

# DMS8002

## Project Proposal

Pete Hindle

### Aim:

To produce a program that will be able to sort through files in a directory and display the information within them in alternate manners

### Reason:

My research within other modules of this course is so far producing a lot of text documents. I would like to create a helper program for the process of examining those documents to help me see the information within them; possibly in new ways, or possibly as ways of examining and representing the data.

This second definition - examining and representing the data - is close to the idea of information visualisation, except that unlike the main aims of that field this project does not seek to re-present the data in the style of graphic design. This might be possible later, but the prime aim of the project should be to create a program that can take input from textual sources, make some sort of comparison, and then output data based on that comparison.

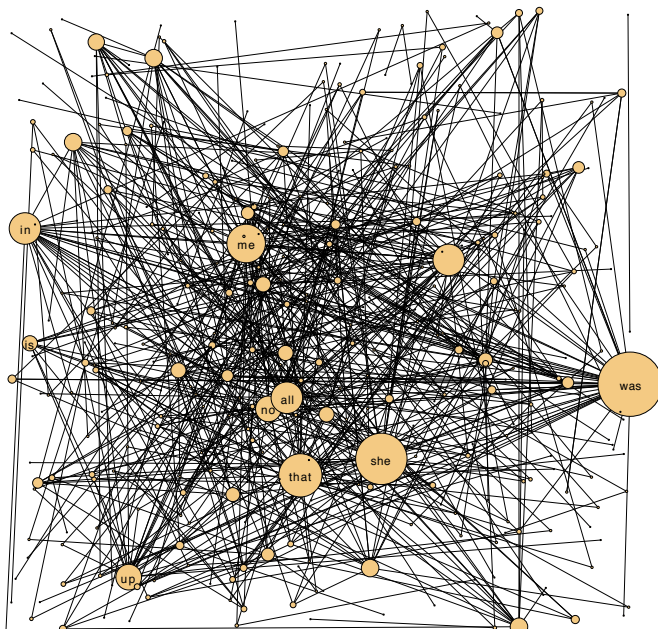
### Technology:

I'll be writing this program using Processing, but not limiting it to that language.

This is because the methods for searching and compiling all the text files within a directory, or even combining all the words within those text files, might be best done by using shell commands or similar. As I'm only intending to have a working prototype for use on my computers, I won't be looking at any Windows-based solutions.

The output of the program will also be easily accessible as a data-stream, and that means that there would be scope - if time allows - to create an Arduino program that can demonstrate the datafield.

### Precursors:

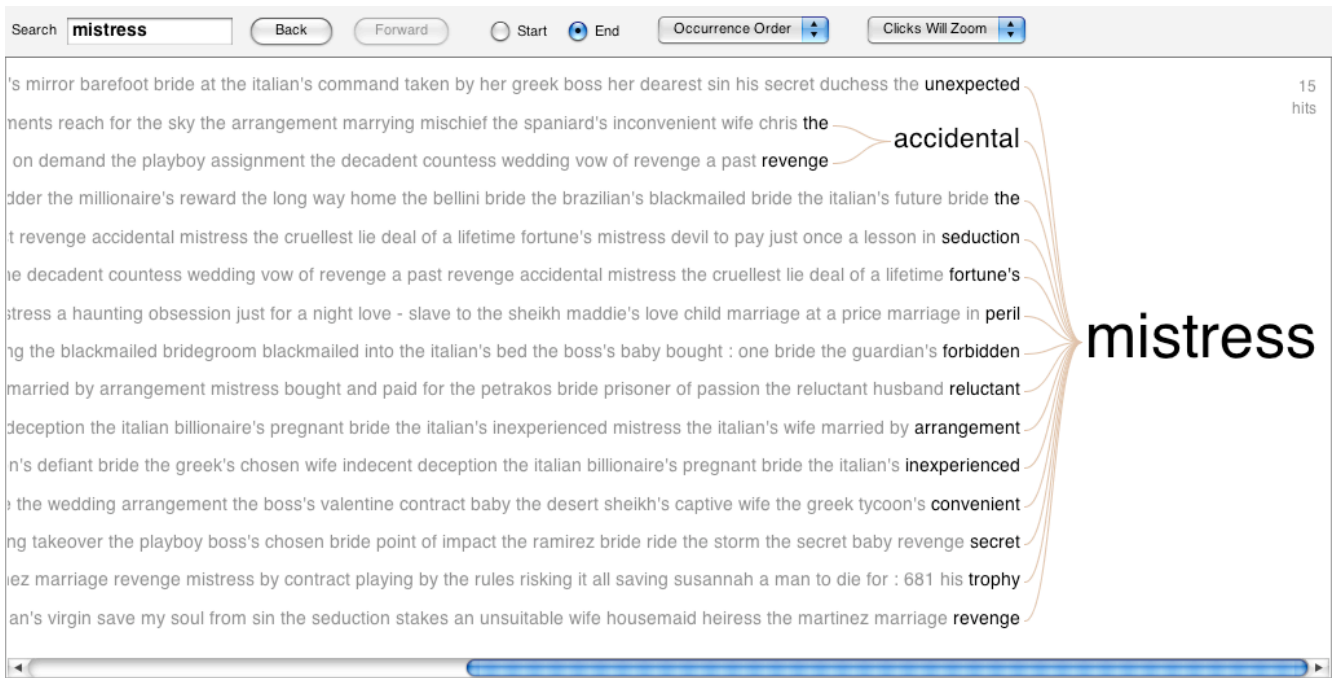


Most importantly for my work within this module, Ben Fry has written up a selection of his techniques within the book "Visualizing Data". The image on the left is taken from one of the examples in his book.

This book also covers in depth the subject of gleaming information from different sources, such as API's, and is extended by support on Fry's website and the main Processing forums.

The core idea of the project is not so

different from many simple visualisation projects, including Wordle, which is now hosted on IBM's ManyEyes site. However, Wordle - and other text-based projects hosted on the ManyEyes site - are not quite as useful for my purposes as they don't allow users to fully



“drill down” into the data.

For instance, the above ManyEyes visualisation of over two hundred Mills and Boon book titles does not allow the user to separate titles. Unlike Wordle-produced diagrams, however, it does allow you to see how many time your chosen word has produced a ‘hit’.

## Planning

Logic Flow:

The logical flow of the program will be divided into three parts:

- Text file finder
- Text analysis
- Output stream

Text file finder logic:

Crawl directories set

Locate text files

Gather files and combine (combining files might be an option)

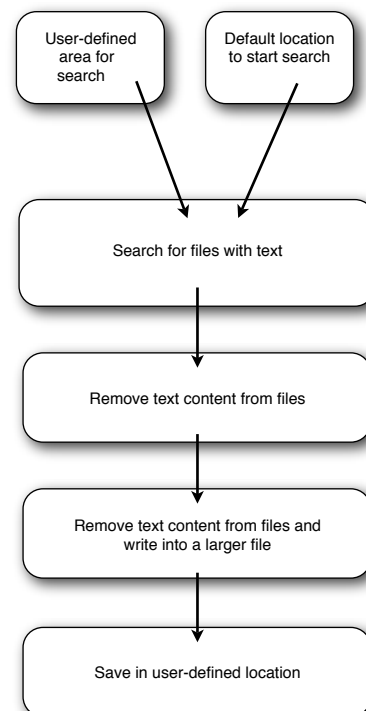
repackage files and save them for next stage

Text analysis logic:

Open and read file

Do some or all of these actions

- count words



- remove joining words
- remove optional words
- find links between words
- count web links

Produce data file which recreates data according to selected criteria

Output stream logic

Open data file

Show data file in smaller chunks

