Info 290 Final Project Proposal
Vanessa McAfee and Joanna Hsu
March 28, 2013

1. Motivation and Project Goals

UC Berkeley offers a daunting number of courses spanning a wide range of topics in about 150 different departments. Given both the interdisciplinary nature of many departments and the fact that subject matter expertise often transcends department boundaries, knowing what courses are available on a certain topic or group of topics is a nontrivial task. The goal of our project is to chart the constellation of courses available at Berkeley, grouping courses and departments by their topics. Specifically, our project will address the following questions:

1.) Given a course description, can we predict what department the course belongs to? Furthermore, how much do teaching topics reflect departmental divisions at Berkeley?

2.) Can we determine which departments are most similar to each other based on their course descriptions?

2. Data Set description

Our data set will be created by scraping data from the Berkeley course catalog. We will create a flat file which includes fields such as: Department, Course Number, Description. Currentlyl the data can be found by issuing queries on the Berkeley course search page: http://osoc.berkeley.edu/catalog/gcc_search_menu. An example course looks like:

Energy Solutions: Carbon Capture and Sequestration -- Chemistry (Department Of) (CHEM) C236 [3 units]
Course Format: Three hours of lecture per week.
Prerequisites: Chemistry 4B or 1B, Mathematics 1B, Physics 7B or equivalents.
Description: After a brief overview of the chemistry of carbon dioxide in the land, ocean, and atmosphere, the course will survey the capture and sequestration of CO2 from anthropogenic sources. Emphasis will be placed on the integration of materials synthesis and unit operation design, including the chemistry and engineering aspects of sequestration. The course primarily addresses scientific and engineering challenges and aims to engage students in state-of-the-art research in global energy challenges. Also listed as Earth and Planetary Science C295Z and Chemical & Biomolecular Engineering C295Z.
(F) Bourg, DePaolo, Long, Reimer, Smit

3. Overview of methods

We will obtain course information (course description, department, etc.) for all the courses listed in Berkeley's online course catalog. After removing duplicates, merging cross-listed courses, and cleaning the data, we will use supervised and unsupervised learning techniques to address our questions. We will build a model that predicts the department a course belongs to based on its course description. We will also use clustering algorithms to group similar courses together. All models will be built using a training set and then evaluated on testing sets. We will generate and present visualizations of the results from our models.

Proposed Project Plan

Tasks	Owner	Due By
Scrape data from course catalog using Beautiful Soup or another tool	Both	4/5
Create data set. Merge data into a flat file with the following labels: Course ID, Department Code, Course Number, College/School Id, Description, Prerequisites, Format, Term offered, Year offered, Instructor(?)	Both	4/5
Generate summary statistics and graphs: A. Most commonly used words across course descriptions B. Average number of words used in the course description (mean, median, mode) C. Average number of different courses offered by department D. Number of cross-listed courses. Think about how to deal with cross listed courses. E. Department with the most unique words?	Both	4/12
More data preparation. Remove stop words. Create training and test sets.	Both	4/12
Supervised Learning/ Predictive Modeling A. Create predictive model for department based on course description words B. Model Evaluation	Joanna	4/26
Unsupervised Learning A. Create clusters of courses based on course description to see which courses are most similar to each other? B. Model Evaluation	Vanessa	4/26
Brainstorm/Create Data Visualizations. Prepare some data visualizations based on interesting findings in the data, for example showing which departments are most similar based on their course descriptions.	Both	4/3
Write final paper	Both	due May 10