Project Proposal for CS 175, Winter 2018

Project Title: title goes here

List of Team Members:
Name1, StudentID1, uci_email_address
Name2, StudentID2, uci_email_address
[ Name3, StudentID3, uci_email_address ]

1. Project Summary
A clear description (2 or 3 sentences) that summarizes your project: e.g., “This project will use XX methods to predict YY using the Z1 and Z2 data sets, with evaluation using classification accuracy and user studies.”

2. Problem Definition
Write at least a paragraph that clearly defines in more detail (beyond the summary) what problem you will be trying to solve. One way to describe this is to think about your project in terms of inputs and outputs: what will the inputs to your system be and what will it produce as output? If you can, add a sentence that mentions 1 or 2 technical approaches (e.g., algorithms) have been used in the past to address this problem. Feel free to add a reference or two (see the class Web site for suggestions, or do a search in Google Scholar using appropriate keywords).

3. Proposed Technical Approach
Write a paragraph or two with a clear description of the methods and algorithms 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.

4. Data Sets
Briefly describe what data set(s) you plan to use in the project (if data set(s) are an important part of your project). Include references to the data (e.g., a URL) if you can. If for example you are doing document classification, you can describe for example how many documents are in the data set, average document length, how many classification labels. If you are using multiple data sets you could put this type of information in a table. You could for example 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. If you are able to access and take an initial look at your data, this is helpful (although not required) – you could for example include a figure or two in this section, e.g., a histogram of document lengths. You can change 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 proposal.
5. Experiments and Evaluation
Provide a brief and clear description of how you will evaluate the results of your project, e.g., accuracy for classification, precision-recall for document ranking. 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 – if you plan to do this it would be good to think through how many users you plan to have participate, at what stage of the project you would do this (e.g., you could do it once to get initial feedback and later for a final evaluation), and so on.

6. 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 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. My expectation is that most students will use Python, given that we have been using Python in class and there are many useful publicly-available tools for text analysis in Python. However, if you prefer to use a language such as Java that is ok too - please indicate this clearly in this section. You may also want to use a tool such as Github to coordinate code development on the project – if you have not used Github before this would be an excellent opportunity to learn to use it.

Note that each team member is expected to write a non-trivial component of code for the project (this could be a specific machine learning algorithm, some code for cleaning noisy data during preprocessing, a script for creating pipelines and running experiments, code for visualizing/displaying results, etc). Each student on a team will need to be able to point to their specific code in the progress report and in the final report. Given this, it is important that in the early weeks of the project that team members identify “who is doing what” – this can change as the project evolves. In this section you describe what code the team will write (your best estimate – this may get updated later as you get into the details of your project) – and in Section 8 below but you will provide your current best guess about who will write what.

7. Milestones
Provide a brief list of milestones. For example, since the project will span 6 weeks of the class (weeks 5 to 10), you could break your milestones into a list of 3 intermediate phases:

- Weeks 5 and 6
- Weeks 7 and 8
- Weeks 9 and 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 experimentation and evaluation in the later weeks. Note that you have a progress report due at the end of week 7.

8. Individual Student Responsibilities
Summarize briefly what each student will be primarily responsible for in the project. For example, you might write something like this
**Name 1:** will write and test the code for Algorithms 1 and 2, will integrate components A and B in the pipeline, will assist in doing experiments and interpreting results, will assist in writing project reports

**Name 2:** will acquire the data sets to test the algorithms, will preprocess the text data (e.g., define the vocabulary for the algorithms), will implement Algorithm 3 and integrate all the components into a pipeline, will write the scripts for evaluating the accuracy of the algorithms, will assist in writing project reports.

[Note these are just suggestions – you can and should organize responsibilities in whatever way makes sense – and inevitably as the project progresses these responsibilities may need to be changed as some tasks may take much more time (or much less time) than originally expected.]