**APPENDICES**

**Appendix 1: Naïve Bayes classifier methods**

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

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.

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.

**Appendix 2: Naïve Bayes Classifier Results**

TABLE A-1. 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 A-1. Classifier accuracy.

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

TABLE A3. Classifier results by department. Depart type abbreviations: ENG=engineering, LANG= language, INT=interdisciplinary, LIB=liberal arts, PROF=professional program, other.

|  |  |  |  |
| --- | --- | --- | --- |
| **Department** | **Type** | **Accuracy** | **Num. 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 |

Appendix 2: Data Set Statistics

Number of Words in Course Description