Assessment 1

Important dates

- Assignment start: Monday, January 30th, 14:00
- Assignment due: Monday, February 27th, 23:59

Overview

In this assignment you will implement a set ADT. Specifically, you will implement what is known as an *ordered* set, since the set relies on an ordering of elements and supports iterating over its elements in that order. The operations to be supported by the set are:

- Adding an element to the set.
- Getting the current size of the set.
- Checking whether a specific element is contained in the set.
- Getting the union of the set and another set.
- Getting the intersection of the set and another set.
- Getting the set difference between the set and another set. (Also known as the relative complement set).
- Iterating over the elements of the set, in sorted order.

There should be no bound on the number of elements that may be inserted into the set, short of running out of memory.

Algorithms

This problem has two parts: implementation and evaluation. Note that this project requires quite a lot of work: start early.

Implementation: You should make **two separate implementations** of the specified ADT. We recommend that you start with a list-based implementation where you represent sets as sorted lists. This has the advantage of optimal complexity for iteration (as for *why* it's optimal, that should go into your report) but has worse complexity for adding elements than many other alternatives. We have provided a working list implementation. Use this; do not waste time writing your own.

The second implementation is your choice. One possibility would be to do a search-tree-based solution. Whatever you choose, you need to support iteration over the elements in sorted order. This suggests that the algorithms for set unions, intersections, and differences could be the same regardless of your representation: All of these can be implemented as variations of the basic merge algorithm. All in all you should emphasize writing a good report; you will get little or no credit for sophisticated algorithms unless they are documented in your report. Starting out with something simple and refining it if you have the time is probably a good idea.

Evaluation: We want you to do an experimental evaluation to compare the two approaches: measure the time that it takes to do the various operations under each solution for various dataset sizes. For instance, how long does it take to add 100 elements, on average? How about 1000?

How about 10000? Be ambitious. How long does set unions take for datasets of various size? Etc, etc. Note that you need to generate some test data for this.

Your evaluation goes in the report. Show figures and numbers illustrating the strengths and weaknesses of the two implementations.

Applications

The code comes with a complete test application that calculates various sets of integers (even numbers, odd numbers, primes and non-primes). This application, called *numbers*, may be used for testing (although you probably want to write some test code of your own, too). For a correct implementation of the set ADT, the output of the numbers application looks like the contents of the file *numbers-expected.txt* (located in the zip file with the assignment code). The application also serves as an example that illustrates how the ADT interface is used. Look here for examples of iteration, insertion, unions, etc.

You will also complete the implementation of a second application, *spamfilter*, which is to be a very simple and naive application for classifying e-mails (text files actually, but let's pretend they are e-mails) as spam or non-spam.

The algorithm used by the spamfilter application should be as follows. It starts out with a number of e-mails known to be spam, and a number of e-mails known NOT to be spam. These are pre-classified examples from which the algorithm will attempt to classify the remaining e-mails. The rule for classifying an e-mail as spam is this: If the e-mail contains a word that occurs in ALL of the spam e-mails and NONE of the non-spam e-mails, then it is spam; otherwise it is not spam. To phrase this in terms of sets, assume that each e-mail has been tokenized into a set of words. (Use the tokenize() function in spamfilter.c for this). Let S1,S2, ..., Sn be the spam word sets, and N1, N2, ..., Nm be the non-spam word sets. An e-mail M is then to be classified as spam if and only if:

$M \cap ((S1 \cap S2 \cap ... \cap Sn) - (N1 \cup N2 \cup ... \cup Nm)) \neq \emptyset$

The spamfilter application should accept three directory names on the command-line, specifying where to find the spam files, the non-spam files, and the files to be classified, respectively. So, assuming these files are to be found in the directories spamfiles, nonspamfiles, and mailfiles, the command-line would look like this:

```
./spamfilter spamfiles nonspamfiles mailfiles
```

The zip file with the assignment code also contains a sample data set for the spamfilter application. If you implement the above algorithm correctly, it should make the classifications in the file *spamfilter-expected.txt* (also located in the zip file with the assignment code) for the data set.

Code

Your starting point is the following set of files:

- set.h - Specifies the interface of the set ADT. Do not modify this file.
- list.h- Specifies the interface of the list ADT from the previous assignment.
- linkedlist.c- An implementation of the list ADT based on doubly-linked lists.
- common.h - Defines utility functions for your convenience. Two functions have been added since the first assignment: list_files(), which recursively traverses a directory and returns a list of file names, and compare_strings(), which is a comparison function that may be used with sets of strings and lists of strings.
- common.c- Implements the functions defined in common.h.
- numbers.c - The numbers test program.
- spamfilter.c - A stub for the spamfilter program (complete this).
- assert_set.c - Functionality for testing if your set operations work as expected.
- Makefile - A Makefile for compiling the code.

You need to add new source files that implements your set ADT, and edit the name of this source file into the Makefile, as the value of the SET_SRC variable.

We've bundled the source code in a zip file along with the sample spamfilter data set. The zip file is in the same folder as this document (a1-pre.zip).

Deliverables

There are two deliverables for this assignment: A report and the source code.

The report must be in pdf format. The source code must be placed in a compressed file (zip, rar, tar.gz or tar.bz2).