Project Proposal

Project Title:
Student Name:

Please fill out this proposal template with a description of your project. Replace the italicized text below in each section with your own text.

General Requirements:
Each project is required to have the following components (as discussed in class):

1. **Database component**: your data for the project should be stored in a database
2. **Machine learning or statistical component**: your project needs to have a component that does machine learning or builds a statistical model, e.g., develop a module for classifying sentiment in tweets, or extracting information from product reviews, etc.
3. **Visualization/human interface component**: part of your project will need to have a way for a user to interact with your system, e.g., via queries, via menu selection, visualizing results – essentially an interface that demonstrates to a user what your system can do .

Note that your projects will likely include significant additional pieces beyond the 3 above, e.g., your project might require scripts and code for obtaining data via Web-scraping or APIs, and then performing significant preprocessing (e.g., extracting relevant information, cleaning the data, etc).

1. Project Summary

Provide a clear description (2 or 3 sentences) that summarizes your project. A good way to summarize your project is to start with a sentence that clearly defines the problem, e.g., "This project will develop a system to do Y…." . You should follow this with a brief summary of your planned technical approach, e.g., "The approach I will take to address this problem is 3-fold: (1), use method A to do X (2), use method B to do Y, (3)….." (as an example)

2. Proposed Technical Approach

Write 2 or 3 paragraphs with a clear, more detailed description of the methods and algorithms that you plan to use on the project. If the system you are building can be thought of as a pipeline with multiple components, a useful approach is to provide a figure that illustrates the pipeline (with blocks for different components) along with brief descriptions of each component (e.g., the names of algorithms or methods you plan to evaluate). Make sure it is clear what your pipeline or system is doing, i.e., what each component will do in terms of taking inputs and producing outputs.

This section should include a brief description of each of the database, machine learning/statistics, and visualization/interface components in your project.

3. Data Sets

Briefly describe what data sets you plan to use in the project. Ideally you should have a social media data set (such as Reddit, Twitter, Yelp) plus a "non social media" data set.

Include specific references to the data (e.g., a URL) if you can. If, for example, you are planning to work with Web/text data, it would be good to do some preliminary assessment of how much is available, what the fields/attributes/metadata are, if there are labels for the data (if you are doing supervised machine learning), and so on. For text data you could say whether you plan to work with data that is already tokenized and already has a predefined vocabulary or whether you plan to investigate different tokenization methods and explore different vocabularies.

You can change/update your data sets during the project if you need to, but you should have identified at least one data set to work with by the time you submit the initial proposal.

4. Experiments and Evaluation

Provide a brief and clear description of how you plan to evaluate the results of your project. For example, if you are doing classification you should consider metrics such as classification accuracy and precision-recall. Will you use cross-validation, or does your data set(s) come with a fixed train-test partition? For unsupervised learning tasks like clustering or topic modeling you may have to do some research to see how evaluation is done on these tasks. For some projects you may have to do some user studies for evaluation, e.g., present users with results from Algorithm A and Algorithm B, using the same input data for each algorithm, without telling the user which algorithm is which, and have them select the one they prefer. Or your evaluation may be more qualitative in that you hope to generate insights about a particular problem.

5. Software

Provide a list of the major pieces of project software that you expect to use, divided into 2 sets: (1) publicly-available code, and (2) code you will write yourself. The list of what public software you will use will probably be incomplete at this point (which is fine) since you may not know yet about all of the software that might be relevant to your project. You may also want to use a tool such as Github to coordinate your code development on the project – if you have not used Github before, this would be an excellent opportunity to learn to use it.

6. Milestones

Provide a brief list of milestones in 2-week "chunks" for your Spring quarter work:

- *Weeks 1-2*
- *Weeks 3-4*
- *Weeks 5-6*

- *Weeks 7-8*
- *Weeks 9-10*

For example, much of the data gathering and preprocessing and coding (development and test) could happen in the earlier weeks, and much of the evaluation and writing in the later weeks. Note that you will be required to provide updates and progress reports on a regular basis during the quarter.