Naïve Bayes Model

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:

|  |  |  |
| --- | --- | --- |
| Label | priorProbs | Label Count |
| TUR | 0.095317726 | 57 |
| JPN | 0.100334448 | 60 |
| ARA | 0.085284281 | 51 |
| DEU | 0.056856187 | 34 |
| KOR | 0.100334448 | 60 |
| TEL | 0.10367893 | 62 |
| HIN | 0.078595318 | 47 |
| SPA | 0.086956522 | 52 |
| ITA | 0.088628763 | 53 |
| ZHO | 0.115384615 | 69 |
| FRA | 0.088628763 | 53 |

If we classified doc in the dev by the classifier with the highest prior probability from the training data set (ZHO with P(Y) = 0.110510622), the **majority class baseline**accuracy would be 0.115384615.

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 |

Lemmatization lowers the optimal alpha by almost 2%. A possible reason is that certain translations of words are used by different native speakers. Dev was

Question 5:

Using the alpha = 0.05

dev: 0.7491638795986622

test: 0.7152317880794702

Using the alpha = 0.05 and lemmatization

dev: 0.7374581939799331

test: 0.6771523178807947

The difference between the the naïve bayes models without lemmatization is approximate 3% different. Since dev was used as the tuning dataset, there may have been slight over fitting, but not anything significant. As for using lemmatization, the percentages are significantly different. This is probably because the lemmatization process could be considered as another form of fitting.