Parker Moore
INFO 4300
Assignment 3
Report

As with most of my programming nowadays, I took an object-oriented approach when designing the implementation of the assignment. Beyond the normal data structures (str, int, float, dict, list, etc), I also created three classes: Page, Brain, and Searchr.

Page contained all the logic for dealing with a single page: generate a unique ID (md5 hash of the prettified HTML), generate output for search results and metadata XML file, and manage all the data which surrounds a page. Brain contains all the logic for parsing the file of URLs, grabbing the HTML from the pages and dealing with that and performing the PageRank algorithm. Searchr contains the logic for building the index based on the Brain and handling queries.

The PageRank implemented was that which was described in lecture, whereby the following expression is calculated over and over until it converges to the principal eigenvector (with d=0.85):

$$\boldsymbol{w}_k = dB\boldsymbol{w}_{k\text{-}1} + (1 - d)\boldsymbol{z}_0$$

The convergence of this expression leads to the principal eigenvector which, at each index, contains the PageRank value for the page which corresponds to that index. The program then goes through this eigenvector and assigns each page its rank. The Searchr then builds the index based on the anchor text and the PageRank value associated with each page and, when a query is input, grabs any page which was linked to with anchor text which contained that any of the query terms, sorts them based on PageRank values from greatest to smallest and outputs the appropriate search output.

Usage

To use the program, I’ve included a Rakefile for convenience. Just run rake to run the program normally. If rake is not installed, get the program running like so:

```
# install dependencies
easy_install lxml
easy_install progressbar
# run the program
python gargle.py
```

Optionally pass in the query or URL file arguments:

```
python gargle.py -u urls.txt
python gargle.py -q my query string
```

Or, combine them (-q must be last):

```
python gargle.py -u urls.txt -q my query string
```