Pseudowords

John^H^H^H^H Dmitri Graham-Cumming
The POPFile Project

Virus Bulletin 2005

Agenda

- About Me
- About POPFile
- About Pseudowords
- Example Message
- All POPFile's Pseudowords
- Which pseudowords are worth the effort?

Me

- Original creator of POPFile
  http://getpopfile.org/
- Independent consultant and writer
  http://www.jgc.org/

- License polymail/LIBSD anti-spam library
  http://www.extravalent.com/

Me

- The Spammers' Compendium
  http://www.jgc.org/tsc/
- Anti-spam Tool League Table
  http://www.jgc.org/astlt/
- jgc's spam and anti-spam newsletter
  - Every two weeks
  - I promise not to spam you!

POPFile

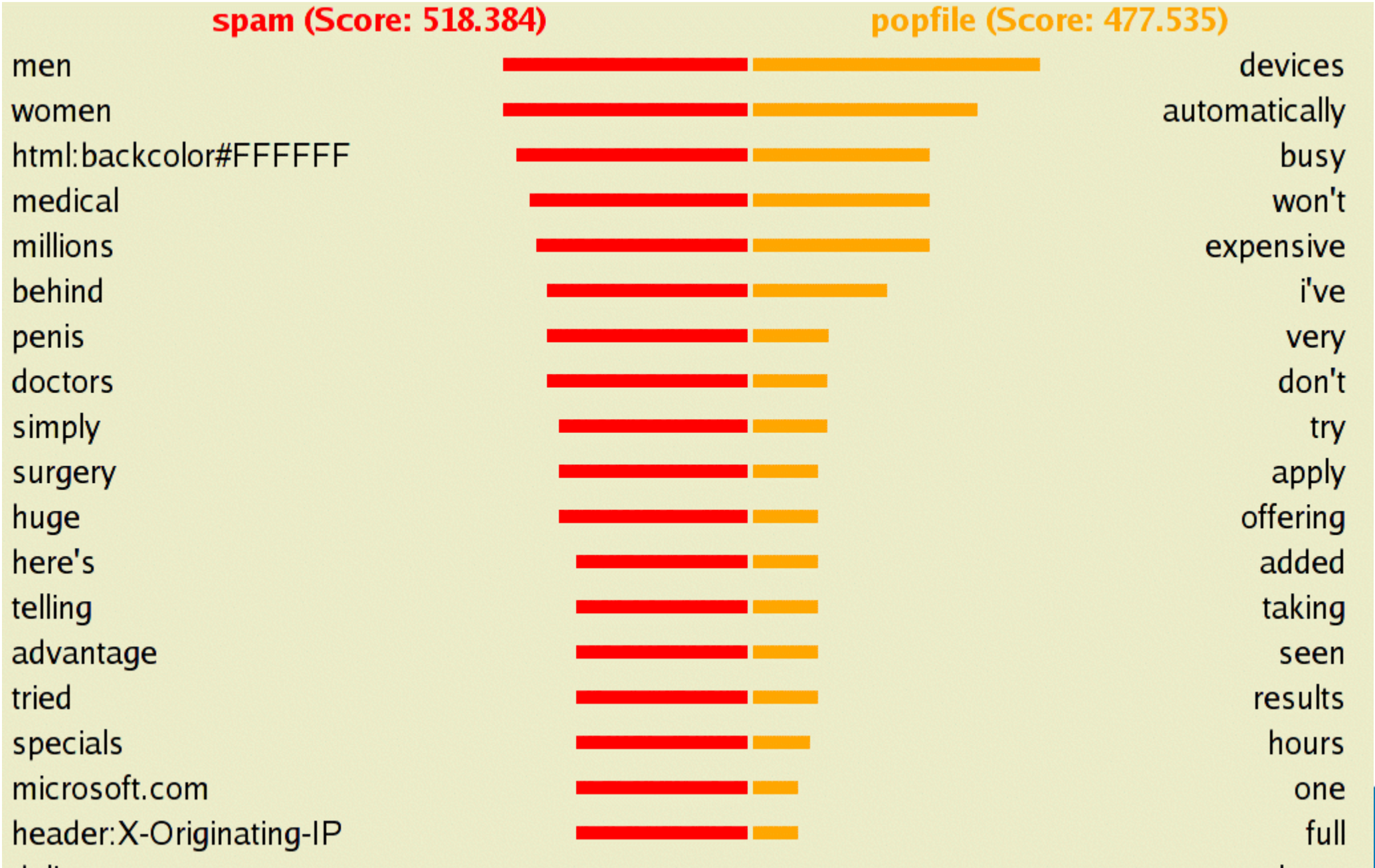
- Automatic email classification proxy
- Open Source and multi-platform
- Classifies email using a Naive Bayesian engine into user-defined categories
- Everyone has a 'spam' category
- POPFile does special analysis to

Example Message

```
Content-Type: text/html; charset="us-ascii"
Content-Transfer-Encoding: 8bit

<html>
<head>
<title>black leyden tradesmenallyn</title>
</head>
<body bgcolor="#FFFFFF">
<P><STRONG>Good morning,</STRONG> </P>
here's that site I was telling you about. They are offering huge disscounts now on Peniss
Enhancmeent
Patches<br><br>
A top team of British scientists and medical doctors have worked to develop the state-of-the-art
<b>Peniss Enhancmeent
Patch</b>
delivery system which automatically increaeses penis sizze up to 3-4 full inchees. The patches ar
the easiest and most
effective way to
increase your peniss size. You won't have to take pills, get under the knife to perform expensiv
and very painful
surgery, use any
pumps or other devices. <b>No one will ever find out that you are using our product.</b> Just
apply one patch on your
body and wear
it for 3 days and you will start noticing dramatic results.<br><br>
Millions of men are taking advantage of this revolutionary new product - <b>Don't be left
```

How POPFile decided



Pseudoword

- A synthetic word that cannot appear in a message
- Constructed from some metadata in the message
- Used to highlight specific features of the message

header:XXX

- Added for each header present in a message
- For example,
  - header:From
  - header:MIME-Version
  - header:X-Originating-IP
- Case preserved

from:XXX, to:XXX, cc:XXX

- Added with word taken from the From, To and Cc headers
- e.g. From: John Graham-Cumming <jgc@jgc.org> creates:
  - from:john, from:graham, from:cumming
  - from:jgc, from:jgc.org

charset:XXX, encoding:XXX

- Created from the character set in the message and the message encoding
- For example, a Japanese message might have:
  - charset:euc-jp
  - encoding:8bit

html:authorization

- Added if the message contains a URL that has authorization
- For example,
  - http://microsoft.com@spammer.biz/
- Spammers use authorization to disguise URLs

html:fontcolorXXX
html:backcolorXXX

- Used to track use of foreground and background colors
- For example,
  - html:fontcolorRed
  - html:backcolor#FFFFFF
- Also track colors used in CSS separately with html:cssfontcolor and html:cssbackcolor

html:colordistanceXXX

- Used to spot the *Camouflage* trick
- Spammer writes text on a background of a very similar color
  - e.g. white text on an off-white background
- Calculate the Euclidean distance between the two (R, G, B) colors

html:fontsizeXXX
html:cssfontsizeXXX

- Used to track font sizes in regular HTML and CSS
- For example,
  - html:fontsize10pt
  - html:cssfontsize+2

html:comment

- Used to track the presence of HTML comments
- Added for each comment:
  - Levi<!-- some random words -->tra

html:cidsrc

- Used to track images embedded in a message
- For example,
  - <img src=cidsrc:65a3f658a4d3e>

html:cssdisplay
html:cssvisibility

- Used to track the CSS DISPLAY and VISIBILITY options
- These are commonly used by spammers to hide random text
- For example,
- <DIV STYLE=DISPLAY: NONE;>banana elephant potato</DIV>

html:emptypair
html:invalidtag

- Used to track two HTML tricks used by spammers
- Empty useless pairs of HTML tags,
  - Viagra
- Invalid random HTML tags
  - Cia<CONCIERGE>lis

html:encodedurl
html:numericentity

- Used to track encoded URLs
- Spammers disguise URLs with encoding
  - http://%21%45%46%76%32.com/
- Track use of HTML numeric entities to disguise words
  - VIAGRA

html:imgremotesrc
html:iframeremotesrc

- Created when a message contains an image loaded from a remote web site
- Or when a <IFRAME> is loaded from a web site
- Intended to track all images including web bugs

html:imgheightXXX
html:imgwidthXXX

- Tracks the height and width of embedded images
- For example,
  - <img height="492" width="193">
  - html:imgheight492
  - html:imgwidth193

html:td

- Simply tracks the presence of the HTML <TD> tag
- Used to spot heavy use of tables in a spam
- Large tables are characteristic of complex spammer trickery

mimeextension:XXX
mimename:XXX

- Used to track MIME types
- Specifically track attachment
- For example,
  - mimename:vb2005.pdf
  - mimeextension:pdf

spamassassin:XXX spamassassinlevel:XXX

- If POPFile is chained with SpamAssassin then POPFile read the SpamAssassin headers

subject:XXX

- Tracks words in the Subject line
- For example,
  - subject:free
  - subject:john
- Case is not preserved

trick:spacedout
trick:dottedwords

- Tracks two specific spammer tricks
- Spacing out words
  - V I A G R A
  - trick:spacedout
- Adding random dots
  - cial.is
  - trick:dottedwords

trick:invisibleink

- Track the classic spammer trick of white text on white
- For example,
  - united robots jury
  - trick:invisibleink

Are they effective?

- Ran 95,405 spams through a fully trained POPFile instance
- Record the weights for each pseudoword
  - html:imgwidth159 1.29
    subject:free 2.49
    from:burton 0
    html:imgheight24 -0.10

Header vs Body

- Header pseudowords more important than the body
  - header 656.55
    body 352.25

Best Header Pseudowords

- The Subject is the most important
  - subject 310.9
    from 256.67
    to 33.09
    cc 16.21
    header 8.99
    mimeextension 4.95
    encoding 4.94
    charset 1.08
    spamassassinlevel -0.27

Digging into the charset

- The charset matters
  - windows-1251 2.19
    windows-1252 1.99
    iso-8859-1 1.83
    gb2312 -0.1
    utf-8 -0.7
    big5 -1.99
    us-ascii -2.14

Most spammy Subject words

- free 2.49
  what 2.19
  only 1.89
  huge 1.89
  judge 1.89
  visa 1.89
  soft 1.89

- *'What are you waiting for?'*

The most useful HTML pseudowords

imgheight	139.49
imgwidth	124.08
fontcolor	38.32
fontsize	16.49
colordistance	15.86
backcolor	13.05

- **The big surprise is that tracking specific image sizes is very helpful**

Tricks

- Only *Invisible Ink* worth the effort
- invisibleink 2.9
  dottedwords -0.6
  spacedout -0.73

Neutral HTML Pseudowords

- Not worth calculating:
  - html:td
    html:encodedurl

Conclusion

- Most important data is in the From, To, Cc and Subject headers
- Tracking HTML image sizes is surprisingly important
- HTML font sizes and colors matter
- *Invisible Ink* and *Camouflage* tricks worth the effort