CS1 5180: Assignment # 3
Handed Out on: Thursday October 30, 2003
Due Date: Thursday November 13, 2003

1 Written Exercises

1. Problem 7.8, p. 262.
2. Problem 8.2 p. 277.
3. Problem 8.14 p. 294.
4. Problems 8.21 and 8.22 p. 307 [Please, don't bother looking at the suggested references unless you are interested].

2 Programming Exercise

This exercise is meant to familiarize you with a topic we will not be covering in class: Clustering (or unsupervised learning). As well, it will help answer one of my questions: Can the citations/references occuring in research papers be clustered into different groups according to the words occurring in the text these citations are embedded in?

The idea is that citations/references may play different roles within a paper (e.g., they may refer to similar work, or the tools used in the research, or they may refer to a more general category of work, etc.). If such information could be captured, tools that use citation graphs to infer similarity between scientific documents could be refined using this information.

The papers you will be working with can be downloaded from: /home/kaml0/usr3/resources/corp/JAIR on one of the school machines. They have already been converted to text format and cleaned.

To answer the question stated above, you will have to:

1. Choose a subset of papers to work with (You don't have to use the entire collection which is quite large)
2. Detect the citations/references in the text.
3. Surround each of these citations/references by a window of 50? 100? 200? 300? words. [Please, experiment with different window sizes].

Also, use an existing Concordancing or KWIC tool to extract these windows or write your own.

4. Form a glossary of all the words (except for stop words) occurring in all the windows.

5. Represent each reference/citation in your data set by a bag of words.

6. Use a clustering algorithm on this data (check the software accompanying Chapter 14 of the text book or use some other software freely available from the Web).

7. Have a look at the various clusters that were formed and discuss whether the groupings make sense to you [Please note that, here, I am not asking you to do a formal evaluation because that would take you too much time since the citations would have to be manually labeled.]