Mailing Lists - Gmane

- Crawl the archive of a mailing list
- Do some analysis / cleanup
- Visualize the data as word cloud and lines

http://www.py4e.com/code3/gmane.zip

Warning: This Dataset is > 1GB

- Sadly the original source of this data (gmane.org) has shut down
- We made a copy of a subset of the data before it was shut down

Use this for your testing:

https://mbox.dr-chuck.net/sakai.devel/4/5

gmane.py

content.sqlite

gmodel.py

content.sqlite

gbasic.py

gline.py

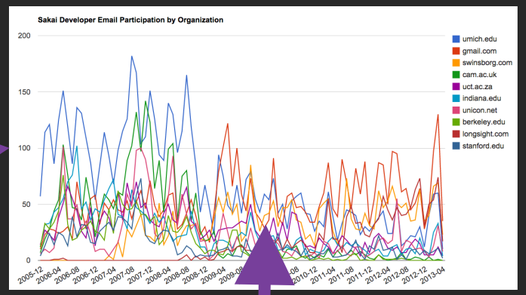
gword.py

gword.js

gword.htm
d3.js

How many to dump? 5
Loaded messages= 51330 subjects= 25033 senders= 1584
Top 5 Email list participants
steve.swinsburg@gmail.com 2657
azeckoski@unicon.net 1742
ieb@tfd.co.uk 1591
csev@umich.edu 1304
david.horwitz@uct.ac.za 1184
...

gline.js



http://www.py4e.com/code3/gmane.zip

gline.htm
d3.js

Summary

- Combining all of the techniques in the class, we can make data management system that can pull data (restarting as necessary), clean it up, and visualize it

Acknowledgements / Contributions

Initial Development: Charles Severance, University of Michigan School of Information

… Insert new Contributors here

...