USER MANUAL

The Habanero Search Engine uses the pugixml parser to parse xml files into an inverted file index that is stored in either an AVL Tree or Hash Table data structure. Once an index is built, it can be saved to a persistent file. This search engine uses a term frequency-inverse document frequency algorithm for ranking search returns on properly formatted Boolean queries. After the returns are displayed, the user may choose a document to view and the contents of the page will be displayed.

User Interface:

The search engine has two modes. The first mode is Maintenance Mode which allows for adding or removing data from the index. The second mode is Interactive Mode which allows for searching over the index and printing statistics about the index.

Maintenance Mode:

The Maintenance Mode allows you to choose a data structure type for the index. You can choose either an AVL Tree or a Hash Table that uses separate chaining. You can then choose to load a persistent index. Then you can choose to either add to the index or to purge the index. You can add to the index from either a single file or from a directory of files. Purging the index deletes the existing index structure.

Interactive Mode:

The Interactive Mode allows you to either view statistics about the index or to search the index. The statistics option displays the number of pages read by the index, the number of words in the index, and the 50 most frequent words. If you choose search, you will be prompted to enter a search.

A search query can contain any string of words. All words following a Boolean expression will be processed by that expression. When a new Boolean expression appears in the query, the words following it will be handled under that expression. The Boolean expressions that can be used are “AND”, “OR”, and “NOT”. “AND” will return documents that contain the words following the “AND”. “OR” will return documents that contains either or both of the word following it. And “NOT” will return documents that do not contain the words following it.

After the list of relevant documents is printed, the Document ID number associated with a returned document can be entered to view the corresponding page.