CS 4740 Introduction to NLP
Fall 2015
NLP System Design

Report: Due via CMS by Fri, Dec 04 11:59pm
No **hardcopy** required.

1 Introduction

The goal in this (non-programming) project is to gain experience in the design of a Natural Language Processing System and the mapping of a real-world text analysis problem to tasks and methods from NLP.

Think about an NLP task or problem that interests you or that you would like to see solved. The project will be better (and MUCH more fun) if this task is something that you are truly interested in. It should involve the analysis and/or generation of text. It might involve human-computer interaction, the analysis of conversations, data from the social web, historical texts, transcribed speech...whatever you want.

Once you have a problem in mind, think about how you would go about developing a system to solve the problem, i.e. perform the NL task. The sections of the Report that you will write will walk you through the various aspects of the problem that you need to address.

2 Report

You should submit a report that contains the following sections. (You can include additional sections if you wish.) There is no page limit, but my guess is that the report will be about 4-5 pages of written text, not including images and data samples. So it could be longer than that depending on the project that you select.

1. **The Problem or Task**

   Clearly describe the problem that you will investigate. What will the input to the system be? What will the output be? **Give at least one concrete example of the input/output behavior.** By the end of this section it should be crystal clear to a reader the task for which you are designing a system.

Why did you choose to investigate this task? Is there a need for NLP systems to solve this problem, or was there a linguistic or cognitive issue that you are interested in that prompted your problem selection?

2. **General Approach or System Architecture**

   Describe at a high level the approach that you will take to solve the problem. Depending on your problem selection, this might best be done via a diagram of the proposed system architecture. How will you break down the task into smaller subtasks? I.e. what modules will be required?

   If there were alternative approaches or architectures that you considered, briefly explain them as well as why you rejected them.

3. **Data and Data Annotation**

   What specific data / text corpora will you need? For what purpose? Here, we want specifics (e.g. all NYTimes articles on the 2016 presidential election) not a general type of data (e.g. news articles, Facebook status statements). **Include samples of the data in the report or as a separate appendix/file.**

   How much textual (or other) data do you think the project will require?

   Where will you obtain the data? Web searches will be useful here. One site that describes many text corpora is from the University of Pennsylvania's Linguistic Data Consortium (LDC). The LDC catalog gives descriptions of the kinds of corpora the LDC has and provides a sample of the text and annotations that comprise each corpus. **Important note: You will not be able to download the data sets** since LDC makes the data available only to organizations who subscribe. But the catalog should give you enough information on the data for your project (and to include in the report as an example).

   Will the project require any manual annotation of training data? If so, explain the annotation types needed and **include examples of the annotation on a small sample of your data**.

4. **Methods and System Development**

   Include a subsection for each major component of the system that describes the methods and techniques that you will employ. Will you be developing any new algorithms? If so, describe them via high-level pseudo code. In cases where you will use machine learning algorithms, be clear about exactly how they will be used, e.g. when applying a classification approach: make clear what the classes are, what features will be used and why, **give examples of actual training instances**.

5. **Implementation**

   What existing packages or libraries will you plan to use? Which aspects of the system will you implement from scratch?

6. **Evaluation**

   How will you evaluate your system or approach? What quantitative metrics will you use? Note that some projects might require manual evaluation, i.e. human evaluation. If so, describe how that would work.

7. **Anything else...**

   Include additional sections as needed to describe important aspects of your particular project that are not addressed above.

8. **Individual Member Contribution**

   Briefly explain the contribution of each group member.

3 Grading Guide

- Problem Selection and Description: [25 pts]
- Proposed Approach/Architecture: [50 pts]
- Data and Implementation Plan: [10 pts]
- Evaluation Plan: [15 pts]

4 What to submit to CMS

- The report (pdf)
- Any additional data samples, system-in-action examples
- If you are submitting more than one file, create an archive for them.