Perceptron

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:

I did two separate iteration checks. The first was to determine where the training data completely separated, and the optimal iteration was generally when the dev accuracy first peaked. In some cases, where the accuracy increases dramatically after the first peak

|  |  |  |  |
| --- | --- | --- | --- |
| Features | Number of iterations to separation | optimal iteration | Test Accuracies using best iteration |
| Lemmatization, uppercase, 2-grams | 14 | 6 | 0.7090301003344481 |
| Lemmatization, uppercased | Does not separate within 30 iterations | 9 | 0.6291390728476821 |
| Lemmatization | 27 | 5 | 0.6490066225165563 |
| uppercased | 22 | 7 | 0.652317880794702 |
| 2-grams | 13 | 6 | 0.6870860927152318 |

Using the combination of Lemmatization, uppercase, and 2-gram, I found that the test accuracy was 0.706953642384106 stopping the learning at iteration 6.

Question 8:

Part A:

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

41 0 2 3 1 1 0 4 3 4 1

1 30 4 3 0 1 0 0 0 0 2

2 1 36 2 3 1 0 2 1 1 2

2 0 1 18 0 0 0 0 6 2 1

2 2 2 0 43 1 0 3 0 0 1

2 0 2 1 0 47 6 0 1 2 1

0 1 1 0 0 13 36 0 0 1 9

3 1 2 2 6 3 1 34 1 5 3

1 0 0 14 0 1 0 1 47 0 0

2 3 0 2 1 2 6 0 1 36 2

2 0 1 1 0 6 4 2 0 0 49

Part B: please see (test weight.tsv for full table)

Part C:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Language | Precision | Recall | F1 | Weight of bias |
| ARA | . 626865671642 | .7 | . 661417322835 | 12 |
| DEU | . 723404255319 | . 829268292683 | . 772727272727 | -10 |
| FRA | . 706896551724 | . 803921568627 | . 752293577982 | -5 |
| HIN | . 548387096774 | . 566666666667 | . 55737704918 | 6 |
| ITA | . 878048780488 | . 666666666667 | . 757894736842 | 4 |
| JPN | . 666666666667 | . 677419354839 | . 672 | -2 |
| KOR | . 672131147541 | . 672131147541 | . 672131147541 | 1 |
| SPA | . 729166666667 | . 573770491803 | . 642201834862 | -5 |
| TEL | . 787878787879 | . 8125 | .8 | 12 |
| TUR | . 636363636364 | . 763636363636 | . 694214876033 | 4 |
| ZHO | .803571428571 | . 692307692308 | . 743801652893 | -17 |

One key thing that I noticed was with language ZHO. In the prior probabilities for but the training and dev documents, ZHO had the highest value. The bias feature for ZHO is unusual because it is on the the lowest feature weights. This occurred more likely because many docs were classified in the training set as ZHO but were not.