**Part 2: Accessing the index**

1. Determine the total number of documents in the corpus. You may find dump-index-manifest useful. Beware, this takes as an argument not the index directory, but the corpus subdirectory within the index directory.

{

"blockCount" : 229,

"blockSize" : 512,

"emptyIndexFile" : false,

"filename" : "/u/riker/u96/xu931/./project1-index/corpus",

"keyCount" : 3204,

"maxKeySize" : 512,

"mergerClass" : "org.lemurproject.galago.core.index.merge.CorpusMerger",

"readerClass" : "org.lemurproject.galago.core.index.corpus.CorpusReader",

"tokenizer" : {

"fields" : [ ],

"formats" : {

}

},

"writerClass" : "org.lemurproject.galago.core.index.corpus.CorpusFolderWriter"

}

The total number of documents in the corpus is 3204.

1. Determine the number of documents containing the word `retrieval'.

132 retrieval

The number of documents containing the word ‘retrieval’ is 132.

1. List the documents containing the word `Rice'.

unk-0 Q0 /homes/cs473/project1/cacm/CACM-1677.html 1 -6.73131542 galago

unk-0 Q0 /homes/cs473/project1/cacm/CACM-0302.html 2 -7.31091066 galago

unk-0 Q0 /homes/cs473/project1/cacm/CACM-1312.html 3 -7.32263718 galago

unk-0 Q0 /homes/cs473/project1/cacm/CACM-2312.html 4 -7.33037915 galago

1. List the top 5 documents returned for the queries `information retrieval' and `machine learning'. Report the documents for the given queries. Note that you'll need to transform queries into json format to use batch-search. You may find the galago query-transform command useful

**Top 5 documents returned for “information retrieval”**

unk-0 Q0 /homes/cs473/project1/cacm/CACM-3134.html 1 -5.27224741 galago

unk-0 Q0 /homes/cs473/project1/cacm/CACM-1699.html 2 -5.32447710 galago

unk-0 Q0 /homes/cs473/project1/cacm/CACM-2288.html 3 -5.45138068 galago

unk-0 Q0 /homes/cs473/project1/cacm/CACM-1681.html 4 -5.51583323 galago

unk-0 Q0 /homes/cs473/project1/cacm/CACM-0891.html 5 -5.60211040 galago

**Top 5 documents returned for “machine learning”**

unk-0 Q0 /homes/cs473/project1/cacm/CACM-2979.html 1 -5.82114865 galago

unk-0 Q0 /homes/cs473/project1/cacm/CACM-2217.html 2 -6.06939866 galago

unk-0 Q0 /homes/cs473/project1/cacm/CACM-2471.html 3 -6.09832118 galago

unk-0 Q0 /homes/cs473/project1/cacm/CACM-1699.html 4 -6.34474011 galago

unk-0 Q0 /homes/cs473/project1/cacm/CACM-1170.html 5 -6.50645165 galago

**Part 3: Evaluation (15%)**

num\_ret 19 1000.00000

num\_rel 19 11.00000

num\_rel\_ret 19 9.00000

P20 19 0.30000

num\_ret 50 NaN

num\_rel 50 NaN

num\_rel\_ret 50 NaN

P20 50 NaN

**Part 4: Evaluation on base.out and estimate.out**

**Base.out**

num\_ret 19 928.00000

num\_rel 19 11.00000

num\_rel\_ret 19 0.00000

P20 19 0.00000

**Estimate.out**

num\_ret 19 1331.00000

num\_rel 19 11.00000

num\_rel\_ret 19 7.00000

P20 19 0.00000

I think estimate is more accurate because the estimate takes the relevant file into account. On the other hand, the method in base.out file assumes that the pi is 0.5, which is not always the case. Therefore, I think estimate method is better.

**Part 5: Other Things to Consider**

When we use time command to run the BIM file, I find out that the real time is less than the system time and user time. Galago batch-search is essentially search every document. If it is not multi-threaded, the program would not finish in a short time because there are totally 3104 documents in the corpus.