

# Utvärdering av några skräppostfilter

Hanna Berglund  
hannab@stp.ling.uu.se

Examensarbete i datorlingvistik  
Språkteknologiprogrammet  
Uppsala universitet · Institutionen för lingvistik och filologi

6 juni 2005

**Handledare, Uppsala universitet  
Bengt Dahlqvist**

## Sammandrag

Detta examensarbete är en undersökning och utvärdering av program som utför klassificering och filtrering av de meddelanden som skickas via e-post. Utvärderingen baserar sig på tester av filterprogrammen med hjälp av en korpus, Uppsala Spam Corpus, bestående av skräppostmeddelanden och hammeddelanden. De filter som testas mot korpusen är SpamAssasin, en probabilistisk e-postklassificerare som skapats vid institutionen för lingvistik och filologi, samt Bishop 0.3.0 och NNSpam. Filtren utvärderas efter deras respektive metods förmåga att korrekt klassificera spam respektive ham. I uppsatsen presenteras de olika metoder som de utvalda filtren använder för att skilja mellan skräppost och ham, i och med detta tar jag även upp ämnet lagstiftning på området, vilket är ytterligare en metod att jobba med problemet skräppost. Jag redogör också för bakgrunden till problemet och dess uppkomst samt en kategorisering av skräppost. I examensarbetet gör jag även en liten enkätundersökning för att se hur problemet med skräppost hanteras hos olika företag.

# Innehåll

Tack . . . . .	iii
<b>1 Inledning</b>	<b>1</b>
1.1 Syfte . . . . .	1
1.2 Uppsatsens upplägg . . . . .	1
<b>2 Bakgrund</b>	<b>2</b>
2.1 Historik . . . . .	2
2.2 Kategoriindelning av skräppost . . . . .	4
2.2.1 Bedrägerier . . . . .	4
2.2.2 Säljbrev/Reklam . . . . .	5
2.2.3 Kedjebrev/Lotteribrev . . . . .	5
2.2.4 Övrig skräppost . . . . .	5
2.3 Metoder för skräppostfiltrering . . . . .	6
2.3.1 Svartlistning . . . . .	6
2.3.2 Vilstning . . . . .	6
2.3.3 Grålistning och challenge-responssystem . . . . .	6
2.3.4 Heuristiska/Regelbaserade metoder . . . . .	7
2.3.5 Statistiska metoder . . . . .	7
2.3.6 Neurala nätverk . . . . .	9
2.3.7 Manuellt . . . . .	9
2.4 Lagar kring skräppost . . . . .	10
2.4.1 Svensk lagstiftning . . . . .	10
2.4.2 Lagar inom EU . . . . .	11
2.4.3 Lagar i USA . . . . .	11
2.4.4 Skillnader mellan amerikansk och svensk lagstiftning . . . . .	11
<b>3 Enkätundersökning</b>	<b>13</b>
<b>4 Uppsala Spam Corpus</b>	<b>15</b>
<b>5 Utvärdering av skräppostfilter</b>	<b>16</b>
5.1 Utvärderingsmetod . . . . .	16
5.1.1 Precision, täckning och korrekthet . . . . .	16
5.2 Urval av filter . . . . .	17
5.2.1 SpamAssassin . . . . .	17
5.2.2 Bishop 0.3.0 . . . . .	19
5.2.3 Neural Networks For Spam Detection . . . . .	20
5.2.4 Probabilistisk e-postklassificerare . . . . .	21
5.2.5 DSPAM . . . . .	21

5.2.6	0Spam . . . . .	23
5.2.7	MailScanner . . . . .	23
<b>6</b>	<b>Resultat</b>	<b>25</b>
6.1	SpamAssassin . . . . .	25
6.2	Bishop 0.3.0 . . . . .	26
6.3	Neural Networks For Spam Detection . . . . .	26
6.4	Probabilistisk e-postklassificerare . . . . .	27
<b>7</b>	<b>Diskussion och slutsatser</b>	<b>28</b>
7.1	Vad säger mina resultat från undersökningarna? . . . . .	28
7.2	Framtid? . . . . .	29
<b>A</b>	<b>Ordförklaringar</b>	<b>30</b>
<b>B</b>	<b>Exempel från Uppsala Spam Corpus</b>	<b>31</b>
<b>C</b>	<b>Länkar till hemsidor för skräppostfilter</b>	<b>34</b>
<b>D</b>	<b>Nätverk i NNSpam</b>	<b>35</b>
	<b>Litteraturförteckning</b>	<b>37</b>

# Tack

Jag vill tacka min handledare, Bengt Dahlqvist, för all hjälp med denna uppsats. Jag vill också tacka Mats Dahllöf, som hjälpt till med uppsatsen i dess slutskede. Tack även till Lotta Nordling, för korrekturläsning samt alla behövliga kaffepauser under arbetets gång. Tack till Moa Olsson, för hjälp med det juridiska, och Per Starbäck för alla tips och råd vad gäller datorfrågor. Tack även till alla andra, ingen nämnd och ingen gkömd, som varit till stöd och hjälp för mig under den här tiden.

# 1 Inledning

## 1.1 Syfte

Syftet med uppsatsen är att utvärdera skräppostfilter och undersöka problemet med skräppost, vilka metoder som finns att bekämpa skräppost samt hur och hur väl dessa metoder fungerar. De bekämpningsmetoder som beskrivs är en översikt av de lagar som finns för att motverka skräppost samt en del som beskriver de tekniska metoder som finns för att bekämpa skräppost. I uppsatsen utvärderas några skräppostfilter med avseende på precision, täckning och korrekthet samt vilka metoder som de respektive filtren använder och hur dessa fungerar, särskilt med avseende på deras sätt att hantera text. För att testa filtren används Uppsala Spam Corpus som sammanställts på institutionen för lingvistik och filologi vid Uppsala universitet. Det redovisas även några till filter som inte utvärderas enligt metoderna ovan, men som beskrivs för att visa på det stora utbud av varierande filter som finns för att klassificera e-post. I uppsatsen redovisas även en mindre enkätundersökning om hur skräppost hanteras av företag.

## 1.2 Uppsatsens upplägg

I bakgrundskapitlet (kapitel två) ges en introduktion till problemet med skräppost och dess uppkomst, för att ge förståelse för ämnet. Därtill beskrivs vilka olika sorters skräppost som förekommer samt de metoder för att bekämpa skräppost som används. Dessa metoder innefattar tekniska metoder som svartlistning och heuristiska metoder samt lagar mot skräppost, som också kan sägas vara en metod för att bekämpa skräppost om än på ett annat sätt. Enkätundersökningen beskrivs i kapitel tre. I kapitel fyra beskrivs Uppsala Spam Corpus som är de testdata jag använt i min utvärdering. I femte kapitlet, utvärdering av skräppostfilter, beskrivs de metoder jag använt för att beräkna resultaten från testkörningarna, accuracy, precision och recall. Där framställs också de olika skräppostfiltren och deras metoder för att klassa vad som är skräppost och vad som inte är det. I kapitel sex beskrivs sedan de resultat som testkörningarna av dessa program med Uppsala spam corpus gett upphov till. Resultaten och vad de kan säga diskuteras slutligen i kapitel sju.

# 2 Bakgrund

## 2.1 Historik

Ordet spam kommer ursprungligen från namnet på en burkskinka tillverkad av Hormel Foods ('Shoulder Pork and hAM'/'SPiced hAM'). Idag betecknar ordet spam skräppostmeddelanden som skickas till personers eller företags e-postadresser eller skräpinlägg till USENET. Men hur fick ordet spam den mening det har idag? Uppkomsten av ordet spam i dagens betydelse kopplas ihop med en sketch av Monty Python's Flying Circus. Sketchen utspelar sig på en restaurang där en kund frågar vad som finns på menyn. Och servitrisen som berättar menyn för honom upprepar ordet hela tiden för att tala om hur mycket spam det är i rätterna, det är "egg and bacon; egg, bacon, and spam; egg, bacon, sausage, spam; spam, bacon, sausage, and spam; spam, egg, spam, spam, bacon, and spam;". Under tiden hon gör detta är det en grupp vikingar som börjar sjunga en sång som går ungefär så här: "Spam, spam, spam, spam, spam, spam, spam, spam, lovely spam!" "Wonderful spam!".

De första användningarna av ordet, i dagens mening, skedde på chatrum på Internet och på de Internetspel som kallas MUDs(multiuser dungeons)(Schwarz and Garfinkel 1998). Några oseriösa användare upprepade samma meddelande om och om igen i ett chatrum, och fyllde på så sätt hela skärmen. Andra användare kallade meddelandena för spam, precis som i Monty Pythons sketch var det fråga om meningslöst upprepande. Alltså betyder spam något som upprepas och upprepas tills det blir väldigt irriterande, precis som dagens skräppostmeddelanden uppfattas.

Det finns många exempel på tidiga former av spam och spamliknande företeelser. Redan i november 1975, insåg Internetpionjären Jon Postel att det fanns grundläggande brister med elektronisk post. Enligt Postel var det möjligt att attackera en dator bara genom att skicka mer post än den kunde hantera. Han hade dock ingen lösning till detta potentiella problem annat än att säga "det skulle vara nyttigt för en värd (host) att kunna avvisa meddelanden från källor den tror uppför sig illa eller bara är irriterande."

Och det blev problem. Under 1980-talet uppstod problemet med elektroniska kedjebrev. Kedjebrev i elektronisk form har samma innehåll som deras föregångare i pappersform, de lovar ofta framgång och rikedom, men på villkor. Om man skickar kedjebrevet vidare till ett givet antal nya mottagare så kommer lyckan att le emot en, men om man bryter kedjan kommer lidande och otur att drabba en. Kedjebrev i pappersform har haft ett större självsanerande eftersom de är betydligt jobbigare att sända vidare, det behövs papper, kuvert och frimärken och dessutom måste man bege sig till postlådan. Att sända vidare ett kedjebrev via en dator är betydligt enklare. Men även på datorer var det omöjligt att skicka breven till obegränsat många mottagare. Efter fyra generationer så hade ett enkelt meddelande på 2 KB växt till att ta upp 20 MB utrymme. Och efter fem eller sex generationer, så kunde det stänga ner den dator på vilken det kopierades, eftersom det inte fanns mer lagringsutrymme kvar. Många universitet och företag försökte undervisa sina användare om faran med att skicka kedjebrev. Men det visade sig vara svårt, dels såg få användare faran med att skicka iväg bara fem eller tio meddelanden och i takt med att datoranvändandet ökade blev det helt enkelt för många användare att utbilda.

Ett av de första kommersiella spammen eller skräppostmeddelandena kom den 12 april 1994 när två advokater från Phoenix, Laurence Canter och Martha Siegel, skickade ett meddelande till mer än 6000 nyhetsgrupper på Usenet, världens största konferenssystem online. Meddelandet innehöll reklam som marknadsförde deras tjänster i samband med det amerikanska 'Green card-lotteriet'. De hade skickat meddelandet några gånger tidigare, men den här gången tog de hjälp av en programmerare som skrev ett enkelt script för att skicka annonsen till så många av Usenets nyhetsgrupper. Även om det var ett litet massutskick med dagens mått mätt så orsakade det starka reaktioner eftersom det bröt mot den underförstådda nätetiketten. Det fanns två anledningar till att Canters och Siegels spammande av Usenet upprörde folk. Det första var omfattningen och fräckheten hos utskicket. Ingen hade postat ett meddelande till alla Usenets nyhetsgrupper tidigare. Det andra var innehållet i meddelandet. Green Card-lotteriet var gratis att delta i, men Canters och Siegels reklammeddelande gav sken av att om man betalade 100 dollar till deras byrå, kunde man som illegal invandrare öka sina chanser att vinna. Canter och Siegel fick mycket kritik och deras ISP blev avstängd. De lät sig dock inte nedslås utan skickade några ytterligare skräppostmeddelanden från andra ISP:er och skrev så småningom även en bok som heter *How to Make a Fortune on the Information Superhighway*. Detta var ett genombrott för skräpposten och efterhand blev Usenets nyhetsgrupper mer eller mindre oanvändbara och sedan följde nästa steg: massutskick till personers e-postadresser. Och fenomenet med skräppost spred sig.

En som läste Canters och Siegels bok var Jeff Slanton. Han arbetade som Gulasidornaförsäljare på US West Direct i Albuquerque i New Mexico och han beslöt sig för att prova på marknadsföring via e-post för att se om det skulle fungera. Under första halvåret 1995 började Slanton att samla in e-postadresser, adresser från mejllistor och namnen på nyhetsgrupper på Usenet, och i juli 1995 skickade han ut sitt första spam. Men innan Slanton skickade ut meddelandet, frågade han ledningen vid sin ISP, Route 66, om de skulle misstycka om han spammade världen från sitt konto. De svarade att de visst skulle misstycka och Slanton meddelade att han ville avsluta sitt konto i slutet av månaden. Och med bara två dagar kvar innan avslutandet sände han ut sitt första massmeddelande. Reklammeddelandet handlade om planerna för de första atombomberna och det hamnade överallt. Det hamnade visserligen hos grupper som kunde vara intresserade av meddelandet, men också hos dem som inte var det och till och med hos grupper där det kunde ses som direkt olämpligt att det hamnade, som hos en stödgrupp för människor med hjärntumörer. Men Slanton kände inga ångerkänslor över sitt tilltag. Slanton fortsatte med sitt spammande och han utforskade många nya tekniker och utnämnde sig själv 'Spam King'. Några av de tekniker Slanton använde sig av var dessa. För att begränsa andelen klagomål via mejl började Slanton skicka skräppost från fiktiva e-postadresser och domäner, till exempel `SpAmKiNg@505-821-1945-new.LOW.rates`. Och för att skydda sig själv och sina kunder från förföljelse så såg Slanton till att hans spammeddelanden bara innehöll telefonnummer som gick till en röstbrevlåda och inte till en riktig telefonlinje. Han visste även att ISPs snabbt skulle stänga hans konton och hålla honom ansvarig för den skräppost som skickats ut, alltså fick han sina kunder att använda disponibla konton och sedan ringa honom med det användarnamn, lösenord och telefonnummer som han förutsatte sig att använda. Hans program för att skicka meddelanden använde sig också av en teknik som gick ut på att de sända ut satsvis med e-postmeddelanden till avlägsna datorers mejlservrar, vilka sedan skickade iväg individuella meddelanden. Detta såg till att Slanton kunde skicka ut betydligt fler meddelanden över ett modem än vad som annars varit möjligt. En ytterligare åtgärd för att minska klagomålen gick ut på att Slanton hävdade att han använde sig av möjligheten till opt-out (se Appendix A för förklaring). Men då han skickade sina spam till både mejllistor och nyhetsgrupper så fanns det ingen opt-outmekanism som användes eller kanske ens existerade. Och Spamkungen fortsatte med sina utskick. Under 1995 sade sig Slanton skicka upp till 15 olika spam per vecka, och ofta blev det hans kunder som fick ta smällen från arga människor som hade fått oönskade meddelanden. Många följde också i hans fotspår, medan andra försökte motarbeta Slanton på olika sätt, något som Slanton själv inte hade något emot då han påstod att uppmärksamheten var bra för hans affärer. Han gjorde till och med en reklamkampanj för sig själv, där han påstod att man kunde få sin adress

borttagen från hans databas för en summa av 5 dollar. Det var lurendrejeri, och enligt Slanton gjorde han det för att skapa en kontrovers. Med Slantons fortsatta spammande blev vissa Internetlistor stängda så att bara medlemmar kunde skicka inlägg medan andra konfigurerades så att de bara accepterade mejl som kom från en moderator. Men det skrämde eller stoppade inte Slanton, enligt honom gjorde sådana lösningar det bara lite mer utmanande för den som var övertygad.

Och man får väl säga att Slanton hade rätt. Idag har skräppostfenomenet spridit sig och blivit en riktig plåga i dagens elektroniska samhälle där det utgör en så stor del som 86 procent av alla meddelanden som skickas. Tidigare var skräpposten ett stort irritationsmoment för några få Internetanvändare. Idag är problemet så stort att majoriteten användare av e-post uppfattar fenomenet som ett problem. Dels beror det på den stora mängd skräppost som användaren måste radera, dels beror det på det besvärande innehållet i många av breven. Problemet är dock större än att bara vara ett irritationsmoment. I samband med skräppostens explosiva ökning har även de finansiella kostnaderna ökat, om än indirekt, för alla användare. Detta beror på att man idag inte betalar något för att skicka e-post, men alla måste vara med och betala för den nätverkstrafik som uppstår till följd av skräpposttrafiken

## **2.2 Kategoriindelning av skräppost**

### **2.2.1 Bedrägerier**

#### **Nigeriabrev**

Nigeriabrev kommer inte alltid från Nigeria. Det är en form av bedrägeri som till en början mestadels drabbade företag. Med den ökade användningen av e-post så har även många privatpersoner drabbats. Det är en form av bedrägeri där själva breven är en del av ett större brott. Hur ser då ett Nigeriabrev ut? De kommer från olika avsändare, men för det mesta har denne avsändare en titel som till exempel doktor, prins eller kamrer. Denna person påstår sig ofta ha blivit rekommenderad att kontakta just dig. Avsändaren har kommit över en stor summa pengar och behöver din hjälp för att få ut pengarna ur ett land. Det finns något hinder som gör att avsändaren själv inte kan föra ut pengarna ur landet, och som tack för hjälpen ska du få en summa av dessa pengar. Det hävdas att man kommer att tjäna miljoner bara på att låna ut sitt bankkonto för en transaktion, men när en person fångats in och svarat så behövs mer hjälp och kanske en mindre inbetalning för att muta en tjänsteman eller så och på det viset kommer det att fortsätta tills offret blir djupt indragen i svindeln. Då är det oftast svårt att dra sig ur också eftersom man betalat in mycket pengar som man då skulle förlora. Varje år luras ungefär 20-talet svenska företag och privatpersoner av dessa ligor av svindlare. I Norden har dessa ligor lurat till sig uppskattningsvis 100 miljoner kronor i Norden. Nigeriabrev dök i Sverige upp redan på 50-talet, men på grund av att avsändarna verkar från andra länder, har nästan ingen någonsin dömts för brott.

#### **Phishing**

Phishing (uttalas som fishing) är en annan form av Internetbedrägeri. Det fungerar så att den som skickar ut meddelandena vill få fram känslig information om Internetanvändare för att sedan använda denna information i brottsliga syften.

Det finns två olika typer av phishing. Den ena fungerar så att användaren får ett meddelande från en falsk avsändaradress. Meddelandet ser riktigt ut, och verkar komma från ett seriöst företag, till exempel en bank. Mottagaren uppmanas att lämna ut ömtålig information, som sitt kontonummer och sin bankomatkod. Detta kan ske på två sätt, antingen genom att mejla tillbaka informationen eller genom att följa en länk till en hemsida. Denna hemsida ser ut att vara företagets riktiga hemsida, men är en falsk sådan, och där får offret fylla i ett formulär med sina känsliga uppgifter som sedan kan användas av svindlarna.

Den andra formen av phishing-meddelande fungerar så att mottagaren får ett e-postmeddelande som innehåller ett bra erbjudande. Det kan exempelvis vara ett erbjudande om att köpa ett billigt datorprogram på nätet. I meddelandet finns en länk till en hemsida. När användaren går in på hemsidan laddas ett program ner till dennes dator utan att det märks. Detta dolda program samlar sedan in information som användarnamn och lösenord till banktjänster och skickar dem vidare.

### 2.2.2 Säljbrev/Reklam

Att spamma är i de flesta fall en affärsverksamhet, och spammarnas mål är en vinst. För att nå denna vinst måste vissa steg utföras. Först måste en potentiell kundkrets finnas. För att spammaren ska nå ut till sina tänkbara kunder krävs att han eller hon får tillgång till en lista med e-postadresser. Det finns två huvudmetoder för att komma åt en sådan lista, adresssamlade<sup>1</sup> och att köpa färdiga listor. Som ett andra steg måste spammaren kunna erbjuda de potentiella kunderna en produkt eller en tjänst. Som det tredje och sista steget i denna kedja måste spammaren sälja och leverera produkten eller tjänsten till en viss procent av de potentiella kunderna.

Anledningen till att skräppost som säljer en vara eller en tjänst är en sådan succé beror på de låga kostnaderna för de två första stegen i kedjan. Eftersom det är så billigt att skaffa fram e-postadresser och skicka ut e-postmeddelandena så kan även en väldigt låg svarsprocent leda till en vinst. Det finns olika typer av säljbrev med produkter och tjänster. De kan delas upp i de som erbjuder möjligheter till investeringar och affärsmöjligheter, som att jobba hemma. De som vänder sig till en 'vuxen' publik, till exempel pornografi och dating-tjänster. Det finns också en finansiell kategori, med offerter om kreditkort, köp av valuta och så vidare. Och kategorin som handlar om kropp och hälsa, här ingår de spam som erbjuder möjligheter till organförstoringar och pharmacy-brev. Pharmacy-brev är en vanligt förekommande form av skräppost. Det är ofta olika potenshöjande läkemedel som saluförs, till exempel Viagra. Sedan finns det också de spam som erbjuder reserelaterade förslag som semestererbjudanden, och de spam som riktar in sig på utbildning, till exempel erbjudanden om examensbevis från olika skolor.

### 2.2.3 Kedjebrev/Lotteribrev

Kedjebrev är ganska luriga och verkar ofta inte erbjuda något utan bara vilja ha din uppmärksamhet. De innehåller ofta meningar som säger saker som "det här är sant, fortsätt läsa, det är lagligt" och så vidare. Detta för att vinna ditt förtroende, men som e-postmottagare bör man vara försiktig, särskilt om det innehåller erbjudanden om att tjäna pengar utan ansträngning. De innehåller ofta hot om att något hemskt kommer att hända om brevet inte skickas vidare. En speciell form av kedjebrev är hoaxvirus, som innebär att någon sänder ut en falsk virusvarning och påstår att den kommer från en säker källa. För att brevet ska verka äkta används tekniska termer och personen som får meddelandet uppmanas att skicka det vidare till alla som han/hon känner.

### 2.2.4 Övrig skräppost

Kinesiskt skräp kallar jag här den sorts skräppost som bara innehåller tecken som inte ingår i vårt latinska alfabet och måste alltså inte komma från Kina utan även från andra länder med ett annat alfabet som exempelvis Korea. Det kan säkert ofta vara fråga om skräppost som tillhör någon av de tidigare kategorierna, men eftersom vi inte kan tolka dessa meddelanden så får de hamna under kategorin övrigt. Ett filter som är känsligt för vilka språk ett legitimt meddelande får vara skrivet i kan sortera bort denna form av skräppost<sup>2</sup>.

---

<sup>1</sup>address harvesting - innebär att e-postadresser samlas in, oftast från allmänna domäner genom att till exempel söka igenom hemsidor eller hitta förutsägbara adressnamn

<sup>2</sup>Informationen till kategoriindelningen är hämtad från följande källor (Svenskt näringsliv 2004), (Lunds universitet 2003), (Högskolan i Kalmar 2003) & (Federal Trade Commission 2003)

## 2.3 Metoder för skräppostfiltrering

### 2.3.1 Svartlistning

Ett vanligt sätt att blockera ovälkomna e-postmeddelanden är genom att hindra inkommande meddelanden i mejlservern på grund av deras ursprung. De olika svartlistorna baserade på DNS-baserade databaser innehåller varierande adresser och nätverk som listats beroende på deras syfte och målsättning. Vanliga databaser innehåller open proxies, öppna reläer (open relays), nätverk eller individuella adresser som är skyldiga till att ha skickat skräppost, nätverk som är kända att bestå av uppringsningsanvändare (dial-up users) och andra mindre vanliga listor. De flesta av listorna innehåller nätverk som mejlserveroperatörer troligtvis inte vill ska nå deras server. Tillägg i listorna görs genom olika metoder som skiljer sig från lista till lista. Men generellt utpekas vissa nätverk och adresser. De blir då utredda av moderatorer, som försöker kontakta ägarna så att de kan korrigera de problem som uppstått. Om problemet inte åtgärdas och nomineringarna fortsätter, så kan de adderas till listan. Dessa listor är vitt spridda och användas av de som tillhandahåller mejlservice, vilket är ett starkt incitament för ISP:er att försäkra sig om att deras användare inte missbrukar nätverket och får hela ISP:ns nätverk eller mejlserver svartlistad.

Eftersom blockering av ett meddelande med hjälp av svartlistning som baseras på ursprung sker innan meddelandet är mottaget och bearbetat av den mottagande mejlservern så kan svartlistning reducera många av kostnaderna som förknippas med oönskad e-post. Men svartlistning är ingen slutgiltig lösning på problemet med skräppost. Det kan till exempel inte filtrera bort skräppost som kommer från en server som inte är svartlistad, och likt många andra lösningar finns risken för att meddelanden som användaren vill ha filtreras bort, särskilt om ett nätverk blir svartlistat på grund av en missbrukande användare eller en mejlserver som konfigurerats felaktigt (Hird 2002).

### 2.3.2 Vitlistning

Vid vitlistning används en metod där enbart e-postmeddelanden från en redan känd kontakt eller säker/pålitlig domän accepteras och släpps igenom filtret. Här ges användaren enbart möjlighet att ha kontakt med en förutbestämd mängd kontakter. Den mest uppenbara bristen med en sådan metod är just att den begränsar kommunikationen till en redan känd krets, vilket är opraktiskt för väldigt många användare då man inte på förhand kan veta om alla man kan vilja eller behöva bli kontaktad av eller kontakta själv.

### 2.3.3 Grålistning och challenge-responssystem

Ett system som använder sig av grålistning fungerar så att det inte automatiskt tar emot alla meddelanden. I stället för att direkt ta emot ett meddelande och därefter bearbeta det så skickar det centrala e-postsystemet ett meddelande till det uppkopplade systemet och säger "Jag kan tyvärr inte ta emot ditt meddelande just nu, prova igen om en liten stund". Ett riktigt e-postsystem kommer att göra precis detta, medan spammarnas e-postsystem än så länge inte tar någon notis om vad mottagarnas e-postsystem svarar och alltså inte provar att skicka meddelandet igen, vilket ett legitimt e-postsystem gör och denna gång får meddelandet komma igenom. Ett sådant här system kan medföra en försening av meddelanden till en början, men har ett meddelande släppts igenom så behöver nästa brev från samma avsändare inte gå igenom samma procedur utan kommer igenom automatiskt. Det vill säga, det vitlistas.

En annan metod på samma tema är ett challenge-respons system, vilket används för att verifiera att det faktiskt finns en avsändare till meddelandet, detta sker genom att skicka en utmaning till sändaren och kräva ett riktigt svar. Kritiker till ett sådant här system menar att det tar upp en massa ytterligare bandbredd, då många skräppostmeddelanden har förfalskade from-adresser. Dessa adresser, som kan tillhöra helt oskyldiga offer för spammare, kan sen bli bombarderade med meddelanden från sådana här system. Det finns andra tillfällen när detta sätt att uppnå en kontakt inte är möjligt att använda, som när man

handlar online och får en bekräftelse av sin order via mejl.

### 2.3.4 Heuristiska/Regelbaserade metoder

Filtreringsmetoder som baserar sig på heuristiska metoder och kontroll av innehållet i ett meddelande. Att använda heuristiska metoder kan hjälpa upp problemet med skräppost genom andra tekniska lösningar, då e-post filtreras på grund av innehållet i meddelandena snarare än på grund av det sätt de har kommit fram på. Det kan användas transparent, utan att användaren behöver ändra sitt beteende eller sin mjukvara. På grund av detta är detta en vanlig metod som implementeras av många operatörer, särskilt för att reducera skräppost som är stötande och som är troliga att innehålla vissa förutsägbara nyckelord som kan användas för filtreringen, till exempel ?viagra?. Genom att placera ett filtersystem vid sidan av sin server, låter det användarna att få tillgång till filtrering utan att behöva någon extra mjukvara, det tillåter också flexibilitet för användarna att ladda ner de meddelanden som inte har blivit klassade som skräppost. Det kräver dock stora resurser på servern för att bearbeta all e-post för alla användarkonton, och det skapar svårigheter med att anpassa filtret till varje särskild användares behov. Att filtrera baserat på innehåll gör också lite för att ta hand om problemen med bandbredd och lagringskapacitet som orsakas av skräppost eftersom att meddelandena fortfarande måste tas emot för att kunna hanteras. Här finns också problemet med vad man ska göra med de meddelanden som klassas som skräppost. Att bara göra sig av med skräppostklassade meddelanden räknas som ett dåligt alternativ, främst på grund av möjligheten att det är fråga om en falsk positiv. Det har också visat sig att filtrering baserat på innehåll inte når upp till, och är inte särskilt troligt att det någonsin kommer att nå upp till, hundraprocentig korrekthet. Och användare vill hellre ha ett filter som missar att klassificera en liten procent av all skräppost (false negatives) än ett filter som felaktigt rankar en liten procent av välkomna mejl som skräppost (false positives). Risken för att märka upp ett meddelande fel och få en falsk positiv medför att det bör finnas en konservativ hållning till filtrering. Filter brukar generellt inte radera e-post som blivit märkta som skräppost, utan istället hantera dem så att de inte stör eller blandas ihop med de meddelanden som har passerat genom filtret, men så att de fortfarande finns åtkomliga för att kunna ses igenom och eventuellt tas tillbaka.

SpamAssassin är ett filter som använder sig av kollaborativ filtrering (Razor- och DCC-kontroller) som letar och identifierar ickekorrekta eller misstänkta huvuden (headers), antal mottagare och så vidare. Det sätt som SpamAssassin och andra liknande filter använder sig av är likartat det som en riktig person skulle använda sig av för att bedöma om ett meddelande var skräppost eller inte. En person skulle titta på 'From:'-adressen och subjektraden, se om det kommer från någon de känner eller är något de förväntar sig eller om det verkar slumpmässigt gjort eller kommersiellt. Om meddelandet passerar denna koll skulle en person snabbt gå igenom innehållet i meddelandet och på så vis skaffa sig bevis för vad det är för slags meddelande. När meddelandet överstiger en viss tröskel så kommer meddelandet att klassas som skräppost och därefter hanteras som ett sådant bör hanteras. Om det däremot klassas som legitimt, kan SpamAssassin också använda sig av automatisk vitlistning (AWL), vilket tillåter systemet att hålla en adress som säker. Detta minne kan sedan användas i framtiden för att ge poäng till meddelanden baserat på andelen legitima meddelanden som sänts från den specifika adressen.

### 2.3.5 Statistiska metoder

Den för närvarande mest populära statistiska metoden för att kämpa mot skräppost är Bayes metod. Thomas Bayes var en engelsk präst som levde 1702 till 1761. Hans uppsats *Essay towards solving a problem in the doctrine of chances* publicerades först efter hans död (1763) och i den ger Bayes en formel som beskriver hur man bedömer sannolikheten för att ett antagande är korrekt med ledning av vad som redan är känt om liknande fall. Denna postuma publicering av Bayes uppsats gjorde honom känd bland matematiker, och på slutet av 1900-talet kom hans verkliga genombrott. Anledningen till att Bayes metod inte blev populär tidigare var att den inte passade för beräkningar åt till exempel samhällsplanerare och

försäkringsbolag, men att den visade sig vara som gjord för datorernas värld (Lotsson 2004). Bayes formel beräknar hur man med ledning från tidigare erfarenheter kan bedöma hur troligt det är att en specifik händelse kommer att inträffa givet vissa faktorer. Bayes formel ser ut på följande sätt:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

För att göra en analys med hjälp av Bayes metod behöver man följande:

- Ett antagande - till exempel, "detta e-postmeddelande är skräppost för det innehåller ordet Viagra".
- Ett eller flera alternativ - till exempel "detta e-postmeddelande är inte skräppost, trots att det innehåller ordet Viagra".
- Ett faktum - till exempel "detta e-postmeddelande innehåller ordet Viagra".
- Underlag - som är ett antal meddelanden att jämföra med. De måste vara uppdelade i spam och icke-spam. För att analysen ska bli meningsfull måste det finnas e-post innehållande ordet Viagra samt utan ordet Viagra. Det är också bra om det finns en bedömning av hur pass representativt underlaget är, det vill säga hur väl proportionen spam/ham stämmer överens med verklighetens proportion. Om de inte stämmer överens blir a priori-sannolikheterna missvisande och kommer att störa den övriga skattningen.

Bayes metod är en jämförelse av sannolikheter. Jag ger här ett exempel där alla data är hämtade från Uppsala Spam Corpus, den korpus som jag använder i mina utvärderingar. Korpusen består av 5005 e-postmeddelanden, varav 3303 är spam och 1702 är ham. Ordet Viagra förekommer i 243 av 3303 spam (7,4 procent) men bara i 1 av de 1702 hammeddelandena (0,1 procent). Sannolikheten för att ett meddelande innehåller ordet 'Viagra' ( $P(\text{Viagra})$ ) blir då 4,9 procent ( $244 / 5005$ ). Härifrån kan man beräkna sannolikheten för om ett meddelande är ett skräppostmeddelande eller ett hammeddelande givet att det innehåller ordet 'Viagra' på följande sätt:

$$P(\text{spam}|\text{viagra}) = P(\text{viagra}|\text{spam})P(\text{spam})/P(\text{viagra}) = (0,074 * 0,66)/0,049 \approx 0,99$$

Man kan på samma sätt beräkna sannolikheten för att det skulle vara ett ham:

$$P(\text{ham}|\text{viagra}) = P(\text{viagra}|\text{ham})P(\text{ham})/P(\text{viagra}) = (0,001 * 0,34)/0,049 \approx 0,007$$

Sannolikheten för att meddelandet är ett skräppostmeddelande respektive ett hammeddelande är alltså 99 procent respektive 0,7 procent. Bayes löste problemet med att underlaget för de olika sannolikheterna är olika. I detta exempel är det ju mer spam än ham, men antalet brev räknas inte. Istället jämförs två olika sannolikheter. Fördelen är att metoden kan användas för att jämföra antaganden som utgår från ojämförbara underlag. Till exempel: Beror de varma somrarna på växthuseffekten eller på naturliga variationer i klimatet? Uppenbarligen beräknar man följderna av växthuseffekten på ett annat sätt än man beräknar naturliga variationer i klimatet. Men om man kan få fram mått på sannolikheterna kan man ändå jämföra de två hypoteserna med hjälp av Bayes metod. Man måste komma ihåg att Bayes metod är ett sätt att utvärdera antaganden, och inte ett sätt att komma fram till en absolut sanning.

Att ha ett skräppostfilter som bara tar hänsyn till ett ord är inte särskilt användbart. Man måste ta hänsyn till alla ord i e-posten eller åtminstone till ett urval. För att räkna med flera ord gör man en kedjeberäkning eller bayesiskt nätverk. Enklaste sättet att göra detta för ett skräppostfilter är genom att ta ett ord i taget och för varje ord som bearbetas får man en ny sannolikhet. Som exempel kan man anta att meddelandet förutom ordet 'viagra' också kan väntas innehålla ordet 'prescription', och att ordet 'prescription' förekommer i 8,7 procent av alla skräppostmeddelanden och i 0 procent av alla ham.

För att då beräkna sannolikheten för att meddelandet givet orden 'viagra' samt 'prescription' är ett skräppostmeddelande används följande formel:

$$P(\text{spam}|\text{viagra}, \text{prescription}) = \frac{P(\text{spam}, \text{viagra}, \text{prescription})}{P(\text{prescription}|\text{viagra}) * P(\text{viagra})}$$

$P(\text{spam}, \text{viagra}, \text{prescription})$  är sannolikheten för de meddelanden som uppfyller kraven att vara skräppost samt innehålla orden "viagra" och "prescription", i detta exempel är det  $90/5005 \approx 0,018 = 1,8$  procent.  $P(\text{prescription} | \text{viagra}) = 90 / 244 \approx 0,37 = 37$  procent och  $P(\text{viagra})$  är enligt tidigare beräkning 4,9 procent.  $P(\text{spam}|\text{viagra}, \text{prescription})$  blir enligt dessa siffror 99 procent ( $0,018 / (0,37 * 0,049)$ ).

### 2.3.6 Neurala nätverk

Neurala nätverks sätt att attackera skräppost bygger på att efterlikna den mänskliga hjärnans sätt att arbeta, att samla erfarenheter och med dessas hjälp göra en bedömning (Miller 2003). Neurala nätverk bygger på igenkänning av mönster.

Definitionen av ett neuralt nätverk är en uppsättning av internt sammankopplade noder eller neuroner. Det mest komplexa och därtill mest kända neurala nätverket är den mänskliga hjärnan. Vår hjärna utför mängder med beslut och för att ta dessa beslut krävs ofta både medveten och undermedveten input. När vi växer upp och lär oss mer om vår omvärld har vi en enastående förmåga att känna igen mönster, både familjära och avvikande, väldigt snabbt och utan särskilt mycket medvetet tänkande. För att ta ett förenklat exempel så springer vi inte vår väg när vi ser en leopardpälss trots att alla leoparder som faktiskt är farliga för oss bär en sådan. Om vi skulle göra det skulle det innebära att vårt tankesystem var binärt och bara tog hänsyn till om 'leopardpälss' utgör en fara eller ej och detta är inte nödvändigtvis det korrekta beslutet. Istället tar vår hjärna hänsyn till andra bitar av information, som att vi befinner oss mitt i en stad, i ett varuhus och att det är en person som har leopardpälss på sig. Skulle vi däremot befinna oss i en djungel, och det som hade på sig pälss morrade och hade skarpa tänder och vi kommit bort från vår turistgrupp skulle beslutet att fly vara betydligt mer korrekt. Detta är ett konstruerat exempel, men det visar att det oftast inte är bara en bit information som bestämmer hur vi kommer att reagera. Och detta är inget som vi gör medvetet utan vi litar på att vår hjärna undermedvetet kommer att känna igen de mönster som behövs för att dra en slutsats.

När man använder datorbaserade neurala nätverk gör man en artificiell approximation som försöker härma den mänskliga hjärnan. Dessa neurala nätverk kan användas i bekämpningen av skräppost. Skräppost är en utmaning inte bara när det gäller det stora antalet meddelanden utan också på grund av sitt innehåll som ofta överlappar det hos många hammeddelanden. I all skräppostfiltrering utgår man från att skräppost ser annorlunda ut än hammeddelande och man försöker, med olika metoder, att hitta de särdrag som utmärker skräppost och skilja mellan de två klasserna av meddelanden. Den taktik som neurala nätverk använder sig av är ett försök att efterlikna det sätt som människor visuellt skiljer mellan skräppost och ham. Även utan att ha sett alla skräppostmeddelanden som någonsin skapats så kan en människa snabbt lära sig att skilja mellan dessa klasser. Den underliggande principen med neurala nätverk är att varje meddelande kan kvantifieras med hjälp av ett mönster. För att identifiera dessa mönster måste nätverket först tränas. Denna träning involverar en datamaskinell analys av innehållet i meddelanden genom att använda stora, representativa delar av både skräppost och ham. På detta sätt kommer nätverket att 'lära sig' vad vi människor menar med skräppost respektive ham.

### 2.3.7 Manuellt

Många användare är också vana att manuellt rensa sin inkorg från skräppostmeddelanden eftersom de kan se direkt på subjektraden att det inte är fråga om ett riktigt e-postmeddelande. Något som fått spammarna

att hitta nya knep för att locka dessa användare att öppna meddelandena i stället för att ta bort dem direkt. Detta genom att tillverka subjektrader och innehåll som ser ut som ett riktigt hammeddelande, till exempel genom att lägga till ett 'Re' (för Reply) i början av subjektraden så att det ser ut som att meddelandet är ett svar på ett tidigare meddelande.

## 2.4 Lagar kring skräppost

### 2.4.1 Svensk lagstiftning

I Sverige trädde den första april 2004 en ny lagändring ikraft som förbjuder utskick av oönskad e-postreklam. Lagändringen sker i Marknadsföringslagen (1995:450) för att där i svensk rätt genomföra artikel 13 i kommunikationsdataskyddsdirektivet vilket jag kommer att beskriva närmare senare i detta kapitel. Lagändringen i §§ 13 b och c medför en så kallad opt in-lösning (se Appendix A). Detta gäller även vid marknadsföring till juridiska personer. Kravet på samtycke kan upphöra att gälla under vissa förutsättningar vid marknadsföring inom ett etablerat kundförhållande. Ett 'etablerat kundförhållande' mellan företag och konsument är till exempel om konsumenten köpt något från företaget, då får företaget skicka e-postreklam till konsumenten om följande förutsättningar är uppfyllda: Att du som kund inte motsatt dig att din e-postadress används i marknadsföringssyfte. Att marknadsföringen avser företagets egna, liknande produkter som den du köpt och att du (kunden) får möjlighet att kostnadsfritt och enkelt kunna hindra att e-postadressen används i marknadsföringssyfte, att företaget slutar skicka reklam till dig. Det ska vara samma företag som har fått uppgifterna om elektronisk adress som använder uppgifterna vid marknadsföring med hjälp av elektronisk post. Enligt lagenändringen ska det i e-postreklam också alltid finnas en giltig adress dit mottagaren kan skicka en begäran om att marknadsföringen ska upphöra. Om näringsidkaren inte iakttar sin skyldighet att ange giltig adress skall marknadsstörningsavgift kunna dömas ut.

Den definition av elektronisk post som används i 3 § i marknadsföringslagen är denna: "ett adresserat eller på annat sätt individualiserat elektroniskt meddelande i form av text, röst, ljud eller bild som sänds via ett allmänt kommunikationsnät och som kan lagras i nätet eller i mottagarens terminalutrustning tills mottagaren hämtar det". Begreppet omfattar enligt ingresspunkt 40 i direktivet uttryckligen SMS-meddelanden, det vill säga en mobil teletjänst som består i överföring av text. Även MMS-meddelanden, det vill säga avancerade former av SMS-meddelanden som tillåter överföring av färgbilder, animationer samt ljud- och videosegment som ett komplement till textmeddelandet, omfattas av begreppet. Definitionen omfattar även en del icke-adresserad kommunikation som sänds ut till allmänheten. Till exempel text-TV och radiotidningar för synskadade.

Marknadsföringslagens begrepp 'produkter' som enligt definitionen i lagens 3 § omfattar varor, tjänster, fast egendom, arbetstillfällen och andra nyttigheter.

Kommunikationsdataskyddsdirektivet som nämndes tidigare avser artikel 13 i Europaparlamentets och rådets direktiv 2002/58/EG av den 12 juli 2002 om behandling av personuppgifter och integritetsskydd inom sektorn för elektronisk kommunikation. Syftet med direktivet om integritet och elektronisk kommunikation är bland annat att säkerställa ett likvärdigt skydd av de grundläggande rättigheterna när det gäller behandling av personuppgifter inom sektorn för elektronisk kommunikation. Direktivets bestämmelser om icke begärd kommunikation i artikel 13 skall nu genomföras inom ramen för förehavande lagstiftningsärendet. Här finns också bestämmelser som avser icke begärd kommunikation vid direkt marknadsföring. Artikel 13 i direktivet om integritet och elektronisk kommunikation har rubriken "icke begärd kommunikation". Enligt definitionen i artikel 2 i direktivet avses med 'kommunikation' all information som utbyts eller överförs mellan ett begränsat antal parter genom en allmänt tillgänglig elektronisk kommunikationstjänst. Dock innefattas inte information som överförs som del av en sändningstjänst

för rundradio eller television till allmänheten via ett elektroniskt kommunikationsnät, utom i den mån informationen kan sättas i samband med den enskilde abonnenten eller användaren av informationen. Artikel 13 reglerar endast vad som benämns som 'direkt marknadsföring'. Vad som avses med direkt marknadsföring i direktivets mening klargörs inte närmare. Av ingresspunkt 40 i direktivet följer emellertid att regleringen i artikel 13 gäller vissa former av icke begärd kommersiell kommunikation. I artikel 13.1 föreskrivs att användningen av automatiska uppringningssystem utan mänsklig medverkan (automatisk uppringningsutrustning), telefaxapparater (fax) eller elektronisk post för direkt marknadsföring kan tillåtas bara i fråga om abonnenter som i förväg har gett sitt samtycke, en opt-in-lösning. Artikel 13.4 innehåller en bestämmelse som innebär ett förbud mot att skicka icke begärd marknadsföring med hjälp av elektronisk post om identiteten på meddelandets avsändare döljs eller hemlighålls eller giltig adress saknas till vilken mottagaren kan anmäla sin önskan om att marknadsföringen upphör. Bestämmelsen är även tillämplig på abonnenter som är juridiska personer. Av artikel 13.5 följer att medlemsstaterna skall, inom ramen för gemenskapslagstiftningen och tillämplig nationell lagstiftning, säkerställa att berättigade intressen för abonnenter som inte är fysiska personer är tillräckligt skyddade när det gäller icke begärd kommunikation. I artikel 13.4 ställer direktivet om integritet och elektronisk kommunikation upp en överordnad förbudsbestämmelse vad gäller användningen av elektronisk post för direkt marknadsföring. Av artikel 13.4 och 13.5 framgår att det under alla omständigheter skall vara förbjudet att skicka elektronisk post, till såväl fysiska som juridiska personer, om identiteten på den avsändare för vars räkning meddelandet skickas döljs eller hemlighålls eller om det inte finns en giltig adress till vilken mottagaren kan skicka en begäran om att sådana meddelanden ska upphöra<sup>3</sup>.

#### **2.4.2 Lagar inom EU**

Hittills har lagstiftningen kring e-postreklam sett olika ut i olika EU-länder. Nu finns det ett direktiv från EU som ska efterföljas av medlemsländerna. Den svenska lagändringen följer EU-direktivet om en gemensam europeisk lagstiftning. Hur det ser ut i de övriga EU-länderna är svårt att säga då det skulle kräva mycket tid och därtill språkresurser för att tyda de olika ländernas lagar, men enligt konsumentverkets hemsida så gäller liknande regler inom hela EU. Så mycket kan sägas att förbudet inte kommer att sätta stopp för all oönskad e-postreklam. Det är i praktiken mycket svårt att förhindra att så kallad skräppost skickas ut från oåtkomliga servrar runtom i världen (Riksdagen 2004).

#### **2.4.3 Lagar i USA**

Can-Spam Act of 2003, är en amerikansk lag om skräppost som trädde i kraft första januari 2004. Kontrollen av attackerna med icke efterfrågad pornografi och marknadsakter kräver att oönskade e-postreklammeddelanden märks upp (dock inte med en standard metod) och inkluderar en opt-out instruktion samt avsändarens fysiska adress. Den förbjuder användandet av vilseledande subjekt-rader och falska huvuden i sådana meddelanden. FTC:n är auktoriserad (men inte obligerad) att skapa ett 'mejla-inte'-register. Statliga lagar som kräver att man märker upp oönskad e-postreklam eller förbjuder sådana meddelanden är helt förebyggande, även om regler om bara att adressera falskt och vilseledande kommer att finnas kvar.

#### **2.4.4 Skillnader mellan amerikansk och svensk lagstiftning**

Det amerikanska lagsystemet är helt annorlunda uppbyggt, allt är väldigt detaljerat beskrivet för fall till fall. I USA har man ett system med statliga lagar som gäller alla delstater, men där varje delstat även kan skapa sina egna federala lagar. Först förklaras bakgrunden till beslutet, sen kommer alla definitioner och därefter kommer 'förbud mot aggressiva och grova kommersiella e-postmeddelanden', andra skydd för användare av kommersiell elektronisk post, bland annat förbud mot falsk eller missvisande

---

<sup>3</sup>Enligt (Regeringen 2003)

överföring av information, affärsverksamhet som medvetet marknadsfört genom e-post med falsk eller missvisande överföring av information, inverkan på andra lagar, 'mejla-inte'-register, studie av följder av e-postreklam, restriktioner på andra överföringar, applikationer till trådlös kommunikation med mera. I svensk lagstiftning är lagarna mer generellt utformade och det är upp till en domstol att tolka i varje enskilt fall om lagen är tillämpliga. I ett amerikanskt fall kan endast de fall som är väldigt klara dömas under lagen och det är eventuellt lättare att hitta kryphål som gör att ett enskilt fall inte kan dömas efter skräppostlagen".

# 3 Enkätundersökning

För att ta reda på hur företag och organisationer hanterar det växande problemet med skräppost gjorde jag en mindre enkätundersökning där jag förhoppningsvis kan urskilja någon tendens angående deras hantering av e-post. Jag ringde upp ett antal (större och mindre) företag och verksamheter och frågade dem tre frågor:

- Om de för tillfället använder något filter/program för att sortera bort oönskad e-post?
- Om de är nöjda med de resultat som det eventuella filtret ger?
- Och slutligen om de har använt något annat filter/program tidigare och varför de bytt?

Det var inte en helt enkel uppgift, dels är det svårt att få tag på rätt person som kan och därtill har tid att svara på ens frågor. Svaren är också väldigt varierande beroende på hur insatt den man frågar verkligen är, vissa ger verkligen ingående och utförliga svar (vilket kan tolkas som att de är väldigt insatta och jag verkligen fått tag på den person som är ansvarig för mejltrafiken och därmed också skräpposthanteringen) medan andra ger väldigt korta svar och ibland svar som är mycket vaga och svårtolkade, där kan jag ibland misstänka att jag inte riktigt fått tag i rätt person alternativt att företaget/organisationen inte har en särskild plan för hantering av oönskad e-post. Att göra en stor enkätundersökning med statistisk relevans skulle kräva ett examensarbete i sig, men det kan ändå vara av intresse att få en inblick i skräppostens inverkan på företag.

Svar	Antal
Eget/Något slags spärrsystem/På gång	3
Inbyggt	1
AntiGen	2
TrendMicros IMSS	1
Computer Associates	1
Virusfilter, men ej skräppostfilter	1
SpamSentiel	1
Eudora	1
Symantec Antivirus	1
Filtrerar ej	1

**Tabell 3.1** Resultat av första frågan i enkäten

Detta var de svar jag fick fram, genom de företag jag fick svar från. Det går inte att direkt utläsa något av dessa resultat annat än att det finns ett behov av filter, särskilt när man tittar på antalet olika filter som finns, där de flesta är kommersiella filter. Generellt var de flesta som svarade på mina frågor nöjda med sina filter, även om det kom fram åsikter som att de släpper igenom för mycket skräppost, märker upp för många ham som spam, eller har språkliga problem som ett företag vilka använde ett amerikanskt filter

som bland annat tittade på ord och inte släppte igenom svenska ord som 'slut' och 'fart'. Några ansåg även att filtret var för nyligt installerat för att kunna avgöra om det fungerar bra eller dåligt. På frågan om de haft något skräppostfilter tidigare som de bytt ifrån och varför svarade de flesta att de inte haft det, några hade använt andra filter innan, men ingen kunde direkt svara på vilka filter som använts. Någon funderade även på att byta sitt nuvarande.

## 4 Uppsala Spam Corpus

För att utvärdera filtret behövs data att testa på, alltså en skräppostkorpus och en hamkorpus. Skräppostkorpusen har samlats in och redigerats av Rikard Kjellberg och Oskar Lundén. De har, tillsammans med mig, samlat in inlägg från nyhetsgrupper för att skapa en hamkorpus (korpus med 'riktiga' e-postmeddelanden). Skräppostmeddelandena är insamlade dels från våra egna e-postkonton och sedan genom en förfrågan om att anställda vid institutionen för lingvistik och filologi skulle bidra med de skräppostmeddelanden de fått till sina inkorgar, och det är de övriga meddelandena. Dessa data används som testmaterial på de skräppostfilter jag valt att utvärdera.

Korpusen består av sammanlagt 5005 meddelanden, varav 3303 är skräppostmeddelanden och 1702 är hammeddelanden. Korpusen är insamlad och bearbetad under vårterminen 2004. Den bearbetning som skett bestod i att meddelandena rensades på reklam och bilder i ascii-format. Meddelandena fick rätt format, i detta fall mbox-format, med hjälp av formail. Formail är ett filter som användes för att tvinga mejlen till samma sorts mejlformat.

# 5 Utvärdering av skräppostfilter

I detta kapitel beskrivs de metoder som används för att utföra utvärderingen samt de filter som ingår i utvärderingen.

## 5.1 Utvärderingsmetod

Den metod jag valt för utvärderingen är kvalitativ och syftar till att ge en bild av hur några utvalda filter som använde de metoder för att bekämpa skräppost som beskrivits tidigare fungerar mot de data jag valt att testa dem emot, Uppsala Spam Corpus.

### 5.1.1 Precision, täckning och korrekthet

Vilka begrepp är viktiga när man bedömer ett skräppostfilters funktionalitet? Det är i regel viktigare att ett filter inte klassificerar fel och sorterar bort legitima meddelanden än att det slinker igenom ett eller annat skräppostmeddelande. För räkna på dessa uppgifter sorteras meddelandena i följande fyra kategorier. Sanna positiva, de meddelanden som klassats som skräppost, och då kan dessa vara sanna positiva (*true positives*,  $tp$ ), de som alltså är korrekt klassade, eller falska positiva (*false positives*,  $fp$ ), de som har blivit felaktigt klassade som skräppost. Deras motsvarighet är då negativa, de meddelanden som inte är skräppost utan alltså ham, och också dessa är indelade i sanna (*true negatives*,  $tn$ ) och falska (*false negatives*,  $fn$ ) beroende på om de är korrekt klassade som icke-spam eller inte.

När man ska räkna på hur bra ett filter fungerar finns det några olika mått för att beräkna detta (Manning and Schütze 2001). Det första jag ska beskriva är korrekthet, som räknar på hur stor del av all skräppost som fångas upp av den lösning man valt och hur stor del av e-posten som blir felaktigt klassad. Korrekthet beräknas med formeln:

$$\frac{tp + tn}{tp + fp + tn + fn}$$

Nästa mått som jag ska beskriva är precision, det är ett mått på hur stor del av skräpposten som (den utvalda delen man valt att räkna på) systemet har klassat korrekt. Precision beräknas med hjälp av följande formel:

$$\frac{tp}{tp + fp}$$

Det sista måttet jag ska använda i mina beräkningar är täckning. Täckning mäter hur stor del av relevanta data som klassificeraren täcker in, i detta fall andelen korrekt klassificerade skräppostmeddelanden i förhållande till all skräppost och beräknas på formeln:

$$\frac{tp}{tp + fn}$$

Det finns begränsningar hos alla dessa metoder. Vad gäller precision och täckning är det lätt att få bra resultat på det ena värdet på bekostnad av det andra. För att få ett hundraprocentigt värde på täckning väljs alla data i mängden, något som ger ett dåligt värde på precision. Korrekthet har också en svaghet, måttet blir väldigt missvisande om andelen sanna är väldigt skilt från andelen falska. Eftersom det är så är det bra att använda alla tre måtten som komplement till varandra.

Alla skräppostfilter måste utvärderas utifrån deras förmåga att klassificera inkommande skräppost korrekt. Det är viktigt att filtret får en hög procent infångade skräppostmeddelanden (ibland refererat till som upptäckningskvalitet) och en låg del falska positiva (ham-meddelanden som blockerades). Utvärdering av ett filter måste mäta både falska negativa (de skräppostmeddelanden som slank igenom filtret) och andelen falska positiva.

I min utvärdering kommer jag förutom dessa mått över filtrens klassificeringsförmåga också att titta på hur de fungerar. Jag ska då titta på vilka olika typer av särdrag som de respektive filtren använder sig av, och hur dessa fungerar. Jag ska främst försöka koncentrera mig på de särdrag som behandlar språkliga företeelser inom skräppostfiltrering, men också beskriva de andra delar som finns.

## 5.2 Urval av filter

Urvalet av filter baserar sig på att försöka testa de metoder för skräppostfiltrering som beskrivits i bakgrundskapitlet. Jag har också valt filter som är open source<sup>1</sup>, med undantag för den e-postklassificerare som utvecklats vid institutionen, eftersom de är lättillgängliga och deras dokumentation är enklare att få tag i på grund av deras fria distribution.

### 5.2.1 SpamAssassin

SpamAssassin<sup>2</sup> är ett open source-filter, det är ett projekt för Apache Software Foundation. SpamAssassin är ett skräppostfilter för Linux, SunOS/Solaris och FreeBSD. Det är ett kostnadsfritt filter som användaren kan ladda ner och installera själv. SpamAssassin är ett mejlfilter som försöker identifiera skräppost genom att använda ett stort antal varierande mekanismer. Det inkluderar bland annat textanalys och Bayesisk filtrering, DNS blocklistor och kollaborativa filtreringsdatabaser. SpamAssassins kombination av filtreringsmetoder verkar vara ett intressant sätt att angripa problemet med oönskad e-post som många användare är nöjda med.

SpamAssassin är alltså till största delen regelbaserat (rule based), och använder ett brett spektra av heuristiska tester på e-posthuvuden (mail headers) och brödtexten (body text) för att identifiera skräppost. Det använder sig av Bayesisk inlärning, svartlistning (DNS blacklists) och automatisk vitlistning (automatic whitelisting). SpamAssassin använder följande taktik för att identifiera skräppost:

Huvudanalys (header analysis): spammare (de som skickar skräppost) använder sig av ett antal trick för att dölja sina identiteter, lura dig att tro att de har skickat ett riktigt e-brev, eller lura dig att tro att måste ha tecknat dig för brevet på något sätt. SpamAssassin försöker hitta dessa med hjälp av olika regler, bland annat regler som tittar på:

Textanalys: skräppost har ofta en karakteristisk stil som kan hjälpa filtret att identifiera meddelandet. Att känna igen och hitta dessa kännetecken i texten är något som SpamAssassin gör mycket effektivt och något som jag kommer att återkomma till och beskriva närmare senare i kapitlet.

---

<sup>1</sup>Open source kan översättas med fri källa, men på grund av att den engelska termen är vanligare kommer jag att använda den

<sup>2</sup><http://spamassassin.apache.org/>

Svartlistor (blacklists): SpamAssassin stöder många användbara existerande svartlistor, som till exempel mail-abuse.org och ordb.org och andra. DNS blocklists, som också refereras till som DNS blacklists eller DNSBLs, är en vanlig form av nätverkstillgängliga databaser som används för att hitta skräppost. Razor: Vipul's Razor är en kollaborativ skräppostspårings databas, som fungerar genom att ta signaturen från skräppostmeddelanden. Eftersom skräppost typiskt fungerar så att det skickar ett identiskt meddelande till hundratals mottagare, så kortsluter Razor det genom att den första personen som mottar ett skräppostmeddelande adderar det till databasen, vilket gör att alla andra automatiskt kommer att blockera det meddelandet.

När ett meddelande har blivit identifierat kan det bli valfritt taggat som skräppost för senare filtrering genom att använda användarens egen e-postapplikation.

Vilka särdrag använder då SpamAssassin för att hjälpa sina användare att slippa problemet med skräppost?

Wide-spectrum: SpamAssassin använder en bred variation av lokala test samt nätverkstest för att identifiera skräppostsignaturer. Det här gör det svårare för spammare att identifiera en aspekt som de kan jobba runt och försöka lura när de ska skapa sina meddelanden.

Fri mjukvara: den är distribuerad under samma termer och villkor som Perl själv.

Det är lätt för användaren att utveckla och bygga på filtret själv. Regler, vikter och för användaren synlig text lagras i textkonfigurationsfilerna i så stor utsträckning som möjligt. Här kan användaren (eller systemadministratören) redigera för att modifiera redan existerande regler eller lägga till nya.

## **Bayesisk filtrering i SpamAssassin**

Bayes-klassificeraren i SpamAssassin försöker identifiera skräppost genom att titta på de tokens som finns, i detta fall ord eller korta sekvenser av tecken som ofta återfinns i skräppost och ham. Om man till exempel ger SpamAssassins inlärningsmetod 100 meddelanden som innehåller frasen 'penis enlargement' och förklarar att de alla är skräppost och när sedan det 101:a meddelandet som kommer in med orden 'penis' och 'enlargement', så kommer den Bayesiska klassificeraren att vara ganska säker på att det nya meddelandet är skräppost och kommer att öka poängen för att klassificera det meddelandet som skräppost.

## **SpamAssassin - textanalys**

SpamAssassin använder många regler som tittar på språkliga faktorer i de meddelanden det söker igenom, vilket jag kommer att återkomma till. Men filtret kan också programmeras att ta hänsyn till vilket eller vilka språk som det är okej för ett inkommande mejl att vara skrivet på. Detta bestäms med kommandot:

```
ok_languages xx [ yy zz ... ] (default all)
```

SpamAssassin försöker då identifiera det språk som en meddelandetext innehåller och sedan ge poäng efter detta. Skulle det däremot vara så att det uppstår en osäkerhet om vilket språk meddelandet är skrivet på så ges inga poäng, detta för att minska risken för felaktig klassificering. Den regel som baseras på språksökningen är UNWANTED\_LANGUAGE\_BODY. SpamAssassin stöder väldigt många språk, från finska till hebreiska och indonesiska, och för att tillåta ett språk anges dess två eller tre bokstäver långa kod i ok\_languages-raden. Så om man till exempel vill godkänna svenska och engelska som språk ser

raden ut så här:  
ok\_languages sv en

SpamAssassin använder också många andra regler, varav många fokuserar på att hitta språkliga indikationer på att ett meddelande tillhör en viss kategori av skräppost, till exempel att det är ett nigeriabrev eller reklam för viagra. Detta kan gå till på flera sätt, det enklaste är att leta efter det specifika ordet 'viagra', men det är ganska lätt för spammare att komma undan en sådan regel och gömma ordet för regeln genom att byta ut bokstaven i mot en 1:a, ordet 'v1agra' matchar inte längre med strängen 'viagra', men en person kan fortfarande läsa vad som egentligen står. För att ändå fånga meddelandet som skräppost finns en regel som letar efter 'gappy versions' av ord som 'viagra', det vill säga strängar som liknar ordet 'viagra', men där vissa bokstäver saknas eller är ersatta med något annat. För att hitta just mejl innehållande reklam för viagra använder SpamAssassin följande regler: SUBJECT\_DRUG\_GAP\_VIA, som kontrollerar om subjektraden innehåller en delad version av strängen 'viagra'. Vid genomsökning av texten i meddelandet används flera regler, bland annat VIA\_GAP\_GRA, som letar efter försök att dölja ordet 'viagra' och DRUGS\_SMEAR1, som också letar efter försök att dölja namn på läkemedel men denna gång genom att klistra ihop två namn till ett.

### 5.2.2 Bishop 0.3.0

Bishop 0.3.0<sup>3</sup> är ett enkelt program för klassificering, skrivet i programmeringsspråket Ruby av Matt Mower och publicerat under GNU Licence.

Bishop 0.3.0 är en klassificerare som använder sig av Robinsons och Robinsons-Fishers algoritmer för att klassificera data. För att kunna testa Bishop 0.3.0 på en korpus behövs följande uppdelning. Man behöver ett antal hammeddelanden respektive ett antal spammeddelanden som används för träning och därefter ett antal blandade ham- och spammeddelanden för själva testkörningen. Programmet använder sig av en tokeniserare för att skilja ut de ord ur meddelandena som Bishop 0.3.0 tränas på. Bishop 0.3.0 kontrollerar hur många gånger varje token har förekommit, och detta sparas i en lista där frekvensen för varje token i alla spam respektive ham finns.

#### Robinsons algoritm

Gary Robison är namnet bakom den tolkning av Bayes algoritm som blivit så populär bland skapare av statistiska filter och som också är den metod som används av programmet Bishop 0.3.0. För varje ord i en korpus beräknas  $p(w)$ , sannolikheten att ett meddelande, vilket som helst, som innehåller ordet  $w$  är ett skräppostmeddelande med hjälp av formeln:

$$p(w) = \frac{b(w)}{b(w) + g(w)}$$

där  $b(w)$  = antalet skräppostmeddelanden som innehåller  $w$  delat med det totala antalet skräppostmeddelanden och  $g(w)$  = antalet hammeddelanden som innehåller  $w$  delat med det totala antalet hammeddelanden. I ett filter beräknas  $p(w)$  för att användas som bas för vidare beräkningar. Men så finns problemet med sällan förekommande ord, till exempel om ett ord förekommer exakt en gång i ett meddelande, och detta är ett spam, blir  $p(w)$  1.0. Det innebär dock inte en garanti för att alla framtida meddelanden innehållande detta ord skulle vara spam. En människa har bakgrundsinformation och vet att så gott som varje ord kan finnas i antingen spam eller ham. Bayes låter kombinera vår generella bakgrundsinformation med de data vi samlat ihop för ett ord. På så sätt kan vi beräkna graden av tro till huruvida detta ord, när det dyker upp igen, är ett spam med följande formel:

---

<sup>3</sup><http://rubyforge.org/projects/bishop/>

$$f(w) = \frac{(s * x) + (n * p(w))}{s + n}$$

Där  $s$  är den styrka vi vill ge bakgrundsinformation,  $x$  är vår gissade sannolikhet, baserad på vår generella bakgrundsinformation, att ett ord vi inte sett förut först kommer att dyka upp i ett spam. Och  $n$  är det antal e-post som innehåller  $w$ . Denna beräkning ger ett enkelt sätt att få fram rimliga sannolikheter snarare än extremvärden, förutsatt att  $s$  och  $x$  ges bra startvärden, företrädesvis genom test för optimering.

### Robinsons-Fishers algoritm

En av de vanligaste metoderna för att kombinera sannolikheter kommer från statistikern R.A. Fisher och hans arbete inom den del av statistiken som kallas metaanalys. Den tidigare beskrivna algoritmen ger varje meddelande ett set av sannolikheter, medan Robinsons-Fishers algoritmen vill kombinera dessa individuella sannolikheter till en övergripande indikator för hur spamigt eller hamigt ett meddelande är. Detta görs genom att ta ett set sannolikheter  $p_1, p_2, \dots, p_n$  och med hjälp av dem beräkna en chi-squarefördelning på formeln:  $-2 \ln p_1 * p_2 * \dots * p_n$  och med det resultatet räkna ut en chi-squaretabell för att beräkna sannolikheten att få ett lika extremt, eller mer extremt, resultat än det beräknade. Denna 'kombinerade' sannolikhet summerar alla individuella sannolikheter och kan kallas  $H$  och beräknas på följande sätt:

$$H = C^{-1}(-2 \ln \prod f(w), 2n)$$

Där  $C^{-1}$  är den inverterade chi-squarefunktionen som används för att beräkna ett  $p$ -värde för en valfri variabel från chi-squarefördelningen. Denna uträkning utgår från en nollhypotes, att  $f(w)$  är korrekta, att ett meddelande är en slumpvis samling av ord oberoende av varandra så att de  $f(w)$  som finns inte är i en uniform fördelning. Men så är inte verkligheten. Till exempel är meddelanden som innehåller ord som 'sex' är troligare att innehålla ord som 'porn' än ett meddelande från en vän som skriver om programmering. Trots det fungerar nollhypotesen genom att den sätts upp för att sedan slå över antingen mot att det är ett spam eller ett ham. Denna beräkning är känslig för värden nära noll, alltså de meddelanden som kommer att klassas som ham, men de ord som indikerar spam har  $f(w)$  som ligger nära ett. För att räkna på detta får man vända på beräkningen och för varje ord beräkna  $1-f(w)$ . Eftersom att  $f(w)$  representerar sannolikheten att ett slumpvis valt meddelande från de meddelanden som innehåller  $w$  är ett spam, så representerar  $1-f(w)$  sannolikheten att det är ett ham. Så utför man samma Fisher-beräkning som innan, fast med  $1-f(w)$  i stället för  $f(w)$ . Detta resulterar i kombinerade sannolikheter nära noll, förkastande nollhypotesen, när många ord som indikerar spam finns med. Kalla denna kombinerade sannolikhet för  $S$ . Nu kan man beräkna  $I$  som är en indikator som hamnar nära ett när det talar för att meddelandet är ett spam och nära noll när det verkar vara ett ham med följande formel:

$$I = \frac{1 + H - S}{2}$$

### 5.2.3 Neural Networks For Spam Detection

Det är också relevant att se hur ett filter som bygger på principen om neurala nätverk fungerar. Det filter jag har valt heter Neural Networks For Spam Detection (NNSpam)<sup>4</sup> och är skapat av Christian Eichenberger och Nicola Fankhauser och publicerat under Gnu Public License och verkade vara ett relativt enkelt att sätta sig in i trots att det använder sig av neurala nätverk som ofta kan verka krångliga och svåra att sätta sig in i hur de fungerar.

För att skapa nätverket behövs först ett antal skräppostmeddelanden som används för att generera en frekvensordlista av de ord som ingår i dessa meddelanden. Av de  $n$  första orden skapas input-neuroner,

<sup>4</sup><http://variant.ch/phpwiki/>

sedan ett lager med osynliga neuroner, lika många som antalet input. Och så två output-neuroner, en för *spam* och en för *ham*. Alla neuroner i input ska vara kopplade till varje dold neuron och varje dold neuron ska vara kopplad till de två output-neuronerna. Som del två måste nätverket tränas, detta med hjälp av de skräppostmeddelanden som användes för att generera frekvensordlistan samt ett antal hammeddelanden. I varje meddelande räknas förekomsten av de ord som finns i ordlistan och neuronerna sätts till ett värde beroende av detta. Nätverket aktiveras och värdena på neuronerna justeras beroende på det önskade och det faktiska värdet på output-neuronerna. Så fortsätter man att träna nätverket på alla meddelanden. Därefter kan nätverket testas på en korpus bestående av både skräppost och ham.

### 5.2.4 Probabilistisk e-postklassificerare

Denna e-postklassificerare skrevs som ett projekt på kursen korpuslingvistik på institutionen för lingvistik och filologi vid Uppsala universitet, vårterminen 2004, av Filip Salomonsson, Fredrik Skogberg och Adam Svanberg.

Klassificeringsmodellen i filtret bygger till största delen på statistiska metoder, med undantag för tokeniseringen och jämförelsen av sannolikheter. Den statistiska modellen har sin grund i Bayes regel, och klassificeringen av meddelandena bygger på antagandet att alla ord i ett meddelande bara är beroende av klassen spam och ham och inte inbördes av varandra. Beräkningen grundar sig också på antagandet att den obetingade sannolikheten för ett mejl är konstant. Tillämpandet av dessa antaganden benämns *naive Bayes* och formel för dess beräkning ser ut som följer:

$$P(B|A) = \prod P(B_i|A)P(A)$$

Och den motsvaras då i detta filter av:

$$P(\text{mail}|\text{class}) = \prod P(\text{ngram}|\text{class})P(\text{class})$$

När sedan sannolikheten för varje ord givet en klass ska beräknas används en metod som kallas Maximum Likelihood Estimation (MLE), som enklast beräknas genom att dividera antalet förekomster av ordet i klassen med totala antalet ord i klassen:

$$PMLE(\text{ngram}|\text{class}) = \frac{C_{\text{ngramclass}}}{N_{\text{class}}}$$

### 5.2.5 DSPAM

DSPAM<sup>5</sup> står för De-Spam, skulle kunna översättas till svenska med att 'avspamma'. Det är ett open-sourcefilter som baserar sig på statistiska metoder. Det är ett projekt som fokuserar starkt på forskning, och flera algoritmer och sätt att bekämpa skräppost har kommit ur projektet. Några av DSPAMs sätt att bekämpa skräppost är dessa, länkade token (chained tokens), nätverk (neutral networking), message inoculation (en metod för att inkapsla ett e-postmeddelande eller en textdel för syftet att träna ett mejlfilter), avancerad teknik för de-obfuscation och en algoritm för störningsreduktion (noise reduction) som kallas Bayesian Noise Reduction.

DSPAM använder sig bland annat av följande särdrag. System-wide administrativt maintenance free filtering. Det vill säga att DSPAM maskerar sig som e-postserverns levererare av post (eller proxyagent om det krävs) vilket möjliggör filtrering på servernivå. Filtret har en enkel inlärningsmekanism, där det enda användaren behöver göra att skicka vidare sin skräppost en 'spam email address' för träning. DSPAM stödjer också flera olika algoritmer för filtrering. För närvarande stödjer DSPAM följande kombinationer av algoritmer, Graham-Bayesian, Burton-Bayesian, Robinson's Geometric Mean och Fisher-Robinson's

<sup>5</sup><http://www.nuclearelephant.com/projects/dspam/>

Chi-Square. Den som är administratör kan välja en eller flera av dessa algoritmer för att använda mot skräppost eller kombinera flera för att utöka filtrets räckvidd. DSPAM använder också två p-värdes algoritmer (p-value algorithms), Grahams och Robinsons, plus flera andra statistiska algoritmer.

### **Länkade token**

Länkade token är även kända som multi-word tokens och n-grams. Det är fråga om en enkel algoritm för att bearbeta data som används för att skapa mer specifika data för statistiska algoritmer att jobba med. Detta genom att länka ihop närliggande token för att generera ett nytt token. Dessa nya token är inte istället för individuella token utan ett komplement. Det finns också några regler för hur man länkar token. Man länkar inte mellan huvudet och kroppen i meddelandet och inte heller mellan individuella huvuden. Och ord kan kombineras med icke-ordtoken. Så här skulle det bli om man länkar de token som finns i meningen 'CALL NOW IT'S FREE!', en mening som består av fyra individuella token (CALL, NOW, IT'S och FREE!), men med hjälp av länkade token får man ytterligare tre token att analysera (CALL NOW, NOW IT'S, IT'S FREE!). Man beräknar sannolikheten för alla token, och på grund av de siffror man får fram kan man avgöra om det är fråga om ett skräppostmeddelande eller inte, och det är här det blir intressant med länkade token eftersom det finns speciella förekomster av ordkombinationer i skräppost.

Man kan identifiera mönster i meddelandets text och dessa mönster kan både fria eller fälla ett meddelande. Ett exempel där ett länkat token ger misstanke om ett skräppostmeddelande är när man har de två token 'color' och '#000000', var för sig ger de inte så mycket information. 'Color' är ett ord som kan förekomma i både skräppost och ham och '#000000' är koden för svart som mycket väl kan dyka upp i ett legitimt HTML-baserat mejl. Men när de står bredvid varandra och beräknas beroende av varandra får man en annan bild. Varför får man då det? Jo, de HTML-generatorer som används för skräppost skiljer sig lite från dem som används till ham. Till exempel så genererar Microsoft Outlook sin HTML kod annorlunda. Som exempel kan en legitim mejlclient använda formen COLOR\*#000000 eller COLOR\*WHITE. Man ska också komma ihåg att länkade token oftast kommer att dyka upp mest när färgen för en font bestäms och att svart är en väldigt populär färg för stor, fet text i skräppost.

### **Bayesian Noise Reduction (BNR)**

Bayesian noise reduction försöker lösa problemet med Bayesiska störningar (noise), vilket i sin enklaste definition rör sig om irrelevanta data som förekommer i meddelanden som ska klassas av ett filter. Störningar är data som antingen avsiktligt eller oavsiktligt hamnat utanför meddelandets kontext och därigenom försvårar en riktig klassificering. Det finns olika former av störningar, några exempel är vanliga störningar som förekommer i de flesta textexempel och som ska identifieras som normala, då det alltid finns små avvikelser från kontexten i mänsklig kommunikation. Ett annat exempel på störningar som däremot inte är normala är 'Arbitrary Word List Attacks', som kan översättas med attacker av långa serier av ord från ordböcker, efternamn eller andra samlingar av ord som infogas i meddelandet. Detta görs för att överösa de bayesiska matriser som beräknar sannolikheten för vilken sorts meddelande det är fråga om och skapa en illusion av oskyldighet.

BNR sker i tre steg. I första steget initieras maskingenererade kontexter som baseras på mönster av värden på token inom en fönsterram. Nästa steg är inlärning och identifiering av intressanta kontexter att jobba med. Och det tredje och sista steget är att sedan hitta statistiska abnormiteter inom den utvalda kontexten.

### 5.2.6 OSpam

Filtret OSpam<sup>6</sup> är gratis och används via nätet, det är kompatibelt för Hotmails såväl som Yahoos e-postprogram. Till skillnad från många filter behövs det inte laddas ner utan kopplas till ditt e-postkonto (inget software program), och man behöver inte ändra några inställningar på sitt e-postkonto.

Till skillnad från andra antiskräpposttjänster så antar OSpam att all post är skräppost tills motsatsen är bevisad genom en serie av användargenererad och automatisk vitlistning. Om ett meddelande passerar en vitlistningsregel tillåts det att stanna kvar i användarens inkorg. Vid ickeauktoriserade meddelanden krävs att avsändaren verifierar adressen en gång för att visa att det inte var fråga om skräppost.

En stor fördel med ett serverbaserat system är att inga program eller mjukvara behöver installeras eller laddas hem, OSpam är kompatibelt med alla operativsystem som stöder en webbläsare (web interface), vilken e-postklient som helst, och alla POP-baserade e-postkonton (POP betyder Post Office Protocol, den vanligaste metoden för att använda e-post). Användaren kan fortsätta att använda samma e-postadress som förut, och de kan fortsätta att kolla sina mejl på det sätt de är vana vid med hjälp av samma e-postklient som förut och de kan behålla de filter som deras e-postklient erbjuder om de vill.

Ett typiskt problem för verifierande e-postfilter är att de ofta blockerar nyhetsbrev, konfirmerande e-postmeddelanden och andra automatiserade mejl där ingen mänsklig avsändare finns som kan bekräfta att det inte är ett skräppostmeddelande. OSpams lösning på problemet är 'vitlistning med hjälp av nyckelord', ett särdrag som tillåter användaren att vitlista ord eller fraser i meddelanden, som exempelvis deras efternamn, adress eller namnet på den nyhetsgrupp de är med i, som inte återfinns i skräppost. Utifall att något av dessa riktiga mejl tas bort och markeras som skräppost kan användaren alltid logga in på sitt konto och läsa eller återta meddelandet.

Så fort ett e-postkonto har registrerats och konfigurerats med OSpam.com, hela systemet är automatiskt, förväntas inget med arbete från användarens sida. Ett hjälpavsnitt (tutorial) finns för att hjälpa användaren att komma igång.

### 5.2.7 MailScanner

MailScanner<sup>7</sup> är ett open-source-system för e-post. Det hanterar över 500 miljoner e-postmeddelanden varje dag, tar bort två miljoner virus och identifierar 75 miljoner skräppostmeddelanden. MailScanner används på 20 000 sajter runt om i världen. MailScanner letar igenom alla e-postmeddelanden efter virus, skräppost och attacker mot sårbarheter i säkerheten. Programmet använder sig av 14 olika virus-sökare. Det uppdaterar sina virussökarprogram varje timme. MailScanner fungerar på följande sätt. När ett meddelande kommer söks det igenom av RBL:s svartlistor först, sedan görs en sökning med SpamAssassin, därefter kontrolleras om meddelandet innehåller virus, och sen kontrolleras de bifogningar som meddelandet håller och sist utförs ännu en koll av innehållet.

MailScanner använder sig av något de kallar för Spam Scanning. Den största delen av sökningen efter skräppost med MailScanner görs med hjälp av SpamAssassin. Då används DSN svartlistor (Blacklists), över 850 heuristiska regler, bayesiska sannolikhetsystem (probability systems), utdelade (distribuerade) nätverksbaserade kontroller som Razor, DCC och Pyzor som kollar frekvensen av mejl (the frequency of messages) runt hela världen för att identifiera skräppost.

MailScanner hanterar den skräppost de finner på följande sätt. Subjektraden är taggad så att användaren lätt kan filtrera meddelandena. Meddelandena kan vara taggade, sända/skickade (delivered), raderade,

---

<sup>6</sup><http://www.ospam.com/index.shtml>

<sup>7</sup><http://www.mailscanner.info>

arkiverade, studsade (bounced), inkapslade (encapsulated), noterade (notified) och/eller avskalade till ren text. Att skala ner ett meddelande till ren text ger användare möjlighet att slippa se stötande bilder som kan förekomma i till exempel pornografiska meddelanden. MailScanner hanterar också problemet med virus.

# 6 Resultat

## 6.1 SpamAssassin

Vid min första testkörning av SpamAssassin fick jag följande resultat. Endast 365 av meddelandena i korpusen klassades korrekt som skräppost. De övriga klassades som ham. Detta ger dessa resultat när det gäller precision, täckning och korrekthet. Precision får 100 procent, täckning 11,0 procent och korrekthet 41,3 procent. Detta var ju inte särskilt positiva resultat och jag fick börja fundera över vad som kunde ha lett till dessa resultat. Vid närmare koll av vad som SpamAssassin hade baserat sina uträkningar på såg jag bland annat att det fanns ett särdrag som så gott som alla meddelanden, både skräppost och legitima meddelanden, hade fått samma poäng för. Detta test resulterade i att samtliga hade fått en negativ poäng, det vill säga en poängsumma som indikerar att det inte är fråga om ett skräppostmeddelande. Det var den regel som kontrollerade om meddelandet passerat genom några UNTRUSTED HOSTS på vägen till sitt slutmål. Jag kontrollerade detta mot korpusen och det visar sig att denna information saknas hos en stor del av meddelandena. På grund av detta valde jag att testköra en gång till men denna gång utan att poängsätta meddelandena med hjälp av denna regel. Jag valde också att sänka gränsen för vilken poäng ett meddelande behöver för att klassas som skräppost från 8.0 till 5.0. Denna gång valde jag att inte testa hela korpusen utan att istället testa halva, för att sedan kunna testa andra halvan och se om SpamAssassin tränar under testkörningen och alltså då skulle presterar ett bättre resultat på den andra halvan av korpusen. Denna gång blev resultaten bättre och visas i tabell 6.1, även om de fortfarande inte är tillräckligt för vad man kan önska av ett skräppostfilter. Precision blev även i detta fall 100 procent, täckning 40,9 procent och korrekthet 61,0 procent. Resultat av nästa del av korpusen blev ungefär detsamma så det verkar inte som att någon form av automatisk inlärning skett. Jag kontrollerade mina resultat och kom fram till att det i de flesta fallen stod 'autolearn=no' i spamAssassins utvärdering av meddelandena. Jag kontrollerade detta mot SpamAssassins hemsida där jag fann följande om autolearn. 'autolearn=no' betyder inte att autolearn inte fungerar, utan att just detta meddelande inte har blivit automatiskt inlärt. Om ett meddelande redan har blivit inlärt lärs det inte in en gång till, och de alternativ de kan läras in som är 'spam' och 'ham'. Och SpamAssassin kräver minst tre poäng från meddelandets huvud och tre poäng från kroppen för att använda den automatiska inläringen till spam eller ham. Den version av SpamAssassin som jag har använt i mitt test har sex olika fall som autolearn kan få, antingen 'spam', 'ham' eller 'no', vilka jag har förklarat ovan. Men den kan också få värdena 'disabled', vilket betyder att det är konfigurerat så att bayes\_auto\_learn 0 eller use\_bayes 0, och har alltså inte använt autolearn. Nästa möjlighet är 'failed', då SpamAssassin försökt använda autolearn men av någon anledning misslyckats. Och den sista möjligheten är 'unavailable' vilket fångar upp alla andra fall som inte fångas upp av de ovan.

Mått	Resultat
Korrekthet	61,0%.
Precision	100,0%.
Täckning	40,9%.

**Tabell 6.1** resultat från SpamAssassin

Vid en närmare koll av vilka regler som använts för att bedöma mejlen i korpusen, och då främst de som avser en språklig kontroll, kom jag fram till följande. En väldigt stor procent av de regler jag anser avse språk har faktiskt använts och genererat poäng. Alla reglerna som avser att subjektraden skulle innehålla ett försök till att dölja ett namn på ett läkemedel, som exempel viagra och valium, hade utnyttjats.

## 6.2 Bishop 0.3.0

Vid undersökningen av Bishop använde jag delar av Uppsala Spam Corpus. Jag gjorde detta då det är ett enkelt filter och jag därmed gjorde antagandet att hammeddelanden blandade på svenska och engelska som jämfördes med spammeddelanden enbart på engelska skulle ge resultat som blev missvisande. Jag gjorde också ett till antagande att jag skulle använda en korpus som i så stor utsträckning som möjligt efterliknade den jag annars använt och skapade den med ungefär 40 procent ham och 60 procent spam. Detta innebar att jag även tränade Bishop med en sådan mängd. Detta visade sig vara ett felaktigt antagande, då viktningen blev övervägande för spam och väldigt många hammeddelanden blev felaktigt klassade medan alla spam blev korrekt klassade. I detta test fick jag värdet 73,1 procent på korrekthet, 68,0 procent precision och 81,9 procent täckning. Jag ändrade då mitt antagande och tränade filtret på lika delar spam och ham, medan jag lät testdata fortsätta att ha den tidigare fördelningen på spam och ham, och fick resultat som stämde betydligt bättre med dem man kan förvänta sig av en klassificerare, korrekthet 97,2 procent, precision 98,2 procent och täckning 97,6 procent. Resultaten visas i tabell 6.2.

Mått	Resultat
Korrekthet	73,1%.
Precision	68,0%.
Täckning	81,9%.

**Tabell 6.2** Resultat från Bishop 0.3.0

## 6.3 Neural Networks For Spam Detection

Detta program såg till en början ut att ha instruktioner som gjorde det lätt att både installera och därefter utvärdera med hjälp av en korpus. Det visade sig dock under arbetets gång att så inte riktigt var fallet och på grund av att tiden inte räckte till så saknas fullständiga resultat från NNspam, men jag ska beskriva det arbete så långt som det utfördes.

På grund av att de shell script som används i NNspam inte är optimerade för större mängder data kunde bara en del av korpusen användas, uppdelade i tre delar en spamdel, en hamdel och en testdel med blandat ham och spam. Det var också så att programmet inte klarade av det format, mbox, som korpusen var i, så det första jag fick göra var att formatera om den del som skulle användas till maildir. Detta gjordes med hjälp av programmet mb2md-3.20.pl. Därefter skulle en frekvensordlista skapas av spammeddelandena. Detta med hjälp av skriptet generate\_combined, tyvärr saknades en lib-fil som behövdes för att använda skriptet, detta löstes genom att kopiera en annan lib-fil och döpa om den. När frekvensordlistan var skapad skulle den anpassas för att användas till att generera nätverket. Jag använde mig av de ord till och med frekvens fyra, och sorterade även bort domännamn och annan liknande information enligt programmets anvisningar. Därefter användes skriptet generate\_network för att skapa nätverket. För att därefter träna och testa det nätverk som skapats behövdes nätverkssimulatore JavaNNS, skapad vid Department of Computer Architecture, Tübingens universitet, av Igor Fisher, Fabian Hennecke, Christian Bannes och Andreas Zell. Det gick bra att installera men går inte så fort att använda på Linuxsystemet. Jag har lyckats träna nätverket mot två meddelanden i spamdelen, men att träna meddelandena ett och ett tar väldigt lång tid, det som skulle behövas för att träna alla meddelanden är att funktionen 'learn all'

skulle kunna användas. Och eftersom att nätverket inte har kunnat tränas finns inga testdata att utvärdera från detta program.

## 6.4 Probabilistisk e-postklassificerare

```
hannab@numerus$ cd ~/salo/shared/blanski
hannab@numerus ./eval2.py -H ham-big.db -S spam-big.db -h
~/hannab/exjobb/ham -s ~/hannab/exjobb/spam
Tokens Types
Ham 1242053 117870
Spam 957525 167167
5008 messages (2353 ham, 2655 spam, 0 unsure, 0 failed)
lambda=0.500000, mindist=0.000000
tp fp tn fn P R
2470 185 1520 833 93.03 74.78
hannab@numerus$
```

Så här ser den utskrift som genererades av den probabilistiska e-postklassificerare som skrivits på institutionen för lingvistisk och filologi ut. Den version jag har använt vid min testkörning var tränad med den träningskorpus som användes vid skapandet av klassificeraren och testades här med Uppsala spam corpus. Klassificeraren gavs testdata (korpuser) i mbox-format, vilket är det format som korpuserna har och ett av de format som klassificeraren stöder. Jag hade också delat upp korpuserna i två delar, en spam och en ham, eftersom klassificeraren kräver testdata i den formen. I utskriften får man lite information om data, hur många token (graford) som de respektive delarna av korpuserna innehåller samt hur många olika typer de innehåller. Här skrivs också ut hur många meddelanden som totalt analyserades och hur de klassades. Det skrivs ut hur många som var sanna och falska positiva samt sanna och falska negativa. Och utifrån dessa värden har precision och recall beräknats. Precision beräknades till 93,0 procent och täckning till 74,8 procent samt korrekthet till 80,0 procent.

Dessa resultat var inte särskilt bra, men inte så oväntade med tanke på att klassificeraren är statistisk och inte tränad på Uppsala Spam Corpus, så för att se om det gick att få bättre resultat tränades klassificeraren på ett antal spam- och hammeddelanden från korpuserna och testades därefter igen, fast denna gång på en mindre del av korpuserna. Det visade sig att resultaten, som väntat, blev bättre och betydligt bättre efter träning. Denna gång fick korrekthet 99,1 procent, precision 99,1 procent och täckning 99,4 procent.

Mått	Resultat
Korrekthet	99,1%.
Precision	99,1%.
Täckning	99,4%.

**Tabell 6.3** Resultat från Probabilistisk e-postklassificerare

# 7 Diskussion och slutsatser

I detta kapitel diskuteras och sammanfattas de resultat som jag fått i undersökningen.

## 7.1 Vad säger mina resultat från undersökningarna?

De resultat som fåtts i undersökningen visas i tabell 7.1 nedan.

Filter	Korrekthet	Precision	Täckning
SpamAssassin	61,0%	100,0%	40,9%
Prob. e-postkl.	99,1%	99,1%	99,4%
Bishop 0.3.0	97,2%	98,2%	97,6%
NNSpam	X	X	X

**Tabell 7.1** Resultat från utvärderingen

Siffrorna från resultaten är väldigt varierande. SpamAssassin hade visserligen inga hammeddelanden som felaktigt klassats som skräppost, men har å andra sidan väldigt dåliga resultat i övrigt. Den probabilistiska e-postklassificeraren som är gjord vid institutionen och Bishop 0.3.0 klarar sig i detta fall betydligt bättre, även om de också har sina begränsningar. Ett skräppostfilter ska i första hand klara av att klassificera e-posten på ett tillfredsställande sätt, men det finns andra aspekter att ta hänsyn till när man ska välja vilket filter man vill ha till sin mejlbox. Det ska passa det operativsystem man arbetar på, i SpamAssassins fall finns det versioner både för Unix/Linux och för Windows. Där finns kostnadsaspekten och kanske viktigast av allt att det ska vara lätt att använda. De övriga program, utöver SpamAssassin, som jag har utvärderat har fått betydligt bättre värden på accuracy, precision och recall. Men de är betydligt mer primitiva i sitt utformande och passar ännu bara som tester eftersom de inte är utformade för att användas mot ett mejlssystem.

När det gäller resultaten på korrekthet, precision och täckning så fick ju SpamAssassin det till synes sämsta resultatet, om man bortser från att det har 100 procent precision som är väldigt bra då det innebär att inga hammeddelanden klassats som spam. Men vad de dåliga resultaten beror på kan man fråga sig? En tänkbar möjlighet är att korpusen skulle ha bidragit till de dåliga resultaten, filtret verkar fungera i 'verkligheten' alltså på de mejl som kommer in till e-postbrevlådor. SpamAssassin var det första filtret jag testade och jag har senare hittat flera exempel i korpusen där information från insamlingen inte lyckats rensas bort och det således finns rester kvar, bland annat taggar från tidigare analyser utförda av SpamAssassin samt meddelanden som innehöll flera meddelanden i ett. Det känns ändå som att dessa bara delvis skulle förklara det sämre resultatet, och att statistik nog är ett bättre sätt att motarbeta skräppost än regler. Det var roligt att de bästa resultaten återfanns hos den klassificerare som skapats vid institutionen och intressant att en naiv metod kan fungera så mycket bättre än en mer komplicerad som de hos SpamAssassin. Det åskådliggör också på hur individuellt det är med skräppost och att en statistisk metod som tränas på en specifik användare, eller i detta fall en korpus, kan lära sig att klassa dennes e-post väldigt bra, medan den inte gör så bra ifrån sig otränad, vilket mina resultat som redovisades i kapitel 6.4

också visade.

Har lagarna kring skräppost på något sätt förändrat bilden av skräppost? Det vill säga, har andelen minskat, har någon dömts på grund av att ha sänt ut skräppost, har bilden av spammare och skräppost ändrats på något sätt nu när det har blivit olagligt? Sedan lagändringen angående marknadsföring via e-post kan man på konsumentverkets hemsida<sup>1</sup> anmäla skräppost som man fått. Man får då fylla i ett formulär som sedan i den mån det går följs upp av konsumentverket för att de åtgärder som är möjliga ska kunna utföras, allt för att lagen ska kunna tillämpas. Det är inget lätt arbete då stor del av all skräppost kommer från länder utanför EU som inte har samma regler. Konsumentverket är en statlig myndighet vilket innebär att det som skickas dit blir en offentlig handling som registreras och sparas.

I Sverige har ännu ingen dömts för att ha skickat skräppost, men i USA dömdes en man, Jeremy Jaynes, och hans syster, Jessica DeGroot, för detta brott. De dömdes i Virginia som har en mycket sträng tillämpning av den amerikanska lagen om skräppost. Jeremy Jaynes dömdes till nio års fängelse och hans syster till böter på 7500 dollar. Jaynes tjänade stora pengar på att sälja produkter och tjänster som han marknadsförde genom att skicka ut miljontals spam, Jaynes tjänade så mycket som upp till 750 000 dollar i månaden. Han är den förste i USA som dömts till fängelse för att ha skickat ut skräppost<sup>2</sup>.

## 7.2 Framtid?

Att skräppost som företeelse skulle upphöra verkar inte troligt, så länge spammare kan tjäna pengar på att skicka ut sin 'reklam' så kommer det att finnas de som är beredda att utföra det oavsett om de gillas av det stora flertalet eller inte, och det kommer inga lagar eller filter att kunna hindra. Men detta är absolut ingen anledning att ge upp kampen och erkänna sig besegrad. Jag tycker att det är bra att lagarna har skärpts för att skydda personer och företag mot skräpposten och all den kostnad och obehag som den medför, det är om inte lösning så ett steg i rätt riktning. När det gäller filter för att stoppa skräppost så finns det nog tyvärr ingen möjlighet att skapa det ultimata filtret som skulle vara lösningen på problemet för all framtid. Spammarna är trots allt människor och människan har en stor förmåga överkomma hinder. När det gäller de som motarbetar spam och deras metoder tror jag på den sortens statistiska filter jag har utvärderat i min uppsats. De fungerar bra och de låter oss skapa filter som fungerar utan att behöva studera skräpposten alltför ingående, en mycket tråkig syssla då stor del av skräpposten innehåller material man helst vill slippa läsa. Jag tror på fria filter snarare än kommersiella, då jag anser att kommersiella filter är något av att spela spammarna rakt i händerna då kommersiella filter inte kan existera utan skräpposten. Medan de som konstruerar filter som användare fritt kan utnyttja inte har det problemet, upphörde problemet med skräppost skulle de inte förlora på samma sätt. Det gäller också som motarbetare att inte sänka sig till spammarnas nivå, att skicka tillbaka meddelanden och liknande. Ett sådant system skulle bara skapa än större oreda, och hur skulle en person som av misstag får sådant skräp i sin inkorg reagera? Jag har också under mitt arbete börjat få en vidare syn på skräppost och vad som är skräppost. Det finns mycket som ligger på gränsen, till exempel företag som anordnar tävlingar på Internet där man ska skicka vidare en inbjudan till tävlingen till så många man känner. Verkar kanske oskyldigt, men när det har inkommit ett tjugotal inbjudningar till ens inkorg från vänner och bekanta så börjar det lätt kännas tjatigt. Eller alla dessa små hälsningar och relativt oskuldsfulla kedjebrev som cirkulerar, vad ska man göra åt dem? Det är saker för både filtertillverkare och e-postanvändare att fundera vidare på.

---

<sup>1</sup>(Konsumentverket 2005)

<sup>2</sup>Hämtat från (Fyrlund 2005) och (BBC 2005)

# A Ordförklaringar

spam – skräppost, oönskad e-post som skickas till stora mängder e-postadresser, ofta innehållande material för att försöka sälja en produkt eller tjänst, alternativt bedrägerier i någon form.

ham – ’riktig’ e-post, som skickats till en användare av till exempel en vän eller till en diskussionsgrupp som användaren prenumererar på. Hit räknas också e-postreklam från företag med ett etablerat kundförhållande till användaren (under förutsättning att användaren inte motsatt sig reklamen).

header (huvud) – den del av ett e-postmeddelande där information om avsändare, mottagare, ämne och metainformation om meddelandet lagras.

body (kropp) – e-postmeddelandets innehåll, som kan vara lagrat i olika format. Det vanligaste formatet har varit rent textformat, men det börjar bli vanligare med HTML-format (HyperText Markup Language), det format som är dominerande för till exempel hemsidor.

opt-in – en policy för att ge tillåtelse, under denna ger användaren sin explicita tillåtelse till att en webboperatör får antingen samla in informationen, använda den i ett specifikt ändamål och/eller dela den med andra när ett sådant användande eller yppande till en tredje part är orelaterat till anledningen att informationen samlades in.

opt-out – en policy under vilken användarens tillåtelse är underförstådd om inte användaren explicit har begärt att hans eller hennes information inte får samlas in, användas och/eller delas ut till en tredje part när ett sådant utdelande inte är relaterat till ändamålet för vilket informationen samlades in.

ISP (Internet Service Provider) – ett företag som säljer direktåtkomst till Internet, oftast genom att ringa upp ett lokalnummer.

Open relays – i öppna reläer låter en mejlserver en annan dator, vilken som helst, att sända e-post genom servern. Öppna reläer har använts av individer och företag för att skicka skräppost.

# B Exempel från Uppsala Spam Corpus

Exempel på skräppost—reklam som marknadsför examina:

Message-ID: <KIBEDHFVATDKAUGWZAJHGGQP@yahoo.com>  
From: "Coy Silver" <tgfbsitmdzw@yahoo.com>  
Reply-To: "Coy Silver" <tgfbsitmdzw@yahoo.com>  
To: mats.dahllof@ling.uu.se  
Subject: let us help you...D.I.P.L.O.M.A  
Date: Sun, 22 Feb 2004 12:31:23 +0200  
MIME-Version: 1.0  
Content-Type: multipart/alternative;  
boundary="--90238864550129458067"  
X-Webmail-Time: Sun, 22 Feb 2004 06:29:23 -0400

Content-Type: text/plain;  
Content-Transfer-Encoding: quoted-printable

GET YOUR UNIVERSITY DIPLOMA

There are no required tests, classes, books, or interviews!

Get a Bachelors, Masters, MBA, and Doctorate (PhD) diploma!

Receive the benefits and admiration that comes with a diploma!

No one is turned down!

Call Today 1-248-927-0446 (7 days a week)

Confidentiality assured!

Exempel på skräppost—Nigeriabrev

>From: eva.gustafsson@lingfil.uu.se Tue, 24 Feb 2004 15:02:50  
From: by way of Eva Gustafsson <eva.gustafsson@lingfil.uu.se> <eva.gustafsson@lingf  
Message-Id: <524.132002.>  
Date: Tue, 24 Feb 2004 15:02:50

Subject: FROM DR IBRAHIM

FROM: Dr Ibrahim Quattara.  
Chairman Contract Review Panel,  
Abidjan, Cote D'Ivoire.  
West Africa.  
Tel:008821633306713

DEAR FRIEND.

LETTER FOR URGENT ASSISTANCE ON FUND TRANSFER

First, I must solicit your strictest confidence in this transaction. This by virtue of its nature as being utterly confidential and TOP SECRET.

I got your contact in our search for a foreign partner who has the Ability and reliability to prosecute a transaction of great magnitude Involving a pending business transaction requiring maximum confidence. We are top officials of the Federal Government contract review panel who are interested in investment in your country with funds which are presently trapped here in Cote D Ivoire. In other to commence this Business we solicit your assistance to enable us transfer into your Account the said trapped funds.

The source of this fund is as follows: During the last regime here of General Robert Guei in Cote D Ivoire some government official's set up companies and awarded themselves contracts which were grossly over Invoiced in various ministries. The government set up a contract review panel and we have Identified a lot of inflated contracts funds which are presently Deposited in a BANK here in Abidjan, Cote D Ivoire .

However, by virtue of our position as civil servants and members of the panel ,we cannot acquire this money in our name. I have therefore ,been delegated as a matter of trust by my colleagues to look for An overseas partner into whose account we would transfer the total sum of USD\$25,500,000.00 [TWENTY FIVE MILLION, FIVE HUNDRED THOUSAND UNITED STATES DOLLARS].

Hence we are writing you this letter. We agreed to share the money thus:

- [1] 20% FOR THE ACCOUNT OWNER [YOU]
- [2] 80% FOR US [THE OFFICIALS]

It is from the 80% that we wish to commence investments in your country as you will also stand as our foreign agent over there. Please note that this transaction is 100% safe and we hope to commence the transaction

latest seven [7] days from the date of the receipt of the following

information bellow.

- [A] YOUR NAME AND BENEFICIARY OF ACCOUNT.
- [B] YOUR PERSONAL TELEPHONE NUMBER AND FAX NUMBERS.
- [C] BANK ACCOUNT/SORT/ABA/ROUTING NUMBERS  
WHICH THE FUND WILL BE TRANSFERED TO.
- [D] YOUR BANK ADDRESS, TELEPHONE NUMBERS/FAX NUMBERS.

The above information will enable us commence the transfer of this funds into your account in your country without delay  
We are looking forward to doing this business with you and solicit your confidentiality in this transaction.

Please acknowledge the receipt of this letter using the above email address; I will bring you into the complete picture of this pending project when I hear from you.

With Kind regards,  
Dr Ibrahim Quattara.

# C Länkar till hemsidor för skräppostfilter

Länkar till de filter som beskrivs i uppsatsen

<http://spamassassin.apache.org/> - SpamAssassin

<http://rubyforge.org/projects/bishop/> - Bishop 0.3.0

<http://variant.ch/phpwiki/NeuralNetworksForSpamDetection> - NNSpam

<http://www.nuclearelephant.com/projects/dspam/> - DSPAM

<http://www.0spam.com/index.shtml> - 0Spam

<http://www.mailscanner.info> - MailScanner

# D Nätverk i NNSpam

Här visas en del av nätverket i NNSpam samt hur de mönsterfiler som nätverket kan tränas och testas mot ser ut.

```
SNNS network definition file V1.4-3D
generated at Mon Apr 25 15:58:28 1994
network name : NNSpam
source files :
no. of units : 2
no. of connections : 0
no. of unit types : 0
no. of site types : 0
```

```
learning function : Std\_Backpropagation
update function   : Topological\_Order
```

unit default section :

act	bias	st	subnet	layer	act func	out func
0.00000	0.00000	h	0	1	Act\_Logistic	Out\_Identity

unit definition section :

no.	typeName	unitName	act	bias	st	position	act func	c
-----	----------	----------	-----	------	----	----------	----------	---

# definition of input units

1		the		0.00000		0.00000		i		1,1,0	
2		from		0.00000		0.00000		i		1,2,0	
3		http		0.00000		0.00000		i		1,3,0	
4		and		0.00000		0.00000		i		1,4,0	
5		received		0.00000		0.00000		i		1,5,0	
6		for		0.00000		0.00000		i		1,6,0	
7		with		0.00000		0.00000		i		1,7,0	
8		you		0.00000		0.00000		i		1,8,0	
9		nbsp		0.00000		0.00000		i		1,9,0	
10		div		0.00000		0.00000		i		1,10,0	
11		subject		0.00000		0.00000		i		1,11,0	

```
12 | | feb | 0.00000 | 0.00000 | i | 1,12,0 |||
13 | | your | 0.00000 | 0.00000 | i | 1,13,0 |||
14 | | this | 0.00000 | 0.00000 | i | 1,14,0 |||
```

Exempel på en mönsterfil för ett spam:

```
SNNS pattern definition file V3.2
generated at Mon Apr 25 15:58:23 1994
```

```
No. of patterns : 1
No. of input units : 653
No. of output units : 2
```

```
# Input pattern for /home/stp00/hannab/mKorpus/spam/spam2.mbox:2,
7 11 4 3 6 5 4 12 17 0 3 3 1 2 2 2 7 2 1 1 1 3 1 2 0 1 1 1 0 0 1 1 2 0 1 0 1
1 1 0 0 1 2 1 1 2 2 1 1 3 1 0 0 0 0 0 1 0 1 1 1 1 0 3 0 2 0 0 0 0 1 0 2 0 0
2 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 5 0 0 0 0 0 0 0 0 0 3 0 0 0 0
0 0 1 2 2 0 2 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 1 0 0 0 0 0 0 1 0 1 1 3 1 0 0
0 1 4 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0
3 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 1 0 3 0 0 0 0 0 1 0 0 0 0 1 0 0 2 0 3 0 1
1 0 0 0 2 0 0 0 0 0 1 2 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 2 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 3 0 0 1 0 0 2 0 1
0 0 0 0 0 0 0 0 1 0 0 1 1 3 1 0 0 0 0 0 0 1 0 2 0 0 2 0 2 1 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 2 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 2 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0
0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 0 0 0 1
```

```
# Output pattern for /home/stp00/hannab/mKorpus/spam/spam2.mbox:2,
1 0
```

# Litteraturförteckning

BBC (2005). Man gets nine years for spamming. Download available from <http://news.bbc.co.uk/1/hi/world/americas/4426949.stm>.

Federal Trade Commission (2003). False claims in spam: A report by the FTC's division of marketing practives. Download available from <http://www.ftc.gov/reports/spam/030429spamreport.pdf>.

Fyrlund, K. (2005). Skickade miljoner spam – fick nio år, Aftonbladet. Download available from <http://www.aftonbladet.se/vss/nyheter/story/0,2789,629950,00.html>.

Hird, S. (2002). Technical solutions for controlling spam, Distributed systems technology centre. Download available from [http://www.security.dstc.edu.au/papers/technical\\_spam.pdf](http://www.security.dstc.edu.au/papers/technical_spam.pdf).

Högskolan i Kalmar (2003). IT-sektionen-Handledningar. Download available from <http://www.it.hik.se/handledning/mail/spam2.php>.

Konsumentverket (2005). Download available from <http://www.konsumentverket.se>.

Lotsson, A. (2004). Därföt står det konstiga ord i spam, ComputerSweden.

Lunds universitet (2003). Datasäkerhet – Svindlare på Internet. Download available from <http://www ldc.lu.se/security/svindleri.shtml>.

Manning, C. D. and Schütze, H. (2001). *Foundations of Statistical Natural Language Processing*.

Miller, C. (2003). Neural Network–based Antispam Heuristics, Symantec. Download available from <http://mn-issa.org/whitepapers/Symantec/AntiSpamrs.pdf>.

Regeringen (2003). Regeringens proposition 2003/04:43. Obeställd e-postreklam.

Riksdagen (2004). Pressmeddelande från riksdagen 2004-02-17, Oönskad e-postreklam förbjuds.

Schwarz, A. and Garfinkel, S. (1998). *Stopping spam*.

Svenskt näringsliv (2004). Phishing och Nigeriabrev. Download available from [http://sn.svensktnaringsliv.se/sn/publi.nsf/Publikationerview/59164599C4AE34F1C1256F700055DB92/\\$File/PUB200011.pdf](http://sn.svensktnaringsliv.se/sn/publi.nsf/Publikationerview/59164599C4AE34F1C1256F700055DB92/$File/PUB200011.pdf).