Mailboxes - basics

Mail boxes as "social" data, Chapter 3 of Mining the Social Web

Outline

- Email Basics
- CouchDB
- MapReduce Basics

email Basics – The old days

- Mail used to be very simple
  - Sort of "all in one"
  - "mail" command
    - View list of received, view text of mail
    - Line editor to create a message
  - Sending – was on the same machine with "sendmail"
  - Messages were stored in one long file the "mbox"
    - A text file, one message after another

email Today

- Multiple Mail "Clients"
  - Outlook, Apple Mail, Thunderbird, Web based mail
  - Client oriented protocols POP, IMAP
  - Every client stores mail differently
- Different mail Gateways
  - SMTP, authenticated SMTP
  - Some still use sendmail

email – A social media

- email is probably the earliest social media
  - Send, receive, reply
  - Threading
  - Distribution lists (one to many)
- Lot's of research has been done on corporate email repositories

Converting to mbox format

- Outlook
  - There is an "Export" option
- Thunderbird
  - Actually is mbox already
- Apple Mail
  - Save As seems to work

Manipulating mbox in Python

- Book covers some good examples
- Basic conversion to JSON

CouchDB

- CouchDB is probably the main point of this Chapter
- A highly parallel Key:Value store
  - Value is always a "document" – best if it is JSON
  - Often think one key, one value – but this can be a many to one – multiple keys, one value
- Map/Reduce capable

Basic idea Map/Reduce

- Provides simple parallelization
- Map function
  - For each record that meets some set of conditions, evaluate the map() function against the record
- Reduce
  - Collect each of the map() results
- Limits
  - Number of compute nodes
  - Distribution of the data over which the map() function operates

Why?

- Class project
  - You'll propose a project in a couple weeks
  - CouchDB might be a good solution for an analysis
  - Twitter – Keys, various parts of the twitter data
    - messages
      - tweeter, retweeter, date, #hashtags
    - people
      - Followers
  - Wikidump
    - To be honest, our small extraction probably too large to make CouchDB a good solution for a wikidump analysis