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']

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

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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' |

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.