IRE Mini Project Phase 1

Roll no: 2019114006

Name: Suyash Vardhan Mathur

Running Instructions

Run the command below to index

```
bash index.sh <path_to_wiki_dump> <path_to_inverted_index> invertedindex_stat.txt
```

Files created

First, a temporary index is created for **each fields**. After this, these temporary index files are combined into one, such that all are sorted alphabetically. The format of the final index is:

```
<word1> <document_id1>:<frequency1> <document_id2>:<frequency2>
```

Further, vocabulary files are also created. These contain each token, and with it, frequency of that token in all docs of a particular field type, as well as the file number of the index file which contains the document information for that particular token. The format for the vocabulary files is:

```
<word1> <field_id1>-<frequency_in_all_fieldid_docs>-<file_number_for_indexed_data>
```

Also, a document ID to document title mapping exists. This is also sorted based upon the IDs, which are encoded in 64-base encoding. The format for this Document mapping is:

```
<document-ID1> <Document Title>
```

The merging of the intermediate indexing to final index is done using K-way merging through Heaps.

Benchmarking

I ran the exact submitted code on ADA using 8 CPUs and 0 GPUs on gnode042. The total runtime is below:

```
suyash.mathur@gnode042:/scratch/suyash.mathur/Wikipedia-Search-Engine$ time bash index.sh ../enwiki-20220720-pages-articles-multistream15.xml-p1582
4603p17324602 dir stat_file.txt

real    5m57.361s
user    5m55.311s
sys     0m0.920s
```

The total index-size, including all Document ID hashing and vocabulary files, comes out to **237492** which is much lesser 1/4th of the given file size. The stat files can be created, and give output:

```
2064728
1580369
```