The Spammer’s Compendium

John Graham-Cumming
popfile.sourceforge.net
January 17, 2003

Quick POPFile Introduction

- POPFile is an open source POP3 proxy that does automatic email classification
- POPFile uses Naïve Bayes and a specially adapted email parser to identify the type of an email
- Email types are user defined (e.g. spam, personal, work, knitting, …)
- Written because I get a lot of email every day… inspired by Jason Rennie's *ifile*.
- POPFile runs on Windows, Mac OS X, Linux, Solaris, FreeBSD, OS/2, …

Copyright (c) 2003 John Graham-Cumming

The Unexpected

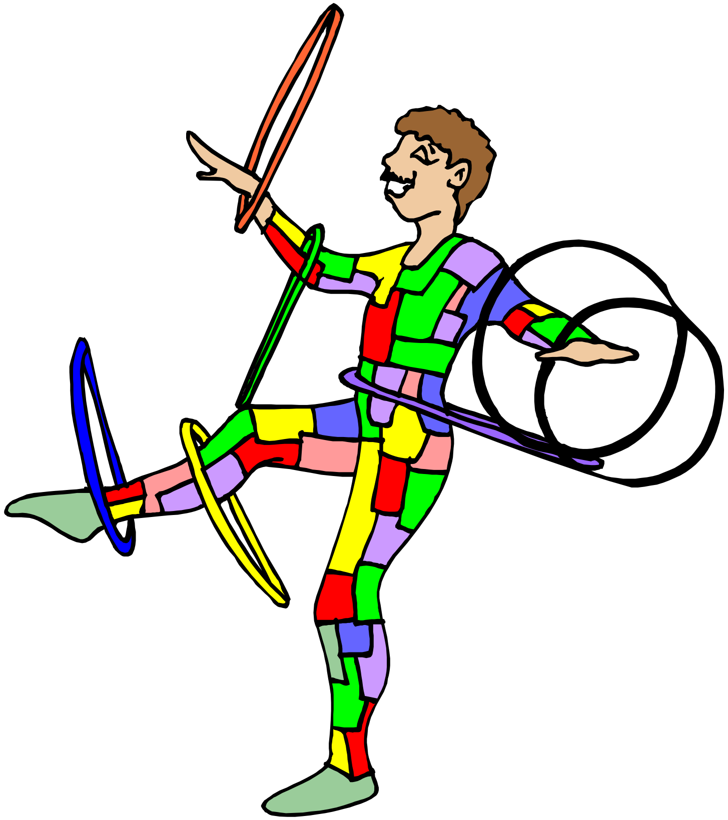
- POPFile has become very popular for spam removal
- Spammers seem to have noticed the power of POPFile and similar programs and have started to respond
- I've ended up reading a lot of spam to make POPFile better
- Here are a few things I've discovered…

Copyright (c) 2003 John Graham-Cumming

Spammers' Tricks

- Examples and Lessons:

  1. *The Big Picture*
  2. *Invisible Ink*
  3. *The Daily News*
  4. *Hypertextus Interruptus*
  5. *Slice and Dice*
  6. *MIME is Money*
  7. *LOST IN SPACE*
  8. *ENIGMA*
  9. *Script*
  10. *leet speak/Ze Foreign Accent*
  11. *Speaking in Tongues*



Copyright (c) 2003 John Graham-Cumming

The Big Picture

- Sending HTML with no words in it

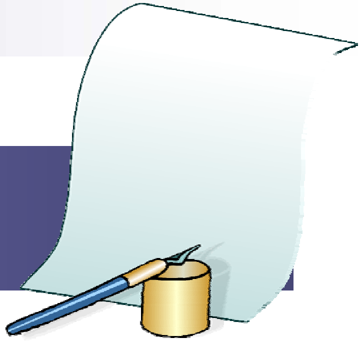
```
<html><img src="http://www.your-info-
station.com/Sla/chalkboard.gif"><div><a
href="http://www.your-info-
station.com/Sla/eb.php?x=52c"><img
src="http://www.your-info-
station.com/Sla/pitch.gif"></a></html>
```

- Lesson: Words should not be restricted to natural languages
  - □ POPFile reads certain HTML artefacts such as domain names to use as words (colored red above)

Common Trick

simple

Invisible Ink



- Add some real random words before HTML

```
3398782801 macabre macabro 9986649111 5484352062 2242352281 1466161152
2146781542 Annex (verb) take possession of, seize, capture 2594269869
```

- Add an email header packed with keywords no one sees

```
X-Mime-Key: search words: suspensory obscure aristocratical meningorachidian
unafeared brahmachari
```

- Write white text on a white background

```
<font color="white" size="-1">search words: suspensory obscure aristocratical
meningorachidian unafeared brahmachari</font>
```

- Lesson: If the user can't see it neither should the Bayes engine
  - POPFile ignores numbers; ignores headers it doesn't care to know about; will track the font background color in HTML
  - Small number of words unlikely to confuse Bayes, if spammer uses same words they become a way to identify the mail

Common Trick

sneaky

Copyright (c) 2003 John Graham-Cumming

The Daily News

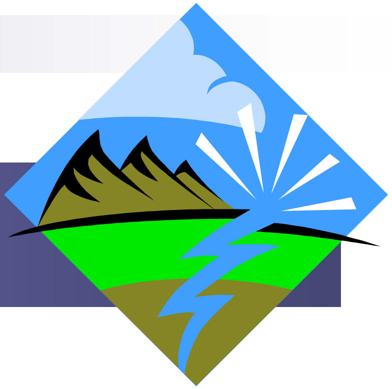
- Insert a bogus HTML tags containing large amount of text

```
<Despite statements last week from chief U.N. inspector Hans Blix that full
cooperation was expected from Iraq, Iraqi Foreign Minister Naji Sabri lashed out at
the United Nations in a 19-page letter to Secretary-General Kofi Annan written in
Arabic. In it, Sabri repeated previous claims that Iraq has no weapons of mass
destruction and that the inspections are just a false pretense for the United States
and Britain to attack his country. Sabri assailed U.N. Security Council resolution
1441, adopted November 8, that called for Iraq to give immediate, unfettered access to
weapons inspectors. Iraq "is being subjected to terrorism for more than 30 years from
international and regional powers," he wrote. "And Iraq's under a daily aggression
represented in the terrorism of the U.S. and Britain through the imposition of the no-
fly zones." Iraq has shot at U.S. and British aircraft repeatedly in the no-fly zones
since they were established after the Persian Gulf War, and coalition aircraft have
fired on Iraqi bases in response. In the most recent action, coalition aircraft struck
a mobile radar system Saturday in the southern no-fly zone, according to the U.S.
Central Command. The Iraqi News Agency said the aircraft fired on civilian and service
facilities. After Iraq fired on U.S. and British planes last week, U.S. officials said
the attacks constituted a "material breach" of Resolution 1441, which could trigger a
meeting of the U.N. Security Council at which the United States could call for
military action against Iraq>
```

- Lesson: need to handle HTML tags
  - Large text block is hidden by HTML browsers as it appears to be an invalid tag
  - Can confuse classifier that does not parse HTML correctly
  - Chosen to be likely to be in your regular email
  - POPFile parses HTML and ignores tags that would be ignored by the browser

Rarely Used

Hypertextus Interruptus*



- Split words using HTML comments

```
milli<!-- xe64 -->onaire
```

- Lesson: Mail parsers are going to need to understand HTML very well and will probably need a rendering engine of some fashion
  - □ Looks correct in browser/email client
  - □ POPFile automatically strips HTML comments and reconstructs words
  - □ This technique can be extended to include other blank HTML tag combinations: e.g. `milli<b></b>onaire`

*coined by Bill Yerazunis

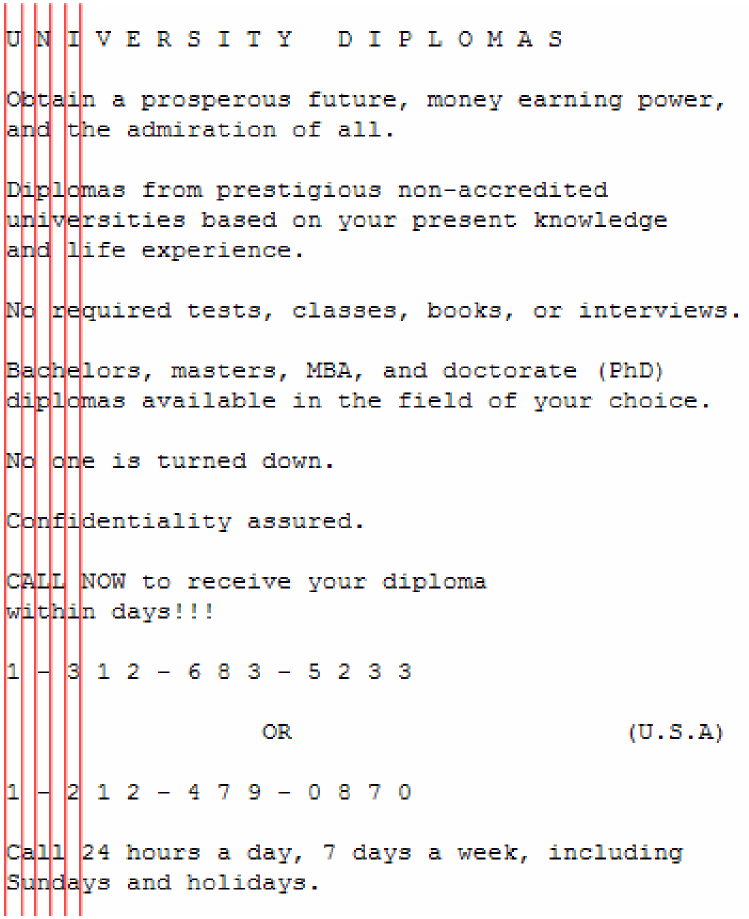
Common Trick

Slice and Dice

- Use <table> tag and monospace font to form text out of fragments

```
<table cellpadding=0 cellspacing=0 border=0><tr>
<td><table cellspacing=0 cellpadding=0 border=0><tr><td><font
       face="Courier New, Courier, mono" size=2>
&nbsp;<br>U<br>&nbsp;<br>O<br>a<br>&nbsp;<br>D<br>u<br>a<br>&nbsp;<br>N
       <br>&nbsp;<br>B<br>d<br>&nbsp;<br>N<br>&nbsp;<br>C<br>&nbsp;<br>C
       <br>w<br>&nbsp;<br>1<br>&nbsp;<br>&nbsp;<br>&nbsp;<br>1<br>&nbsp;
       <br>C<br>S<br></font></td></tr></table></td>
<td><table cellspacing=0 cellpadding=0 border=0><tr><td><font
       face="Courier New, Courier, mono" size=2>
&nbsp;&nbsp;&nbsp;<br>&nbsp;N&nbsp;<br>&nbsp;&nbsp;&nbsp;<br>bta<br>nd&
       nbsp;<br>&nbsp;&nbsp;&nbsp;<br>ipl<br>niv<br>nd&nbsp;<br>&nbsp;&n
       bsp;&nbsp;<br>o&nbsp;r<br>&nbsp;&nbsp;&nbsp;<br>ach<br>ipl<br>&nb
       sp;&nbsp;&nbsp;<br>o&nbsp;o<br>&nbsp;&nbsp;&nbsp;<br>onf<br>&nbsp
       ;&nbsp;&nbsp;<br>ALL<br>ith<br>&nbsp;&nbsp;&nbsp;<br>&nbsp;-
       &nbsp;<br>&nbsp;&nbsp;&nbsp;<br>&nbsp;&nbsp;&nbsp;<br>&nbsp;&nbsp
       ;&nbsp;<br>&nbsp;-
       &nbsp;<br>&nbsp;&nbsp;&nbsp;<br>all<br>und<br></font></td></tr></
       table></td>
<td><table cellspacing=0 cellpadding=0 border=0><tr><td><font
       face="Courier New, Courier, mono" size=2>
&nbsp;&nbsp;&nbsp;<br>I&nbsp;V<br>&nbsp;&nbsp;&nbsp;<br>in&nbsp;<br>the
       <br>&nbsp;&nbsp;&nbsp;<br>oma<br>ers<br>lif<br>&nbsp;&nbsp;&nbsp;
       <br>equ<br>&nbsp;&nbsp;&nbsp;<br>elo<br>oma<br>&nbsp;&nbsp;&nbsp;
       <br>ne&nbsp;<br>&nbsp;&nbsp;&nbsp;<br>ide<br>&nbsp;&nbsp;&nbsp;<b
       r>&nbsp;NO<br>in&nbsp;<br>&nbsp;&nbsp;&nbsp;<br>3&nbsp;1<br>&nbsp
       ;&nbsp;&nbsp;<br>&nbsp;&nbsp;&nbsp;<br>&nbsp;&nbsp;&nbsp;<br>2&nb
       sp;1<br>&nbsp;&nbsp;&nbsp;<br>&nbsp;24<br>ays<br></font></td></tr
       ></table></td>
<td><table cellspacing=0 cellpadding=0 border=0><tr><td><font
       face="Courier New, Courier, mono" size=2>
&nbsp;&nbsp;<br>&nbsp;E<br>&nbsp;&nbsp;<br>a&nbsp;<br>&nbsp;a<br>&nbsp;
       &nbsp;<br>s&nbsp;<br>it<br>e&nbsp;<br>&nbsp;&nbsp;<br>ir<br>&nbsp
       ;&nbsp;<br>rs<br>s&nbsp;<br>&nbsp;&nbsp;<br>is<br>&nbsp;&nbsp;<br
       >nt<br>&nbsp;&nbsp;<br>W&nbsp;<br>da<br>&nbsp;&nbsp;<br>&nbsp;2<b
       r>&nbsp;&nbsp;<br>&nbsp;&nbsp;<br>&nbsp;&nbsp;<br>&nbsp;2<br>&nbs
       p;&nbsp;<br>&nbsp;h<br>&nbsp;a<br></font></td></tr></table></td>
```

- Takes horizontal English text and divides it vertically yielding only word fragments that a parser can understand
- Lesson: spammers are using smart people to think up these tricks



UNIVERSITY DIPLOMAS

Obtain a prosperous future, money earning power,
and the admiration of all.

Diplomas from prestigious non-accredited
universities based on your present knowledge
and life experience.

No required tests, classes, books, or interviews.

Bachelors, masters, MBA, and doctorate (PhD)
diplomas available in the field of your choice.

No one is turned down.

Confidentiality assured.

CALL NOW to receive your diploma
within days!!!

1-312-683-5233

OR (U.S.A)

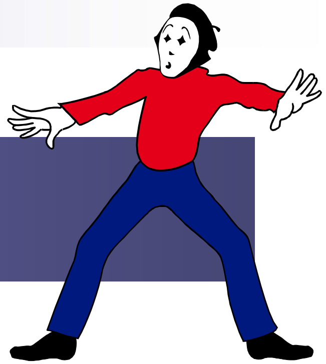
1-212-479-0870

Call 24 hours a day, 7 days a week, including
Sundays and holidays.

Rarely Used

Copyright (c) 2003 John Graham-Cumming

MIME is Money



- A two part MIME document with the spam message in the HTML section and bogus text in plain text section

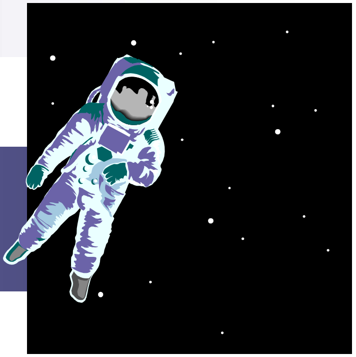
```
------=_NextPart_001_2D3DF_01C29D73.26716240
Content-Type: text/plain;
The modes of letting vacant farms, the duty of supplying buildings and permanent
   improvements, and the form in which rent is to be received, have all been
   carefully discussed in the older financial treatises. Most of these questions
   belong to practical administration, and are, moreover, not of great interest in
   modern times. Certain plain rules, may, however, be stated. The claims of
   successors to the late tenant should not be overlooked; it is better for the
   tenure to be continued without break, and therefore the question of new letting
   ought rarely to
occur.
------=_NextPart_001_2D3DF_01C29D73.26716240
Content-Type: text/html;
<p><b><font face=Arial>Now is the perfect time to get a mortgage, and we have a
   simple and free way for you to get started.</font></b></td>
```

- Lesson: may need to score MIME parts separately to check for discrepancies

clever

Rarely Used

LOST IN SPACE



- Space out words to make them unrecognizable to word parsers

```
M O R T G A G E
```

- Other characters can be used instead

```
F*R*E*E V’I’A’G’R’A O*N*L*I*N*E
```

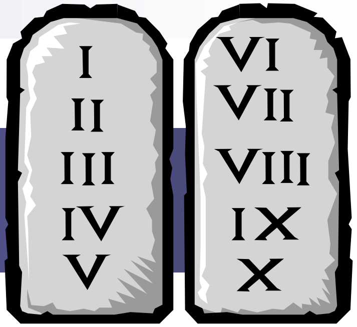
- Lesson: parsers need to stay up with spammers’ tricks that exploit the difference between human and machine pattern matchers
  - Humans are very good at seeing MORTGAGE, FREE, VIAGRA and ONLINE
  - POPFile automatically merges words separated by common characters into the real words
  - POPFile sees mortgage, free, viagra and online in the example

Common Trick

mundane

Copyright (c) 2003 John Graham-Cumming

ENIGMA



- To hide URLs spammers use various encoding techniques: decimal, hex and octal

```
http://7763631671/obscure.htm
http://0xCeBF9e37/obscure.htm
http://0316.0277.0236.067/obscure.htm
http://3468664375@3468664375/o%62s%63ur%65%2e%68t%6D
```

- Lesson: Filter writers need to learn about HTML encodings of all forms and can learn from the web filtering vendors.
  - POPFile recognizes all these variants and turns them into a canonical form
  - POPFile removes username/password from URLs

sneaky

Rarely Used

Copyright (c) 2003 John Graham-Cumming

Script

- Placing entire spam in a Javascript that changes the email contents on load

```
<HTML><HEAD><SCRIPT LANGUAGE="Javascript"><!-- var
Words="%3CHTML%3E%0D%0A%3CHEAD%3E%0D%0A%3CTITLE%3E%3C/TITLE%3E%0D%0A%3CMETA%20HTTP-
EQUIV%3D%22Content-
Type%22%20CONTENT%3D%22text/html%3B%20charset%3DBig5%22%3E%0D%0A%3CMETA%20HTTP-
EQUIV%3D%22Expires%22%20CONTENT%3D%22Sat%2C%201%20Jan%202000%2000%3A00%3A00%20GMT%22%3E%0D%0A%3
CMETA%20HTTP-EQUIV%3D%22Pragma%22%20CONTENT%3D%22no-
cache%22%3E%0D%0A%3C/HEAD%3E%0D%0A%3CFRAMESET%20ROWS%3D%22100%25%2C0%22%20FRAMEBORDER%3DNO%20BO
RDER%3D%220%22%20FRAMESPACING%3D0%3E%0D%0A%3CFRAME%20SRC%3D%22http%3A//203.204.53.231/a1_K_2/e1
2w_k2/a_w_a_0__2k-
1_second%22%20NAME%3D%22AMENU%22%20SCROLLING%3DAUTO%20MARGINHEIGHT%3D0%20MARGINWIDTH%3D0%3E%0D%
0A%3CFRAME%20SRC%3D%22%22%20SCROLLING%3DNO%20noresize%3E%0D%0A%3C/FRAMESET%3E%0D%0A%3CNOFRAMES%
3E%0D%0A%3C/NOFRAMES%3E%0D%0A%3C/HTML%3E%0D%0A“ function SetNewWords() { var NewWords; NewWords
= unescape(Words); document.write(NewWords); } SetNewWords(); // --> </SCRIPT> </HEAD> <BODY>
</BODY> </HTML>
```

- Lesson: decoding Javascript is important
  - □ POPFile does not have a good solution to this… yet!
  - □ POPFile will read the contents of Javascript variables looking for encoded HTML
  - □ POPFile already handles all the text encoding seen above

simple

Fairly Rare

leet speak/Ze Foreign Accent

- Replace letters that look like numbers with numbers

```
V1DE0 T4PE M0RTG4GE
```

- Use accented characters in English

```
Fántástìç -- eárn mõnéy thrôugh
unçõlleçted judgments
```

- Lesson: need to undo these mappings

Common Trick

mundane

Copyright (c) 2003 John Graham-Cumming

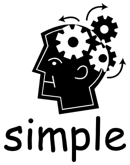
Speaking in Tongues

crecrephas
wukutugucr
ovazichonu

- Adding long random words

```
crecrephaswukutugucrovazichonuprixis
luwephimajoq
```

- Lesson: needs techniques to keep corpus of words “clean”
  - □ Either displayed at end of message and ignored by user or not displayed at all
  - □ Chosen to be recognized as a “word” by classifiers
  - □ Might be chosen to make a message look base64 encoded
  - □ Single occurrence only skews Bayes calculation slightly
  - □ POPFile plans a corpus aging technique to eliminate nonsense words



Common Trick

Copyright (c) 2003 John Graham-Cumming

Big Lessons from POPFile

- Thomas Bayes knew what he was talking about!
- Parsing mail messages is the hard work
- Spammers’ technical ability should be respected
- Training a Bayesian filter only on errors is effective and intuitive to average users
- Spam to cell phones cost real money when you pay per byte

Copyright (c) 2003 John Graham-Cumming

How to get POPFile

- Download POPFile from

  http://popfile.sourceforge.net

- Help support POPFile by using it, reporting bugs, updating the manual, translating the instructions into other languages, …

Copyright (c) 2003 John Graham-Cumming