Instructions for the SI 630 Project Proposal

Version 1.0

Be sure to put your name here in the submitted version

1 Introduction

The course project is intended to provide an opportunity for students to dive deeper into one problem or topic of their choice and write a very small scale study on the topic. The project proposal is the very first step to identify some *text-based* question that you think is interesting and to begin thinking about how to evaluate it and where to get the data for it.

The learning goals for the course project are as follows.

1. Learn advanced NLP develop skills through practice
2. Gain specialized knowledge in a particular topic or problem
3. Provide and end-to-end experience from data to results in answering a question using NLP
4. Practice forming a research question and designing a series of experiments to answer that question
5. Learn about new NLP methods through literature search
6. Learn about a particular topic or problem through literature search
7. Practice preparing high-quality technical reports in typesetting (latex)
8. Produce a concrete artifact that can be shown to interested parties (employers).

The project proposal should *not* focus on *how* you are going to solve the problem or what methods you will use. Right now, you're still learning a lot about NLP and the kinds of skills you will gain through the course and later homework will greatly inform how you solve your problem. Instead, the proposal is focused around the core question(s) you want to pursue. You can (and should) safely assume that in later weeks, you'll learn how to answer your questions.

Please remember that the 630 instructors are here for you and will gladly offer suggestions and advice on projects. We want your projects to succeed, to be fun to work on, and to spark your intellectual curiosity!

We've collected a list of advice on how to do the course project from past semesters https://docs.google.com/spreadsheets/d/1_L-kh237dD8BT-EqQzT-2z-ftVoi5UlMo0c1Vd230MI/edit?usp=sharing, which we hope will help you in your planning. You'll get a chance to add to this list at the end of the semester too!

2 Formatting

For the project proposal, you will to use the ACL LaTeXtemplate.[1] This format is helpful for creating a standard layout for all project documents and to help you gain familiarity with how to write in a technical markup language. You're hopefully getting familiar with this format through the project notes. LaTeXhas a great reference management system as well, BibTeX, that will make your life much easier. Google scholar and Semantic Scholar can directly explore bibtex as well. If you are not familiar with LaTeXyet, Overleaf has some very helpful documentation to get you started.[2]. As a side bonus, all of your write-ups will look very fancy and professional.

As one helpful point we've seen from past proposals, you should learn how to properly cite references in your latex code. Here's how to cite something. First, get the BibTex entry for your reference. If you're using Google Scholar, there's an "Import into BibTeX" link below that will give

[1] https://www.overleaf.com/read/crtcwgxzjskr

[2] https://www.overleaf.com/learn/latex/Tutorials

you the text you want. A bibtex entry might look like this:

```
@inproceedings{zhou2020condolence,
  title={Condolence and Empathy in Online Communities},
  author={Zhou, Naitian and Jurgens, David},
  booktitle={Proceedings of EMNLP},
  pages={609--626},
  year={2020}
}
```

Second, add the BibTex entry to your `.bib` file, which is your bibliography—but the contexts of this file is *not* what shows up in your references. That first part of the entry is the *bib key* which is what you'll need to use when making a citation. For example, if we added that bibtex entry to our `.bib` file, we would then cite the work using its bib key by adding

`\cite{zhou2020condolence}`

where I want to cite the reference. Adding that `cite` will cause that citation to appear in the text—like this (Zhou and Jurgens, 2020)—and then a reference for it will show up in my paper's References section. If you want to refer to the author name in a citation like "Zhou and Jurgens (2020) showed that..." then use the `citet` command instead.

3 Project Scoping and Constraints

We provide the following items as guidance and constraints based on observations from previous semesters.

- Feel free to be ambitious in your project. We will help you scope your project, if needed, after you propose it.

- Multi-person teams need to have more complex projects. These kinds of projects should not just be like "Person A will collect the data and Person B will train the classifier." As you'll see in later homeworks, that kind of process will be too easy and neither person will learn much from the project. We are looking for projects with big enough scope that *every person* in the group has a chance to get involved and deepen their skills.

- **Projects cannot use data obtained from Kaggle or other similar sites where there are existing examples of how to use the data.** This is bolded because every year some students still use Kaggle data (or similar). We want you to practice going end-to-end with your project and these kinds of datasets are often accompanied by code that has too much done for you. Any project proposal using this kind of data may automatically get a zero.

The following projects are not allowed for proposals:

- Sentiment or emotion analysis

- Stock market prediction

- Open-ended text generation[3]

4 What to Cover

Your project proposal is expected to address the following points in a coherent document that reads like a pitch for something you want to work on—not a list of bullet points. Project proposals written like a list of bullet points will receive a reduced grade. We typically expect that final reports are written in a coherent structure with sections denoting coherent elements, and the proposal is intended as scaffolding to get you there. You will likely re-use parts (or most) of the text for the proposal when you write your update and final report, so the time you put in now will help you out later.

Project Goals (5 points) For the proposal, you should have a rough draft of an Introduction section that clearly states what are the goals of the project is—what question are you trying to answer?—and provides some broader context for why these are important goals. You should also include a statement on why solving this problem matters–who would care if you solved it and what effect would solving it have? As an eye to the future, in the proposal, think about who you would want to see your work and why your results would matter to different groups of people.

NLP Task Definition (10 points) The section describes what specific problem or NLP Task you plan to solve and goes into much more detail than what's specified in the Project Goals in the Introduction. You can also add details about what your problem is not to help guide the reader's expectations. For the proposal, describe very clearly what problem you will solve. It helps to be specific about what kind of input your system will use and what specifically it will produce as output.

[3] This project topic is cool but just too hard to evaluate; if you *really* want to try some open-ended generation, come talk with us first.

For the proposal, you do *not* need to know how to solve your problem! We are asking you to think of a question you could answer with NLP and then to figure out how to describe that as a specific task. Tasks have inputs and outputs, which you can use to explain your system. For example, does your system only take text as input, or will it need other things (networks? twitter bios? biomedical data?) and what kinds of things will it produce (scores? classifications?). You'll use this system to solve a task that lets you answer your question.

Data (10 points) This section goes into detail what data you will use to train and evaluate a model for your particular NLP Task. Ideally, you should already have the data on hand or spend a few minutes getting it. Projects are much more successful when most of the time is spent solving the problem rather than searching for the right data. If you don't have data at proposal time, please be very specific on how you plan to get it. We might be able to recommend something but we encourage you to come to talk to us during office hours or after class. **We strongly encourage requests for data to be made as public Piazza posts so we can crowdsource data that folks might be looking for.**

If you have the data, you should include a few examples and can also include some very rough statistics (e.g., how many instances you have). Tables and figures are very useful for showing examples. Please make sure the text is legible! Proposals with figures that are hard to read or where the font is too small will receive reduced credit.

Since not everyone has a burning question they're dying to answer with NLP, we will post some example tasks/datasets on Piazza. For these tasks, the problem, data, and evaluation criteria are already provided– though you should be sure to describe them here as a part of your proposal. Important note: **projects cannot be derived from any Kaggle competition or similar type of website**, without prior approval. Essentially, if the instructors search online for your dataset and problem and can find a fully-worked example on a blog/github, the idea can't be used. Submissions using this kind of data will be turned back with a grade of zero. If you're stuck for looking for data try using one of the many dataset search engines from Microsoft or Google. In general, we recommend finding a topic area that you're passionate about, thinking of a few questions that involve text from that area, and then looking for datasets. If you're stuck, ask on Piazza and we can collectively try to find some- thing.

Related Work (5 points) You should have at least three papers related to your current problem and a few sentences describing what they did to solve the problem. The related work section should describe how other people have thought about the problem you're working on. How did they approach it? What makes their problem different from yours? Why do you think your approach will be better? Be sure to cite the papers using (see the references.bib file for the ACL Overleaf template as an example). If you're not sure which papers to cite, try searching for your topic and keywords on Google Scholar or Semantic Scholar; their "related papers" functionality can help you find interesting papers for ideas.

We also want to point out that the papers you choose for related work often can help frame your own task. What kinds of motivation did these papers use? What kind of data did they use—can you incorporate that data too? What kinds of evaluation metrics did they use for their models—would those metrics work for your case? We ask you to look for related works to help you ground your approach within the larger space of NLP and get a bit of insight into how others have thought about their problems. Later on when you're working on solving your problems, you may revisit these papers to see how they built their models too!

Evaluation (20 points) You **must** include an specific evaluation metric as a part of the proposal. In essence, **you need to define how you will measure how good your NLP system is at your particular task or at solving your problem**. Manual analysis of output can be helpful, but you need to know quantitatively how to measure goodness.[4] If you're not sure about a good evaluation, please post of Piazza and we can work it out there.

The second part of the evaluation plan is what baselines you'll compare your system against. One baseline should be random performance. A

[4] As a side note, projects involving topic modeling on a corpus usually have a harder time coming up with a precise metric (i.e., how good is one topic model versus another?), so I would avoid those unless you have an extrinsic tasks that you can test on (i.e., topic modeling is used as a method to facilitate solving some other kind of task). Projects involving text generation too often have a harder time with evaluation, so it's worth looking at related work in your proposal's topic area to see how they evaluated.

second baseline should be something reasonable that doesn't require much knowledge or learning. For example, if you're doing a classification, always choosing the most frequent class is a useful baseline. A baseline is essential here because it helps you figure out what is the simplest way to solve your task. Baselines provide a useful comparison for putting your model in context—several students in past semesters have been surprised by how well baselines performed

Work Plan (5 points) You'll end with a plan for how you'll accomplish your project. We hope this section can help you think at a high-level about which tasks are necessary to get to the point where you have a working model. Of course, plans are often made and then changed when new information or challenges emerge, so we won't hold you you to this. However, the act of writing a plan can greatly help you figure out how think about the process and, in general, projects that are proposed with more concrete work plans tend to be more successful. In later project documents, you'll be asked to reflect on this plan and consider what went right (or wrong) on the timing; this kind of reflection can help lead to better time-estimation skills.

Multi-person Team Justification If you have more than one person on your project, you should justify why the work requires the number of people you have. In addition, you should provide a concrete explanation of what each team member is expected to do, keeping in mind that all team members need to be doing some kind of NLP for the project. An n person project should have $n * 1$ peoples worth of work. If you are not sure your project is big enough in scope, please come talk to us. Team-based project proposals that are underscoped may have points taken off if the proposal is overly simple.

References

Naitian Zhou and David Jurgens. 2020. Condolence and empathy in online communities. In *Proceedings of EMNLP*, pages 609–626.