**DATASCI W266: Natural Language Processing**

Fall 2016

Final Project Guidelines

In addition to two traditional programming assignments, students will complete a final project that represents a significant implementation and/or application of NLP techniques. Here’s what to expect:

* You can work on the project in a group of up to 3 students. Solo projects are also fine, but we strongly encourage you to collaborate.
* The project can be on anything within the scope of the class, although it need not be something covered explicitly in lecture.
* At the end, you’ll turn in a writeup in the style of a research paper, and give a short presentation to the class. You’ll also submit your code.

**Note on groups:** You’re welcome to form a group with peers from another section; however, we ask that such groups give their presentation to *both* sections. If so, you’re welcome to re-use slides, and need only turn in a single writeup for the group.

Timeline and Logistics

The project is expected to be significantly more work than each of the assignments, as you’ll need to do planning and research as well as implement much of the code framework to process data, run experiments, generate plots, etc. How to manage all of this? Start early!

We’ll set several milestones to check in along the way:

* **Project proposal** (and groups), due at end of Week 7 (Sunday October 16 @ 11:59 PM)
* **Milestone** due Week 11
* **Final Presentations** in class Week 15
* **Project writeup** due last day of classes (Week 15)

Project Scope and Ideas

The project is very open-ended, but we do require it to be an NLP project. This means that, at its core, your project should be about processing text, speech, or other form of language. It need not be completely novel in the way that academic research is, but should attempt to do something non-trivial or non-obvious with the data and/or algorithm. For example, you could:

* Apply existing (well-known or otherwise) NLP algorithm(s) to a new dataset
* Develop a new NLP algorithm and apply it to a well-known dataset
* Implement an algorithm from a recent paper and apply it to a new domain or dataset
* Use NLP algorithms in a descriptive way to find trends or patterns of practical or linguistic interest

As for scope: an ambitious project for a group of 2-3 would be similar in scope and depth of experiments to a conference paper.

For project ideas, take a look at the final projects from [Stanford cs224n](http://nlp.stanford.edu/courses/cs224n/) (all years), or for deep-learning projects, [cs224d](http://cs224d.stanford.edu/) ([2015](http://cs224d.stanford.edu/reports_2015.html), [2016](http://cs224d.stanford.edu/reports_2016.html)). A few examples, as a place to start:

* Abstractive summarization of news articles
* Restaurant menu extraction from user reviews
* Question-answering with neural attention or memory models
* Image captioning (language generation)
* Analysis of gender roles or power dynamics in movie dialogues
* Explore techniques for interpreting “black box” / neural models for language

Also see below for NLP conferences - [ACL](http://acl2016.org/index.php?article_id=68) and [EMNLP](http://www.emnlp2016.net/accepted-papers.html) are the top two, and you can find plenty of interesting ideas from the recent proceedings!

Project Proposal

Each group should submit a proposal by the end of Week 7; we’ll read each one and give you detailed feedback. The proposal should be a short writeup (150-200 words), similar to an abstract. Be sure to answer or address, as best you can:

* What do you plan to do?
* Why is it important, and why is it challenging?
* What dataset(s) will you use, and how will you plan to obtain the data?
* What algorithms might you use? Are good implementations available, or will you need to write your own? *(Don’t worry if you can’t answer this well at this stage of the course.)*

You should also include 2-3 references to relevant papers for the technique and/or dataset you’ll be using. Some resources that may be helpful:

* *(start here)* The [Association of Computational Linguistics (ACL) Anthology](http://aclweb.org/anthology/), which indexes papers from most of the top NLP conferences.
* The [ACM Digital Library](http://dl.acm.org/) is also a great resource, indexing across a number of CS conferences. In particular, [SigKDD](http://www.kdd.org/) (Knowledge Discovery and Data Mining) and [WSDM](http://www.wsdm-conference.org/) (Web Search and Data Mining) might be useful.
* [NIPS](https://nips.cc/Conferences/2015) and [ICML](http://icml.cc/2016/) are the top machine learning conferences; these are good places to look if you want to use neural networks.
* [Google Scholar](https://scholar.google.com/) and the [arXiv](https://arxiv.org/) are great to find specific papers.

This year’s NLP conferences might be good for inspiration as well, or if you just want a sense of what people in the field are working on: [EMNLP (2015)](http://aclweb.org/anthology/D/D15/), [NAACL](http://aclweb.org/anthology/N/N16/), [ACL](http://aclweb.org/anthology/P/P16/).

Milestone

You’ll submit a partial implementation of your project by the end of Week 11. This should include a baseline model, which can be something very simple: random predictor, most-common-class, bag-of-words model, etc. The important part is that you’ve been able to obtain, load, and play around a bit with your data, and get some baseline results that you can compare to when you start doing real experiments.

We’ll announce more about the milestone format as we get closer to the due date. It will be fairly lightweight, as we don’t want this submission to distract from work on the actual project!

Presentations and Final Submission

We’ll release more details on presentations as the date is closer, but expect that your group will give a short (~ 10 min) presentation of your project during live session in the last week.

As a final deliverable, we ask that you make a short writeup. The easiest way to do this is as an IPython notebook with inline figures and code snippets, but you’re also welcome to use LaTeX. Your writeup should have the usual structure of a research paper, including background, methods, and results and discussion sections, as well as an abstract. For either format, aim for something between an ACL [short paper](http://aclweb.org/anthology/P/P16/#2000) and [long paper](http://aclweb.org/anthology/P/P16/#1000) in length.