Grouping Catalog Entries
Natural Language Processing

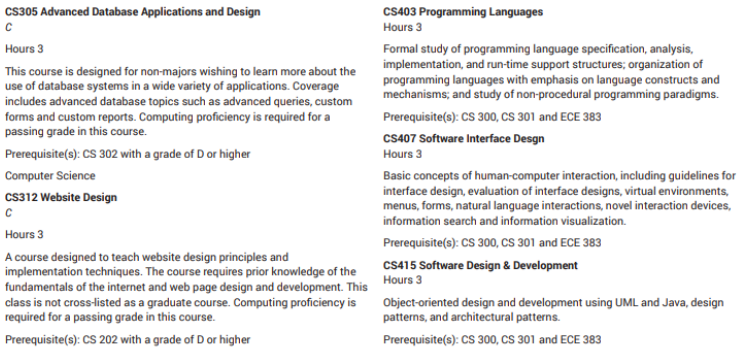
Jacob Light
Stanford University
jdlight@stanford.edu
GitHub: https://github.com/jacob-light/coursenet

1 Introduction

Applications of text to data methods are growing in popularity in the social sciences. One challenge social scientists face in transforming unstructured text into structured data is encoding patterns obvious visually but that may not be obvious to a computer. In my research, I use data from college and university course descriptions to study how the skills college students develop evolve over time. One basic challenge I face working across time and institutions is standardizing the names of departments. For example, computer science courses can be denoted as "Computer Science," "Electrical Engineering and Computer Science," "CS," etc.

In this project, I propose a deep learning method to standardize the names of departments across institutions. Although the application sounds specific, the capacity for a deep learning algorithm to detect similarities in courses across institutions can be extremely informative of similarities/differences in the content of courses (and, therefore, the skills graduates develop) at different colleges. The input for this project is a dataset of more than 3 million courses I have scraped from course catalogs published by a nationally diverse sample of colleges and universities. I will build a model that clusters departments based on course descriptions. The output will map departments to clusters.

An example of the course catalog data is in the image below.



CS305 Advanced Database Applications and Design
C

Hours 3

This course is designed for non-majors wishing to learn more about the use of database systems in a wide variety of applications. Coverage includes advanced database topics such as advanced queries, custom forms and custom reports. Computing proficiency is required for a passing grade in this course.

Prerequisite(s): CS 302 with a grade of D or higher

Computer Science

CS312 Website Design
C

Hours 3

A course designed to teach website design principles and implementation techniques. The course requires prior knowledge of the fundamentals of the internet and web page design and development. This class is not cross-listed as a graduate course. Computing proficiency is required for a passing grade in this course.

Prerequisite(s): CS 202 with a grade of D or higher

CS403 Programming Languages
Hours 3

Formal study of programming language specification, analysis, implementation, and run-time support structures; organization of programming languages with emphasis on language constructs and mechanisms; and study of non-procedural programming paradigms.

Prerequisite(s): CS 300, CS 301 and ECE 383

CS407 Software Interface Desgn
Hours 3

Basic concepts of human-computer interaction, including guidelines for interface design, evaluation of interface designs, virtual environments, menus, forms, natural language interactions, novel interaction devices, information search and information visualization.

Prerequisite(s): CS 300, CS 301 and ECE 383

CS415 Software Design & Development
Hours 3

Object-oriented design and development using UML and Java, design patterns, and architectural patterns.

Prerequisite(s): CS 300, CS 301 and ECE 383

Source: University of Alabama, 2019-20

2 Dataset and Features

The raw data for this project will come from college course catalogs from a sample of colleges and universities in the United States. Each catalog contains thousands of courses offered by the institution during an academic year. An entry in the course catalog dataset is a tuple of course id,

CS230: Deep Learning, Fall 2020, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

department name, and a text course description of approximately 3-5 sentences. Each department contains upwards of 30 courses, frequently more than 50.

3 Method

3.1 Pre-processing

From my dataset of course descriptions, I draw the full list of courses in each department[1] in the most recent year for which I have data. Typically, the courses were offered during the 2019-2020 academic year. In instances where a department has been eliminated or the name of a department has changed, I restrict to courses offered in the latest year I observe the department in my dataset. The resulting dataset contains approximately 350,000 courses. Next, I collapse the data from the course to the department level. I concatenate course descriptions for each course offered in a department to create a single description string for each department. I remove all stop words and punctuation and any words that contain numerals. Finally, I stem each of the words (Leopold and Kindermann [2002]).

The input to my model is a vectorized transformation of the concatenated department description. To vectorize the description, I use the feature extraction module of the sklearn package. I run the tf-idf vectorizer over the full dataset (Isa et al. [2008], Soucy and Mineau [2005]). To control the size of the feature vector, I restrict to the 2,500 words and bigrams with greatest tf-idf score to define my feature space. Thus, an input for my model is a 2500×1 vector corresponding to the frequency of each word/bigram in a department's concatenated course description. In robustness analysis, I test the gain from expanding to a $5,000 \times 1$ feature space.

I am working with labeled data. Each department has been manually labeled as one of 36 department categories (as defined in Blom et al. [2015]). Although the categories are generally distinct, a few of the categories may contain overlapping courses (for example, "Early and Elementary Education" versus "Education Fields, Other").

3.2 Implementation

The base model is a sequential fully connected model with 6 hidden layers. Each of the intermediate hidden layers uses a relu activation function. The final layer of the model uses a softmax activation to select among the 36 unique departments. Accordingly, I use the categorical cross-entropy loss function to evaluate the performance of the model . In robustness tests, I estimate the model using a neural network with 8 hidden layers to assess the performance gain from a larger network. Increasing the size of the network, without additional modification to hyperparameters in the model, does not lead to dramatic improvement in performance on training data.

A priori, I worry that the model will tend to over-fit narrow words or proper nouns rather than general skills students develop through classes in a department. For example, while a model could easily infer from words like "Proust" or "Joyce" that a department is "Literature and Languages Fields," a more useful model would weight words/phrases associated with the skills students develop, such as "composition" or "critical reading." To reduce overfitting, I include L1 regularization in each of the hidden layers. L1 regularization will tend towards sparsity in the words/phrases selected by the model, which should reduce overfitting. The scaling parameter of the L1 regularization penalty is a hyperparameter that I am tuning.

In subsequent sections, I summarize different specifications of my main model after tuning. Beyond the specifications described below, results from tuning two hyperparamters, the L1 regularization penalty λ and the mini-batch size, are omitted. The model performed best with a relatively small mini-batch size, which adds noise to the model, and a relatively high regularization penalty.

4 Preliminary Results

Figure 1 below summarizes the loss from 40 epochs of the base model. The model converges relatively quickly, after approximately 20 epochs. After quick convergence, the performance of the model plateaus. With L1 regularization, I would not expect the loss to converge to 0.

[1] Hereafter, by department, I refer to a unique department name string and institution pair.

To evaluate and troubleshoot the loss, I compare the performance of the model on the test, dev, and training datasets. My prior is that Bayes error is close to 0; most departments should be easily identifiable from the courses offered. Unavoidable error may come from extremely small departments, departments that fit into multiple or none of the categories, or poorly scraped courses. Table 1 summarizes the model accuracy for each dataset. The model performs well but has substantial room to improve. The avoidable error in the model is approximately 15-20%. I expect that I can reduce some of this error by increasing the size of the neural net or the number of features included in the feature vector. Over-fitting to the training dataset also appears to be a problem, although less of a problem than bias.

In Table 2, I compare the performance of my base model to two alternative models. The first alternative model keeps all elements of the model identical but adds two additional large hidden layers early in the network. The second alternative model increases the feature space of the base model from 2,500 words/bigrams to 5,000 words/bigrams. I compare model accuracy on the training, dev, and test sets in Table 2. To ensure comparability across models, the training, dev, and test observations are identical in each of the three models. [NTD - what happens]

As a debugging exercise and out of curiosity, it may be useful to examine performance of the model on different categories of classes. Table 3 summarizes the accuracy of predictions on the test set by true department category. The model performs well on the most common categories of departments, such as Literature and Languages, Accounting, Biology, Business, and Computer Science. The performance of the model is weaker in subjects with less coverage in my dataset: Journalism, for example, is a narrow major offered distinct from a broader Communications or English major at only a small number of schools. The few Journalism observations in the dataset, therefore, are often misclassified as English or Communications classes. Similar misclassification may arise in other instances of small majors closely connected to other larger major categories.

As a final debugging exercise, I test whether heterogeneity in institution characteristics contributes to misclassification error. My sample contains a mix of two- and four-year colleges and universities. Four-year institutions range from non-selective public colleges to highly selective state flagship and private universities. Two-year institutions are the least selective programs, often public community colleges that offer training in general skills or specialized trades. The mission statements of two- and four-year colleges differ: four-year institutions provide a wider range of skills for students to enter a national labor market, whereas two-year institutions focus on skills of local/regional relevance. To the extent that the missions of two- and four-year institutions differ, we might worry that heterogeneity within department categories may confound classification in the model.

Table 4 summarizes the accuracy of the base model by department for two- and four-year institutions separately. Performance of the model is generally comparable across institution category or slightly better for four-year institutions. Instances when the model performs better for two-year institutions typically correspond to departments where the terminal degree is an Associates Degree, such as in protective services (law enforcement) and in other education fields (teaching support staff, pre-K instruction).

5 Conclusion

The model described in this paper classifies course descriptions into department categories. The direct value of this model is to standardize classification when the same category can be described using different names (e.g. “Computer Science” versus “CS” versus “Computer Science and Informatics”...). However, the success of the classification exercise suggests that there may be common skills or tasks that identify a particular major. Understanding the specific skills that characterize a student’s degree has important implications for the labor market and public policy. Characterizing and summarizing these skills is left to future study.

References

Erica Blom, Brian C Cadena, and Benjamin J Keys. Investment over the business cycle: Insights from college major choice. 2015.

Dino Isa, Lam H Lee, VP Kallimani, and Rajprasad Rajkumar. Text document preprocessing with the bayes formula for classification using the support vector machine. *IEEE Transactions on*

Knowledge and Data engineering, 20(9):1264–1272, 2008.

Edda Leopold and Jörg Kindermann. Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46(1-3):423–444, 2002.

Pascal Soucy and Guy W Mineau. Beyond tfidf weighting for text categorization in the vector space model. 5:1130–1135, 2005.

Figure 1: Model loss - base model, 40 epochs

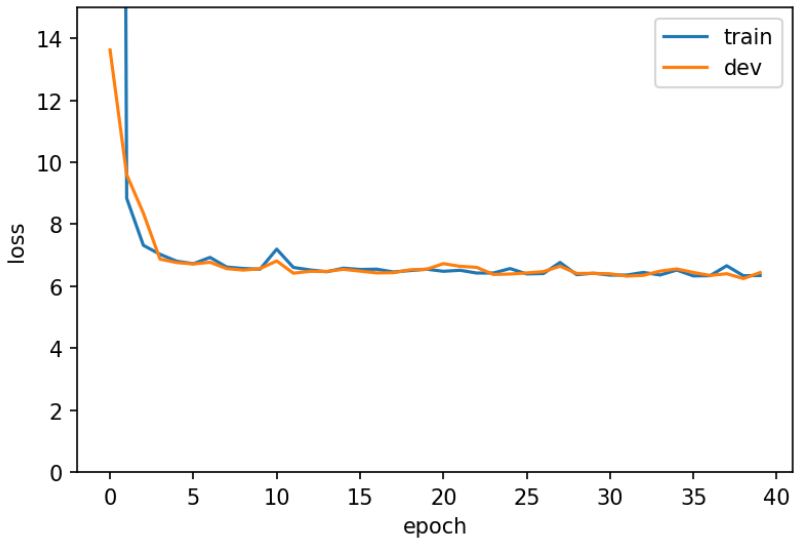


Table 1: Accuracy - base model, 40 epochs

split	accuracy
train	0.802912
dev	0.771930
test	0.735493

Table 2: Accuracy - compare across models

	split	model description	accuracy
0	train	Base Model	0.802912
0	train	Additional Layers	0.793380
0	train	Expand features space	0.800736
1	dev	Base Model	0.771930
1	dev	Additional Layers	0.761134
1	dev	Expand features space	0.767881
2	test	Base Model	0.735493
2	test	Additional Layers	0.731444
2	test	Expand features space	0.753036

Table 3: Prediction accuracy by department

department	count	accuracy
Accounting	137	0.99
Agriculture	249	0.61
Architecture	83	0.43
Biology Fields	562	0.94
Business Fields, not Finance	889	0.78
Chemistry and Pre-Med	211	0.67
Communications Fields	488	0.84
Computer-Related Fields	479	0.83
Early and Elementary Education	1012	0.83
Economics	183	0.87
Education Fields, Other	300	0.28
Engineering Fields	793	0.64
Environmental and Natural Resource Fields	73	0.23
Family and Consumer Sciences	185	0.54
Finance	105	0.96
Journalism	16	0.00
Leisure Studies	442	0.87
Liberal Arts and History Fields	823	0.87
Literature and Languages Fields	2090	0.94
Mathematics and Statistics	274	0.86
Natural Science Fields, Other	435	0.34
Nursing	157	0.94
Other Fields	270	0.76
Pharmacy	41	0.68
Physics	312	0.83
Political Science and International Relations	368	0.85
Pre-Law and Legal Studies	111	0.76
Protective Services	107	0.91
Psychology Fields	213	0.81
Public Affairs, Health, Policy	73	0.59
Social Science Fields, Other	986	0.56
Social Work	84	0.82
Sociology	153	0.79
Technical Engineering Fields	297	0.58
Technical Health Fields	580	0.64
Visual and Performing Arts	1225	0.89

Table 4: Compare prediction accuracy - 2- vs 4-year institutions

department	4-year Accuracy	2-year Accuracy
Accounting	0.96	1.00
Agriculture	0.69	0.51
Architecture	0.76	0.80
Biology Fields	0.81	0.73
Business Fields, not Finance	0.89	0.87
Chemistry and Pre-Med	0.87	0.97
Communications Fields	0.73	0.71
Computer-Related Fields	0.79	0.81
Early and Elementary Education	0.74	0.64
Economics	0.90	0.84
Education Fields, Other	0.20	0.26
Engineering Fields	0.81	0.71
Environmental and Natural Resource Fields	0.41	0.50
Family and Consumer Sciences	0.63	0.26
Finance	0.88	0.71
Journalism	0.00	NaN
Leisure Studies	0.90	0.85
Liberal Arts and History Fields	0.76	0.69
Literature and Languages Fields	0.91	0.96
Mathematics and Statistics	0.88	0.88
Natural Science Fields, Other	0.51	0.51
Nursing	0.85	0.70
Other Fields	0.76	0.82
Pharmacy	0.71	1.00
Physics	0.80	0.86
Political Science and International Relations	0.84	0.89
Pre-Law and Legal Studies	0.58	0.86
Protective Services	0.88	0.90
Psychology Fields	0.88	0.91
Public Affairs, Health, Policy	0.57	0.25
Social Science Fields, Other	0.83	0.89
Social Work	0.86	0.73
Sociology	0.93	0.96
Technical Engineering Fields	0.52	0.44
Technical Health Fields	0.83	0.91
Visual and Performing Arts	0.87	0.95