Gmane: Mailing list

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

The Dataset is > 1GB

large → takes mins to read

not normalized raw, just for spider → cleans up the og outdated db

(Mbox, dr-chuck.net) → [gmane.py] → Content.sqlite gmodel.py

pulling data from API
to db   resumable

[gbasic.py]

dump db analysis
text info
contents

Content.sqlite ← small

not normalized raw

gworldify → gword.js   gword, http d3.js

[gline.py]

(gline.js)

Visualization kws cloud

[Visual] ↑ gline.htm d3.js

gmane.py $\longrightarrow$ content.sqlite   1GB

mapping.sqlite $\rightarrow$ gmodel.py   clean up $\rightarrow$ gword.js
                                            few min to
gmodel read the mapping   run $\searrow$ htu
                            index.sqlite $-$ gword.py

gbasic.py   gline.py

                             gline.js

                                                    Visual
                                                    graph
                                                      ↑
                                                gline.htm
                                                    d3.js