Question 1:

|  |  |  |  |
| --- | --- | --- | --- |
| Language | Prior | Count | Count/Total Count |
| TEL | 0.099329109 | 533 | 0.099329109 |
| HIN | 0.065598211 | 352 | 0.065598211 |
| SPA | 0.083861349 | 450 | 0.083861349 |
| KOR | 0.103801714 | 557 | 0.103801714 |
| FRA | 0.088147596 | 473 | 0.088147596 |
| JPN | 0.103801714 | 557 | 0.103801714 |
| ARA | 0.092061126 | 494 | 0.092061126 |
| ITA | 0.096161014 | 516 | 0.096161014 |
| TUR | 0.093924711 | 504 | 0.093924711 |
| ZHO | 0.110510622 | 593 | 0.110510622 |
| DEU | 0.062802833 | 337 | 0.062802833 |

Question 2:

Question 3:

|  |  |
| --- | --- |
| alpha | accuracy |
| 0.01 | 0.725752508 |
| 0.05 | 0.735785953 |
| 0.1 | 0.747491639 |
| 0.2 | 0.74916388 |
| 0.5 | 0.732441472 |
| 1 | 0.68729097 |
| 2 | 0.581939799 |
| 5 | 0.382943144 |

Based on the alpha tuning, alpha = 0.2 gives the best accuracy.

Question 4:

|  |  |
| --- | --- |
| alpha | accuracy |
| 0.01 | 0.719063545 |
| 0.05 | 0.737458194 |
| 0.1 | 0.737458194 |
| 0.2 | 0.732441472 |
| 0.5 | 0.7090301 |
| 1 | 0.678929766 |
| 2 | 0.575250836 |
| 5 | 0.377926421 |

Question 5:

dev: 0.7491638795986622

test: 0.7152317880794702

Using the alpha = 0.05 from lemmatization

dev: 0.7374581939799331

test: 0.6771523178807947

Question 6:

At the 24th iteration (iteration 23), the training data is completely separated because the accuracy is 1.0. While the 13th iteration (iteration 12) produces the highest accuracy when testing on the dev set, the 8th iteration would be better in terms

Question 7:

|  |  |  |  |
| --- | --- | --- | --- |
| Features | Number of iterations to separation | Best iteration | Test Accuracies using best iteraton |
| Lemmatization, uppercase, 2-grams | 11 | 10 | 0.6939799331103679 |
| Lemmatization, uppercased | 30 | 9 | 0.6291390728476821 |
| Lemmatization | 27 | 5 | 0.6490066225165563 |
| uppercased | 22 | 7 | 0.652317880794702 |
| 2-grams | 13 | 6 | 0.6870860927152318 |
| No features | 24 | 4 | 0.6705298013245033 |

Question 8:

Part A:

['ARA', 'DEU', 'FRA', 'HIN', 'ITA', 'JPN', 'KOR', 'SPA', 'TEL', 'TUR', 'ZHO']

'ARA' [42 1 0 0 0 1 1 4 5 5 1]

'DEU' [ 1 34 4 0 0 1 0 0 0 0 1]

'FRA' [ 2 2 41 1 1 0 0 2 1 1 0]

'HIN' [ 2 1 1 17 0 0 0 0 5 3 1]

'ITA' [ 3 2 4 1 36 2 0 4 0 2 0]

'JPN' [ 4 1 2 0 0 42 9 1 1 2 0]

'KOR' [ 2 1 1 2 0 8 41 0 0 1 5]

'SPA' [ 6 2 2 1 4 1 1 35 1 6 2]

'TEL' [ 2 0 0 7 0 1 0 1 52 1 0]

'TUR' [ 0 3 1 1 0 2 5 0 0 42 1]

'ZHO' [ 3 0 2 1 0 5 4 1 1 3 45]

Part B:

max weights for ARA: ['ALOT OF', 'EVERY THING', 'AND A', ', AND', 'REASON .', 'MANY REASON', 'TO HELP', 'OF THAT', 'IN ADDITION', 'ANY THING']

min weights for ARA: ['. IF', '. BUT', 'THAT PEOPLE', 'TO ENJOY', 'PEOPLE DO', 'ARE THE', 'HA BEEN', 'ABLE TO', 'A MORE', 'MORE IMPORTANT']

max weights for DEU: [', THAT', '. BUT', 'THE STATEMENT', 'AND THEREFORE', ', BECAUSE', 'I WOULD', 'ONE HAND', '. FURTHERMORE', 'HA TO', 'ALL IN']

min weights for DEU: [', AND', 'WE CAN', 'THAT IS', 'TIME TO', ', I', 'TO MAKE', 'IF WE', 'THEM TO', ', YOUNG', 'WHEN I']

max weights for FRA: ['. INDEED', 'IS A', 'IN FACT', 'INDEED ,', 'TO CONCLUDE', 'EVEN IF', 'TO TAKE', 'EXPERIENCE .', 'NOWADAYS ,', ', OR']

min weights for FRA: ['THE IDEA', 'THE PEOPLE', 'IN MY', 'AGREE THAT', 'WHICH MAKE', 'THERE ARE', 'TO GET', 'WHEN I', 'SHOULD BE', 'NOT ONLY']

max weights for HIN: ['A COMPARED', 'IN TODAY', 'OLD AGE', 'RISK AND', 'SAY THAT', 'BUT HE', 'NUMBER OF', 'A WELL', 'OF LIFE', 'TO ME']

min weights for HIN: ['. FINALLY', 'BASED ON', 'OUR LIFE', '. IF', 'HARD TO', '. BECAUSE', 'AND THE', 'BECAUSE THE', 'PEOPLE WILL', 'IN THE']

max weights for ITA: ['I THINK', 'THINK THAT', 'POSSIBILITY TO', 'TO IMPROVE', 'THAT IN', 'PEOPLE THAT', 'THE PAST', ', IN', 'A SPECIFIC', 'IN ITALY']

min weights for ITA: ['. BECAUSE', '. BUT', 'BECAUSE OF', '. THERE', 'OVER THE', 'ONE HAND', '. YOU', 'WHICH IS', '. ALSO', 'GOING TO']

max weights for JPN: ['IN JAPAN', 'JAPAN ,', ', AND', '. THEREFORE', 'I DISAGREE', '. IF', '. FROM', 'IF PEOPLE', 'JAPAN .', 'RECENTLY ,']

min weights for JPN: ['ALL THE', 'OF TIME', 'ARE A', 'TO GIVE', 'LIFE ,', 'FOR A', 'IS A', 'IN MY', 'THE YOUNG', 'ADVERTISEMENT ,']

max weights for KOR: ['IN KOREA', 'KOREA ,', 'THESE DAY', '. HOWEVER', 'EVEN THOUGH', 'HOWEVER ,', 'ALSO ,', '. ALSO', 'REASON THAT', ', THERE']

min weights for KOR: ['FOR ME', 'OF THE', 'THINK THAT', 'OF VIEW', 'IMPORTANT FOR', 'AT THE', 'THEY MAY', ', BECAUSE', ', TO', 'IN JAPAN']

max weights for SPA: [', IS', 'PEOPLE IS', 'OTHER HAND', 'THAT ARE', 'A BETTER', ', ETC', 'IN THEIR', 'TRYING TO', 'IDEA THAT', ', AND']

min weights for SPA: ['FROM THE', '. IT', 'ACCORDING TO', 'AND SO', 'I WANT', 'PEOPLE .', 'BETTER THAN', 'WHICH ARE', 'ON .', 'TODAY ,']

max weights for TEL: ['I STRONGLY', 'THE ABOVE', 'ALL THE', '. FINALLY', 'MAY BE', 'WHEN COMPARED', 'IN THE', 'THE STATEMENT', '. EVERY', '. BY']

min weights for TEL: ['HOWEVER ,', ', YOU', 'DO NOT', 'I THINK', '. HOWEVER', 'A WELL', ', AND', 'AND A', '. FIRST', 'YOU WILL']

max weights for TUR: ['. BECAUSE', 'CAN NOT', 'START TO', 'OF THIS', 'MAKE U', 'IN TURKEY', 'MUCH MORE', 'THE IDEA', 'THIS WAY', 'OF THESE']

min weights for TUR: [', AND', ', THE', 'TO KNOW', 'ENJOY THEIR', ', BUT', 'AGREE WITH', 'LEARN FACT', 'A GOOD', 'A LOT', 'HIS LIFE']

max weights for ZHO: ['. TAKE', 'OPINION ,', 'ENJOY THE', "WO N'T", ', THE', 'TO TRY', 'TIME ON', 'DIFFERENT PEOPLE', 'PEOPLE MAY', ', AND']

min weights for ZHO: ['EVEN IF', '\*\*\*bias\_term\*\*\*', 'THAT ARE', 'RISK AND', 'ABLE TO', 'TRYING TO', 'THE TIME', 'SITUATION .', 'HAVE TO', 'AND THAT']

bias feature for ZHO: -17

Part C:

precision for ARA: 0.626865671642

recall for ARA: 0.7

F1 for ARA: 0.661417322835

bias feature for ARA: 12

precision for DEU: 0.723404255319

recall for DEU: 0.829268292683

F1 for DEU: 0.772727272727

bias feature for DEU: -10

precision for FRA: 0.706896551724

recall for FRA: 0.803921568627

F1 for FRA: 0.752293577982

bias feature for FRA: -5

precision for HIN: 0.548387096774

recall for HIN: 0.566666666667

F1 for HIN: 0.55737704918

bias feature for HIN: 6

precision for ITA: 0.878048780488

recall for ITA: 0.666666666667

F1 for ITA: 0.757894736842

bias feature for ITA: 4

precision for JPN: 0.666666666667

recall for JPN: 0.677419354839

F1 for JPN: 0.672

bias feature for JPN: -2

precision for KOR: 0.672131147541

recall for KOR: 0.672131147541

F1 for KOR: 0.672131147541

bias feature for KOR: 1

precision for SPA: 0.729166666667

recall for SPA: 0.573770491803

F1 for SPA: 0.642201834862

bias feature for SPA: -5

precision for TEL: 0.787878787879

recall for TEL: 0.8125

F1 for TEL: 0.8

bias feature for TEL: 12

precision for TUR: 0.636363636364

recall for TUR: 0.763636363636

F1 for TUR: 0.694214876033

bias feature for TUR: 4

precision for ZHO: 0.803571428571

recall for ZHO: 0.692307692308

F1 for ZHO: 0.743801652893