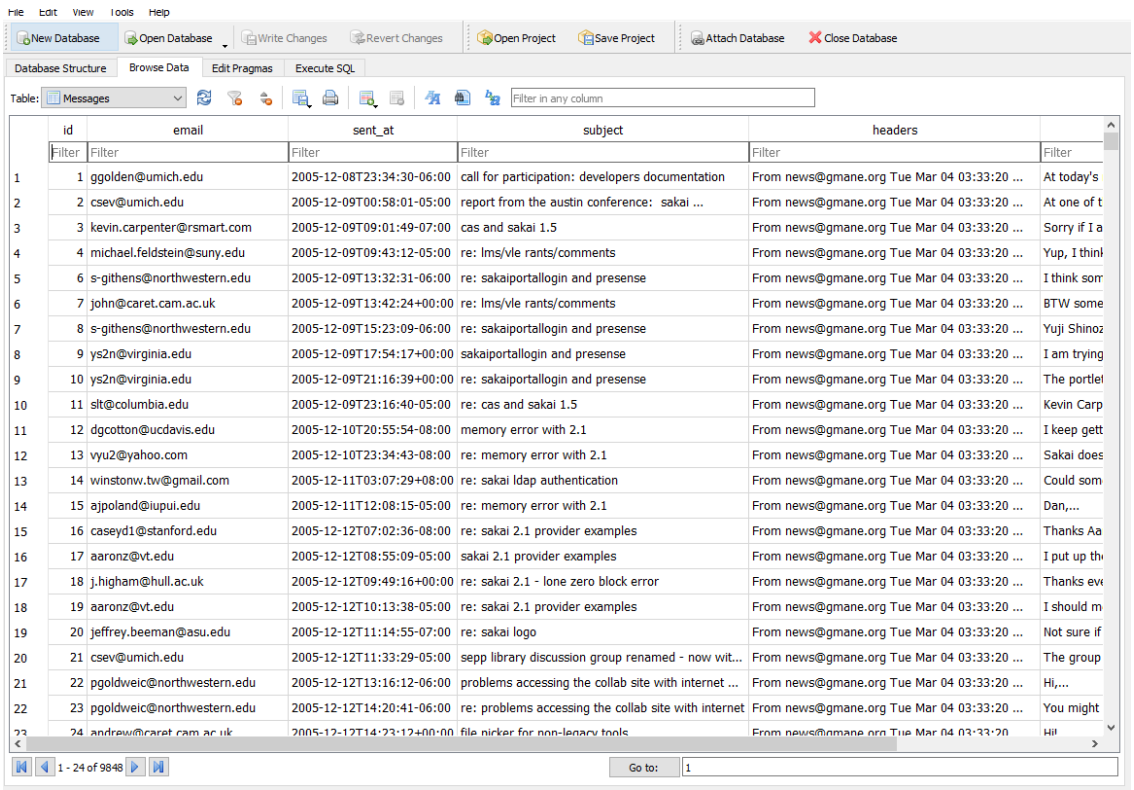
For this project, I retrieved a large set of email data from the web, and then I stored it in a database. I ran gmane.py, a python script, to retrieve email data and then store it into a database called content.sqlite.

```
C:\Users\ilmaa\Desktop\gmane>gmane.py
How many messages:60000
http://mbox.dr-chuck.net/sakai.devel/1/2 2662
    ggolden@umich.edu 2005-12-08T23:34:30-06:00 call for participation: developers documentation
http://mbox.dr-chuck.net/sakai.devel/2/3 2434
    csev@umich.edu 2005-12-09T00:58:01-05:00 report from the austin conference:  sakai developers break into song
http://mbox.dr-chuck.net/sakai.devel/3/4 3055
    kevin.carpenter@rsmart.com 2005-12-09T09:01:49-07:00 cas and sakai 1.5
http://mbox.dr-chuck.net/sakai.devel/4/5 11721
    michael.feldstein@suny.edu 2005-12-09T09:43:12-05:00 re: lms/vle rants/comments
Unable to retrieve or parse page http://mbox.dr-chuck.net/sakai.devel/5/6
Error HTTP Error 522: Origin Connection Time-out
http://mbox.dr-chuck.net/sakai.devel/6/7 3586
    s-githens@northwestern.edu 2005-12-09T13:32:31-06:00 re: sakaiportallogin and presense
http://mbox.dr-chuck.net/sakai.devel/7/8 10600
    john@caret.cam.ac.uk 2005-12-09T13:42:24+00:00 re: lms/vle rants/comments
http://mbox.dr-chuck.net/sakai.devel/8/9 4892
    s-githens@northwestern.edu 2005-12-09T15:23:09-06:00 re: sakaiportallogin and presense
http://mbox.dr-chuck.net/sakai.devel/9/10 3206
    ys2n@virginia.edu 2005-12-09T17:54:17+00:00 sakaiportallogin and presense
http://mbox.dr-chuck.net/sakai.devel/10/11 4993
    ys2n@virginia.edu 2005-12-09T21:16:39+00:00 re: sakaiportallogin and presense
http://mbox.dr-chuck.net/sakai.devel/11/12 3298
    slt@columbia.edu 2005-12-09T23:16:40-05:00 re: cas and sakai 1.5
http://mbox.dr-chuck.net/sakai.devel/12/13 3096
    dgcotton@ucdavis.edu 2005-12-10T20:55:54-08:00 memory error with 2.1
http://mbox.dr-chuck.net/sakai.devel/13/14 4307
    vyu2@yahoo.com 2005-12-10T23:34:43-08:00 re: memory error with 2.1
http://mbox.dr-chuck.net/sakai.devel/14/15 4935
    winstonw.tw@gmail.com 2005-12-11T03:07:29+08:00 re: sakai ldap authentication
http://mbox.dr-chuck.net/sakai.devel/15/16 4374
    ajpoland@iupui.edu 2005-12-11T12:08:15-05:00 re: memory error with 2.1
http://mbox.dr-chuck.net/sakai.devel/16/17 2475
    caseyd1@stanford.edu 2005-12-12T07:02:36-08:00 re: sakai 2.1 provider examples
http://mbox.dr-chuck.net/sakai.devel/17/18 3931
    aaronz@vt.edu 2005-12-12T08:55:09-05:00 sakai 2.1 provider examples
http://mbox.dr-chuck.net/sakai.devel/18/19 3828
    j.higham@hull.ac.uk 2005-12-12T09:49:16+00:00 re: sakai 2.1 - lone zero block error
http://mbox.dr-chuck.net/sakai.devel/19/20 2839
    aaronz@vt.edu 2005-12-12T10:13:38-05:00 re: sakai 2.1 provider examples
http://mbox.dr-chuck.net/sakai.devel/20/21 8505
    jeffrey.beeman@asu.edu 2005-12-12T11:14:55-07:00 re: sakai logo
http://mbox.dr-chuck.net/sakai.devel/21/22 2625
    csev@umich.edu 2005-12-12T11:33:29-05:00 sepp library discussion group renamed - now with open membership
http://mbox.dr-chuck.net/sakai.devel/22/23 2960
    pgoldweic@northwestern.edu 2005-12-12T13:16:12-06:00 problems accessing the collab site with internet explorer...
http://mbox.dr-chuck.net/sakai.devel/23/24 4715
    pgoldweic@northwestern.edu 2005-12-12T14:20:41-06:00 re: problems accessing the collab site with internet
http://mbox.dr-chuck.net/sakai.devel/24/25 2705
    andrew@caret.cam.ac.uk 2005-12-12T14:23:12+00:00 file picker for non-legacy tools
http://mbox.dr-chuck.net/sakai.devel/25/26 4378
```



File Edit View Tools Help

New Database | Open Database | Write Changes | Revert Changes | Open Project | Save Project | Attach Database | Close Database

Database Structure | Browse Data | Edit Pragmas | Execute SQL

Table: Messages | Filter in any column

	id	email	sent_at	subject	headers	
	Filter	Filter	Filter	Filter	Filter	Filter
1	1	ggolden@umich.edu	2005-12-08T23:34:30-06:00	call for participation: developers documentation	From news@gmane.org Tue Mar 04 03:33:20 ...	At today's
2	2	csev@umich.edu	2005-12-09T00:58:01-05:00	report from the austin conference: sakai ...	From news@gmane.org Tue Mar 04 03:33:20 ...	At one of t
3	3	kevin.carpenter@rsmart.com	2005-12-09T09:01:49-07:00	cas and sakai 1.5	From news@gmane.org Tue Mar 04 03:33:20 ...	Sorry if I a
4	4	michael.feldstein@suny.edu	2005-12-09T09:43:12-05:00	re: lms/vle rants/comments	From news@gmane.org Tue Mar 04 03:33:20 ...	Yup, I think
5	6	s-githens@northwestern.edu	2005-12-09T13:32:31-06:00	re: sakaiportallogin and presense	From news@gmane.org Tue Mar 04 03:33:20 ...	I think som
6	7	john@caret.cam.ac.uk	2005-12-09T13:42:24+00:00	re: lms/vle rants/comments	From news@gmane.org Tue Mar 04 03:33:20 ...	BTW some
7	8	s-githens@northwestern.edu	2005-12-09T15:23:09-06:00	re: sakaiportallogin and presense	From news@gmane.org Tue Mar 04 03:33:20 ...	Yuji Shinoz
8	9	ys2n@virginia.edu	2005-12-09T17:54:17+00:00	sakaiportallogin and presense	From news@gmane.org Tue Mar 04 03:33:20 ...	I am trying
9	10	ys2n@virginia.edu	2005-12-09T21:16:39+00:00	re: sakaiportallogin and presense	From news@gmane.org Tue Mar 04 03:33:20 ...	The portlet
10	11	slt@columbia.edu	2005-12-09T23:16:40-05:00	re: cas and sakai 1.5	From news@gmane.org Tue Mar 04 03:33:20 ...	Kevin Carp
11	12	dgcotton@ucdavis.edu	2005-12-10T20:55:54-08:00	memory error with 2.1	From news@gmane.org Tue Mar 04 03:33:20 ...	I keep gett
12	13	vyu2@yahoo.com	2005-12-10T23:34:43-08:00	re: memory error with 2.1	From news@gmane.org Tue Mar 04 03:33:20 ...	Sakai does
13	14	winstonw.tw@gmail.com	2005-12-11T03:07:29+08:00	re: sakai ldap authentication	From news@gmane.org Tue Mar 04 03:33:20 ...	Could som
14	15	ajpoland@iupui.edu	2005-12-11T12:08:15-05:00	re: memory error with 2.1	From news@gmane.org Tue Mar 04 03:33:20 ...	Dan,...
15	16	caseyd1@stanford.edu	2005-12-12T07:02:36-08:00	re: sakai 2.1 provider examples	From news@gmane.org Tue Mar 04 03:33:20 ...	Thanks Aa
16	17	aaronz@vt.edu	2005-12-12T08:55:09-05:00	sakai 2.1 provider examples	From news@gmane.org Tue Mar 04 03:33:20 ...	I put up th
17	18	j.higham@hull.ac.uk	2005-12-12T09:49:16+00:00	re: sakai 2.1 - lone zero block error	From news@gmane.org Tue Mar 04 03:33:20 ...	Thanks eve
18	19	aaronz@vt.edu	2005-12-12T10:13:38-05:00	re: sakai 2.1 provider examples	From news@gmane.org Tue Mar 04 03:33:20 ...	I should m
19	20	jeffrey.beeman@asu.edu	2005-12-12T11:14:55-07:00	re: sakai logo	From news@gmane.org Tue Mar 04 03:33:20 ...	Not sure if
20	21	csev@umich.edu	2005-12-12T11:33:29-05:00	sepp library discussion group renamed - now wit...	From news@gmane.org Tue Mar 04 03:33:20 ...	The group
21	22	pgoldweic@northwestern.edu	2005-12-12T13:16:12-06:00	problems accessing the collab site with internet ...	From news@gmane.org Tue Mar 04 03:33:20 ...	Hi,...
22	23	pgoldweic@northwestern.edu	2005-12-12T14:20:41-06:00	re: problems accessing the collab site with internet	From news@gmane.org Tue Mar 04 03:33:20 ...	You might
23	24	andrew@caret.cam.ac.uk	2005-12-12T14:23:12+00:00	file picker for non-legacy tools	From news@gmane.org Tue Mar 04 03:33:20	Hi!

1 - 24 of 9848 | Go to: 1

Next, I clean up the email data in content.sqlite by running gmodel.py, which normalizes the email data into structured tables and sends the data into the database index.sqlite.

```
C:\Users\ilmaa\Documents\code3\gmane> gmodel.py
Loaded allsenders 232 and mapping 29 dns mapping 1
1 2005-12-08T23:34:30-06:00 ggolden22@mac.com
251 2005-12-22T10:03:20-08:00 tpamsler@ucdavis.edu
501 2006-01-12T11:44:14+02:00 marquard@uct.ac.za
751 2006-01-24T17:05:52-06:00 omer@rice.edu
1001 2006-02-02T16:12:58-05:00 ggolden22@mac.com
1251 2006-02-16T12:08:10-05:00 ggolden22@mac.com
1501 2006-02-24T10:44:32-08:00 lydial@stanford.edu
1751 2006-03-13T21:48:03+00:00 ian@cam.ac.uk
2001 2006-03-28T21:44:58-05:00 slt@columbia.edu
2251 2006-04-06T13:52:37-04:00 azeckoski@unicon.net
2501 2006-04-18T16:09:03-04:00 daisy.flemming@gmail.com
2751 2006-04-26T13:08:22-07:00 caseyd1@stanford.edu
3001 2006-05-04T13:39:13-04:00 csev@umich.edu
```

DB Browser for SQLite - C:\Users\ilmaa\Desktop\code3\gmane\index.sqlite

File Edit View Tools Help

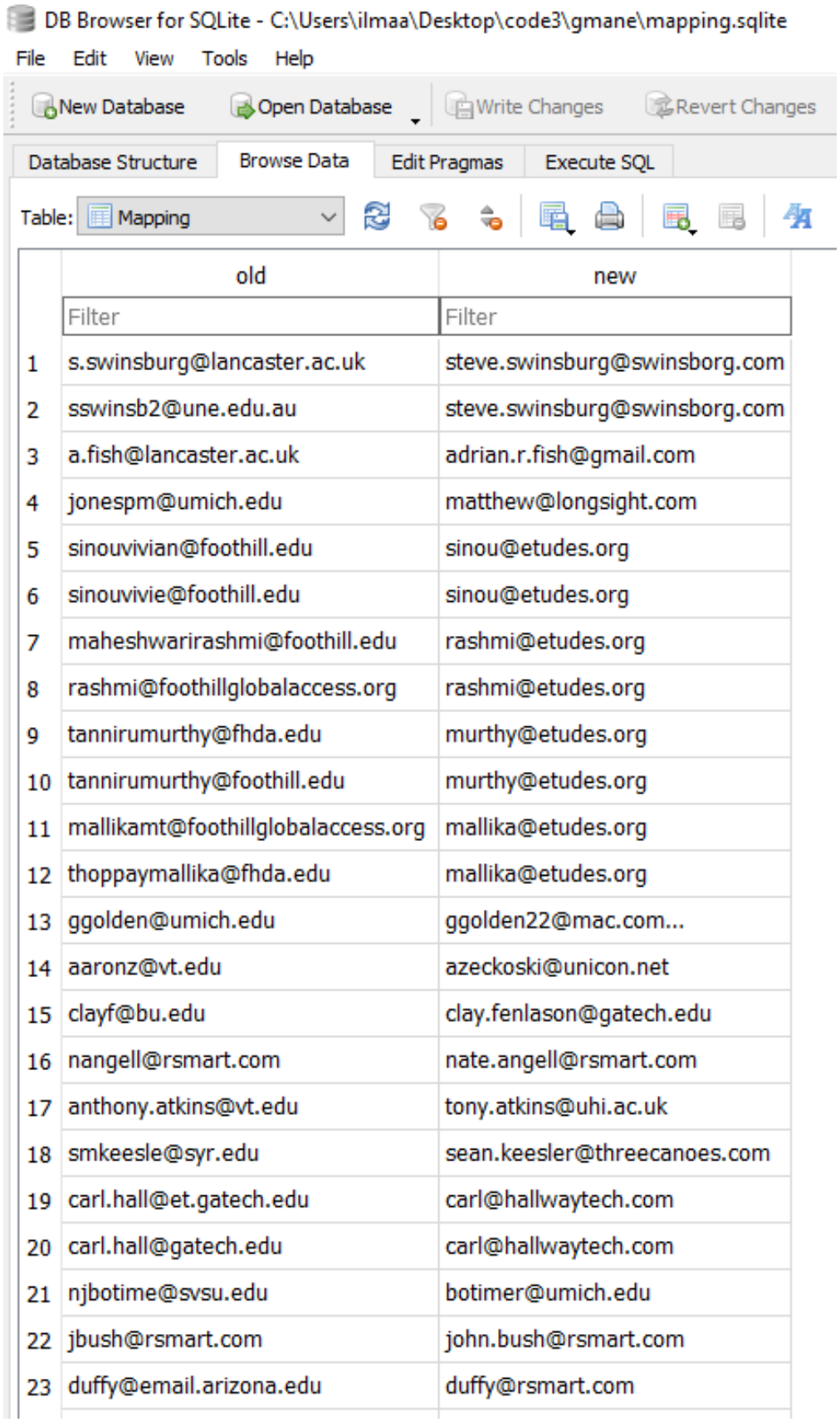
New Database | Open Database | Write Changes | Revert Changes | Open Project | Save Project | Attach Databas

Database Structure | Browse Data | Edit Pragmas | Execute SQL

Table: Messages | Filter in any column

	id	guid	sent_at	sender_id	subject_id	headers	body
	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	1	<b7eaf76e-3366-4ae3-...	2005-12-09 05:34:30	1	1	BLOB	BLOB
2	2	<bb7ce631-ccf0-40a3-...	2005-12-09 05:58:01	2	2	BLOB	BLOB
3	3	<16703710.1134144397398.javamail.tomcat5@s...	2005-12-09 16:01:49	3	3	BLOB	BLOB
4	4	<9622225.1134139564303.javamail.tomcat5@m...	2005-12-09 14:43:12	4	4	BLOB	BLOB
5	5	<22585576.1134135343809.javamail.tomcat5@...	2005-12-09 13:32:29	5	4	BLOB	BLOB
6	6	<29494755.1134157064976.javamail.tomcat5@s...	2005-12-09 19:32:31	6	5	BLOB	BLOB
7	7	<20623058.1134135893412.javamail.tomcat5@...	2005-12-09 13:42:24	5	4	BLOB	BLOB
8	8	<14721465.1134163692331.javamail.tomcat5@...	2005-12-09 21:23:09	6	5	BLOB	BLOB
9	9	<31872472.1134150954615.javamail.tomcat5@...	2005-12-09 17:54:17	7	6	BLOB	BLOB
10	10	<1304695.1134163143767.javamail.tomcat5@m...	2005-12-09 21:16:39	7	5	BLOB	BLOB
11	11	<3598962.1134188359277.javamail.tomcat5@sh...	2005-12-10 04:16:40	8	7	BLOB	BLOB
12	12	<590808.1134277104504.javamail.tomcat5@ma...	2005-12-11 04:55:54	9	8	BLOB	BLOB
13	13	<27775220.1134286694278.javamail.tomcat5@s...	2005-12-11 07:34:43	10	9	BLOB	BLOB
14	14	<18541911.1134241825684.javamail.tomcat5@s...	2005-12-10 19:07:29	11	10	BLOB	BLOB
15	15	<20021180.1134321068392.javamail.tomcat5@s...	2005-12-11 17:08:15	12	9	BLOB	BLOB
16	16	<28266638.1134399879087.javamail.tomcat5@s...	2005-12-12 15:02:36	13	11	BLOB	BLOB
17	17	<1755944.1134395859482.javamail.tomcat5@m...	2005-12-12 13:55:09	14	12	BLOB	BLOB
18	18	<10733551.1134381175177.javamail.tomcat5@...	2005-12-12 09:49:16	15	13	BLOB	BLOB
19	19	<13486694.1134400555513.javamail.tomcat5@s...	2005-12-12 15:13:38	14	11	BLOB	BLOB
20	20	<25165005.1134413878170.javamail.tomcat5@...	2005-12-12 18:14:55	16	14	BLOB	BLOB
21	21	<d3dc4255-5936-4582-9aa1-05b298302dfa@um...	2005-12-12 16:33:29	2	15	BLOB	BLOB
22	22	<6.0.1.1.2.20051212130514.0277fce8@casbah.it...	2005-12-12 19:16:12	17	16	BLOB	BLOB
23	23	<6.0.1.1.2.20051212140958.027c5e10@casbah.i...	2005-12-12 20:20:41	17	17	BLOB	BLOB

1 - 24 of 600 | Go to: 1



	old	new
	Filter	Filter
1	s.swinsburg@lancaster.ac.uk	steve.swinsburg@swinsborg.com
2	sswinsb2@une.edu.au	steve.swinsburg@swinsborg.com
3	a.fish@lancaster.ac.uk	adrian.r.fish@gmail.com
4	jonespm@umich.edu	matthew@longsight.com
5	sinouvivian@foothill.edu	sinou@etudes.org
6	sinouvivie@foothill.edu	sinou@etudes.org
7	maheshwarirashmi@foothill.edu	rashmi@etudes.org
8	rashmi@foothillglobalaccess.org	rashmi@etudes.org
9	tannirumurthy@fhda.edu	murthy@etudes.org
10	tannirumurthy@foothill.edu	murthy@etudes.org
11	mallikamt@foothillglobalaccess.org	mallika@etudes.org
12	thoppaymallika@fhda.edu	mallika@etudes.org
13	ggolden@umich.edu	ggolden22@mac.com...
14	aaronz@vt.edu	azeckoski@unicon.net
15	clayf@bu.edu	clay.fenlason@gatech.edu
16	nangell@rsmart.com	nate.angell@rsmart.com
17	anthony.atkins@vt.edu	tony.atkins@uhi.ac.uk
18	smkeesle@syr.edu	sean.keesler@threecanoes.com
19	carl.hall@et.gatech.edu	carl@hallwaytech.com
20	carl.hall@gatech.edu	carl@hallwaytech.com
21	njbotime@svsu.edu	botimer@umich.edu
22	jbush@rsmart.com	john.bush@rsmart.com
23	duffy@email.arizona.edu	duffy@rsmart.com

I then run the gbasic.py to calculate histogram data on the retrieved email messages. I computed the top 25 email list participants and the top 25 email list organizations.

```
Top 25 Email list organizations
umich.edu 532
indiana.edu 179
mac.com 174
berkeley.edu 174
cam.ac.uk 170
uct.ac.za 123
ucdavis.edu 121
gmail.com 109
yale.edu 105
unicon.net 104
ufp.pt 98
stanford.edu 82
etudes.org 80
columbia.edu 72
virginia.edu 65
mtu.edu 53
gatech.edu 53
rsmart.com 49
earthlink.net 46
rutgers.edu 45
ucmerced.edu 41
aeroplanesoftware.com 37
northwestern.edu 35
unl.edu 35
hull.ac.uk 33
```

```
Top 25 Email list participants
csev@umich.edu 257
ggolden22@mac.com 174
azeckoski@unicon.net 87
nuno@ufp.pt 85
ian@cam.ac.uk 83
jholtzman@berkeley.edu 78
marquard@uct.ac.za 73
slt@columbia.edu 72
ray@berkeley.edu 70
ajpoland@indiana.edu 67
cheryl.wogahn@yale.edu 64
jimeng@umich.edu 64
lance@indiana.edu 62
sinou@etudes.org 57
jleasia@umich.edu 56
swgithen@mtu.edu 53
clay.fenlason@gatech.edu 53
dhorwitz@uct.ac.za 50
markjnorton@earthlink.net 46
ys2n@virginia.edu 45
tpamsler@ucdavis.edu 42
vrajgopalan@ucmerced.edu 39
dave.ross@gmail.com 37
zach@aeroplanesoftware.com 37
jpgorrono@ucdavis.edu 36
```

Next, I showcase the most common words used in the retrieved email through gword.py and visualize it with gword.htm.

At the very end, I produce a timeline visualization of the emails retrieved through running the gline.py and visualize it with gline.htm.