Spidering and Modeling Email Data – Introduction

This week we do the first half of a project to download, process, and visualize an email corpus from the Sakai open source project from 2004-2011:

<http://mbox.dr-chuck.net/>

This is a large amount of data and requires significant cleanup to make sense of the data before we visualize it.

**Important:** You do not have to download all of the data to complete this project. Depending on your Internet connection, downloading nearly a gigabyte of data might be impossible. All we want to do is to have you download a small subset of the data and run the steps to process the data.

Here is the software we will be using to retrieve and process the email data:

<https://www.py4e.com/code3/gmane.zip>

If you have a fast network connection with no bandwidth charge - you can download all the data. If you try to download all the data it may take well over 24 hours to pull the data. The good news is that because there are separate crawl, clean, model, and visualization steps, you can start and stop the crawl process as often as you like and run the other processes on the data that has been downloaded so far.

