förinställda värden. För att kunna analysera vad som är ham och vad som är spam måste filtret ha både spam och ham som den kan analysera för att få ett bättre resultat, visat i *figur 2.1*. Det är med det här värdet som filtret sedan baserar sina beslut på.
När databaser har skapats med information om ham och spam kan sedan beräkningar påbörjas och filtret kan användas. När ett nytt e-postmeddelande kommer, delas det in i ord och de som är mest relevanta för att bestämma om det ska räknas till spam eller ham blir utpekade. Det är från de här orden bayesiska filtret sedan räknar ut om det är spam eller inte.
En fördel med bayesisk filtrering är att den lär sig själv. Eftersom bayesisk filtrering kan lära sig själv är det också flexibelt. Ett bayesiskt filter är svårt att lura då den lär sig själv snabbt om nya knep för att ta sig förbi de olika regelverken som filtret har [10].



Figur 2.1: *Bayesisk filtrering, Spam och Ham utgör uträkningsdatabasen.*

2.4.1 Matematiska algoritmen

Enligt (*The Bayesian Spam Filter with NCD*, 2013*) [12] beskrivs sannolikheten att ett e-post är spam beroende på innehållet av särskilda ord med hjälp av följande formel:

$$\Pr(S \mid W) = \frac{\Pr(W \mid S \; x \; \Pr(S))}{\Pr(W \mid S) \; x \; Pr(S) + \Pr(W \mid H) \; x \; \Pr(H)}$$

Förklaring av formeln:
Pr(S | W) är sannolikheten att ett e-post är spam med vetskapen att den inehåller det analyserade ordet.
Pr(S) är den totala sannolikheten att ett e-post är spam
Pr(W | S) är sannolikheten att det analyserade ordet förekommer i spam

Pr(H) är sannolikheten att ett e-post inte är spam
Pr(W | H) är sannolikheten att det analyserade order förekommer i ham

Ett ords värde ligger mellan 0.0 och 1.0. Ett ords värde över 0.5 betyder att ett meddelande innehållande ordet sannolikt är spam. Medan ett värde mindre än 0.5 indikerar att meddelandet innehållande ordet sannolikt är ham.
Värdet är uträknat på förekomsten av ordet i ett antal spam- och ham-meddelanden. Om ett ord förekommer 50 gånger i ett spam-meddelande och bara två gånger i ett ham-meddelande kommer ordet att få ett värde på ungefär 0.9, vilket betyder att ett meddelande innehållande ordet troligtvis är spam. Programvaror som använder bayesisk filtrering sparar dessa ord och dess värde i en databas. Databasen fylls på med ord från både spam- och ham-meddelanden och ju fler meddelanden som granskas desto bättre beslut fattar programvaran.
Fastställandet om e-post är spam eller ham kan inte baseras på endast ett ord, då det troligvis ger missvisande resultat. En genomgång av flera ord måste således utföras för att säkerställa ett så korrekt resultat som möjligt [12].

2.5 Vad är spam?

Sedan början av e-posteran har annonsörer försökt marknadsföra sig genom att skicka digitala meddelanden till miljontals användare. Spam förknippas ofta med oönskade meddelanden, det är information som skickas digitalt till en eller flera mottagare utan att tidigare kontakt har förekommit. Spam är oftast skickat med motivet att skicka ut reklam, men det kan även vara bland annat phishing eller andra sorters bedrägerier. Flera länder har diskuterat hur spam ska lagstiftas och hur det ska hanteras, men eftersom det är ett globalt problem är det svårt att hitta effektiva sätt att stoppa det på. Spam är utspritt till den nivån att 94% av alla e-postmeddelanden beräknas vara spam. På grund av detta utvecklas hela tiden nya programvaror för att förhindra att spam skickas till olika användare [13].

2.5.1 Historik

Det debatteras om var ursprunget av termen spam kommer ifrån, men den mest accepterade versionen är att den kommer från Monty Python, en brittisk komikergrupp. Sketchen utspelades på ett café där de serverade den konserverade köttprodukten spam i nästan varje rätt. När servitrisen förklarade menyn och ordet spam kom tillbaka hela tiden brast de ut i sång om spam. Det sägs då vara därifrån ordet kommer då spam återkommer utan att någon vill ha det. Texten i sången jämförs med spam-meddelanden då den innehåller repetition av värdelös text [1].

2.5.2 Ekonomiska aspekter av spam

Spammare föredrar elektronisk reklam, då det är smidigare och kostnaden hamnar hos mottagaren. Detta eftersom mottagaren måste ta emot och bearbeta meddelandet, men även lagra stora volymer av meddelanden som mottagaren inte vill ha. För större företag är detta inte en billig teknik. Det är beräknat att kostnaden för spam till företag runt om i världen uppgår till 100 miljarder per år. Spammare tjänar in pengar genom att mottagarna fullföljer det som står i meddelandet, det kan vara till exempel att klicka på en reklamlänk som ger avsändaren reklamintäkter eller att någon köper en produkt från sidan som skickas med i meddelandet. För att spam ska vara effektivt för avsändaren måste det jämföras om det är värt att skicka ut spam med både kostnaden för utrustning och risken att skicka spam vilket kan medföra rättsliga åtgärder. Av all spam som

skickas är det ungefär 0.0001% som beräknas bli fullföljd av mottagaren, vilket inte låter särskilt mycket. Men tack vare antalet mottagare av spam blir det oftast lönsamt för spammare [13].

2.6 E-postmeddelande

Kapitlet beskriver kortfattat hur ett e-postmeddelande är uppbyggt, detta för att få en förståelse hur det fungerar.

2.6.1 E-postmeddelande och dess uppbyggnad

Ett e-postmeddelande är uppbyggt i två delar, en header och en body. Headern innehåller information om hur meddelandet ska skickas. En header består åtminstone av tre delar:

Från: Avsändarens e-postadress.
Till: Mottagarens e-postadress.
Datum: Datumet då meddelandet skickades.

Headern kan även, men måste inte innehålla följande:
Mottagen: Information om de olika servrarna som bearbetar meddelandet och datumet som det bearbetas.
Svara till: En svarsadress.
Ämne: Meddelandets ämne.
Meddelande-ID: en unik identifierare för meddelandet.

Meddelandets body är själva meddelandet som avsändaren vill skicka och är separerat från headern med en linjebrytning [14].

3 Metod

Kapitlet beskriver hur testerna utförs. En genomgång av de hård- och mjukvaruinställningar som används under testerna ingår i det här kapitlet. Eventuella problem och fallgropar som kan uppstå vid testerna kommer att belysas här.

3.1 Val av metod

Den praktiska undersökningen kommer beröra den experimentella metoden. Tester genomförs flera gånger i olika system, för att få fram ett resultat hur väl bayesisk filtrering utnyttjas. Testerna utformas för att säkerställa ett tillförlitligt resultat genom att använda standardinställningar, samt identiska e-postmeddelanden som utformas på ett sätt som gör att innehållet kan kontrolleras och på så sätt få ett önskvärt resultat. Informationen från header-delen i e-postmeddelanden kommer att granskas.

3.2 Genomförande

Testerna utförs i en virtuell miljö, men kan även appliceras på fysiska servrar. Systemen som används är Ubuntu 12.04 och Windows Server 2008 R2. Testerna utförs fyra gånger. Med hjälp av virtuella maskiner kan man återställa systemet från en tidpunkt innan testerna, och på så vis göra om testerna från grunden för att säkerställa ett tillförlitligt resultat.

3.2.1 Testmiljö

För att utföra testerna användes följande mjukvara. För installation av programvara i Ubuntu används pakethanteraren Apt. Installation i Windows gjordes med ISO-filer.

Ubuntu Server 12.04

- Postfix 2.9.6
- Dovecot 2.0.19
- Spamassassin 3.3.2
- Bind 9.8.1-P1
- Python 2.7.3

Windows Server 2008 R2

- Exchange 2010
- GFI MailEssentials 2012, 2013/02/26
- Python 2.7.4

3.2.2 Konstruktion av e-post

Antalet uppsättningar av e-post uppgår till fyra olika typer. Två typer är utvalda kontrollmeddelanden, 15 ham- och 15 spammeddelanden. De två sista uppsättningarna är två stora samlingar, en med skräppost från riktiga spamkällor och en med ham baserat på legitima meddelanden. Samlingarna innehåller 2400 mail vardera och majoriteten av dem är skrivna på engelska. Skräppostsamlingen kommer från en användares riktiga skräppostkatalog, den kan anses trovärdig eftersom e-posten kommer från verkliga källor. Samlingen med ham är hämtat från en onlinekälla innehållande tillförlitlig e-post för det här syftet [15]. De används till att lära upp programmens bayesiska filtreringsfunktion. Kontrollmeddelandets syfte är att kontrollera hur programvarornas filtreringfunktion beter sig både före och efter utskicket av den stora

samlingen upplärningspost. Spammeddelanden är utvalda för att ha olika karaktärer som innehåller ord som tyder på att meddelandet är skräppost. Hammeddeladen är även dem utvalda att ha olika karaktär, vissa meddelanden har valts för att de innehåller ord som kan misstolkas för spam, detta för att se hur programmet reagerar. All e-post som används återfinns i bilaga B.

3.2.3 Utskick av e-post för test 1

För att på ett enkelt sätt skicka ett stort antal e-postmeddelanden används ett skript skrivet i Python. Skriptet läser in varje meddelande i den katalog det befinner sig i och skickar det till en angiven adress genom den inställda smtp-servern. Skriptet tillåter användaren att välja delar av det sparade meddelandet som ska skickas som mail.
Ett problem med testmiljön är att utskicket av mail skickas inom samma subnnät som mottagaren, dock ej från samma domän.

Börja med att skicka 15 spam och 15 ham med hjälp av skriptet. Skriptet meddelar om den lyckats skicka meddelanden eller ej.

3.2.4 Mottagande av e-post

Programvaran på servern kontrollerar e-posten genom att gå igenom alla meddelanden och letar efter signaturer som visar på eventuell skräppost. Är programmet tillräckligt säker på att det är skräppost, flaggas det som spam.
Klienten hämtar e-posten och placerar dem i kataloger avsedda för spam eller ham.
I headern på varje e-postmeddelande återfinns information om vad programmet har gjort, vad det har reagerat på, vilka ord som granskats och poängsatts. Är poängen tillräckligt hög klassas det som skräppost och läggs i spam-katalogen. Det gäller endast för SpamAssassin då MailEssentials bara loggar om det är spam eller ej.

Undersök inkorgen på klienten och e-postheadern där information om granskningen finns.

3.2.5 Upplärning av programvara.

För att programvaran ska kunna upptäcka skräppost bättre och mer tillförlitligt anpassas det genom upplärning. En genomgång av ett stort antal e-postmeddelanden, redan klassificerade som skräppost, medför att programmet bygger upp sin databas som det använder vid sin kontroll. I och med att språket i e-postmeddelandena är på engelska kommer även databasen innehålla engelska ord, således fungerar programvaran inte så väl med andra språk i sin nuvarande form.

Utför nya utskick från katalogen med de stora antalet skräppost. Ett stort antal e-postmeddelanden kommer att skickas till användaren. Skräpposten sparas i spamkatalogen. Peka programmet till att använda e-posten i spamkatalogen för uppbyggnad av sin databas av spam. Gör likadant med ham e-post. Peka programmet att meddelanden som ligger i inkorgen är ham.

3.2.5.1 Upplärning i SpamAssassin

sa-learn --no-sync [--spam or --ham] --mbox [folder]
Kommandot säger till SpamAssassin att hämta ut tokens från meddelanden i katalogen som specificerats.
sa-learn --sync

Kommandot synkroniserar databasen och lägger till de tokens som lärdes upp enligt ovanstående kommando. *–no sync* växlen används för synkronisering av databasen, det kan ta lång tid och därför synkroniseras spam och ham samtidigt i testerna.

3.2.5.2 Upplärning i MailEssentials

MailEssentials använder sig av GFI MailEssentials Bayesian Analysis Wizard för att skapa en ny databas som bayesiska filtret kommer att använda sig av. Genom att skicka upplärningsmailen till två olika konton, ett för spam och ett för ham kan sedan guiden köras och en fil skapas med den nya databasen. Databasen läggs sedan in i rätt katalog där den gamla databasfilen tas bort.

3.2.6 Kontroll av upplärning

Kontrollera de olika programvarornas databaser.

3.2.6.1 SpamAssassin

Med hjälp av kommandot *sa-learn --dump magic* visas hur många tokens som databasen innehåller samt hur många spam- och ham-mail programmet analyserat för att bygga databasen.

3.2.6.2 GFI MailEssentials

Med hjälp av användargränssnittet “MailEssentials Configuration” och under “Spamfilters” ses olika filter för spam. Om “Bayesian Analysis” väljs visas antal ham och spam som ligger i databasen.

3.2.7 Utskick av kontrollmail för test 2

Det andra testet utförs på samma sätt som test 1, samma 15 spam och 15 ham skickas. Kontrollmailen ingår inte i upplärningen och har inte påverkat databasen. Testet kommer att visa på om och hur programvarorna förändrat sin analys och sitt beslut om ett e-postmeddelande är spam eller ej, efter upplärning och uppbyggnad av databas. Systemet har ej återställts mellan testerna, före och efter upplärning.

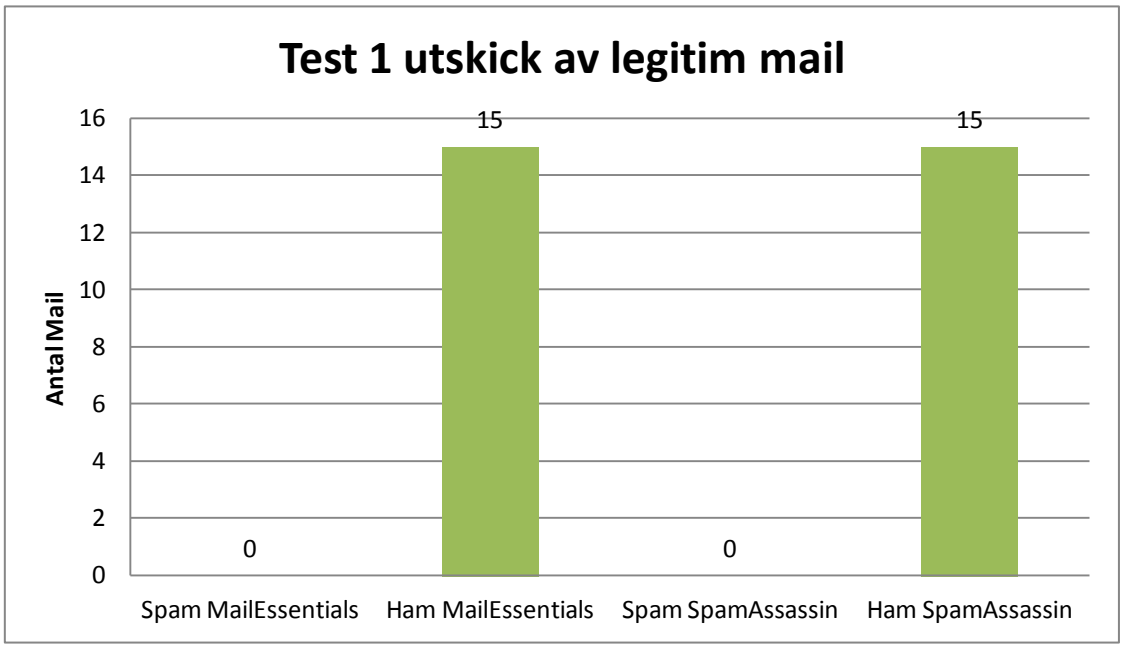
4 Resultat och analys

4.1 Test 1, innan upplärning av filter

I första testet skickades 30 e-postmeddelanden där programvarorna ej var upplärda och deras databaser ej populerade. Resultatet visas i två grafer. Alla e-postmeddelanden återfinns i bilaga B.

4.1.1 Legitim mail

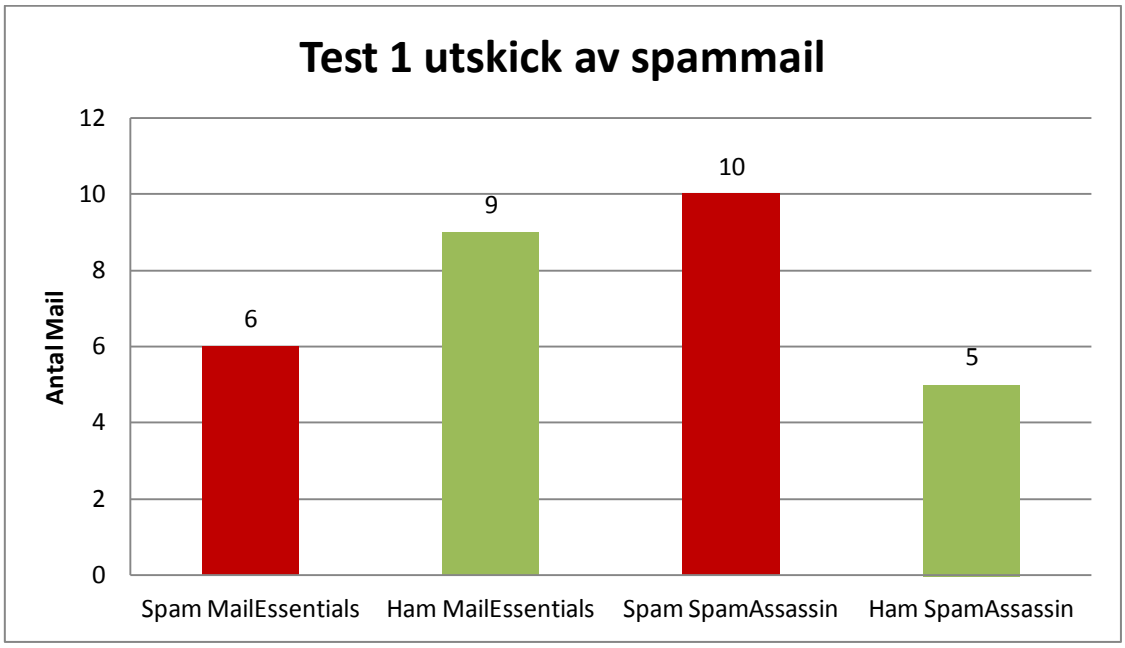
15 legitima e-postmeddelanden analyseras av programvarorna. *Figur 4.1.1* visar resultatet. Resultatet av hur programvarorna hanterar de 15 legitima mailen visar på att de efter analys klassat alla 15 e-postmeddelanden som legitima. Det vill säga inga falsk negativa resultat.



Figur 4.1.1

4.1.2 Spammail

15 spammeddelanden analyseras av programvarorna. *Figur 4.1.2* visar resultatet. Resultatet av hur programvarorna hanterade de 15 spammailen visar på en skillnad mellan programvarorna. SpamAssassin klassar 10 av mailen som spam och 5 (Mail nr. 3, 4, 7, 11,15) som ham medan MailEssential klassar sex av mailen som spam och nio (Mail nr. 1, 2, 3, 5, 10, 11, 12, 14, 15) som ham. Båda programvarorna ger därmed falska positiva resultat.



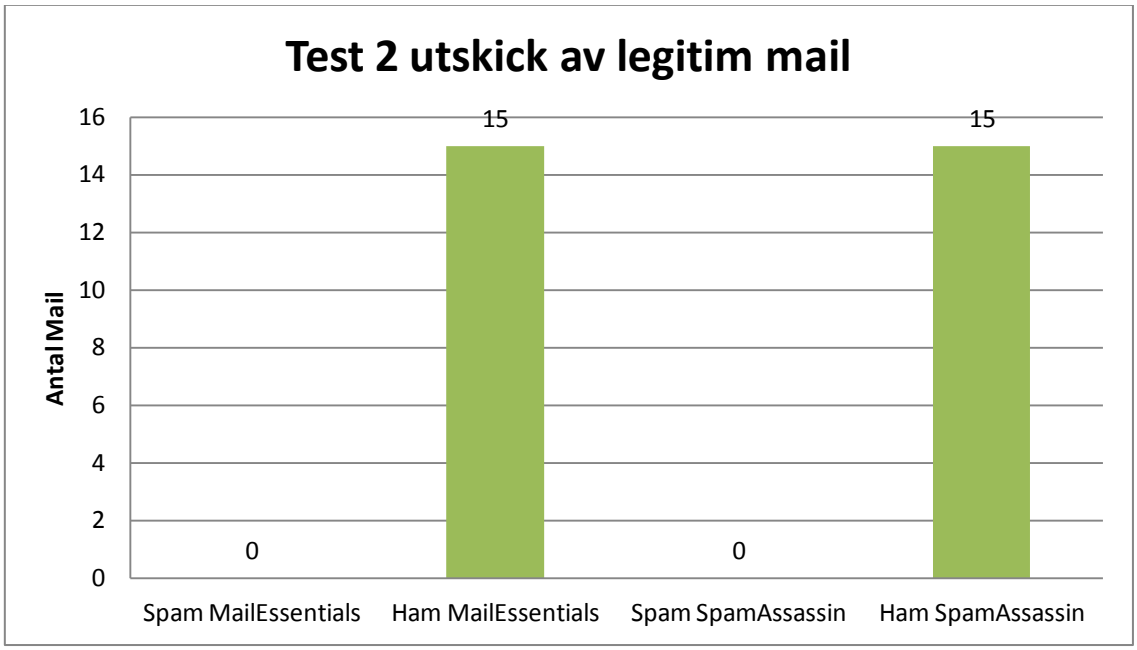
Figur 4.1.2

4.2 Test 2, efter upplärning av filter

I andra testet skickas 30 e-postmeddelanden efter det att programvarorna är upplärda och deras databaser populerade. Programvarorna har olika databaser specifika för vardera program. Det går ej att granska databaserna då de är krypterade. Det går dock att kopiera databaser inom samma programvara. E-postmeddelanden återfinns i bilaga B.

4.2.1 Legitim mail

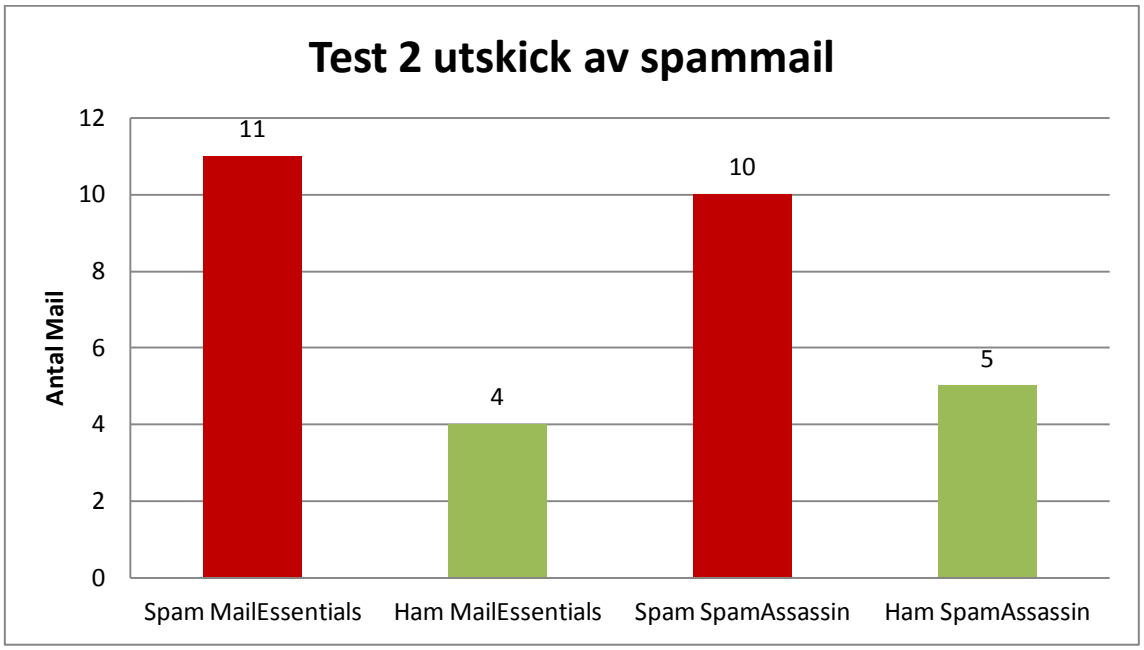
15 legitima e-postmeddelanden analyseras av programvarorna. *Figur 4.2.1* visar resultatet. Resultatet av hur programvarorna hanterade de 15 legitima mailen visar på att de efter analys klassat alla 15 e-postmeddelanden som legitim.



Figur 4.2.1

4.2.2 Spammail

15 spam meddelanden analyseras av programvarorna. *Figur 4.2.2* visar resultatet. Resultatet visar även här en skillnad mellan programvarorna, SpamAssassin ger samma resultat som test 1, 10 spam och 5 ham (Mail nr. 3, 4, 7, 11, 15). MailEssential ger en förbättring av resultatet med 11 spam och 4 ham (Mail nr. 3, 4, 12, 15). Båda programvarorna ger falska positiva resultat men MailEssentials har minskat dessa med fem jämfört med test 1, innan upplärning.



Figur 4.2.2

4.3 Resultatanalys

Resultatet visar att det finns en viss skillnad mellan programvarornas förmåga att filtrera skräppost med hjälp av bayesisk filtrering. Före upplärning av programvaran gav SpamAssassin ett resultat på fem falska positiva mail av 15 spam, medan MailEssentials gav ett resultat på nio falska positiva mail av 15 spam. Det kan förklaras med att SpamAssassin har en större standarddatabas än MailEssentials. Efter upplärning av de 2400 ham- och 2400 spam-mailen förbättrade MailEssentials sitt resultat och sänkte de falska positiva till 4 av 15 mail, SpamAssassin visade ingen förbättring utan gav samma resultat som innan. Resultatet visar att MailEssentials har byggt upp sin databas bättre efter upplärning.

Nedanstående mail skiljer programvarorna åt efter upplärning. MailEssentials anser att det är spam medan SpamAssassin klassade det som ham (dock med ett nära värde för spam). Det har en tydlig karaktär av spam, men innehåller många ham-tokens.

Subject: You can earn more! We offer a personal decision from medicine clinic.

We invite you to work in the remote assistant position.

This work takes 2-3 hours per week and requires absolutely no investment.
The essence of this work for incoming client requests in your city.
The starting salary is about 2500 EUR per month + bonuses.

You get paid your salary every 2 weeks and your bonuses after fulfilling each task!

We guarantee work for everyone. But we accept applications this week only! Therefore, you should write a request right now. And you will start earning money, starting from next week.

Please indicate in the request:
Your name:
Your email address:
City of residence:

Please send the request to my email Merlin@quintcareerseu.com,and I will answer you personally as soon as possible

Sincerely,
Merlin Marquez

Nedanstående mail tolkar SpamAssassin som legitimt.

Subject: Try it Risk Free for 60 Days. Erect Plus - Amazing Results With Clinically Supported & Doctor Approved Male Erection Enhancement Pills. mnk5rl2

Erect Plus - Amazing Results With Clinically Supported & Doctor Approved Male Erection Enhancement Pills. Delivery in 2-3 Days Worldwide. Try it Risk Free for 60 Days. http://batweb.ru

Nedanstående mail tolkar MailEssentials som legitimt.

Subject: The most popular goods NorcoCialis ProfessionalViagra FemaleLevitra Professional

Floor price and All assortment for treatment Osteoporosis. Weight Loss. Cancer
How to solve problems - read here.

We do not require recipes,Best sellers for today ViagraEphedrineLevitraDiazepam

Your chance! The fine price today
Nexium may also be given to prevent gastric ulcer caused by infection with helicobacter pylori (H. pylori), or by the use of nonsteroidal anti-inflammatory drugs (NSAIDs).
Generic Name: Esomeprazole (ee so MEP ra zol).Brand Names: Nexium
http://toystoreweb.com.ua/

Delivery across the USA for 2-3 days who that can faster...?
The qualitative goods and anonymous delivery
http://toystoreweb.com.ua

De mail som gav falska positiva resultat innehåller färre typiska ord eller fraser vanligt förekommande i spam. Vilket gör att det bayesiska filtret ger mailet ett lägre sammanlagt poäng.

5 Diskussion och slutsats

Syftet med arbetet var att få en insyn i hur effektivt programvarorna utnyttjar bayesisk filtrering. Det finns många olika filter i de programvaror som testas men arbetet fokuserar och avgränsas till bayesisk filtrering och hur effektivt de olika programvarorna utnyttjar det.

Resultaten av testerna förbryllar oss något när det gäller SpamAssassin. I testerna som gjordes fick vi samma resultat både innan och efter upplärningsfasen. Efter en extra kontroll av databasen, före och efter upplärning kunde vi konstatera att databasen var populerad med nya tokens och antal analyserade e-postmeddelanden. Ett antagande att testerna i SpamAssassin gav de resultaten är att det krävs fler analyserade meddelanden för att databasen ska bli optimerad. Eftersom standardinställningar har använts i båda programvarorna kan det skilja sig på hur hög eller låg poänggränsen är för vad som klassificeras som spam.

En begränsning i analysen av resultaten var den brist på information MailEssentials gav oss, både från tillverkaren och loggar. Programmet lämnar ingen information om hur och varför filtret klassificerade ett meddelande som spam, endast information om att det är spam. En dialog med specialister hos tillverkaren har förts, där har svaret varit att informationen vi vill komma åt inte är tillgänglig för användaren. Programvaran SpamAssassin ger mer information om varför meddelanden klassificeras som spam med vilka ord eller fraser som den reagerar på. Eftersom MailEssentials inte ger samma information kan vi heller inte använda oss av SpamAssassins information eftersom det inte finns någonting att jämföra med.

Fördelar med bayesisk filtrering är att det anpassas efter de e-postmeddelanden som den analyserar. Förändras karaktären på ett spammeddelande kan filtret anpassa sig utefter detta, genom att uppdatera sin databas. Filtret kan och bör utformas enligt användarens behov, meddelanden som vi klassificerar som spam behöver inte betyda att andra användare gör det.

Nackdelar med bayesisk filtrering är att ett meddelande som inte innehåller karaktäristiska fraser och ord för spam, kan lätt bli felaktigt bedömt som legitimt då meddelandet inte innehåller större mängd typiska spamord. Det finns även risker för att databasen blir korrupt. Meddelanden som analyseras felaktigt exempelvis hammeddelanden som analyseras som spam kommer att ge oönskade värden i databasen.

Under arbetets gång har vi konstaterat att ett filter inte är fullt tillräckligt för att motverka spam. Det krävs ytterligare åtgärder i form av exempelvis blacklists och greylists för att uppnå ett för oss tillräckligt skydd. Programvarorna har som standard flera filter aktiverade, men eftersom arbetet endast behandlar bayesisk filtrering avaktiverades dessa.

Vi blev förvånade över att programvarorna gav falska positiva resultat på olika e-postmeddelande som vi trodde skulle klassificeras på samma sätt. Det visar att de inte nyttjar bayesisk filtrering på samma sätt. Eftersom vi inte har insyn i det kan vi inte undersöka det vidare utan kan bara konstatera att skillnader finns.

Vi anser att vikten av ett skräddarsytt bayesiskt filter är högst relevant med resultaten i åtanke. Innan upplärning av databasen i MailEssentials har den en standarddatabas upplärd på cirka 45000 e-postmeddelanden. Även med det höga antalet meddelanden, fick vi ett bättre resultat med 4800 meddelanden eftersom de var liknande till de 15 vi använde i testerna. De visar sig att det är viktigt att skräddarsy miljön efter eget behov, det handlar inte bara om antalet analyserade meddelanden utan även innehållet har en stor betydelse.

5.1 Slutsats

Vi kan konstatera att efter ha testat båda programvarornas bayesiska filter gav MailEssentials ett bättre resultat, om än marginellt. Vi hade förväntat oss ett bättre resultat, både i MailEssentials och framförallt SpamAssassin då det inte visade på förbättring efter upplärning. Vi känner en viss besvikelse över mängden information MailEssentials lämnar, det försvårade analysen för oss. Samtidigt är vi medvetna om konkurrensen på marknaden och att MailEssentials inte vill lämna ut information som kan utnyttjas av andra företag.

De frågeställningar arbetet syftade till, har vi besvarat efter de förutsättningar programvarorna gav. MailEssentials gav ett bättre resultat men det var svårt att analysera varför de klassificerades som spam då det inte gav tillräcklig information. På frågan om det räcker att med endast bayesisk filtrering uppnå ett fulltständigt skydd mot spam har vi konstaterat att det ej räcker. Det krävs ytterliggare metoder som arbetar tillsammans för att säkerställa motverkandet av spam.

När valet av programvara för att bekämpa spam sker, anser vi att det ej enbart räcker med att jämföra dess förmåga att utnyttja bayesisk filtrering effektivt. Det är viktigt att utvärdera vilka behov som finns samt befintlig miljö, exempelvis kräver MailEssentials en Exchangemiljö.

5.2 Vidare forskning

Arbetet har utförts i en laborationsmiljö. Det skulle för oss vara intressant att testerna utföras i en verklig situation eller miljö. Vi anser att resultaten ej skulle skilja sig, men på grund av MailEssentials uteblivna information kan vi ej vara helt säkra. Ett exempel på något som skulle kunna påverka resultaten är att mailen skickades inom samma subnät. Det skulle vara intressant att skicka alla mail från olika domäner och se om det ger andra resultat. Tester med olika filter där slutsater kan dras om vilka filter som arbetar mest effektivt tillsammans.

6 Referenser

[1] Justin M. Rao and David H. Reiley (Summer 2012). *The Economics of Spam* [Online]. Available: http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.26.3.87

[2] Microsoft(2011, June 1). *Windows Server - Active Directory*[Online]. Available: http://technet.microsoft.com/en-us/library/cc780036(WS.10).aspx#w2k3tr_ad_over_qbjd

[3] SpamAssassin(2009, November 13). *What is SpamAssassin?* [Online]. Available: http://wiki.apache.org/spamassassin/SpamAssassin

[4] GFI(2013, April 30). *Overview*[Online]. Available: http://www.gfi.com/exchange-server-antispam-antivirus#overview

[5] GFI(2013, April 30). *Features*[Online]. Available: http://www.gfi.com/exchange-server-antispam-antivirus#features

[6] Jaap Wesselius(2009, October 22). *Introduction to Exchange Server 2010*[Online]. Available: https://www.simple-talk.com/sysadmin/exchange/introduction-to-exchange-server-2010/

[7] Deshpande, V.P. (2007, June 20). *An Evaluation of Naive Bayesian Anti-Spam Filtering Techniques*[Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4267579

[8] 360is(2013). *Postfix - What Is It*[Online]. Available: http://www.360is.com/06-postfix.htm

[9] Security space(2012, January 01). *Mail Server Survey*[Online]. Available: http://www.securityspace.com/s_survey/data/man.201112/mxsurvey.html

[10] GFI(2013). *Why Bayesian filtering is the most effective anti-spam technology*[Online]. Available: http://www.gfi.com/whitepapers/why-bayesian-filtering.pdf

[11] Ubuntu(2013). *The Ubuntu story*[Online]. Available: http://www.ubuntu.com/about/about-ubuntu

[12] Michal Prilepok, Jan Platos, Vaclav Snasel, and Eyas El-Qawasmeh(2013). *The Bayesian Spam Filter with NCD**[Online]. Available: http://ceur-ws.org/Vol-837/paper18.pdf

[13] Michael T. Goodrich, Roberto Tamassia, "Application Security" in *Introduction to Computer Security*, International ed. Irvine: Pearson, 2011, ch. 10, pp. 497-500

[14] Kioskea(2013). *Structure of an email* [Online]. Available: http://en.kioskea.net/contents/117-structure-of-an-email-headers-and-bodies

[15] Apache(2013). *Index of /publiccorpus* [Online]. Available:

spamassassin.apache.org/publiccorpus/

7 Bilagor

7.1 Bilaga A - Pythonskript för utskick av mail.

```
1.  ####################################
2.  #                                  #
3.  #   Skript för att skicka mail     #
4.  #                                  #
5.  ####################################
6.
7.
8.  import re
9.  spam = 0
10. #Skriptet går igenom 3000 mail-
    filer och letar efter ordet som match har deklarerats
11. #och sparar allt innehÃ¥ll efter det till message, som skickas som mail.
12.
13.
14. while spam < 3000:
15.     spam = spam + 1
16.     match = 'X-Spam-Prev-Subject'
17.     message = ''
18.     x = 0
19.     nr = 1000
20.
21.
22. #läs in mail
23.     with open(str(spam)) as f:
24.         for line in f:
25.             x += 1
26.             if line.startswith(match):
27.                 temp = re.sub(match,"",line,count=1)
28.                 message = 'Subject: ' +  temp + '\n'
29.                 nr = x
30.             if x > nr:
31.                 message += line
32.     f.close()
33.
34.
35. #skicka mail
36.     import smtplib
37.     sender = 'spam@spammers.com'
38.     receivers = 'stefan@skynet.lab'
39.     count = 0
40.
41.
42.     if not message:
43.         print 'empty'
44.     else:
45.         count += 1
46.         smtpObj = smtplib.SMTP('mail.skynet.lab')
47.         smtpObj.sendmail(sender, receivers, message)
48.         print 'mail nr: ' + str(spam)
```

7.2 Bilaga B - Mail

7.2.1 Spam

1.

Subject: Improve your "Ph.D" resume in less than 40 days

Bachelors, Masters and PhD's available in your field! Eliminates classrooms and travelling.They are fully verifiable!
Could you ever imagine that already tomorrow you can handle a degree of some famous and prestigious university?
You can EARN TWICE MORE, confirm your real abilities and knowledge with the diploma. It is YOUR CHANCE

We accept your phone calls 24 hours a day, not to postpone your Improvement of qualification.We send the certificate to all countries
Dial 1-603-509-2001 or Outside USA.: +1-603-509-2001 in your phone and contact our high-end staff that will explain you the whole procedure. If you didn't manage to contact us directly, then you still can leave us your message with the help of voicemail service by leaving your full NAME and TELEPHONE NUMBER (including country code). This is a really unique chance and you cannot miss that!

2.

Subject: Enlarge Your Penis. (Big Penis) 33b9

Enlarge Your Penis. (Big Penis)

Save your time and money!

The Safest & Most Effective Methods Of Penis Enlargement
http://bigpenismallstore.ru

3.

Subject: Hello "tann"!

I'm Anastasia! I am a very womanly, tolerant and romantic lady, who has a heart that never hardens, and a temper that never tired.

I stay lady in every situation because this is my essence and I am proud of it. I am frugal, resourceful and positive.
I would like to meet a man with whom I will be caring, lovely and loyal and who will appreciate this. I don`t want to decide who is boss in our family I just want to be a woman next to my husband.

If I'm your woman, do not hesitate to meet with me. I'm waiting for you:
http://baum01mn.page.tl

4.

Subject: The intrigue about you

Hello! Are you looking for your soulmate? I am a girl, I'm quite happy with my life, but I'll be happy when I find my boyfriend

I am 23 y.o., Never married.
Height 5' 7", my weight 114 lbs.

I have amazing black eyes and long and thick blonde.
My occupation - teacher.
I am pleased to meet you http://jasonwganz25s.webs.com/?qG=57
If I'm interested, write me.

5.

Subject: Buy Cheap Generic Viagra (Sildenafil), Cialis, Levitra Online. Secure Checkout, Visa & Mastercard Accepted. an9vhcuk

Find Cheap Viagra Overnight?

Buy 100mg x 10pills $29,95 Only! Buy Cheap Generic Viagra (Sildenafil), Cialis, Levitra Online. Secure Checkout, Visa & Mastercard Accepted.

Save With The Lowest Price And Get Overnight Delivery.
http://remedycutmedspills.ru

6.

Subject: Learn how people in your profession can earn a 30% increase!

We invite you to work in the remote assistant position.

This work takes 2-3 hours per week and requires absolutely no investment.
The essence of this work for incoming client requests in your city.
The starting salary is about 2500 EUR per month + bonuses.

You get paid your salary every 2 weeks and your bonuses after fulfilling each task!

We guarantee work for everyone. But we accept applications this week only!
Therefore, you should write a request right now. And you will start earning money, starting from next week.

Please indicate in the request:
Your name:
Your email address:
City of residence:

Please send the request to my email Dorian@quintcareerseu.com,and I will answer you personally as soon as possible

Sincerely,
Dorian Santana

7.

Try it Risk Free for 60 Days. Erect Plus - Amazing Results With Clinically Supported & Doctor Approved Male Erection Enhancement Pills. mnk5rl2

Erect Plus - Amazing Results With Clinically Supported & Doctor Approved Male Erection Enhancement Pills. Delivery in 2-3 Days Worldwide. Try it Risk Free for 60 Days. http://batweb.ru

8.

Subject: Swiss luxury watch brands - Rolex, Breitling, TAG Heuer, Cartier, Panerai, Patek Philippe, Hublot.EXPRESS DELIVERY, we accept VISA / MASTERCARD hqwfbip71l

High Quality Global Replica
Top quality replica watches of Swiss luxury watch brands - Rolex, Breitling, TAG Heuer, Cartier, Panerai, Patek Philippe, Hublot.EXPRESS DELIVERY, we accept VISA / MASTERCARD
http://petfan.ru

9.

Subject: Rolex Replica Watches & More bkzhjjs

Rolex Replica Watches & More
Choose from hundreds of perfect replica watches: Rolex, Cartier, Breitling, Omega & many more. Enjoy our monthly promotions! Visit us now!
http://lidbug.ru

10.

Subject: Generic Xanax As Low As $1.33 Per Pill!! NO RX NEEDED!! Reliable Online Pharmacy - Fast & Secure Shipping!! VISA & E-CHECK ACCEPTED!! Order Today & Save!! kd3d2w3n

Xanax Generic 1mg 60 Pills $169!! NO RX!
Buy Xanax (Alprazolam) Generic As Low As $1.33 Per Pill!! NO RX NEEDED!! Reliable Online Pharmacy - Fast & Secure Shipping!! VISA & E-CHECK ACCEPTED!! Order Today & Save!!
http://mensfor.ru

11.

Subject: You can earn more! We offer a personal decision from medicine clinic.

We invite you to work in the remote assistant position.

This work takes 2-3 hours per week and requires absolutely no investment.
The essence of this work for incoming client requests in your city.
The starting salary is about 2500 EUR per month + bonuses.

You get paid your salary every 2 weeks and your bonuses after fulfilling each task!

We guarantee work for everyone. But we accept applications this week only!
Therefore, you should write a request right now. And you will start earning money, starting from next week.

Please indicate in the request:
Your name:
Your email address:
City of residence:

Please send the request to my email Merlin@quintcareerseu.com,and I will answer you personally as soon as possible

Sincerely,
Merlin Marquez

12.

Subject: Fwd: The most popular goods NorcoCialis ProfessionalViagra FemaleLevitra Professional

Floor price and All assortment for treatment Osteoporosis. Weight Loss. Cancer
How to solve problems - read here.

We do not require recipes,Best sellers for today ViagraEphedrineLevitraDiazepam

Your chance! The fine price today
Nexium may also be given to prevent gastric ulcer caused by infection with helicobacter pylori (H. pylori), or by the use of nonsteroidal anti-inflammatory drugs (NSAIDs).
Generic Name: Esomeprazole (ee so MEP ra zol).Brand Names: Nexium
http://toystoreweb.com.ua/

Delivery across the USA for 2-3 days who that can faster...?
The qualitative goods and anonymous delivery
http://toystoreweb.com.ua/

13.

Subject: Replica Chanel Watches, Replica Shoes,Bags,Replica Handbags ...we specialize in Replica watches, Replica handbags, Replica shoes,replica bags and so on qffu7

Replica Chanel Watches, Replica Shoes,Bags,Replica Handbags ...we specialize in Replica watches, Replica handbags, Replica shoes,replica bags and so on
http://tarhit.ru

14.

Subject: Phentermine 37.5mg 90 Pills $289!! NO RX Required!!! Fast & Secure Shipping!! VISA & eCHECK Accepted, Order Today & Save!! 0tqyyqi

Buy Phentermine 37.5mg 90 Pills $289!! Buy ORIGINAL Phentermine 37.5mg

(Adipex) From $119 As Low as $2.92/Pill!!! NO RX Required!!! Fast & Secure Shipping!! VISA & eCHECK Accepted, Order Today & Save!!
http://blueonlinerxhealth.ru

15.

Subject: I wish u to be healthy

Your babe will notice that for sure http://id2.mu-dongdo.net/easy.html

7.2.2 Ham

1.

Subject: More on promiscuity and word choice Re: Selling Wedded Bliss

It was a extreme contrived example because you glosded over the
point of the earlier 3-person example.
But OK, Mr Math, let it be N men and women, for any N>2. They
all pair off. Then, some number H, N>H>0, of men has sex with
all the other N-1 women he hasn't yet had sex with.

Pick any N and H that might be interesting. Any choice of
values results in meaningful differences between the sexes'
"promiscuity", as commonly understood. It should be more
obvious with extreme choices of numbers, but it is also true
for any choice of N and H, if unrealistic totals distract you.

Further, and I was hoping this would be clear without saying
so outright, this model actually approximates the cliche
"common wisdom" about per-gender sexual behavior, if you
reverse the male and female roles.

Those stereotypes are: that more men than women seek multiple
partners -- men being "more promiscuous" than women -- and that
surplus of male interest is satisfied by a smaller number of
hyperpromiscuous women (often derisively labelled "sluts").

2.

Subject: Help

|I can't reproduce this error.

For me it is very repeatable... (like every time, without fail).

This is the debug log of the pick happening ...

18:19:03 Pick_It {exec pick +inbox -list -lbrace -lbrace -subject ftp -rbrace -rbrace} {4852-4852 -sequence mercury}
18:19:03 exec pick +inbox -list -lbrace -lbrace -subject ftp -rbrace -rbrace 4852-4852 -sequence mercury

18:19:04 Ftoc_PickMsgs {{1 hit}}
18:19:04 Marking 1 hits
18:19:04 tkerror: syntax error in expression "int ...

Note, if I run the pick command by hand ...

delta$ pick +inbox -list -lbrace -lbrace -subject ftp -rbrace -rbrace 4852-4852 -sequence mercury
1 hit

That's where the "1 hit" comes from (obviously). The version of nmh I'm using is ...

delta$ pick -version
pick -- nmh-1.0.4 [compiled on fuchsia.cs.mu.OZ.AU at Sun Mar 17 14:55:56 ICT 2002]

And the relevant part of my .mh_profile ...

delta$ mhparam pick
-seq sel -list

Since the pick command works, the sequence (actually, both of them, the one that's explicit on the command line, from the search popup, and the one that comes from .mh_profile) do get created.

kre

ps: this is still using the version of the code form a day ago, I haven't been able to reach the cvs repository today (local routing issue I think).

3.

subject: Re: [SAdev] 2.40 RELEASE PROCESS: mass-check status, folks?

I plan to

1. figure out the freqs tonight, suggest what tests to drop
2. wait for comments
3. drop tests that nobody cares about tomorrow
4. sed out the dropped tests from the mass-check logs

This step is unneccesary -- unless you've changed the scripts much, any test in the logs which aren't in the rules files will just be ignored I think. You do seem to have changed the logs-to-c script and removed the bit where you could specify immutable tests at the top -- I took a brief glance through the code and couldn't fully make out how it had changed. I think we want to be able to specify immutable test scores though in there

somewhere -- or is that now handled by the tflags stuff? For
the last couple releases, any test which occurred infrequently
(by thumb-in-the-wind subjective criteria) I set to have
immutable scores, as well as a handful of other rules.

5. kick off the GA

BTW I'll be away this weekend at Linuxbierwanderung, so Craig,
you might
have to run the GA. ;)

Shouldn't be a problem. Assuming I can get the darned thing to
compile :)

C

4.

Subject: Re: Electric car an Edsel...
Bob,

This guy's an idiot.

I design loads for systems with the 50 kV capacitors. One of those has 864
of such capacitors and stores only 10 megajoules, which means 11 kilojoules
each. They weigh 125 kg.

You need high energy per unit mass, and the capacitive system I picked
maximizes that.

It is precisely the system that Maxwell is touting for electrical braking
and power augmentation for regenerative use in automobiles. You also need
voltage you can use in a DC motor, which is why though the actual
capacitors in the system are charged to 2.5 volts, the system has them
arranged in series to boost the voltage.

Ignore him. He's a waste of my time.

5.

Subject: [zzzzteana] Uncle Mark seeks parole

http://news.bbc.co.uk/1/hi/entertainment/showbiz/2308581.stm

Tuesday, 8 October, 2002, 07:55 GMT 08:55 UK
Lennon killer seeks parole again

The man who shot dead former Beatle John Lennon is making another bid for
early release from prison - the day before what would have been Lennon's
62nd birthday.
Mark David Chapman, 47, was jailed for life after he admitted killing the

superstar outside his New York apartment building in 1980.
It is the second time in two years that Chapman has sought parole from Attica state prison.
At a 2000 hearing, he argued that he was no longer a danger to society and had overcome the psychological problems which led him to shoot the ex-Beatle.
Chapman had said that a voice in his head told him to shoot the star.
Shot dead
Lennon was shot four times as he emerged from a limousine outside his New York City apartment on 8 December 1980.
He and his wife Yoko Ono were returning from a late-night recording session during which time they had been working on Walking on Thin Ice.
Only hours before the shooting, Chapman - who had come to New York from Hawaii - was photographed with the singer outside the same building as Lennon signed a copy of his album Double Fantasy for him.
The killer said Lennon had been just "a picture on an album cover" to him before the shooting.
'Deserved death'
Chapman has said that he should have received the death penalty for his crime.
Lennon's widow told the 2000 parole hearing that she would not feel safe if Chapman were released.
Lennon's songwriting partnership with Paul McCartney propelled the Liverpool-based pop group to international stardom and unparalleled commercial success.
The Beatles front man, peace campaigner, and all-round iconoclast, would have been 62 on Wednesday.

6.

Subject: Re: Unseen window versus Sequences Window
On Wed, 2 Oct 2002, "Chris" == Chris Garrigues wrote:

Chris: I'm not sure I'll get to it any time soon.

Well, you have a pretty good idea of when I might get to it (the phrase hell freezes over comes to mind)... so whenever you can will certainly be fine.

Thanks for considering it. In the meantime I've set the "minimum entry lines" to 1. It certainly isn't going to make me go back to the old version.

--Hal

7.

Subject: study reference numbers
Here are numbers that come from a study of a couple of thousand swedes, with no reference to sexual preference that I could find. The point would seem to be that (1) sexual activity follows a power curve, with a few people, a la

Wilt Chamberlain, having an extaordinarily large number of sexual contacts, ev en in a short period of time and (2) a tendency for men to have more partners than women. I have no idea, being a statistical ignoramus, whether the fact that there seem to be more men than women at the extremely promiscuous end of the sex-partners-distribution curve means that you'd get even more extreme results in a group of men who chiefly have sex with other men.

http://polymer.bu.edu/~amaral/Sex_partners/Content_sex.html

<<For example, the mean number of partners since sexual initiation for women is approximately 7 and in a sample of less than 1500 Swedish women we find an individual with 100 partners. For men, the mean is approximately 15 and we find an individual with 800 partners, which is almost 50 times larger than the mean! So, somehow Don Juan was not such an extraordinary case but just one data point in a wide spectrum of behaviors that can be observed. >>

As for gay men: There is indeed anecdotal evidence of cases of extreme promiscuity among gay men. You can read about it in Randy Schiltz's And the Band Played On. He writes about bathhouse culture pre-HIV; he also discusses how, in the gay politics of the time, there was a sub-culture of what you might call radical gay men who argued (and acted on the argument) that having many partners was an essential part of what being gay actually was. It was an explicitly political statement: monogamy is an artifact of straight culture.
That view seems to have died, in more ways than one. Part of the point of Schiltz's book was to condemn the role that it and bathhouse culture played in spreading the AIDs epidemic that eventually killed Schiltz, among so many others. This doesn't let Eugen off the hook--but it is accurate to say that there was a cult of promiscuity that was particular to the gay community.

Tom

8.

Subject: Alsa/Redhat 8 compatability

Matthias Saou (matthias@rpmforge.net) wrote*:
I really think that with my ALSA packages, ALSA on Red Hat Linux has never been so easy! ;-)

I had been hand building those alsa packages for probably 6 months or more, so I could use my laptop's ESS chip best (hard disk recording and such). Wow, maybe 8 to 10 months, time flys. I didn't look forward to doing that tedious build every time I changed a kernel or something. Matthias has made this a thing of the past! Dude, it's a no brainer, use apt-get or just download and install.

What's more, OSS is fully "imitated", older apps using OSS are as happy as a clam. My personal taste tells me newer apps which use alsa sound better (alsaplayer, xmms with alsa module) to my ear.

The 2 things that trick people are the modules.conf file, they didn't have that automatic matrix page when I started with alsa, you don't know how good you have it.

And on my old installs (only the first time I installed it on a clean box), alsa always was "muted" until you fire up a mixer and turn up the music. It's possible Matthias even took care of that too. Anyway it is/was only a one time thing on 1st install, that was a long time ago for me.

Those ALSA dudes had been trying to get it into the 2.4 kernel but missed it, but it's been in 2.5 for a while now, so alsa is the future of sound in the linux kernel.

--
That's "angle" as in geometry.

9.

Subject: Job Application

I am writing to apply for the programmer position advertised on your website. As requested, I am enclosing a completed job application, my certification, my resume and three references.
The opportunity presented in this listing is very interesting, and I believe that my strong technical experience and education will make me a very competitive candidate for this position. The key strengths that I possess for success in this position include:
I have successfully designed, developed, and supported live use applications
I strive for continued excellence
I provide exceptional contributions to customer service for all customers
With a BS degree in Computer Programming, I have a full understanding of the full life cycle of a software development project. I also have experience in learning and excelling at new technologies as needed.
Please see my resume for additional information on my experience.
I can be reached anytime via email at Rick.Johnson@yahoo.com or my cell phone, 909-555-5555.
Thank you for your time and consideration. I look forward to speaking with you about this employment opportunity.
Sincerely,
Rick Johnson

10.

Subject: Re: Microsoft buys XDegress - more of a p2p/distributed

Mr. FoRK writes:
"Files can be cached on multiple systems randomly scattered around the Internet, as with Napster or Freenet. In fact, the caching in XDegrees is more sophisticated than it is on those systems: users with high bandwidth connections can download portions, or "stripes," of a file from several cached locations simultaneously. The XDegrees software then reassembles these stripes into the whole file and uses digital signatures to verify that the downloaded file is the same as the original. A key component of this

digital signature is a digest of the file, which is stored as an HTTP header
for the file."

This "more sophisticated than [Napster or Freenet]" part seems
to be the same behavior implemented in many other P2P CDNs,
such as:

- Kazaa
- EDonkey/Overnet
- BitTorrent
- Gnutella (with HUGE extensions)
- OnionNetworks WebRAID

...though the quality of the "digest" used by each system varies
wildly.

- Gordon

11.

Subject: Internet Archive bookmobile

http://www.sfgate.com/cgi-bin/article.cgi?file=/gate/archive/2002/09/26/bonoact.DTL

Opening arguments are set to begin early next month in Eldred vs. Ashcroft, a landmark U.S. Supreme Court case that will decide the future of copyright law, including how and when artists and writers can build upon the work of others.
....

To heighten public awareness of the importance of the case an Internet bookmobile is set to depart San Francisco next Monday on a trip that will bring it to the steps of the Supreme Court building in Washington, D.C., before arguments wrap up. The van, which will be stopping at schools, libraries and senior centers along the way, is equipped to provide free high-speed access to thousands of literary and artistic works that are already in the public domain.

I had the opportunity to visit the Internet Archive about a month ago, and saw the bookmobile under construction. It's a neat idea. Take an SUV, put a small satellite dish on the top, and put a computer, printer, and binding machine in the back. Voila -- people can search for a book, then print out a copy right there on the spot. Quite literally on-demand printing. Total fixed cost of the computer/printer/binding machine, if bought new (the IA had the equipment donated) is under $10k.

One of the goals is to show libraries across the country that they could, if

they wished, add these "virtual holdings" (public domain materials) to their existing library, at a fixed cost most libraries can afford. This should be compelling to small libraries in remote areas.

- Jim

12.

Subject: WiFi Trek badges

Brian sez: Vocera Communications has developed what is essentially a Star Trek: TNG-style lapel communicator device that uses WiFi to transmit voice across networks.

> The Vocera Communications System consists of Vocera Server Software, residing on a customer premise server, and Vocera Communications Badges, which operate over a wireless LAN (802.11b). The badge - which weighs less than 2 ounces - includes a microphone and speaker, LCD readout to display text messages, and an 802.11b wireless radio. It can be clipped to a shirt pocket or collar, or worn on a lanyard.

Link[1] Discuss[2] (_Thanks, Brian[3]!_)

[1] http://www.vocera.com/news/press9.shtm
[2] http://www.quicktopic.com/boing/H/F4SLvqGh6XW
[3] http://brian.carnell.com

13.

Subject: Oh my...

Hello fork,

So they have Aaron Schwartz on NPR's Weekend Edition talking about Warchalking. I'll agree, its funny, his voice is squeaky and I'm jealous that he got on radio and I didn't...

But really, WTF is the big deal about warchalking? I have yet to see any of it, anywhere.

LInk will probably pop up on www.npr.org later today.

--
Best regards,
bitbitch

14.

Subject: [ILUG] Using Normal IDE Device with a Dell Latitude CPx
Hi,

I've got an normal 3.5" CD-RW IDE drive that I'd like to be able to use with a Dell Latitude CPx laptop that I've got. Does anyone know any way to enable this, for example through the use of a special cable for the Modular Bay (where CD-ROM or floppy drive is normally).

There is also the possibility of using a docking station, but Dell's docking solution for the Latitude series doesn't seem to allow for the use of an IDE drive, only SCSI... Unless someone knows of a "non-Dell" solution that's compatible.

Anyone any ideas?

Thanks,

Darren.

15.

Subject: AA Meetings the Hottest Place to Meet Women With Big Bucks

AA Meetings the Hottest Place to Meet Women With Big Bucks

And, as always, you can take a page out of Fight Club and start showing up at all sorts of support groups. Look what it did for Marla and Jack...

"JACK You can't have *both* parasites. You take blood parasites and --
MARLA I want brain parasites.
She opens another dryer and does the same thing again. PG 19
JACK Okay. I'll take blood parasites and I'll take organic brain dementia and --
MARLA I want that.
JACK You can't have the whole brain!
MARLA So far, you have four and I have two!
JACK Well, then, take blood parasites. Now, we each have three.
MARLA So, we each have three -- that's six. What about the seventh day? I want ascending bowel cancer.
JACK *I* want ascending bowel cancer.
MARLA That's your favorite, too? Tried to slip it by me, huh?
JACK We'll split it. You get it the first and third Sunday of the month.
MARLA Deal."