1. How many documents are there for each language? Add a line of code to display the learned prior distribution over classes. What is it? Ensure that it is correct.

|  |  |
| --- | --- |
| **language** | **count** |
| ARA | 494 |
| DEU | 337 |
| FRA | 473 |
| HIN | 352 |
| ITA | 516 |
| JPN | 557 |
| KOR | 557 |
| SPA | 450 |
| TEL | 533 |
| TUR | 504 |
| ZHO | 593 |

**A line for printing prior distribution**:

for y in self.priorProbs:

print ((y, self.priorProbs[y]))

**Prior Probabilities:**

('FRA', 0.08814759597465524)

('TEL', 0.09932910920611256)

('JPN', 0.10380171449869549)

('TUR', 0.09392471114424152)

('DEU', 0.06280283265001864)

('ITA', 0.09616101379053299)

('ARA', 0.0920611256056653)

('ZHO', 0.11051062243756989)

('SPA', 0.08386134923592993)

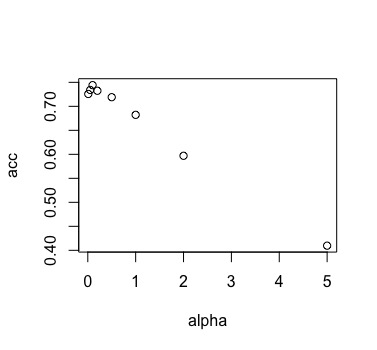
('KOR', 0.10380171449869549)

('HIN', 0.06559821095788297)

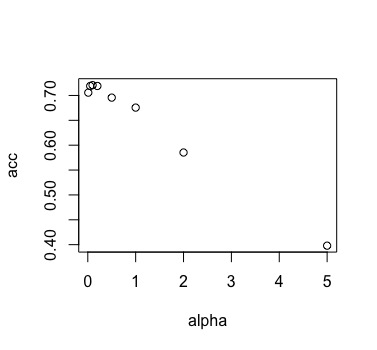
2.How many documents are there for each language in the dev set? What would be the **majority class baseline** accuracy on the dev set?

|  |  |
| --- | --- |
| **language** | **count** |
| ARA | 51 |
| DEU | 34 |
| FRA | 53 |
| HIN | 47 |
| ITA | 53 |
| JPN | 60 |
| KOR | 60 |
| SPA | 52 |
| TEL | 62 |
| TUR | 57 |
| ZHO | 69 |

**majority class baseline** accuracy: 0.6822742474916388

3. Tune the value of α by measuring performance on the dev set. Try α ∈ {0.01, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0}. Create a table or plot of each α value and the corresponding dev accuracy

|  |  |
| --- | --- |
| **alpha** | **accuracy** |
| 0.01 | 0.725752508 |
| 0.05 | 0.734113712 |
| 0.1 | 0.744147157 |
| 0.2 | 0.732441472 |
| 0.5 | 0.719063545 |
| 1 | 0.682274247 |
| 2 | 0.596989967 |
| 5 | 0.409698997 |

4. Lemmatization is a way to counteract the sparsity introduced by morphological inflection

|  |  |
| --- | --- |
| **alpha** | **accuracy** |
| 0.01 | 0.705685619 |
| 0.05 | 0.719063545 |
| 0.1 | 0.720735786 |
| 0.2 | 0.719063545 |
| 0.5 | 0.695652174 |
| 1 | 0.675585284 |
| 2 | 0.585284281 |
| 5 | 0.397993311 |

So for both lemmatize or baseline, the best α turns out to be 0.1, lemmatization doesn’t affect the best value of α, but does affect the accuracy. That’s because by lemmatization, feature number decreases, and thus some spelling mistakes that would help a lot with prediction get corrected.

5. Final Result

|  |  |  |
| --- | --- | --- |
| **alpha** | **baseline** | **lemma** |
| 0.1 | 0.700331126 | 0.675496689 |

The best model is baseline model with α = 0.1.