224N Projects

Soft due date for project proposals: **Oct 20** but we'll accept them until **Oct 27**.
Report / code due date: 11:59pm **Dec 6**. You may use late days, but we will not accept any submissions after **Dec 10**.

This project is an opportunity for you to work for longer on a bigger NLP system in an area of your choice! The projects will be judged on creativity in defining the problem to be investigated, and the methods used, thoroughness in considering and justifying your design decisions, and quality of your write-up, including your testing of the system and reporting of results. You will not be penalized if your system performs poorly, providing your initial design decisions weren't obviously unjustifiable, and you have made reasonable attempts to analyze why it failed, and to examine how the system might be improved.
The final project can be a group project. Indeed, we'd strongly encourage you to work as a group, so you can attempt something larger and more interesting. The amount of work should be
appropriately scaled to the size of the group (though the expected scaling is sublinear), and you should include a brief statement on the responsibilities of different members of the team. Team members will normally get the same grade, but we reserve the right to differentiate in egregious cases. In general we would like group sizes of 2 or 3. If you are considering a bigger group, you must talk to us and convince us that a group of greater than 3 is manageable given the inherent parallelizability of the task, and the time available to organize and implement the system. Solo projects are allowed. For the final project, group size is considered in the grading, but even someone working alone has to complete a good project to get a good grade.
You are free (and, where appropriate, encouraged) to make use of existing code and systems as part of your project, but you should make sure their use is properly acknowledged, and make clear what additional value your project is adding.

Project Abstract

The first deadline is to submit a project abstract. This will not be graded, but is there to encourage you to get organized early, and work out what focused project you are working on. It is also a chance for dialog between the instructors and the team. You can tell us what you plan to do, anything you have achieved so far, and what you hope to achieve in the rest of the quarter. We can give you extra references, and also information on whether we think the scope of the project is too small or too big. So please do think about where you are and have

something focused and concrete ready to submit. This milestone is to put some uniform structure into the process, but beyond that, we really encourage you to stop by one of our office hours to discuss projects in person.

That allows longer and more productive discussion of projects. Talking about project plans is a particularly good place to get useful feedback and information from the course staff. Just due to greater experience or a different viewpoint, this can often help a lot. The project abstract should fit on one page and should be organized around these 4 sections:
• Problem being investigated
• Approach/plan
• Any achievements so far and/or relevant references found
• Plan of work for the remainder of the quarter
Please send it as a plain text email (i.e., not an attached document) to

cs224n-aut1516-staff@lists.stanford.edu

Please put the email addresses of everyone in the group in the message (and cc them) - this will make it easier for us to send the feedback to everyone in the group.

Data

A large amount of natural language data of various sorts is available at Stanford. This includes collections from major publishers such as the Linguistic Data Consortium (http://www.ldc.upenn.edu/), and some smaller collections, such as text categorization and information extraction training and test sets. The biggest amount of this data is in English, but there is also some data in major foreign languages (Chinese, German, French, Arabic, Spanish, . . .), and some parallel text. You can access some of this data under AFS at /afs/ir/data/linguistic-data/, but there are other collections, such as most speech data, which are not online, so do ask or look around at catalogs, such as the LDCs to see what else exists. The site that discusses corpora available at Stanford is at:

http://www.stanford.edu/dept/linguistics/corpora/

Please note that nearly all of this data is licensed to Stanford, and you are not permitted to copy it to other machines or give it to other people. There is also a lot of data on the web (free corpora
and bake-off data, books, blogs, and web pages). If you could use some resources such as tagged or parsed text, or aligned multilingual materials, and you are not sure what there is, let us know. Preferably as soon as possible. You can find some links to corpora and existing tools at:
http://nlp.stanford.edu/links/statnlp.html
http://nlp.stanford.edu/fsnlp/

Grading and Scope

Always a difficult one to define! But roughly you should be aiming for each member of the team to do at least as much work as on one of the homeworks. You should aim to do something that is small but interesting (i.e., not just an exercise in programming). This may only be a fairly modest extension of an existing technique, but there should be a clear focus in terms of what you hope to achieve, or hope to show. It's perfectly okay to extend something you did in an earlier homework.

Your project write-up should be adequate, but doesn't need to scale linearly in size. One person might want to write 6 pages. A three person project may well find that a 10 page write-up is quite sufficient. Think of the write-up as something like a conference paper, focussed on research questions and achievements, though you may want to include a bit more detail on methods used, examples, etc. You could even look at example computational linguistics conference papers: see the site at http://aclweb.org/anthology-new/. As usual, the quality of your write-up is very important.

It's hard to define exactly what the write-up should cover, because it depends on the project, but generally we're looking for:

- Investigation of a research question. There should be a clear question or application, and hypotheses about the answer or a good approach.
- A clear and complete discussion of the algorithms/method used, and a high level description of the implementation.
- A discussion of the testing you did and results you obtained.
- A discussion of alternatives or things you tried to improve performance, and how they fared.
- While it is okay to try things just to see what happens, this in part includes thoughtful revisions,
- and argument as to why things were good to try, rather than just making random changes.
- Clear results showing the performance of the system. (Make use of tables, graphs, etc.)
- Brief but adequate discussion of related work in the literature (as in an academic conference paper)
- Qualitative discussion on linguistic assumptions of the model and their validity.
- Clean, intelligible code (not for the actual report but for your final submission).

Submission

The project should be electronically submitted by the date shown in the title. We want to receive three copies of your final report on paper (to facilitate our grading in the very short period before spring quarter grades are due). Paper materials may be submitted in the usual way (to submission box, in class, or to a TA). We would be pleased if you could include with your electronic submission a web-readable (HTML, PDF, DOC) version of your report, which we will post so others can look at it, and see the kinds of projects people did. To submit your program, put all the needed files in one directory on an AFS machine (eg. corn or myth) and type:

/afs/ir/class/cs224n/bin/submit
choose fp for the question “Which assignment are you submitting?”

You should make sure that you include the source code for your programs, a Makefile or build.xml that will build them, and an electronic copy of the report. If you have any special requirements that make this impractical (e.g., dependencies on large external packages), let us know. Finally, we will have brief presentations of projects during the scheduled exam period for the class. This is a good opportunity to learn all the exciting things people have been doing! It is required that you present your project in this slot. But it’s also a fun event and will not be graded.