**Project Title:** Authors Reviewing Restaurants(Project Proposal for CS 175, Winter 2016)

**List of Team Members:**Shaun McThomas, 138288643, smcthoma@uci.edu  
Sean King, StudentID2, uci\_email\_address  
Evie Smith, StudentID3, uci\_email\_address

**1. Project Summary**The goal of this project is to create yelp like reviews in the voice and tone of well know authors.

A clear description (2 or 3 sentences) that summarizes (a) the goal of the project (what is the primary aim of the project) and (b) how you plan to achieve the goal (mention the methods, algorithms, data sets, experimental evaluation methods you plan to use). (Tip: you could write this last after you have written the other sections – it is intended to give the reader a quick idea of what your project is about).

**2. Problem Description and Background**In modern times, words of famous authors are sometimes lost in a spew of new social media terms. One of the most viewed text models is in reviews of various thing. This project is create a method of

Write about 1 paragraph defining (in more detail than in the summary) what problem your project will address. For example if your project is multi-label document classification then you would clearly define what multi-label document classification is. Mention if you can what methods/algorithms (that you know of) that have been used in the past to address this problem. Add one or more references to a text or a paper that discusses the problem if you can (see the class Web site for suggestions, or do a search in Google Scholar using appropriate keywords).

**3. Data Sets**Briefly describe what data set(s) you plan to use in the 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. Mention for example whether you plan to work with data that already has a predefined vocabulary or whether you plan to define your own vocabulary. If you are able to access and take an initial look at your data, feel free to also 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.

**4. Proposed Technical Approach**Provide a description of the methods and algorithms you will use on the project. If the system you are building can be thought of as a pipeline with multiple components feel free to provide a figure that illustrates the pipeline with blocks for different components and brief descriptions of each component. If your project involves comparing different classification algorithms for document classification then in this section you would list and briefly mention the classification algorithms you plan to use in your project (e.g., naïve Bayes, logistic regression, support-vector machines, neural networks, etc). If there are additional details that are relevant you can mention them, e.g., what type of naïve Bayes model or what type of neural network model.

**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 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.

**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. This list will probably be incomplete at this point (which is fine) since you may not know yet about all of the publicly-available 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.

**7. Milestones**Provide a brief list of milestones, e.g., since the project will span 6 weeks of the class (weeks 5 to 10), you may want to 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 beginning of week 8 (on Feb 22nd).

**8. Individual Student Responsibilities**Summarize briefly what each student will be primarily responsible for in the project. For example, if you are implementing 2 algorithms and there are 3 students, then one arrangement could be:

**Name 1:** will write and test the code for Algorithm 1, will assist in doing experiments and interpreting results, will assist in writing project reports  
**Name 2:** will write and test the code for Algorithm 2, will assist in doing experiments and interpreting results, will assist in writing project reports  
**Name 3:** will acquire the data sets to test the algorithms, will preprocess the text data (e.g., define the vocabulary for the algorithms), 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.]