**Python for Everyone: Notes**

**Chapter 16:** Retrieving and Visualizing Data

* Personal Data mining
  + Our goal is to make you better programmers – not to make you data mining experts
* GeoData
  + Makes a google map from user entered data
  + Uses the google geodata api
  + Caches data in a database to avoid rate limiting and allow restarting
  + Visualizied in a browser using the google maps api
* Page rank
  + Write a simple web page crawler
  + Compute a simple version of google’s page rank algorithm
  + Visualize the resulting network
* Search engine architecture
  + Web crawling
  + Index building
  + Searching
* Web crawler
  + A web crawler is a computer program that browses the world wide web in a methodical, automated manner. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search enginge that will index the download pages to provide fast search
* Web crawler
  + Retrieve a page
  + Look through the page for links
  + Add the links to a list of “to be retrieved” sites
  + Repeat
* Web crawling policy
  + A selection policy that states which page to download
  + A revist policy that states when to check for changes to the pages
  + A politenesss policy that states how to avoid overloading the website
  + A parallelization policy that states how to coordinate distributed web crawlers
* Robots.txt
  + A way for a website to communicate with web crawlers
  + An informal and voluntary standard
  + Sometimes folks make a spider trap to catch bad spiders
* Google architecture
  + Web crawling
  + Index building
  + Searching
* Search indexing
  + Search engine indexing collects, parses, and stores data to facilitate fast and accurate information retrieval. The purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. Without an index, the search engine would scan every document in the corpus, which would require considerable time and computing power
* Mailing list – gmane
  + Crawl the archive of a mailing list
  + Do some analysis / clean up
  + Visualize the data as word cloud and lines