Reigning over high-volume Debian e-mails

Who?    Pablo Ariel Duboue\(^1\)

When?   Debian Conference 2012

\(^1\)DrDub – pablo.duboue@gmail.com
Outline.

What Is It?
How It Works?

Mailing List -> Twitter
Bringing Back Kernel Traffic.

Discussion
Motivation.

- Debian community depends heavily on emails
  - Personal communication
  - Email-based tools
- Each Debian contributor has developed their own set of tools and habits.
- Machine learning and automatic text classification have for many years promised a better solution to this problem.

This meeting

- Share a tool I have developed for my personal use and another for community use
- Brainstorm other techniques / solutions and see if we can adapt the tool I created for a larger audience.
Architecture.
### Demo.

<table>
<thead>
<tr>
<th>Time period</th>
<th></th>
<th>Score Threshold:</th>
<th></th>
<th>Mailing List:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 day</td>
<td></td>
<td>-10.0</td>
<td></td>
<td>debconf-discuss</td>
</tr>
<tr>
<td>2 days</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>week</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>month</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Total messages: 104 (page 1 of 3)**

<table>
<thead>
<tr>
<th>Mailing List</th>
<th>Date</th>
<th>Score</th>
<th>Subject</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>debconf-discuss</td>
<td>Jul 12, 2012 12:51:49 PM</td>
<td>0.5</td>
<td>Re: [Debconf-discuss] [penta] Evaluating new conference software: Comas (reloaded)</td>
<td>Gunnar Wolf<a href="mailto:gwolf@gwolf.org">gwolf@gwolf.org</a></td>
</tr>
<tr>
<td>debconf-discuss</td>
<td>Jul 12, 2012 12:51:49 PM</td>
<td>0.5</td>
<td>Re: [Debconf-discuss] [penta] Evaluating new conference software: Comas (reloaded)</td>
<td>Gunnar Wolf<a href="mailto:gwolf@gwolf.org">gwolf@gwolf.org</a></td>
</tr>
<tr>
<td>debconf-discuss</td>
<td>Jul 12, 2012 13:31:28 PM</td>
<td>0.5</td>
<td>[Debconf-discuss] Assassins: the tension is increasing….</td>
<td>Christian PERRIER<a href="mailto:bubulle@debian.org">bubulle@debian.org</a></td>
</tr>
<tr>
<td>debconf-discuss</td>
<td>Jul 12, 2012 12:23:45 PM</td>
<td>0.5</td>
<td>Re: [Debconf-discuss] [penta] Evaluating new conference software</td>
<td>Gunnar Wolf<a href="mailto:gwolf@gwolf.org">gwolf@gwolf.org</a></td>
</tr>
<tr>
<td>debconf-discuss</td>
<td>Jul 12, 2012 12:03:24 PM</td>
<td>0.5</td>
<td>Re: [Debconf-discuss] [Debconf-announce] DebConf12 Mass OpenPGP Keysigning: 18:00 Thursday, July 12 in Roberto Terán</td>
<td>Daniel Kahn Gillmor<a href="mailto:dkg@fifthhorseman.net">dkg@fifthhorseman.net</a></td>
</tr>
<tr>
<td>debconf-discuss</td>
<td>Jul 12, 2012 11:55:54 AM</td>
<td>0.5</td>
<td>Re: [Debconf-discuss] About the &quot;Seeds of Resistance&quot; event - In front of the hacklab, July 12, 12:00PM</td>
<td>Gunnar Wolf<a href="mailto:gwolf@gwolf.org">gwolf@gwolf.org</a></td>
</tr>
<tr>
<td>debconf-discuss</td>
<td>Jul 12, 2012 11:39:41 AM</td>
<td>-1.5</td>
<td>[Debconf-discuss] Lost camara</td>
<td>Roger Orellana<a href="mailto:rjorellana@gmail.com">rjorellana@gmail.com</a></td>
</tr>
<tr>
<td>debconf-discuss</td>
<td>Jul 12, 2012 10:18:55 AM</td>
<td>0.5</td>
<td>Re: [Debconf-discuss] About the &quot;Seeds of Resistance&quot; event - In front of the hacklab, July 12, 12:00PM</td>
<td><a href="mailto:mangoderosa@riseup.net">mangoderosa@riseup.net</a><a href="mailto:mangoderosa@riseup.net">mangoderosa@riseup.net</a></td>
</tr>
<tr>
<td>debconf-discuss</td>
<td>Jul 12, 2012 10:00:00 AM</td>
<td>0.5</td>
<td>Re: [Debconf-discuss] About the &quot;Seeds of Resistance&quot; event - In front of the hacklab, July 12, 12:00PM</td>
<td>Tom Marble<a href="mailto:tmarble@info9.net">tmarble@info9.net</a></td>
</tr>
</tbody>
</table>
Demo.
Demo.
Database Schema.
Classifier.

- dbacl is a high performance statistical classifier.
  - Handles e-mails natively
  - Can make use of multiple words if there is enough data
  - Written in C
  - Relatively easy to modify, if need be
  - It can play chess!

- The system waits until there are 10 emails annotated as positive and 10 negative to start classifying.
  - Messages marked as +2 or -2 are always included in the train set
  - Messages marked as +1 or -1 are sampled
Rule Engine.

- Besides classification, it is possible to write simple rules.
  - For example, add +1.0 to all emails coming from debian.org emails.
  - Or add +10.0 to all emails mentioning your name/nick/last name

- Currently executable Perl code.
  - Stored in the DB
  - Agh!

- Global and per-list based.
UI.

- AJAX-based
  - NextApp Echo is a Swing-like Ajax Framework for Java.
- The UI is actually written in Scala
  - Using Jetty and a cross compiled SQLite3
- While functional, it has shortcomings
  - Does not expose the threads in the DB
  - Has trouble with encoded emails
Roadmap.

- Fix UI issues
- Fix rule language (lua?)
- IMap integration?
- First installable version
Meet @debian_es.

- Twitter bridge for the debian-user-spanish mailing list
- About 8 months on line.
  - ~1k tweets
  - ~100 followers
- Script available at:
  - http://duboue.net/download/debian_es_bot.pl
  - 177 lines (Perl)
  - Public Domain
- Tweets once per thread, only when a thread has at least 3 replies by at least 2 different people.
Kernel Traffic, Automatic Summaries for LKML.


5 years of summaries, annotated in XML

Very good data for machine learning

Next step: aligning summaries and emails

Anybody has LKML archives? (on-line are missing many emails)
Joey’s Thread Patterns.

- http://joeyh.name/blog/entry/thread_patterns/
- "take it to private email"
- "think before you post"
- "blindingly obvious answer"
- "uninteresting message"
Discussion.

- Which tools do you use?
- How do you cope?
- Tricks and tips?
- Would you use the smart mailing list reader?
- Join in! #listosoreader on FreeNode or /msg DrDub on OFTC