Classifying Courses at Scale: a Text as Data Approach to Understanding Student Course-Taking

Annaliese Paulson
Kevin Stange
Allyson Flaster

Technical Report: Version 1.0
December 2022

I. Overview

This technical report summarizes the training steps for *Classifying Courses at Scale,* a project that uses machine learning and natural language processing to classify courses from student transcripts into a standardized hierarchical taxonomy of course-taking called the College Course Map (CCM). Applying our CCM classification models to unstandardized administrative transcript data allows researchers to compare student course-taking patterns across multiple institutions and postsecondary systems. This technical report summarizes the model training steps. Future versions of the technical report will include model evaluation, error analysis, and demonstrations of applications of the models to administrative data.

II. College Course Map

The CCM is a product of the US Department of Education's National Center for Education Statistics (NCES) that is featured prominently in their longitudinal surveys of postsecondary students. We use the CCM to train machine learning models to standardize student transcripts, drawing on restricted data from NCES's Postsecondary Education Transcript Studies (PETS). The PETS collect postsecondary transcripts of students associated with each of four NCES surveys: High School Longitudinal Study of 2009, Baccalaureate and Beyond Longitudinal Study of 2008-2012, Beginning Postsecondary Students Longitudinal Study of 2004-2009, and Beginning Postsecondary Students Longitudinal Study of 2012-2017. We refer users to NCES for documentation on survey design.

In particular, the four PETS datasets we draw on collect course subjects, catalog numbers, and titles for all courses students enroll in throughout their postsecondary career. Because raw administrative transcripts are unwieldy to work with, these courses are then annotated and standardized by human coders using the 2010 College Course Map (CCM) typology (Bryan & Simone, 2012). Loosely based on the the Classification of Instructional Programs, the CCM maps each course into a six digit code where the first two digits define a course's broad cluster (e.g. 45 - social sciences), the third and fourth digits define a course's subcategory (e.g. 45.06 - economics), and the fifth and sixth digits define a course's specific subject code (e.g. 45.0603 econometrics and quantitative economics).

In the PETS data, courses were annotated by human coders on the basis of course titles, subject codes, and catalog numbers, and - if necessary - course descriptions from postsecondary institutions' course catalogs. If these pieces of information were insufficient for classification, then annotators relied on the additional context provided by the other courses on students' transcripts. However, in the data available to researchers through PETS, course descriptions drawn from course catalogs are not distributed. For our purposes, we rely only on the course titles and subject codes distributed with PETS.

In Table 1, we use publicly-available data to illustrate the importance of using CCM codes to standardize transcript data across institutions. The table shows several examples of subject codes and course titles from public institutions in Texas in the format of the restricted PETS data. Note that-- for courses covering the same content-- there are a variety of course numbering systems and ways of titling courses across transcripts (see BIO 205L and BIOL 3415, for example). Absent the CCM codes, it would be difficult to determine the topical overlap of such classes.

Table 1
Courses from Public Texas Institutions in PETS Format

Subject Code	**Course Title**	**College Course Map Code**
BIO 205L	LAB EXPRMNT BIO-CELL & MOL BIO	26.0204 - Molecular Biology
BIOL 3415	INTRO TO MOLECULAR BIOL	26.0204 - Molecular Biology
ANTH 2414	BIOLOGICAL ANTH	45.0202 - Physical and Biological Anthropology
BIOL 1333	INTRO BIOL	26.0101 - Biology/Biological Sciences General
ANT 5073	ADV. BIOLOGICAL ANTHROPOLOGY	45.0202 - Physical and Biological Anthropology

III. Dataset Construction

In our supervised machine learning task, the first step is to develop a comprehensive corpus of labeled training data that contains both a gold-standard label for each record and the features we will use to predict that label. To create our full corpus of labeled transcript data for use in our training models, we append the four PETS studies together, creating a full dataset of all course-taking of students across all four surveys. We remove any courses that do not have a CCM code in the PETS datasets or that have a missing course title. Because we are interested in predicting the CCM code of a course given a subject number and course title, we keep only unique subject number, course title, and CCM code triplets. Finally, a small number of CCM codes are particularly uncommon and we remove any codes that occur fewer than 25 times.

IV. Feature Transformation

After creating our full corpus, the next step requires us to transform raw unstructured text into a machine computable format, creating a mapping between text as provided to us in our corpus and a numeric representation of that text. To create features to train our model, we transform two pieces of text available in the PETS data, course titles and subject numbers. We work with course titles as they are provided in PETS. However, because of our modeling strategy, the model is unlikely to receive reliable information from the course number present in the subject numbers. As a first preprocessing step for subject numbers, we remove all numeric characters from the subject number strings (e.g. transforming MAT 150A to MAT A). For both course titles and subject numbers, we then lowercase all strings, and split sentences into individual tokens on word boundaries. To create training and test datasets, we randomly sample 90 percent of our corpus for training and 10 percent of our corpus for testing, stratifying on six digit College Course Map code to ensure representation of all codes across both the training and test datasets.

We then remove any tokens that occur fewer than 50 times across the training corpus. To convert these cleaned strings to a machine computable format, we then create two bag of words matrices - two D by V matrices where D is the number of documents (or course records in our corpus) and V is the number of tokens that occur 50 or more times in our training corpus. For a given document, d, and token in our vocabulary, v, cell $\{d,v\}$ of the bag of words matrix contains the number of times token v occurs in document d.

Table 2 shows an example of a bag-of-words matrix using the course titles from the Texas data in Table 1. For instance, because the token “lab” occurs one time in the document “LAB EXPRMNT BIO-CELL & MOL BIO”, the corresponding cell {“LAB EXPRMNT BIO-CELL & MOL BIO”, “lab”} contains a value of one. Similarly, the token “bio” occurs twice in this document and the corresponding cell contains a value of two. As noted above, we create two bag-of-word matrices for subject codes and course titles. For both the training and test datasets, we then concatenate the subject code and course title matrices together to create our final datasets.

Table 2
Bag of Words Matrix

	lab	exprmnt	bio	cell	&	mol	intro	to	molecular	biol	biological	anth
LAB EXPRMNT BIO-CELL & MOL BIO	1	1	2	1	1	1	0	0	0	0	0	0
INTRO TO MOLECULAR BIOL	0	0	0	0	0	0	1	1	1	1	0	0
BIOLOGICAL ANTH	0	0	0	0	0	0	0	1	0	0	1	1
INTRO BIOL	0	0	0	0	0	0	1	0	1	1	0	0

V. Methods

Given our derived set of features and the associated CCM labels, we can train a model to predict the appropriate CCM code. After training, this model can then be used to predict CCM codes on unseen data. For each of the two, four, and six digit CCM codes, we train a regularized multinomial logistic regression model on the training dataset to predict the appropriate CCM code, given the concatenated bag-of-words matrix as features. This produces a logistic regression model with a set of C by V coefficients where C is the number of CCM codes and V is the number of tokens in our concatenated bag-of-words. Cell $\{c, v\}$ contains the learned coefficient for CCM code c associated with token v.

Throughout the process of cross-validation and evaluation we use the model's accuracy as our evaluation metric. To reduce the overall complexity of the model and mitigate risks of the model overfitting the training data, we use an L2 norm to penalize coefficients that are very far from zero. We identify the best performing penalty term for regularization using five-fold cross-validation. After identifying the best performing hyperparameter for our L2 norm, we refit the model on the full 90 percent training data and evaluate the accuracy of predictions on the unseen 10 percent training data. This provides us with a final model that can be used to predict the most likely CCM code for each course at the two, four, and six digit level and provide a predicted probability for that code.

To identify the predicted CCM code for a course, we take the subject code and course catalog title and transform it into a bag-of-words following the steps described in Section IV. We then multiply this bag-of-words by the matrix of coefficients learned by our logistic regression model. Multiplying the D by V bag-of-words matrix by the V by C matrix learned in our model results in a D by C matrix where the value of cell $\{d, c\}$ contains the predicted value of document d associated with CCM code c. In this matrix, the code c with the largest value is the predicted CCM code for document d. The predicted probability of code c for document d is calculated by taking the softmax across all codes in C with respect to c for document d. The predicted probability provides an approximate measure of how confident the model is in its prediction, allowing analysts to incorporate this information into their analyses. If an analyst's research question relies on analyzing data on courses with relatively low predicted probabilities, the analyst may consider doing a manual validation of codes, using the predicted codes as a guide.

References

Bryan, M., Simone, S. (2012). *2010 College Course Map Technical Report*. National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2012/2012162rev.pdf.