Report

1. Introduction: Project Goals 🡪 Joanna

-Problem: motivation, data set

1. Data set

1. Obtaining and cleaning data (challenges, related work) J

2. Data exploration V

3. Supervised learning (results, include related work, failures, parameters, details) J

4. Unsupervised learning (include related work, results, failures, parameters, details) V

5. Conclusion (both) – remaining ideas

6. Appendix: code

**Introduction**

**Choosing courses at UC Berkeley can be daunting. This world-class university offers about 5,000 different 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 addresses the following questions:

1.) Given a course description, can we predict what department the course belongs to?

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

**Data Set**

**Obtaining data**

**We scraped the course catalog in two different ways. First, we searched for “fall” and “spring” courses on the searchable version of the course catalog (**[**http://general-catalog.berkeley.edu/catalog/gcc\_search\_menu**](http://general-catalog.berkeley.edu/catalog/gcc_search_menu)**). Search results were copied and pasted into a text file, and then parsed into a dictionary with regular expressions.**

**The second way we obtained data was scraping the courses from each department’s web page (**[**http://general-catalog.berkeley.edu/gc/curricula.html**](http://general-catalog.berkeley.edu/gc/curricula.html)**) using beautiful soup (code: course\_catalog\_scrape.py). This method required a lot of hacky code to bypass inconsistencies across departments. Regular expressions were used to parse the data.**

**In both cases, the final form of the data was a dictionary of dictionaries stored in a json file. The sample entry in the dictionary shows the structure:**

**u'YIDDISH 102': {u'course\_num': u'102',**

**u'dept\_code': u'YIDDISH',**

**u'dept\_name\_full': u'Yiddish',**

**u'desc': u'Further intensive study of Yiddish, building on the foundation established in 101. Advanced grammar and introduction to the reading of original texts.',**

**u'title': u'Intermediate Yiddish',**

**Data cleaning and exploration**

**We scraped a total of XY# courses from the course catalog. Next, we cleaned and shortened our data set. First, we removed courses that were exact duplicates of others. Next, we removed 277 courses with a course number greater than 300, which correspond to teaching courses and other unusual departmental offerings. Finally, we removed courses with a course description exactly the same as another course’s description and fewer than 10 words. These reflected generic course descriptions such as “Beginner’s course” or “Supervised independent study” that many departments use and are meaningless for classification or grouping topically similar courses. After removing these courses, our data contained XY# of courses from 157 different departments. Figure 1 shows the distribution of number of courses each department.**

**Department Stats**

We used MapReduce to get a sum of the courses by department and found that there are 157 departments represented in the dataset (code: dept\_count.py). The number of courses listed in each department ranges from 1 to 212; the department with the most courses was for Public Health. Interestingly the School of Public Health does not list courses by subject or program area, so courses in Epidemiology and Health Management and Policy are listed under the same code. Other Schools have courses listed by their specific program such as the College of Engineering which has codes for each engineering discipline. This fact could influence the accuracy of our classifier which tried to predict which department a course belongs to. About 50 departments had 10 or fewer courses. The average number of courses per department was 37, with a median of 26 and a standard deviation of 38. The mode number of classes was 6. A histogram showing the number of courses per department is shown below.

**Course Description Stats**

The average number of words in a course description was 50 words with a median of 46 and standard deviation of 31. The mode was 48. The longest course description was 188 words. An example is S ASIAN 146 which has a very detailed description of the course:

S ASIAN 146: This course is designed to provide a dual chronological and thematic approach to the study of one of the greatest empires in human civilization: the Mughal Empire. Although the bulk of this course will focus on the Mughal Empire during its heyday between the 1550s and the early 1700s, careful attention will be paid to the larger historical and geographical contexts that both enabled the emergence and, ultimately, decentralization of Mughal power. In so doing, this course will not only study South Asia's complex history on its own terms but also examine the intricate web of political, economic, and social links that connected South Asia to the rest of the world. Simultaneously, this course will also pay particular attention to a series of common misconceptions that dog the study of pre-modern Islamic polities. Among them, the supposedly lesser role played by women in politics; the dogmatic and central role of Islam in "Muslim" states; and the economic and political superiority of Western Europe. Crucial to these questions also is an examination of the historiography and historiographical traditions that have come to define contemporary understanding of the Mughal Empire. (188 words)

Some courses had as few as one word and about 330 courses had 10 or fewer words. Some courses with such few words failed to explain what the course was about, such as “Research.” But some short descriptions do have some information about the course. Examples are shown below.

CHM ENG 299: Research. (1 word)

CHEM 199: Enrollment is restricted by regulations listed in the General Catalog. (10 words)

SLAVIC 234: Linguistic history and dialectology of Slovenian, Bulgarian, Macedonian, and Serbian/Croatian. (10 words)

A histogram showing the distribution of course description word count is shown below.

After examining the number of words in each description, we examined what the most common words in the dataset were. There were about 13,000 words in the course descriptions (unique\_words.py). Not surprisingly, the most common words were common stop-words such as: and, the, of. After removing 277 stop words, the top 20 words and their frequencies are shown below. Note that some very common words like “course”, “students”, and “semester” do not convey much meaning.

|  |  |  |  |
| --- | --- | --- | --- |
| course 3470  students 2373  topics 1432  research 1140  analysis 941 | study 924  design 916  semester 730  studies 720  social 717 | methods 690  listed 677  include 660  theory 641  including 638 | introduction 631  systems 621  development 610  issues 587  history 570 |

**Additional Findings**

We also discovered a few other interesting facts about the course dataset by exploring the data. About 680 courses are cross-listed with other departments; and cross listed courses might have different titles and description. We also found that some courses are very similar except for maybe just one or two word differences. Finally, a major finding was that certain special topics courses (like INFO 290 Data Mining) are actually not included in the online course catalog and there is no description. This is a weakness for this dataset since many of these special topics courses have descriptions elsewhere on the web but we were not able to access them from the main data set.

1. Supervised learning

METHODS

We developed a Naïve Bayes classifier (code: Bayes\_classifier v2.py) to classify a course to a department based on its course description. We saved 10% of our data to use as a test set, and used the remaining 90% as a training set (code: training\_set.py, your continued cleanup file). To ensure that we could test the accuracy of our classifier on courses of all departments, we used stratified random sampling to ensure that every department was represented in both the test and training sets.

The probability (P) of any course being in a department D given its course description of words w1, w2, w3….wi is:

P in D = P(w1 | D) \* P(w2 | D) \* … P(wi | D) \* P (D)

In other words, the probability of any course being in a department D depends on the conditional probabilities of each of the words in its course description being in that department and the probability that the course is in that department based on the size of the department alone. For each test tuple, the classifier calculates P in D for all 150 departments and then assigns the tuple to the department with the highest probability.

In Python, we built a dictionary of conditional probabilities with {word, department} as {key, value} pairs. The probability of P(wi | D) was calculated as the percentage of courses within D that contained at least one instance of wi. For example, if the word “modern” appeared in 3 out of 10 classes in the English department, the conditional probability of (“modern” | English) would be 30%. If a word in the testing tuple is not in the training set, we set the probability P(wi | D) as 1, effectively dropping this word from being used in the classification process. If a word is in the testing tuple and in the training set, but has never been used in any of the course descriptions in department for which we are calculating the probability, then P(wi | D) is technically 0. Since this would result in the entire P in D being zero, we used a Laplacian correction to inflate the numerator and denominator to yield a non-zero probability. For example, if the testing tuple word “ancient” had never appeared in any of the 10 courses that represent the English department in the training set, but had appeared elsewhere in the training set, the P(“ancient” | English) would be (0+inflation)/(10+ inflation). We varied our inflation parameter by orders of magnitude from 10-7 to 1. We repeated all trials with and without the stop words.

RESULTS

We achieved the best results when inflation was .00001 and stop words were dropped (Table X, Figure X). Using these parameters, 248 out of 629 (39.4%) testing tuples were classified to the correct department.

TABLE X Classifier accuracy.

| **Inflation parameter** | **Stopwords** | **Correctly classified** | **Accuracy** |
| --- | --- | --- | --- |
| 0.0000001 | in | 240 | 0.382 |
| 0.0000001 | out | 244 | 0.388 |
| 0.000001 | in | 241 | 0.383 |
| 0.000001 | out | 246 | 0.391 |
| 0.00001 | in | 243 | 0.386 |
| 0.00001 | out | 248 | 0.394 |
| 0.0001 | in | 238 | 0.378 |
| 0.0001 | out | 233 | 0.370 |
| 0.001 | in | 139 | 0.221 |
| 0.001 | out | 148 | 0.235 |
| 0.01 | in | 34 | 0.054 |
| 0.01 | out | 39 | 0.062 |
| 0.1 | in | 3 | 0.005 |
| 0.1 | out | 4 | 0.006 |
| 1 | in | 2 | 0.003 |
| 1 | out | 2 | 0.003 |

FIGURE X. Classifier accuracy.

Macintosh HD:Users:joannahsu:Documents:Courses:projectCourse:Results.pdf

The inflation parameter was very important in determining the accuracy of the classifier. When the inflation parameter is set too high, the conditional probabilities become very large, especially in departments with very few classes. The resulting overinflation results in almost all testing tuples being classified into the miniscule Telugu or Science and Technology departments. Removal of stop words only marginally improved classification accuracy. This may not be surprising given that the stop words probably are evenly represented across departments. However, removing the 277 stop words decreased the time the classification took.

TABLE X+1. Classifier results by department. Depart type abbreviations: ENG=engineering, LANG= language, INT=interdisciplinary, LIB=liberal arts, PROF=professional program, other.

|  |  |  |  |
| --- | --- | --- | --- |
| Department | Type | Accuracy | Number of courses in department |
| A RESEC | INT | 0.500 | 22 |
| AEROSPC | ENG | 1.000 | 7 |
| AFRICAM | LIB | 0.545 | 109 |
| AMERSTD | LIB | 0.333 | 24 |
| ANTHRO | LIB | 0.500 | 114 |
| ARABIC | LANG | 0.000 | 3 |
| ARCH | SCI | 0.667 | 93 |
| ART | ART | 0.800 | 47 |
| ASAMST | LIB | 0.400 | 42 |
| ASIANST | LIB | 0.000 | 10 |
| AST | SCI | 0.000 | 5 |
| ASTRON | SCI | 0.000 | 35 |
| BANGLA | LIB | 1.000 | 4 |
| BIO ENG | ENG | 0.444 | 85 |
| BIOLOGY | SCI | 1.000 | 5 |
| BIOPHY | SCI | 0.000 | 4 |
| BUDDSTD | LIB | 0.000 | 28 |
| CELTIC | LIB | 0.333 | 28 |
| CHEM | SCI | 0.222 | 91 |
| CHICANO | LIB | 0.250 | 34 |
| CHINESE | LANG | 0.667 | 32 |
| CHM ENG | ENG | 0.000 | 46 |
| CIV ENG | ENG | 0.500 | 139 |
| CLASSIC | LIB | 0.333 | 26 |
| CMPBIO | SCI | 1.000 | 4 |
| COG SCI | SCI | 1.000 | 15 |
| COLWRIT | LIB | 0.667 | 25 |
| COM LIT | LIB | 0.000 | 16 |
| COMPBIO | SCI | 0.000 | 2 |
| COMPSCI | ENG | 0.571 | 70 |
| CRIT TH | LIB | 0.000 | 3 |
| CY PLAN | LIB | 0.000 | 53 |
| DEMOG | INT | 0.000 | 22 |
| DEV STD | INT | 0.000 | 9 |
| DEVP | INT | 0.000 | 10 |
| DUTCH | LANG | 0.500 | 20 |
| EA LANG | LANG | 0.000 | 23 |
| EAEURST | LIB | 1.000 | 6 |
| ECON | INT | 0.000 | 65 |
| EDUC | LIB | 0.889 | 88 |
| EL ENG | ENG | 0.500 | 112 |
| ENE RES | INT | 0.000 | 31 |
| ENGIN | ENG | 0.200 | 42 |
| ENGLISH | LIB | 1.000 | 5 |
| ENV DES | ENG | 0.000 | 27 |
| ENV SCI | SCI | 0.000 | 7 |
| ENVECON | INT | 0.000 | 29 |
| EPS | SCI | 0.500 | 34 |
| ESPM | INT | 0.375 | 148 |
| ETH GRP | LIB | 0.500 | 10 |
| ETH STD | LIB | 0.000 | 38 |
| EURA ST | LIB | 1.000 | 4 |
| EWMBA | INT | 0.000 | 79 |
| FILIPN | LANG | 1.000 | 6 |
| FILM | LIB | 0.750 | 35 |
| FOLKLOR | LIB | 1.000 | 3 |
| FRENCH | LANG | 0.444 | 76 |
| GEOG | SCI | 0.000 | 39 |
| GERMAN | LANG | 0.429 | 66 |
| GMS | INT | 0.000 | 4 |
| GPP | INT | 0.000 | 3 |
| GREEK | LANG | 0.000 | 9 |
| GWS | LIB | 0.333 | 53 |
| HIN-URD | LIB | 1.000 | 10 |
| HISTART | LIB | 0.000 | 44 |
| HISTORY | LIB | 0.727 | 100 |
| HMEDSCI | SCI | 0.500 | 18 |
| IAS | LIB | 0.250 | 34 |
| IND ENG | ENG | 0.167 | 58 |
| INFO | ENG | 0.571 | 64 |
| INTEGBI | SCI | 0.417 | 123 |
| ISF | INT | 0.000 | 14 |
| ITALIAN | LANG | 0.200 | 44 |
| JAPAN | LANG | 0.500 | 26 |
| JEWISH | LANG | 0.000 | 8 |
| JOURN | LIB | 0.800 | 43 |
| KHMER | LANG | 1.000 | 6 |
| KOREAN | LANG | 1.000 | 27 |
| LAN PRO | LANG | 1.000 | 3 |
| LATAMST | LIB | 0.000 | 12 |
| LATIN | LANG | 0.000 | 8 |
| LD ARCH | INT | 0.250 | 68 |
| LEGALST | LIB | 0.000 | 36 |
| LGBT | LIB | 0.500 | 11 |
| LINGUIS | LIB | 0.500 | 39 |
| LNS | INT | 0.400 | 45 |
| M E ST | LIB | 0.000 | 11 |
| MALAY/I | LANG | 0.000 | 5 |
| MAT SCI | SCI | 0.333 | 49 |
| MATH | ENG | 0.667 | 84 |
| MBA | PROF | 0.250 | 110 |
| MCELLBI | SCI | 0.529 | 157 |
| MEC ENG | ENG | 0.429 | 136 |
| MED ST | LIB | 0.000 | 3 |
| MEDIAST | LIB | 0.000 | 24 |
| MFE | ENG | 0.000 | 20 |
| MIL AFF | LIB | 0.000 | 7 |
| MIL SCI | SCI | 1.000 | 9 |
| MUSIC | ART | 0.889 | 88 |
| NAT RES | INT | 0.000 | 3 |
| NATAMST | LIB | 0.000 | 32 |
| NAV SCI | SCI | 0.000 | 5 |
| NE STUD | LIB | 0.333 | 25 |
| NEUROSC | SCI | 0.500 | 15 |
| NSE | ENG | 0.000 | 6 |
| NUC ENG | ENG | 0.600 | 43 |
| NUSCTX | SCI | 0.500 | 44 |
| NWMEDIA | LIB | 0.000 | 9 |
| OPTOM | PROF | 1.000 | 42 |
| PACS | LIB | 0.000 | 26 |
| PB HLTH | INT | 0.773 | 210 |
| PERSIAN | LANG | 0.000 | 4 |
| PHDBA | LIB | 0.500 | 36 |
| PHILOS | LIB | 0.200 | 43 |
| PHYS ED | PROF | 1.000 | 18 |
| PHYSICS | SCI | 0.625 | 69 |
| PLANTBI | SCI | 0.250 | 69 |
| POL SCI | LIB | 0.455 | 101 |
| POLECON | INT | 0.000 | 18 |
| PORTUG | LANG | 1.000 | 6 |
| PSYCH | SCI | 0.167 | 50 |
| PUB POL | INT | 0.600 | 45 |
| PUNJABI | LANG | 1.000 | 4 |
| RELIGST | LIB | 0.000 | 12 |
| RHETOR | LIB | 0.000 | 50 |
| S ASIAN | LIB | 0.500 | 18 |
| S SEASN | LIB | 0.000 | 20 |
| SANSKR | LANG | 1.000 | 6 |
| SCANDIN | LANG | 0.600 | 44 |
| SCMATHE | PROF | 0.000 | 6 |
| SEASIAN | LIB | 0.000 | 8 |
| SLAVIC | LANG | 0.444 | 87 |
| SOC WEL | LIB | 0.286 | 61 |
| SOCIOL | SCI | 0.125 | 70 |
| SPANISH | LANG | 0.250 | 33 |
| STAT | ENG | 0.500 | 44 |
| STS | LIB | 0.000 | 2 |
| TAGALG | LANG | 1.000 | 4 |
| TAMIL | LANG | 1.000 | 6 |
| TELUG | LANG | 1.000 | 2 |
| THAI | LANG | 1.000 | 6 |
| THEATER | ART | 0.125 | 71 |
| TIBETAN | LANG | 0.500 | 14 |
| UGBA | INT | 0.143 | 69 |
| UGIS | INT | 0.333 | 51 |
| VIETNMS | LANG | 1.000 | 6 |
| VIS SCI | SCI | 0.000 | 26 |
| VIS STD | LIB | 0.000 | 7 |
| XMBA | PROF | 0.000 | 27 |
| YIDDISH | LANG | 0.000 | 3 |

FIGURE X+1

Macintosh HD:Users:joannahsu:Documents:Courses:projectCourse:accuracy.pdf

The classifier ranged in success from 0% to 77% across departments. The classifier performed best with the largest department, public health, classifying 17 out of 22 test tuples correctly. With one exception, departments with more than 100 courses had at least 40% classification accuracy. However, because test tuples from small and medium-sized departments could be classified both very accurately or very inaccurately, there was no linear correlation between number of courses in a department and classification accuracy. Average classification accuracy also did not vary between types of departments (liberal arts, languages, interdisciplinary, math and engineering, science).

It is remarkable that our classifier can classify a test tuple to one of 150 departments correctly with 40% accuracy. However, there is still room for improvement! One major way to further improve the classifier would be to use term frequency and inverse document frequency to teach the classifier to weigh certain words more.

CONCLUSION

Hunch: lots of subject areas within large departemtns like public health vs. engineering departments