

osdir.com

mailing list archive
F.A.Q. -since 2001! Search**Subject: Latest CVS update, Ocrad for Windows - msg#00030****List:** mail.spam.spambayes.devel

mail.spam.spambayes.devel Navigation:
 Date: [Prev](#) [Next](#) [Date Index](#) Thread: [Prev](#) [Next](#) [Thread Index](#)

[Website Performance Monitoring](#)

I updated the OCR capabilities a bit more today. I added more intelligent assembly of split images into a single image after noticing that the spammers don't simply chop up multi-part GIF images horizontally. I also added a couple extra options (ocrad_scale and ocrad_charset) which control the image scaling factor (default is 2) and character set (default is "ascii") Ocrad uses. Scaling the image by a factor of 2 was a pretty obvious win:

false positive percentages
 0.000 0.000 tied
 0.000 0.000 tied
 0.000 0.000 tied
 0.000 0.000 tied
 0.000 0.000 tied

won 0 times
 tied 5 times
 lost 0 times

total unique fp went from 0 to 0 tied
 mean fp % went from 0.0 to 0.0 tied

false negative percentages
 4.213 4.213 tied
 1.404 0.843 won -39.96%
 3.371 2.809 won -16.67%
 2.528 2.247 won -11.12%
 4.213 3.652 won -13.32%

won 4 times
 tied 1 times
 lost 0 times

total unique fn went from 56 to 49 won -12.50%
 mean fn % went from 3.14606741573 to 2.75280898876 won -12.50%

Scaling by a factor of three was even better in the false negative department but regressed a bit in the false positive category so I checked Options.py in with a default scaling factor of 2. A couple things could stand to be further tested:

* I have no idea how good Ocrad's scaling algorithm is. It's possible that PIL or NetPBM's scaling code is better. If so, it would make sense to scale the images before feeding to Ocrad.

* The images I've see so far were all plain English, so I blindly made ascii the default charset. The other choices were iso-8859-9 and iso-8859-15. I simply assumed ascii would be the most appropriate default, but didn't test it.

Finally, I put together a really simpleminded Ocrad-for-Windows release based upon the ocrad.exe binary that Tony built. Check the Files section of the SpamBayes project site:

http://sourceforge.net/project/showfiles.php?group_id=61702

and grab ocrad-cygwin.

There are a few caveats:

1. I don't do Windows. (No, really, I don't, strange as that may seem.) This is no fancy-schmancy point-and-shoot Windows installer. It's just a simple zip file with the Ocrad 0.15 distribution, Tony's .exe file and the patch he applied to the source.

2. I don't do Windows. The code I've written so far has been done entirely on my Mac. I've made no obvious concessions to portability. That said, I hope portability issues won't be daunting for any early adopters.

3. I don't do Windows. If you have problems it won't do you any good to mail me directly. Post about problems on the SpamBayes bug tracker:

http://sourceforge.net/tracker/?group_id=61702&atid=498103

4. If you do Windows you will need PIL to take advantage of the recent changes:

<http://www.pythonware.com/products/pil/>

(unless you want to put hair on your chest and build NetPBM on Windows). Fredrik Lundh provides prebuilt Windows versions of PIL. Grab the one appropriate for the version of Python you have installed.

5. If you do Windows (or any other platform for that matter), feedback to the lists about successes and failures would be helpful.

Cheers,

Skip

Recent Msgs:

[commits.gnome/2014-01/msg06031.html](#)
[postgresql-pgsql-hackers/2014-01/msg02002.html](#)
[users-felix-apache/2014-01/msg00051.html](#)
[wp-trac/2014-01/msg01630.html](#)
[dev-directory-apache/2014-01/msg00074.html](#)
[general/2014-01/msg46037.html](#)
[erlang-programming-patches-discuss/2014-01/msg00045.html](#)
[spreed-user/2014-01/msg00165.html](#)
[dev-httpd/2014-01/msg00168.html](#)
[ubuntu-bugs/2014-01/msg21414.html](#)

Thread at a glance:

Previous Message by Date:

[Re: Patch for ocrad to run on Windows?](#)

Next Message by Date:

[How about a 1.1a3 release?](#)

Previous Message by Thread:

[Patch for ocrad to run on Windows?](#)

Next Message by Thread:

[Re: Latest CVS update, Ocrad for Windows](#)**Date Index of Msgs:**