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:

Part B:

Part C: