**6. How many iterations are required to separate the training data? Which number of iterations is likely to represent the best tradeoff between fitting the data and not overfitting?**

With tuning parameter when adjust the weight, 28 times to converge

15 times are the best tradeoff between fitting the data and not overfitting.

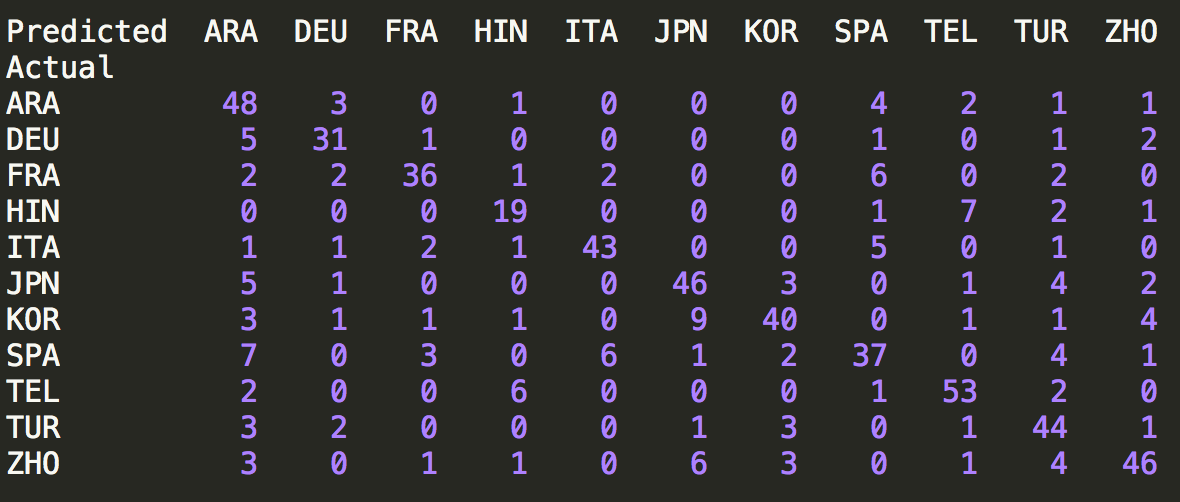
**7. Feature Engineering: Add 3 different features**

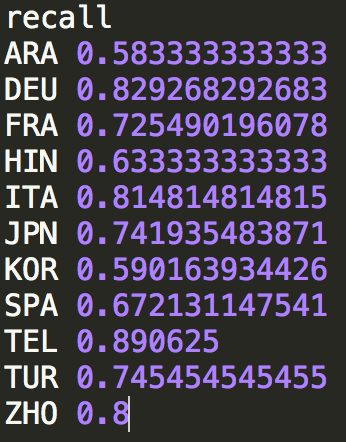
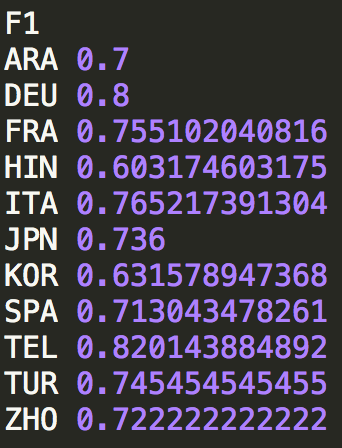
* word n-grams (n = 2)
* lemmas
* lowercase normalization
* unigram

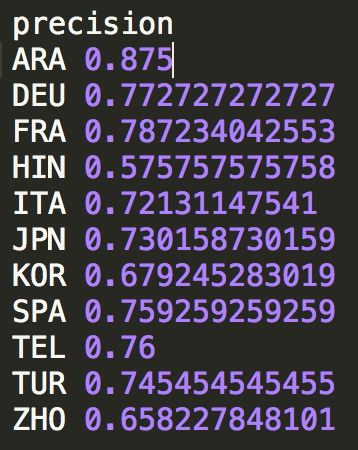
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | trainAcc | devAcc | testAcc | params | iteration(I) |
| All in | 0.9996272828922848 | 0.7374581939799331 | 0.7317880794701986 | 934293 | 11 |
| Minus unigram | 0.9998136414461424 | 0.7558528428093646 | 0.7152317880794702 | 930481 | 12 |
| Minus lemmatizer | 0.9986954901229966 | 0.7441471571906354 | 0.7235099337748344 | 968862 | 11 |
| Minus 2-gram | 0.9998136414461424 | 0.7006688963210702 | 0.6837748344370861 | 152961 | 15 |
| Minus lower | 0.9994409243384271 | 0.7324414715719063 | 0.7284768211920529 | 933357 | 11 |
| Best Model  Unigram+2-gram+lemmatizer+  lower | 0.9996272828922848 | 0.7374581939799331 | 0.7317880794701986 | 934293 | 11 |

8. Error Analysis: Best Model

a. Confusion Matrix



c. Recall, Precision, F1



b. 10 Highest vs 10 Lowest Weights & Bias Weight

HIN

|  |  |  |
| --- | --- | --- |
| Highest | Lowest | Biased – ‘ajdif7af’ |
| then 2.600000000000001  which 2.3000000000000007  but 2.1000000000000005  its 2.0000000000000004  of life 2.0000000000000004  field 1.9000000000000006  start 1.9000000000000006  today 1.9000000000000004  behind 1.8000000000000005  towards 1.8000000000000003 | finally -2.700000000000001  For example -2.3000000000000007  know -1.9000000000000004  and the -1.9000000000000004  nowadays -1.7000000000000004  reasons -1.6000000000000003  may not -1.5000000000000002  there -1.5000000000000002  to get -1.4000000000000001  such a -1.3 | -0.10000000000000003 |

DEU

|  |  |  |
| --- | --- | --- |
| Highest | Lowest | Biased – ‘ajdif7af’ |
| statement 2.2000000000000006  often 2.2000000000000006  possibility 1.9000000000000006  special 1.9000000000000004  question 1.8000000000000005  beeing 1.8000000000000005  the statement 1.7000000000000004  there 1.7000000000000004  get 1.7000000000000002  But 1.6000000000000003 | we -2.9000000000000012  particular -1.6000000000000003  however -1.5000000000000002  However -1.5000000000000002  we are -1.5000000000000002  major -1.4000000000000001  Secondly -1.3  world -1.3  and they -1.3  study -1.3 | -0.6 |

JPN

|  |  |  |
| --- | --- | --- |
| Highest | Lowest | Biased – ‘ajdif7af’ |
| japan 4.000000000000002  Japan 3.800000000000002  I think 3.1000000000000014  If 2.800000000000001  in Japan 2.500000000000001  japanese 2.400000000000001  I disagree 2.3000000000000007  I agree 2.1000000000000005  Japanese 2.1000000000000005  especially 2.1000000000000005 | an -2.9000000000000012  every -2.500000000000001  last -2.400000000000001  maybe -2.2000000000000006  and to -2.2000000000000006  will -2.1000000000000005  be a -2.1000000000000005  you -1.9000000000000004  time -1.9000000000000004  all the -1.8000000000000005 | 0.5 |

FRA

|  |  |  |
| --- | --- | --- |
| Highest | Lowest | Biased – ‘ajdif7af’ |
| indeed 3.4000000000000017  is a 2.9000000000000012  Indeed 2.600000000000001  during 2.2000000000000006  by 2.1000000000000005  instance 2.0000000000000004  even if 2.0000000000000004  think that 2.0000000000000004  differents 1.8000000000000005  one hand 1.7000000000000004 | the idea -2.400000000000001  the people -2.1000000000000005  this is -2.0000000000000004  I -1.9000000000000006  the statement -1.9000000000000004  If -1.7000000000000004  in life -1.6000000000000003  information -1.6000000000000003  times -1.6000000000000003  should -1.6000000000000003 | -0.19999999999999998 |

TUR

|  |  |  |
| --- | --- | --- |
| Highest | Lowest | Biased – ‘ajdif7af’ |
| can not 3.3000000000000016  Because 2.9000000000000012  being 2.700000000000001  turkey 2.400000000000001  about 2.400000000000001  conditions 2.3000000000000007  easily 2.3000000000000007  idea 2.2000000000000006  Turkey 2.2000000000000006  As a 2.2000000000000006 | study -2.2000000000000006  a lot -2.1000000000000005  will -2.1000000000000005  often -2.1000000000000005  case -1.9000000000000006  agree with -1.8000000000000005  out -1.8000000000000005  statement -1.6000000000000003  learn fact -1.6000000000000003  could -1.6000000000000003 | 2.7755575615628914e-17 |

ARA

|  |  |  |
| --- | --- | --- |
| Highest | Lowest | Biased – ‘ajdif7af’ |
| alot 3.4000000000000017  and the 2.9000000000000012  any 2.700000000000001  alot of 2.600000000000001  statment 2.400000000000001  from 2.1000000000000005  thier 2.1000000000000005  Also 2.0000000000000004  many reason 1.9000000000000006  will 1.9000000000000004 | often -2.2000000000000006  possible -2.2000000000000006  of the -2.1000000000000005  seems -1.9000000000000006  can be -1.9000000000000004  been -1.8000000000000005  If -1.8000000000000003  much -1.7000000000000004  knowledge -1.7000000000000004  it is -1.7000000000000002 | 0.30000000000000004 |

ITA

|  |  |  |
| --- | --- | --- |
| Highest | Lowest | Biased – ‘ajdif7af’ |
| think that 3.1000000000000014  probably 2.500000000000001  people that 2.400000000000001  infact 2.400000000000001  that in 2.2000000000000006  possibility to 2.1000000000000005  I think 2.1000000000000005  man 2.0000000000000004  possibility 2.0000000000000004  problems 1.9000000000000004 | may -2.3000000000000007  over -2.0000000000000004  do not -2.0000000000000004  Because -1.9000000000000006  get -1.9000000000000004  But -1.7000000000000004  would -1.7000000000000004  anything -1.7000000000000004  him -1.6000000000000003  which -1.6000000000000003 | 0.30000000000000004 |

ZHO

|  |  |  |
| --- | --- | --- |
| Highest | Lowest | Biased – ‘ajdif7af’ |
| still 3.0000000000000013  will 2.400000000000001  may 2.3000000000000007  is a 2.2000000000000006  china 2.1000000000000005  but not 2.1000000000000005  Take 2.1000000000000005  just 2.1000000000000005  three 2.0000000000000004  maybe 1.9000000000000006 | the other -2.700000000000001  an -2.600000000000001  and that -2.0000000000000004  have a -1.9000000000000006  its -1.9000000000000004  expensive -1.8000000000000005  have to -1.8000000000000003  working -1.7000000000000004  try to -1.7000000000000004  if they -1.7000000000000004 | -0.5 |

KOR

|  |  |  |
| --- | --- | --- |
| Highest | Lowest | Biased – ‘ajdif7af’ |
| korea 3.700000000000002  out 3.0000000000000013  Korea 2.800000000000001  However 2.700000000000001  however 2.400000000000001  various 2.2000000000000006  company 2.1000000000000005  famous 2.1000000000000005  just 2.1000000000000005  these day 2.0000000000000004 | take -2.500000000000001  think that -2.3000000000000007  in the -2.1000000000000005  by the -2.0000000000000004  issue -1.9000000000000004  how -1.9000000000000004  the same -1.8000000000000005  only -1.7000000000000002  japan -1.6000000000000003  a person -1.6000000000000003 | -0.30000000000000004 |

SPA

|  |  |  |
| --- | --- | --- |
| Highest | Lowest | Biased – ‘ajdif7af’ |
| going to 2.400000000000001  people is 2.0000000000000004  the city 2.0000000000000004  are going 1.9000000000000006  that are 1.8000000000000005  example of 1.7000000000000004  not know 1.7000000000000004  going 1.7000000000000004  things 1.7000000000000004  enviroment 1.7000000000000004 | which -3.0000000000000013  might -2.2000000000000006  from -2.2000000000000006  especially -2.1000000000000005  still -2.1000000000000005  by -2.0000000000000004  even if -2.0000000000000004  the fact -1.9000000000000006  on -1.9000000000000004  successful -1.9000000000000004 | 0.10000000000000003 |

TEL

|  |  |  |
| --- | --- | --- |
| Highest | Lowest | Biased – ‘ajdif7af’ |
| finally 3.0000000000000013  some 2.2000000000000006  may 2.2000000000000006  by 2.2000000000000006  the statement 2.2000000000000006  conclude 2.1000000000000005  may not 2.1000000000000005  cannot 2.0000000000000004  strongly 2.0000000000000004  about the 2.0000000000000004 | I think -2.800000000000001  just -2.3000000000000007  have to -2.1000000000000005  first -2.0000000000000004  or -1.8000000000000003  big -1.7000000000000004  however -1.7000000000000004  of life -1.7000000000000004  think that -1.7000000000000004  say -1.7000000000000002 | 0.5 |

What are some of the patterns you observe? Do the bias feature weights behave like priors in naïve Bayes—why or why not?

According to the confusion matrix, between some languages there are high chances of predicting wrongly to each other, such as KOR and JPN, SPA and ITA, referring the similarity between those languages.

Also from the top 10 word lists, country name appears in the list for many languages, so apparently it’s an important percept.

The bias feature in perceptron seems different from the prior probabilities, because the latter in NB won’t change and behaves as a stable factor for calculating the joint probability, while the bias feature in Perceptron is for error tuning purpose.

9. Bonus

python perceptron\_a.py –a 30

The test result of average method is not as good as the former model.