Appendix C

Program Development Log

Sunday 22/3/98

Created main application class, *IndexerApp*, as well as *HTMLFile*, the original Index class, and *PageArray*.

Monday 23/3/98

Created the *PageReference* class.

Tuesday 24/3/98

Created the *Globals* class.

Wednesday 25/3/98

Created the *BadHTMLFileException* class.

Saturday 28/3/98

Created the *HTMLcode* class to ease the addition of different tags to the program for recognition.

Tuesday 14/4/98

Added *BadPhraseException* and *CouldNotStartException* classes.

Even if the user wants a Z-A sort done, page references will still be sorted in an ascending manner.

Made "of" a bad prefix.

Now the program doesn't fail when the *DateFormatSymbols* class is unavailable.

Wednesday 15/4/98

Removing Stop Words from the start or end of a phrase may leave trailing spaces. These are now removed.

Friday 17/4/98

Because the phrase buffer isn't initiated when left tag brackets are tested as a word break (because the 'full' flag may be off), "Egg <H1>" would produce two entries for the word "Egg". This is fixed.

Monday 20/4/98

Can now use a Key Word list. Necessitated extensive changes to Index construction procedure.

Tuesday 21/4/98

Added ability to connect to URLs and read therefrom, and a preference to turn this feature on or off. Stopped specifying a font in the output of index entries in the interests of inter-platform compatibility. Made colours 'web-safe'.

Wednesday 22/4/98

Can now access relative URLs.

Thursday 23/4/98

Added a preference to stop following remote links beyond a certain point, in order to prevent indexing of the entire network.

Friday 10/7/98

Used to be one big index, with all entries added to it, which is then sorted. Now, each document has its own index, and when the document is finished with, the index is merged with the complete one which is built up as each file is used. Then it is sorted. The same total index is produced, but the individualistic approach allows us to use some of the algorithms which use relative frequencies in differing documents to determine importance. Better encapsulation too.

Added *Sparck Jones* and *Salton* weights.

Verified that the program will work when the start file is a URL of a remote file.

After this run, based upon the results, I added five more 'noise words': "Go", "Inc", "Sorry", "Thank", and "URL", since these words cropped up a lot but are of little value.

Tuesday 14/7/98

Identified program bugs and more stopwords to add.

Saw Supervisor, he mentioned adding ability to stop searching beyond a certain host, I mentioned adding GUI.

Wednesday 15/7/98

Identified more bugs and stopwords to add.

Saturday 18/7/98

Added new stopwords, now almost 250 entries.

Added GUI, but it's not done yet.

Handle special characters: &quot, and particularly &amp. Necessitated considerable additions to the main parsing routine.

Now two letter words are only allowed if they're both capitals.
Number prefixes banned, "One" but not "First".

Sunday 19/7/98

Comparing analysis times with and without the stoplist revealed that as the stoplist has grown, the gap has widened dramatically. Using a sample 65,000 word document, the analysis times were 295,705 msecs without the stoplist, but 420,523 msecs with the list. The effect of the change has been very great, on a later run, the times recorded were 260,019 and 222,743 respectively. As a result of the change, analysis times with the stoplist have been cut by 47%, and even without the stoplist, times have been cut by 12%.
Program now recognises anchors.

Wednesday 22/7/98

Added ability to stop searching beyond a certain host. Removed stop list from Globals, giving it a class of its own, to avoid having to recreate a new Globals before each run.

Friday 24/7/98

Now, all weights are calculated before any entry is output. Allows us to restrict which entries are output. Added options for this.
Added *Constants* interface to centralise final variables.

Saturday 25/7/98

Fixed word list. Since we changed to individual indexes which are then merged, we forgot to initialise each new index to the word list one. As a result they started out blank and using the word list would result in **zero** entries found.
Other changes, gave IndexerApp routines better names.

Sunday 26/7/98

Got a better special characters list from Netscape IFC files.
Saw Supervisor. Showed him draft version of dissertation. Demonstrated program on Windows NT PC. Needed to improve GUI.

Sunday 2/8/98

Sorted out GUI.
Fixed a bug where we overran the special chars list.
Fixed a bug with strings containing only quotes.

Monday 3/8/98

Added superfluous options for "individual words" and also implemented a

"frequency method".

Tuesday 4/8/98

Moved the Sort routines into a separate class, *Sorter*.

Monday 10/8/98

It's now ok to have special characters as letter headings in the index file.

Tuesday 18/8/98

Saw Supervisor. Got program to work on Unix environment. Discussed comments on draft copy and need to put more emphasis on colour issues/HCI. Considered possibility of distributing sample web page and asking users to pick out the ***n*** most important words.

Added "@" to the list of word break characters.

"New\rZealand" is now correctly converted to "New Zealand" and indexed as a single phrase rather than as two. As is the following: "New \r Zealand".

Program no longer writes out '\r' characters but always calls the writer.newLine() method to avoid problems on PCs.

Monday 24/8/98

Program will now recognise the initials in proper names as part of the name and not part of a new sentence, so "Andrew G.D. Regan" will be indexed as one entry.

Stop Word list now just over 400 entries.

Added the ability to recognise prefixes and first names, so that "Andrew G.D. Regan" would appear as "Regan, Andrew G.D."

Tuesday 25/8/98

Created the *Prefixes* class to store the first names and prefixes to avoid overburdening the *CompleteIndexTable* class.