Readme

Hanxiong Wu

wh854,N16030110

1. Run environment

* Python 2.7.8
* Standard Python library
* Linux

1. Developing & testing environment

* OS: Ubuntu 14.04 32bit (on VMware workstation)
* Memory: 1GB
* Disk: 20GB
* Processor: 1(Intel Core i7-4510U)@2.00GHz-2.60GHz

1. Input format

Place the uncompressed data files to be indexed in ‘data’ directory under the path of all the codes. Also, please create an ‘final’ directory for output.

You can just use the file structure my attachment provides, it’s already valid.

1. Output
2. Indexer.py:

This program will read the data files and generate a list of postings as file ‘index.txt’, with the DocID-URL table ‘docID’ in directory ‘final’. Actually, it combine the posting with same word into an inverted list in some degree: that is, the result is roughly an inverted index lists but may contains a lot of duplicate lists. Those duplicate lists can be merged by running ‘sort.py’ program.

This program runs about 5min on NZ2 with my environment.

1. sort.py

This program will read the ‘index.txt’ generated by ‘Indexer.py’ program, then call the UNIX sort method to sort the index, and go through the sorted index to merge the duplicate lists(they are adjacent so I don’t need to read the whole file into memory).

Finally it generate inverted index ‘index’ and lexicon table ‘lexicon’ in directory ‘final’.

This program runs about 0.6min on NZ2 with my environment.

1. test.py

This program is a simple test to detect if lexicon table and inverted index are matched. You DON’T need to run it to get the index files required.