

**CS1 5180: Assignment # 2**  
**Handed Out on: Thursday October 9, 2003**  
**Due Date: Thursday October 30, 2003**

**1 Written Exercises**

1. Problems 5.5 and 5.6 on pp. 176-177.
2. Problems 6.2 and 6.4 on pp. 225-226.

**2 Programming Exercises**

1. Build upon your answer for problem 5.6 p. 177 above to design and implement a collocation discovery tool that allows you to discover collocations in a flexible relationship (i.e., collocations that are not restricted to fixed phrases) and that does take variance into account. Test this tool on the Shakespeare's plays that you already used in Assignment # 1 (i.e., form a large corpus made of the six plays you used and apply your tool to that large corpus). Discuss your results. In particular, have you discovered some old (or, at least, Shakespearean) use of language that is not current anymore?
2. As you know, Shakespeare's true identity is highly debated, and it is not clear whether all the plays attributed to him were indeed all written by him. Basing your study on the six plays that you previously used, implement the Kilgarriff and Rose method briefly discussed on p. 171 of the text and described in more depth in their paper [downloadable from <ftp://ftp.itri.bton.ac.uk/reports/ITRI-98-07.ps>] to find out whether all six plays come from the same source or not. Discuss your results. [Use any available tool that can help you in this endeavour].
3. Problem 6.10 p. 227. In particular, download texts in several European languages from <http://europa.eu.int/> and use Language Identification Tools available from <http://odur.let.rug.nl/vannoord/TextCat/competitors.html>.