CS1 5180: Assignment # 1
Handed Out on: Monday September 22, 2003
Due Date: Thursday October 9, 2003

1 Written Exercises

1. Identify every major phrases (noun phrases, verb phrases, adjective phrases , or adverbial phrases) in the following sentences. For each, indicate the head of the phrase (the central constituent that determines the syntactic character of the phrase) and any complements of the head. Be sure to distinguish between complements and optional modifiers (or adjuncts).

   The man played his fiddle in the street.

   The people dissatisfied with the verdict left the courtroom.

2. Classify the following verbs by specifying what different complement structures they allow, using the forms defined on p. 105. Provide example sentences in each case. Think of all possible uses of each verb.

   give, know, assume, insert

   Give examples of additional complement structures that are allowed by these verbs but not listed on p. 105 (when possible). Provide examples that match these structures.

3. Problems 2.3, 2.4 and 2.5 on pp. 59-60.

4. Problems 2.9 and 2.10 on p. 78.

2 Programming Exercise

Download three of Shakespeare's comedies and three of his tragedies from the Oxford Text Archive at http://ota.ahds.ac.uk. (e.g., Comedies: A midsummer-night's dream, Comedy of Errors, Much ado about nothing; Tragedies: Hamlet, Othello, King Lear; However, feel free to download others if you prefer).

The purpose of this exercise will be to find out whether simple computational tools can help us distinguish between Shakespeare’s tragedies and comedies. As well, this exercise will familiarize you with many of the different issues that arise in Statistical NLP as well as with the tools available for dealing with these problems.

In the Oxford Text Archive, both ASCII and SGML formats are provided. Choose the format that you feel will be more useful and discuss your choice in your report. A lot of the software you need for this exercise can be retrieved from the five ”Useful Links” listed on the course’s homepage, though, if you prefer implementing some of these tools rather than using the available ones, feel free to do so.

Write a report describing the techniques you implemented and mentioning the tools you used (include the source you got them from). In all cases discuss the pros and cons of these methods (in terms of success rate and efficiency).

1. Begin the exercise by eliminating as much as the low-level problems occuring with text as discussed in Section 4.2 of the textbook. This includes stripping the text of its punctuation (note, however, that “.”’s sometimes stand for abbreviations. Rather than removing the “.” in this case, replace the abbreviation by its expanded form (e.g., Ia. ⇒ Iago), if available. Make sure, however, that your program does not wrongly expand words that shouldn’t be expanded), standardizing words’ case (similarly, change the case of a letter, only, if you encounter the same word spelled with that letter in a different case), stemming. Consider and try to take into account the other issues discussed in Section 4.2 of the textbook as well as other situations you may encounter.

2. Compute word frequencies and/or other useful counts such as those described in Section 1.4 of the textbook for each play. Attempt to filter stop words (the, and, it, is, etc.), proper names, and any other such problematic lexemes.

3. Write software that takes as input the six lists of words (one list per play) and their frequencies (or other pertinent counts) generated in in the previous section of the assignment, and outputs

   (a) the list of words that occur more frequently in the three tragedies than in the three comedies and

(b) the list of words that occur more frequently in the three comedies than in the three tragedies.

(Display only the portion of the list which you feel shows some pattern of interest)

4. Are you able to distinguish Shakespeare's comedies from his tragedies using this simple method? Discuss your results commenting on what type of characteristics can be learned from the methods you tried and what type can't. If the current method failed, but you have ideas on what would work, discuss these ideas. For example, you can try to download additional plays and see if your results can be improved. [Demonstrating the value of your ideas (even if they do not fully get you to your goal) will earn you extra credits!]

5. Have Fun!