Week 2 Assignment: Indexing a file

The aim of this exercise is to index a text file, by line number. We can think of the input being a list of text strings, and below we've provided an outline Erlang module that reads text files into this format, as well as a couple of example files to process.

In solving this problem you'll need to think about the different stages of processing of the data: you begin with a list of lines, each of which will need to be broken into words, and those lines (and words) will need to be associated with the corresponding line numbers. So, thinking about useful intermediate stages –and helper functions –should help you to make progress in solving the problem.

The output of the main function should be a list of entries consisting of a word and a list of the ranges of lines on which it occurs.

For example, the entry

```
{ "foo" , [{3,5},{7,7},{11,13}] }
```

means that the word "foo" occurs on lines 3, 4, 5, 7, 11, 12 and 13 in the file.

To take the problem further, you might like to think about these ways of refining the solution.

- Removing all short words (e.g. words of length less than 3) or all common words (you'll have to think about how to define these).
- Sorting the output so that the words occur in lexicographic order.
- Normalising the words so that capitalised ("Foo") and non-capitalised versions ("foo") of a word are identified.
- Normalising so that common endings, plurals, etc. are identified.
- (Harder) Thinking how you could make the data representation more efficient than the one you first chose. This might be efficient for lookup only, or for both creation and lookup.
- Can you think of other ways that you might extend your solution?

ASSIGNMENT GUIDELINES

The reviewers will be asked to give you feedback on the following aspects of your assignment, so you should consider these when writing:

- Document the solution: make sure that each function has a clearly identified purpose, and that, when appropriate, test data is provided to illustrate its purpose.
- If submitting a partial solution, make clear what has and hasn't been solved, to provide a clear picture of the scope of the solution.
- If taking the problem further, explain what has been done; otherwise, comment on how the remaining parts of the problem might be approached.