Salvatore Avena, Rohan Patel

User Interface

Command line interface that will allow a user to query keywords against an inverted index or add a document to the index. When starting the program for the first time, the inverted index will be generated as described below and saved to file. The user will then be repeatedly prompted with a menu with the following options: to search/query; to add a document; to quit. The program will only terminate once the user has chosen to quit.

Inverted Index

Upon running the program, an inverted index will be generated off of a preselected input directory. This will occur before the user is prompted with the menu. This inverted index will contain a series of entries containing the search term and a posting list. This posting list will contain a linked list of documents; each document will contain the document ID for the page it is referencing and a payload containing the frequency of the term (number of times that term occurs in the document).

When the inverted index is generated, all terms will be preprocessed before searching. This preprocessing will involve removing stopwords (i.e. the, is, in, of, etc.), converting all letters to lowercase, end stemming the words (i.e. words → word, searching → search, etc.).

This inverted index will be stored to disk. Upon running the program, it will first search a specified file path to see if it contains the inverted index file, if not it will generate the inverted index as specified above.

If a user elects to add a document, the corresponding information will be added to the index and an updated index will be saved to file.

Rank and Retrieval

When a set of keywords are entered by the user, they will be preprocessed before searching in the same way as above.

After searching the index for a set of search terms, the program will take all the inverted index entries containing those terms and run a MapReduce job on them to arrange the documents in order from highest to lowest relevance.