NLP Final Project
Spring 2008

1 Due dates

- One- to two-page proposal due: Thursday, March 27.
- Progress report due: Thursday, April 10
- Class presentations: April 29, May 1, and May 6th (if necessary)
- Written report due: Tuesday, May 6

2 Introduction

I would like you to create a final project that involves the design and application of a major program or programs that is an example of a natural language processing system. You have a fair amount of flexibility in what you do for your project, but you will need to do some research to find something appropriate.

I can imagine several different sorts of projects that would be suitable for this course. Your project might be software intensive in that it presents a significant piece of code that builds on current knowledge in NLP and linguistics, and goes beyond what we have discussed in class. Another alternative is to do a careful comparative study of the different NLP techniques to solve a problem, presenting computational evidence that will help determine which techniques are the most helpful.

An option that some of you might find interesting would be to add capability to the nltk library. As you may be aware, the source code is available on-line. Of course you may also be more interested in applying NLP techniques to solving applied problems. Feel free to build on existing code, but be sure to give credit when you use the work of others.

I encourage you to work together in groups of two to three to work on this project as you will be able to accomplish more and several eyes on code tend to result in better quality code and fewer errors. (See information on pair programming, for example, at Wikipedia.) If you do work in groups (and I hope you do), I ask you to include with your final write up a description of who did what. My intention is to give all members of a team the same grade, but will provide individual grades if there is evidence that some people did considerably more than others. Of course I expect more ambitious projects when more people are involved.

As usual, your final project should not include work done for another course unless you have permission from both me and the other instructor.

3 Deliverables

3.1 One- to two-page proposal

Your one-page proposal must address the following issues:

1. What is the problem that you will be attacking?
2. Why is this interesting as an NLP problem?
3. What are relevant references to existing approaches to this problem?

4. What technical methods or approaches will you use?

5. On what data will you run your system?

6. How will you evaluate the performance of your system?

The section on evaluation is important. Generally I will expect you to accumulate data on how your program works and use at least some statistical analysis to back up your conclusions. Please come see me if you would like some help in how to do this.

Lots of data is available on the web. We've talked about the Penn Treebank and the Brown corpus. There are many other sites with data available. You will likely find the info at
`http://nlp.stanford.edu/links/statnlp.html` helpful.

I would be happy to discuss your proposal with you in advance. Once you have submitted your proposal and I have approved it, I will consider it a contract between us. Changes in direction and scope will need to be negotiated with me. In some cases we may have to negotiate changes in the proposal before I sign off.

The scale of a project is always difficult to specify. A reasonable target is to assume that each participant will put in about as much effort as on two regular homework sets plus a bit more on the write-up.

In case of a group project, I expect only one proposal to be turned in.

3.2 Progress report

A one- to two-page progress report should be sufficient. Refer to your project proposal and explain what has been accomplished so far and what remains to be done. I will expect a written report and a brief oral report in class by each of the teams.

3.3 Class presentations

Each group will give a formal 30 minute report on their project, to be followed by a group discussion of the project. A successful presentation will discuss the prior work in the area, what has been accomplished (with a demo), and an evaluation of how successful the project was. Typically this will involve gathering data and comparing your results with other techniques for solving similar problems.

3.4 Final report

The final report should include the code developed as an appendix to the report. The code should also be turned in electronically along with instructions on how to use it.

The final report will generally be about ten pages long, including a section on prior work, your approach to the problem, analysis of your data, and conclusions, which likely will include suggestions for further work (e.g., suggested modifications to your approach). Projects by a single person might be a bit shorter, while those from a group might be a bit longer. Think of it as a conference paper that focusses on the research questions and results, with some discussion of the techniques used. You may find it useful to look at some of the research papers at `http://acl.ldc.upenn.edu/`, particularly those from the CoLing conference, as examples.

4 Topics

This is always hard to specify. The best projects are often those that you think up (and care about!) yourselves. A list of projects suggested by others can be found starting on page 3 of `http://www.stanford.edu/class/cs224n/handouts/cs224n-fp.pdf`. (I have found that just about everyone teaching this course points to the same list of projects, so I won't bother to copy them.)

If you are interested in contributing to a home-grown project, Professors Eliott and Valenza of CMC have a long-standing project evaluating the authorship of texts attributed to Shakespeare. Several of their papers are available at `http://www.claremontmckenna.edu/facultysites/govt/FacMember/WElliott/select.htm`. They have a large set of Pascal code that they were running on a VAX. I suspect that a lot of what they do by hand could be run using some standard NLP libraries. Please let me know if you are interested and I can arrange for you to meet with the authors.