Stat 110 Final Project Report

<u>Controlling Spam</u>

Lucy Liu

403-088-823

March 17, 2004

Introduction

Spam, the junk mail of the online world, is a problem that troubles virtually everyone who uses the internet. Since the cost of sending e-mail is so low, advertisers do not hesitate to send out spam messages en masse. As a result, anyone with an e-mail account will have found himself at one time or another buried under the amount of spam. Many methods have been developed to counter spam, with varying rates of success. A number of those methods are summarized in the table below.

Method	Advantages	Disadvantages
Complaining to spammers' ISPs	Goes to the root of the problem, making spam more costly	Time-consuming, requires some computer expertise
Mail server blacklists	Cuts off spam at the server level	Incomplete; not always accurate
Signature-based filtering	Has few false positives	Only catches spam from "big names"; can be bypassed with random characters
Bayesian (statistical) filtering	Accuracy of more than 99%, few false positives	Requires "learning" process
Rule-based (heuristic) filtering	Easy to set up; can be highly effective	Can be bypassed by adaptive spammers; significant rate of false positives
Challenge-response filtering	Blocks spam without fail (so far)	Delays or discourages genuine e-mail
Laws	Stops unregulated spam for good	Currently not well-enforced
FFBs	Increases cost to spammers by raising bandwidth usage	Blacklists need to be reliably maintained; morally ambiguous
Slow senders	Increases cost to spammers by slowing rate of mailings	Requires new code and protocols
Penny per mail	Makes spam less affordable to send	Requires many bureaucratic changes to implement
Secret address	Requires no computer tricks	Impractical; still subject to dictionary attacks
Junk address	Blocks off spam when it can be expected	Can't be used for all situations where spam may be generated
Network filtering	Seems to be completely accurate	Only works on 50% of e-mails

In this report, we will focus on a test of the Bayesian method, using a limited sample of e-mail messages. Bayesian filtering sorts out spam through analyzing word frequency statistics in e-mail. When an e-mail is received, its contents are scanned into words, or "tokens", and the top fifteen tokens are compared against known spam words. The combined probability of those fifteen tokens give an accurate estimate of the e-mail's chance of being spam. A good Bayesian filter takes into account message headers, is careful about the criteria for determining tokens, and is biased against preventing false positives.

Data

The data for the training word frequencies were taken from e-mails collected in several accounts over a course of twelve days, with a few additional non-spam e-mails from earlier in the year. The training e-mails, particularly the spam messages, were selected to be as recent as possible so as to best reflect the current trend of words used in spam, which may have been adapted to evade spam filters. Two sets of data were collected: a count of the frequencies of words that appeared in the spam messages, and words that appeared in the non-spam mails. The lists were further sorted into words that appeared only in spam, only in non-spam, and in both.

Notes:

- Data collected is not case-sensitive; i.e. all characters are read in lower case.

- Word-separator characters: space, '-', '_', '@', '. ', ', ', ' ,', '/', ':', '?', '=', '(', ')', '; ', '"', ' ' (character for a space in HTML code), '
' (line break in HTML code), '<p>' (paragraph break in HTML code)

- Characters not ignored in words: '&', '#', '$', '%', ''', '.' when not immediately before a space, ';' when not immediately before a space

- E-mails in HTML are examined by viewing their source code. All HTML code (text between < and >) that is not a separator is stripped, except for <href> and <img> tags, which include web addresses.

- The words counted are taken from the subject and message body of the e-mails. Words from both categories are given the same priority.

Top 15 most frequent words:

nonspam only	relative frequency	frequency	**spam only**	relative frequency	frequency
i	19/839	0.022646	http	66/1610	0.040994
me	8/839	0.009535	refillguide.net	20/1610	0.012422
feel	7/839	0.008343	21404	18/1610	0.011180
my	7/839	0.008343	acq	18/1610	0.011180
hi	6/839	0.007151	img	18/1610	0.011180
received	6/839	0.007151	www.smartbargains.com	18/1610	0.011180
communication	5/839	0.005959	our	17/1610	0.010559
i'll	5/839	0.005959	click	12/1610	0.007453
error	4/839	0.004768	txd	12/1610	0.007453
lab	4/839	0.004768	txh	12/1610	0.007453
letter	4/839	0.004768	here	11/1610	0.006832
message	4/839	0.004768	search	8/1610	0.004969
monday	4/839	0.004768	credit	7/1610	0.004348
physics	4/839	0.004768	s.gif	7/1610	0.004348

Frequency of words appearing in both

	Spam	Non-spam
1	.00062	.00119
2	.00062	.00238
2004	.00062	.00119
4	.00062	.00119
a	.00497	.01549
able	.00062	.00119
about	.00062	.00119
all	.00373	.00238
an	.00311	.00596
and	.01367	.00956
any	.00248	.00358
anyway	.00062	.00119
are	.00373	.00358
around	.00062	.00119
as	.00373	.00238
at	.00373	.01073
available	.00124	.00119
be	.00311	.00954
because	.00062	.00477
been	.00248	.00119
by	.00497	.00119
can	.00186	.00715
directly	.00062	.00119
do	.00186	.00119
don't	.00062	.00238
due	.00186	.00119
e	.00248	.00358
email	.00186	.00477
enjoy	.00062	.00238
for	.00683	.01192
forward	.00062	.00477
free	.00373	.00119
from	.00497	.00358
get	.00248	.00238
go	.00124	.00119
good	.00062	.00119
has	.00186	.00238
have	.00867	.00954
hours	.00124	.00238
however	.00062	.00119
if	.00311	.01788
in	.01118	.01669
is	.01118	.00954
it	.00311	.01669
it's	.00062	.00119
just	.00062	.00238
keep	.00062	.00119
like	.00062	.00119
look	.00062	.00119
lucy	.00062	.00477
mail	.00124	.00358
make	.00124	.00358
message	.00248	.00477
more	.00373	.00119
most	.00124	.00119
much	.00124	.00238
need	.00062	.00358
nendil	.00373	.00119
no	.00124	.00119
not	.00248	.00119
now	.00124	.00119
of	.01429	.00954
off	.00186	.00119
office	.00124	.00358
on	.00373	.01073
one	.00124	.00119
only	.00124	.00119
or	.00683	.00834
other	.00062	.00477
out	.00186	.00119
please	.00186	.00596
questions	.00124	.00238
request	.00062	.00119
right	.00124	.00119
save	.00062	.00119
see	.00186	.00238
send	.00186	.00477
set	.00062	.00358
sincerely	.00062	.00119
so	.00124	.00477
something	.00062	.00358
stop	.00062	.00119
take	.00062	.00119
term	.00124	.00358
than	.00186	.00238
that	.00373	.01311
the	.02547	.03099
their	.00062	.00119
then	.00062	.00119
this	.00435	.02026
those	.00062	.00358
time	.00124	.00477
to	.03168	.03218
today	.00124	.00238
tomorrow	.00124	.00119
ucla	.00062	.00119
until	.00124	.00238
up	.00062	.00358
want	.00062	.00119
was	.00124	.00358
we	.00932	.00238
website	.00186	.00238
were	.00062	.00119
what	.00062	.00119
when	.00124	.00119
who	.00124	.00238
will	.00311	.00477
with	.00373	.00596
you	.01491	.05125
your	.01553	.00834

Non-spam e-mail headers:

Date: Sun, 23 Nov 2003 15:15:41 -0800
From: "MANCIA,DIANA YVETTE" <dmancia@ucla.edu>
To: Nendil <nendil@ucla.edu>
Subject: Re: Hi!

Date: Sun, 30 Nov 2003 15:02:39 -0800
From: "MANCIA,DIANA YVETTE" <dmancia@ucla.edu>
To: nendil@ucla.edu
Subject: Hi!

Date: Tue, 20 Jan 2004 00:14:22 -0800
From: "KURNADI, PRISCILLA PRISKA" <pkurnadi@ucla.edu>
Subject: Physics 4BL: OH, website update

Date: Fri, 13 Feb 2004 09:45:56 -0800
From: "KURNADI, PRISCILLA PRISKA" <kurnadi@physics.ucla.edu>
Subject:

Date: Mon, 16 Feb 2004 00:10:48 -0800
From: "FARZINNIA, NEDA" <neda@stat.ucla.edu>
Subject: Appointments

Date: Mon, 16 Feb 2004 07:38:24 -0800
From: Troy Carter <tcarter@physics.ucla.edu>
To: "Nendil" <nendil@ucla.edu>
Subject: Re: Rec. letter, again

Spam e-mail headers:

Date: Wed, 11 Feb 2004 02:03:18 -0800
From: PayDayRightAway <7361zd@deeuseless.com>
To: nendil@ucla.edu
Subject: Don't wait until next payday

Date: Thu, 12 Feb 2004 13:00:06 -0500
From: "Federico Hodge" <federicohodge@physicianrefill.net>
To: nendil@twin-elements.com
Subject: Ph@rmacy Medicati0n Sa1e

Date: Tue, 17 Feb 2004 13:17:05 GMT
From: "Bedding Discounts" <pudsdapikudefjewfwfwd@citymailserver.com>
To: NENDIL@TWIN-ELEMENTS.COM
Subject: LUCY - Bed & Bath Liquidation: Up to 75% Off!

Date: Thu, 19 Feb 2004 12:09:27 +0000 (GMT)
From: "Gus Cain" <kgtlshvu@sesmail.com>
To: sales@liuart.com
Subject: Drive Thousands of Shoppers to Your Web Office

Date: Thu, 19 Feb 2004 22:05:41 UT
From: "DTS Group"   <reply@dare-to-win5.com>
To: nendil@liuart.com
Subject: We have recently reviewed your resume

Date: Fri, 20 Feb 2004 09:29:44 +0200
From: "Dino Teague" <dinoteague@rxrecommend.net>
To: nendil@twin-elements.com
Subject: L0se 15 P0unds

Date: Sat, 21 Feb 2004 19:55:59 -0600 (CST)
From: Lee <lee@more-personal-ads.com>
To: nendil@ucla.edu
Subject: ADVERT: BruinSingles.com

Date: Sun, 22 Feb 2004 17:05:04 -0500
From: "Larry Thacker" <kdset114@mail.1starnet.com>
To: nendil@twin-elements.com
Subject: jbr Card hdnjg Declined, app

Data Analysis

First, the list of words that appear in both spam and non-spam is used to perform a Chi-square test in order to determine whether the frequency of the common words are statistically the same in spam and non-spam. A Chi-square test is performed by the formula of _(observed freq. – expected freq.)^2/(expected freq.) In this case, the expected frequency is the frequency of the words in non-spam, and the spam frequencies are the observed values being tested against them.

The data and calculations for the Chi-square tests are attached on the next page. The Chi-square value at approximately n = 100 was found to be 25.3, which is far lower than the value of 118.5 for the 0.1% certainty range. Therefore, we cannot reject the null hypothesis, which is that the two groups of words do not differ significantly in frequency. (Practically speaking, however, it would be better to judge the words individually, as words like "on" or "the" are of course equally likely to show up in both kinds of mail, but words such as "free" or "unsubscribe" appear much more often in spam mail.)

Next, four pieces of e-mail supplied by the professor are compared against the existent training lists of words. The status of them being spam or not is determined by comparing the number of words in each message that show up in the spam-only list, and likewise for the non-spam-only list. If one is higher than the other, the message is sorted into that category.

e-mail 1: **IMPORTANT MESSAGE FROM NSF ITR PROGRAM** (spam)
e-mail 2: **Re: [Fwd: Runner Messaging Alert Summary]** (not spam)
e-mail 3: **Italian-crafted Rolex ¬C only $65 - $140!! Free SHIPPING!!2**! (spam)
e-mail 4: **Poultry data** (not spam)

	Frequency of non spam-only words	relative frequency	Frequency of spam-only words	relative frequency	Judgement
e-mail 1	0.12143	17/140	0.05000	7/140	Not spam
e-mail 2	0.19101	17/89	0.07865	7/89	Not spam
e-mail 3	0.09000	18/200	0.08500	17/200	Not spam
e-mail 4	0.17143	6/35	0.11429	4/35	Not spam

	Spam	Non-spam
Spam	0	**2**
Non-spam	**0**	**2**

	Spam actual probability (/1610)	actual frequency	Non-spam theoretical probability (/839)	theoretical frequency	(actual – theoretical)^2	(actual – theoretical)^2/ (theoretical)
1	1	0.00062	1	0.00119	3.249E-07	0.000273
2	1	0.00062	2	0.00238	3.098E-06	0.001302
2004	1	0.00062	1	0.00119	3.249E-07	0.000273
4	1	0.00062	1	0.00119	3.249E-07	0.000273
a	8	0.00497	13	0.01549	0.0001107	0.007145
able	1	0.00062	1	0.00119	3.249E-07	0.000273
about	1	0.00062	1	0.00119	3.249E-07	0.000273
all	6	0.00373	2	0.00238	1.823E-06	0.000766
an	5	0.00311	5	0.00596	8.123E-06	0.001363
and	22	0.01367	8	0.00956	1.689E-05	0.001767
any	4	0.00248	3	0.00358	1.21E-06	0.000338
anyway	1	0.00062	1	0.00119	3.249E-07	0.000273
are	6	0.00373	3	0.00358	2.25E-08	0.000006
around	1	0.00062	1	0.00119	3.249E-07	0.000273
as	6	0.00373	2	0.00238	1.823E-06	0.000766
at	6	0.00373	9	0.01073	0.000049	0.004567
available	2	0.00124	1	0.00119	2.5E-09	0.000002
be	5	0.00311	8	0.00954	4.134E-05	0.004334
because	1	0.00062	4	0.00477	1.722E-05	0.003611
been	4	0.00248	1	0.00119	1.664E-06	0.001398
by	8	0.00497	1	0.00119	1.429E-05	0.012007
can	3	0.00186	6	0.00715	2.798E-05	0.003914
directly	1	0.00062	1	0.00119	3.249E-07	0.000273
do	3	0.00186	1	0.00119	4.489E-07	0.000377
don't	1	0.00062	2	0.00238	3.098E-06	0.001302
due	3	0.00186	1	0.00119	4.489E-07	0.000377
e	4	0.00248	3	0.00358	1.21E-06	0.000338
email	3	0.00186	4	0.00477	8.468E-06	0.001775
enjoy	1	0.00062	2	0.00238	3.098E-06	0.001302
for	11	0.00683	10	0.01192	2.591E-05	0.002173
forward	1	0.00062	4	0.00477	1.722E-05	0.003611
free	6	0.00373	1	0.00119	6.452E-06	0.005422
from	8	0.00497	3	0.00358	1.932E-06	0.000540
get	4	0.00248	2	0.00238	1E-08	0.000004
go	2	0.00124	1	0.00119	2.5E-09	0.000002
good	1	0.00062	1	0.00119	3.249E-07	0.000273
has	3	0.00186	2	0.00238	2.704E-07	0.000114
have	14	0.00867	8	0.00954	7.569E-07	0.000079
hours	2	0.00124	2	0.00238	1.3E-06	0.000546
however	1	0.00062	1	0.00119	3.249E-07	0.000273
if	5	0.00311	15	0.01788	0.0002182	0.012201
			……			
with	6	0.00373	5	0.00596	4.973E-06	0.000834
you	24	0.01491	43	0.05125	0.0013206	0.025768
your	25	0.01553	7	0.00834	5.17E-05	0.006199
					Sum	0.229850

Sum _ n 25.28348

Conclusion

It seems that our filtering method still needs some improvement. The error in accuracy comes from several factors. In the matter of data collecting, the number of e-mails used to build the training lists is quite low – only 6 non-spam and 8 spam messages. Ideally, hundreds of messages would be processed for the database. Also, the e-mails I receive already pass through some number of spam filters, both through software and mail servers, and so it would not be a good representation compared to someone whose e-mail does not pass through the same filters. Finally, the Bayesian filters for everyone would be personalized, depending on the kind of e-mails they are likely to receive, and so the system set up for one person's e-mail may not be as effective for another person, initially.

There is also the matter of what to do with the data once they are gathered. In the studies of Paul Graham, whose articles on which this project is based, the top 15 words of interest in the e-mail are used to calculate the probability of it being spam, whereas I examined all of the words equally. He also involves procedures such as looking at all of the headers of an e-mail, and weighing words in the message body and subject differently, etc. Due to the fact that the data in this study was largely organized by hand instead of computer, I had to simplify some of the process.

It is notable that, though the results for this imitation filter provides false positives (misidentifying spam as non-spam), it does not report false negatives (sorting non-spam as spam) which is "punished" more harshly in Graham's method because filtering out non-spam mail can have more significant consequences. As I have seen from the small mistakes initially made by the filtering program I installed, every Bayesian filter involves a learning process, one which this "filter" is still going through.