1.

|  |  |  |  |
| --- | --- | --- | --- |
| Code | Language | Train | Dev |
| ARA | Arabic | 494 | 51 |
| DEU | German | 337 | 34 |
| FRA | French | 473 | 53 |
| HIN | Hindi | 352 | 47 |
| ITA | Italian | 516 | 53 |
| JPN | Japanese | 557 | 60 |
| KOR | Korean | 557 | 60 |
| SPA | Spanish | 450 | 52 |
| TEL | Telugu | 533 | 62 |
| TUR | Turkish | 504 | 57 |
| ZHO | Chinese | 593 | 69 |

See lang\_num.sh in folders dev and train for getting the number.

Baseline will be guessing every passage as the most frequent language which is Chinese. The probability of guessing correctly in dev is 69/sum of all languages = 69/598 = 11.54%. This probability can be used as the majority class baseline accuracy.

2.

The training data is never truly separated. It never reaches 100% accuracy in 30 iterations. The accuracy is still slowly increasing in the last few iterations. 10 iterations seem to be the best. The train set accuracy is 67.70%, the dev set accuracy is 55.85% and the final test set accuracy is 52.98%. The dev set accuracy is the same after the next iteration. After optimizing performance by selecting the 100 most common features, the highest dev accuracy drops to 50%.

3.

|  |  |  |  |
| --- | --- | --- | --- |
| Features (100 most common features) | Iterations (max dev accuracy) | Train Accuracy (max dev accuracy) | Max Dev Accuracy |
| Unigram | 28 | 61.03% | 50.83% |
| Bigram | 8 | 98.06% | 51.67% |
| Trigram | 4 | 99.16% | 41.64% |
| Char Unigram | 28 | 26.89% | 23.91% |
| Char Bigram | 29 | 39.39% | 36.62% |
| Word Length | 28 | 11.96% | 12.88% |
| Sentence Length | 20 | 15.41% | 13.38% |
| Uni\_bi\_trigram + uni\_bi\_char | 28 | 61.03% | 50.84% |
| Lower + unigram | 30 | 62.72% | 51.34% |
| Uni\_bi\_char | 30 | 23.07% | 20.90% |
| Uni\_bi\_trigram | 29 | 62.99% | 52.17% |

The highest accuracy on dev set is having unigram and bigram and trigram feature. Most of the features still have space of improvement by running more iterations, but the weights of maximum dev accuracy can be already overfitting. It takes too long to run over 30 iterations. If I have a super computer, I will run all features and much more iterations to check if the weights can still improve or are already overfitting. Selecting 100 most common features is a good way to reduce training time. However, it reduces the accuracy. The uncommon features can be useful in making decisions.

4.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ARA | DEU | FRA | HIN | ITA | JPN | KOR | SPA | TEL | TUR | ZHO | Total |
| ARA | 31 | 0 | 7 | 0 | 2 | 2 | 5 | 2 | 4 | 4 | 3 | 60 |
| DEU | 0 | 12 | 7 | 3 | 2 | 3 | 2 | 6 | 0 | 3 | 3 | 41 |
| FRA | 3 | 1 | 27 | 1 | 4 | 2 | 4 | 3 | 0 | 2 | 4 | 51 |
| HIN | 3 | 0 | 4 | 3 | 0 | 0 | 1 | 3 | 10 | 3 | 3 | 30 |
| ITA | 0 | 3 | 8 | 1 | 23 | 3 | 1 | 7 | 0 | 3 | 5 | 54 |
| JPN | 1 | 0 | 2 | 0 | 0 | 30 | 19 | 0 | 0 | 3 | 7 | 62 |
| KOR | 3 | 1 | 1 | 1 | 0 | 13 | 34 | 1 | 2 | 2 | 3 | 61 |
| SPA | 10 | 0 | 8 | 0 | 2 | 6 | 10 | 15 | 0 | 7 | 3 | 61 |
| TEL | 10 | 0 | 3 | 6 | 0 | 2 | 5 | 2 | 26 | 5 | 5 | 64 |
| TUR | 3 | 1 | 2 | 0 | 1 | 4 | 7 | 0 | 2 | 28 | 7 | 55 |
| ZHO | 3 | 0 | 4 | 2 | 1 | 3 | 9 | 2 | 0 | 5 | 36 | 65 |
| Total | 67 | 18 | 73 | 17 | 35 | 68 | 97 | 41 | 44 | 65 | 79 | 604 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ARA |  |  |  |  |  |  |  |  |  |  |
| 10 highest-weighted features | g\_bias | . | to | the | and | of | , | that | in | is |
| weights | 5957 | 5928 | 5920 | 5899 | 5864 | 5739 | 5715 | 5683 | 5644 | 5571 |
| 10 lowest-weighted features | Parking | Houses | 2026 | cowded | babies | 100000 | continuous | roof | compromise | Speciality |
| weights | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| bias | 5957 |  |  |  |  |  |  |  |  |  |
| DEU |  |  |  |  |  |  |  |  |  |  |
| 10 highest-weighted features | . | g\_bias | to | the | of | a | and | is | in | , |
| weights | 5795 | 5795 | 5766 | 5766 | 5730 | 5666 | 5665 | 5626 | 5559 | 5508 |
| 10 lowest-weighted features | learnt | specalized | expand | konwledge | minus | plus | method | opionion | stand | critics |
| weights | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| bias | 5795 |  |  |  |  |  |  |  |  |  |
| FRA |  |  |  |  |  |  |  |  |  |  |
| 10 highest-weighted features | . | to | g\_bias | of | the | , | and | is | a | that |
| weights | 5879 | 5879 | 5879 | 5839 | 5821 | 5781 | 5757 | 5693 | 5626 | 5581 |
| 10 lowest-weighted features | chair | wheel | binary | ascending | follows | descending | guess | halfs | divide | crucial |
| weights | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| bias | 5879 |  |  |  |  |  |  |  |  |  |
| HIN |  |  |  |  |  |  |  |  |  |  |
| 10 highest-weighted features | g\_bias | . | to | the | and | of | is | in | that | a |
| weights | 5984 | 5982 | 5950 | 5919 | 5852 | 5852 | 5771 | 5636 | 5517 | 5450 |
| 10 lowest-weighted features | beyound | throws | peolpe | diagree | witht | taker | predatory | cutting | ascpects | han |
| weights | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| bias | 5984 |  |  |  |  |  |  |  |  |  |
| ITA |  |  |  |  |  |  |  |  |  |  |
| 10 highest-weighted features | the | . | g\_bias | to | of | and | a | that | , | is |
| weights | 5838 | 5838 | 5838 | 5823 | 5803 | 5798 | 5778 | 5749 | 5739 | 5656 |
| 10 lowest-weighted features | SUCCESS | Working | ONESELF | BETTER | TO | WANT | thinghs | avaleable | investments | middle-class |
| weights | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| bias | 5838 |  |  |  |  |  |  |  |  |  |
| JPN |  |  |  |  |  |  |  |  |  |  |
| 10 highest-weighted features | . | , | g\_bias | to | and | of | the | that | is | I |
| weights | 5952 | 5952 | 5952 | 5919 | 5874 | 5745 | 5738 | 5555 | 5525 | 5430 |
| 10 lowest-weighted features | Apart | thes | cunstermer | overlook | downplay | advitisement | acontroversial | Sicen | semm | Prikura |
| weights | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| bias | 5952 |  |  |  |  |  |  |  |  |  |
| KOR |  |  |  |  |  |  |  |  |  |  |
| 10 highest-weighted features | . | g\_bias | , | to | the | and | of | that | is | in |
| weights | 5975 | 5975 | 5918 | 5912 | 5888 | 5817 | 5725 | 5654 | 5527 | 5335 |
| 10 lowest-weighted features | wear | colthes | puma | Adiddas | Boss | Nike | overadvertising | pierod | sung | sam |
| weights | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| bias | 5975 |  |  |  |  |  |  |  |  |  |
| SPA |  |  |  |  |  |  |  |  |  |  |
| 10 highest-weighted features | g\_bias | the | to | . | a | , | that | and | of | is |
| weights | 5868 | 5867 | 5853 | 5839 | 5833 | 5828 | 5784 | 5776 | 5746 | 5634 |
| 10 lowest-weighted features | loking | Firt | acompanied | secondary | wide | scared | undervalue | regarded | aptitudes | constanlly |
| weights | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| bias | 5868 |  |  |  |  |  |  |  |  |  |
| TEL |  |  |  |  |  |  |  |  |  |  |
| 10 highest-weighted features | g\_bias | the | to | . | and | in | of | is | that | , |
| weights | 5994 | 5962 | 5948 | 5933 | 5854 | 5848 | 5821 | 5586 | 5483 | 5450 |
| 10 lowest-weighted features | travels | appears | heard | lightening | accurately | shifts | automaticaly | puting | achievs | assurance |
| weights | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| bias | 5994 |  |  |  |  |  |  |  |  |  |
| TUR |  |  |  |  |  |  |  |  |  |  |
| 10 highest-weighted features | to | g\_bias | . | the | of | and | is | , | a | in |
| weights | 5955 | 5955 | 5926 | 5893 | 5872 | 5867 | 5772 | 5581 | 5517 | 5502 |
| 10 lowest-weighted features | avoiding | planing | suprised | locations | ussage | calories | restrict | restrictions | allowing | Addition |
| weights | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| bias | 5955 |  |  |  |  |  |  |  |  |  |
| ZHO |  |  |  |  |  |  |  |  |  |  |
| 10 highest-weighted features | . | g\_bias | to | the | , | of | and | is | that | in |
| weights | 5901 | 5901 | 5900 | 5872 | 5872 | 5760 | 5742 | 5540 | 5445 | 5406 |
| 10 lowest-weighted features | mentions | expence | goodness | repeats | weakness | trickness | presents | witnesses | disscusion | behavir |
| weights | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| bias | 5901 |  |  |  |  |  |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 |
| ARA | 0.462687 | 0.516667 | 0.488189 |
| DEU | 0.666667 | 0.292683 | 0.40678 |
| FRA | 0.369863 | 0.529412 | 0.435484 |
| HIN | 0.176471 | 0.1 | 0.12766 |
| ITA | 0.657143 | 0.425926 | 0.516854 |
| JPN | 0.441176 | 0.483871 | 0.461538 |
| KOR | 0.350515 | 0.557377 | 0.43038 |
| SPA | 0.365854 | 0.245902 | 0.294118 |
| TEL | 0.590909 | 0.40625 | 0.481481 |
| TUR | 0.430769 | 0.509091 | 0.466667 |
| ZHO | 0.455696 | 0.553846 | 0.5 |

My training model is poor at judging SPA and HIN. It is relatively good at judging ITA and ZHO. My model confuses JPN and KOR a lot.