Kyle Tunis
Algorithms for Natural Language Processing
Fall 2017
Assignment 4

1.

Language	**Train**	**Dev**
KOR	557	60
JPN	557	60
SPA	450	52
ZHO	593	69
TEL	533	62
ITA	516	53
ARA	494	51
FRA	473	53
HIN	352	47
DEU	337	34
TUR	504	57
Total	**604**	**598**

The majority class baseline accuracy would be 11.05

2. The training data is classified with 100% accuracy after 32 iterations. However, the performance of the model on the test data is maximized after 21 iterations. This likely represents the best point where the data is well fitted, but not overfitted.
3. I implemented a perceptron which takes as percepts binary unigrams, binary trigrams, bigram counts, and all words are lowercased.

Test	**Iterations to converge**	**Best number of iterations**	**Accuracy at best number of iterations**
Baseline	32	21	0.6937086092715232
Full	17	11	0.7235099337748344

Without bigrams	14	9	0.7152317880794702
Without lowercasing	12	10	0.7201986754966887
Without trigrams	22	11	0.7102649006622517

4.

Language	KOR	
Pred\Gold	KOR	Not KOR
KOR	48	24
Not KOR	13	389
Precision	0.6666666666666666	
Recall	0.7868852459016393	
F1	0.7218045112781954	
Top 10 Features	[('korea', 54), ('in korea', 48), ('enjoy their', 47), ('however ,', 39), ('their own', 38), ('such as', 38), ('. however', 37), ('even though', 36), ('these days', 36), ('their life', 35)]	
Bottom 10 Features	[('use exazageration', 1), ('company use exazageration', 1), ('use exazageration ,', 1), ('exazageration , a', 1), ('famous person ,', 1), ('person , and', 1), ('and repeating', 1), (', and repeating', 1), ('and repeating ,', 1), ('repeating , people', 1)]	
Bias Weight	3	

Language	JPN	
Pred\Gold	JPN	Not JPN
JPN	49	10
Not JPN	13	388
Precision	0.8305084745762712	
Recall	0.7903225806451613	

F1	0.8099173553719008
Top 10 Features	[('I', 69), ('in japan', 64), ('japan', 48), ('japan ,', 43), ('. i', 41), ('there are', 40), ('i think', 38), (', and', 36), (', the', 35), ('. if', 33)]
Bottom 10 Features	[('advertisements that can', 1), ('be trusted', 1), ('can be trusted', 1), ('trusted by', 1), ('be trusted by', 1), ('trusted by a', 1), ('of costommers', 1), ('lot of costommers', 1), ('costommers .', 1), ('of costommers .', 1)]
Bias Weight	5

Language	SPA	
Pred\Gold	SPA	Not SPA
SPA	36	17
Not SPA	25	401
Precision	0.6792452830188679	
Recall	0.5901639344262295	
F1	0.631578947368421	
Top 10 Features	[(', but', 39), (', and', 38), (', is', 38), ('going to', 37), ('person', 37), ('things', 35), ('the idea', 35), ('of the', 32), ('that the', 32), ('that you', 31)]	
Bottom 10 Features	[('others', 1), ('the others', 1), ('then you will', 1), ('you will reach', 1), ('reach your', 1), ('will reach your', 1), ('reach your own', 1), ('own limits', 1), ('your own limits', 1), ('own limits .', 1)]	
Bias Weight	5	

Language	ZHO	
Pred\Gold	ZHO	Not ZHO
ZHO	47	11
Not ZHO	18	390

Precision	0.8103448275862069
Recall	0.7230769230769231
F1	0.7642276422764227
Top 10 Features	[('just', 51), ('still', 51), (', but', 44), (', the', 41), ('?', 36), ('may', 34), ('maybe', 34), (', and', 33), ('is a', 33), ('always', 32)]
Bottom 10 Features	[('can help each', 1), ('other . so', 1), ('so in this', 1), ('this world', 1), ('in this world', 1), ('world every', 1), ('this world every', 1), ('world every one', 1), ('every one can', 1), ('one can enjoy', 1)]
Bias Weight	2

Language	TEL	
Pred\Gold	TEL	Not TEL
TEL	45	6
Not TEL	19	392
Precision	0.8823529411764706	
Recall	0.703125	
F1	0.7826086956521738	
Top 10 Features	[('so ,', 37), ('and also', 37), ('the statement', 36), ('to the', 35), ('about the', 34), ('in the', 33), ('all the', 33), ('we can', 33), ('statement', 33), ('he', 33)]	
Bottom 10 Features	[('buying car', 1), ('towards buying car', 1), ('buying car to', 1), ('car to have', 1), ('luxuries', 1), ('have luxuries', 1), ('to have luxuries', 1), ('luxuries life', 1), ('have luxuries life', 1), ('luxuries life .', 1)]	
Bias Weight	7	

Language	ITA	
Pred\Gold	ITA	Not ITA

ITA	39	15
Not ITA	15	398
Precision	.75	
Recall	0.7222222222222222	
F1	0.7358490566037735	
Top 10 Features	[(':', 54), ('i think', 43), ('you can', 36), (', for', 35), ('think that', 33), ('I', 33), ('life', 32), (', in', 32), ('think', 32), ('a specific', 31)]	
Bottom 10 Features	[('and also for', 1), ('improvment', 1), ('the improvment', 1), ('for the improvment', 1), ('improvment of', 1), ('the improvment of', 1), ('of your', 1), ('improvment of your', 1), ('your personal life', 1), ('personal life !', 1)]	
Bias Weight	10	

Language	ARA	
Pred\Gold	ARA	Not ARA
ARA	43	11
Not ARA	17	394
Precision	0.7962962962962963	
Recall	0.7166666666666667	
F1	0.7543859649122806	
Top 10 Features	[('alot', 49), ('alot of', 46), (', and', 41), ('from', 41), ('any', 40), ('will', 35), ('will be', 33), ('. also', 32), ('every', 31), ('some people', 31)]	
Bottom 10 Features	[('and many unexpected', 1), ('dangerous', 1), ('unexpected dangerous', 1), ('many unexpected dangerous', 1), ('dangerous life', 1), ('unexpected dangerous life', 1), ('life matters', 1), ('dangerous life matters', 1), ('matters .', 1), ('life matters .', 1)]	
Bias Weight	13	

Language	FRA	
Pred\Gold	FRA	Not FRA
FRA	41	17
Not FRA	10	396
Precision	0.7068965517241379	
Recall	0.803921568627451	
F1	0.7522935779816514	
Top 10 Features	[('think that', 42), ('even if', 39), (', it', 36), ('to take', 36), (', to', 35), (', the', 35), ('indeed', 35), ('when you', 34), ('to be', 33), ('young people', 33)]	
Bottom 10 Features	[('level , if', 1), ('if you are', 1), ('you are a', 1), ('are a young', 1), ('person', 1), ('young person', 1), ('a young person', 1), ('young person or', 1), ('person or an', 1), ('or an older', 1)]	
Bias Weight	3	

Language	HIN	
Pred\Gold	HIN	Not HIN
HIN	17	15
Not HIN	13	420
Precision	0.53125	
Recall	0.5666666666666667	
F1	0.5483870967741935	
Top 10 Features	[('which', 49), ('then', 43), ('has', 41), ('. so', 40), ('of life', 38), ('can be', 36), ('and concept', 33), ('say', 32), ('its', 30), ('. now', 30)]	
Bottom 10 Features	[('effective means', 1), ('- effective means', 1), ('effective means of', 1), ('of transport like', 1), ('like buses', 1), ('transport like buses', 1), ('like buses ,', 1), ('trains etc', 1), (', trains etc', 1), ('trains etc .', 1)]	
Bias Weight	7	

Language	DEU	
Pred\Gold	DEU	Not DEU
DEU	27	6
Not DEU	14	410
Precision	0.8181818181818182	
Recall	0.6585365853658537	
F1	0.7297297297297297	
Top 10 Features	[(', that', 59), (', because', 48), ('the statement', 37), ('. but', 35), ('do not', 35), ('. this', 34), ('get', 33), ('of a', 32), ('statement', 31), ('important to', 31)]	
Bottom 10 Features	[('hart times .', 1), ('times . but', 1), ('but what will', 1), ('will come', 1), ('what will come', 1), ('will come to', 1), ('to us', 1), ('come to us', 1), ('to us ,', 1), ('us , young', 1)]	
Bias Weight	4	

Language	TUR	
Pred\Gold	TUR	Not TUR
TUR	45	37
Not TUR	10	392
Precision	0.5487804878048781	
Recall	0.8181818181818182	
F1	0.656934306569343	
Top 10 Features	[('can not', 58), ('. because', 56), ('the life', 41), ('much more', 37), ('being', 37), ('lots of', 34), ('turkey', 34), ('. if', 34), ('an', 33), ('can be', 33)]	
Bottom 10 Features	[('more expensive and', 1), ('expensive and less', 1), ('less useful', 1), ('and less useful', 1), ('less useful than', 1), ('useful than it', 1), ('than it is', 1), ('it is seem', 1), ('seem to you', 1), ('to you .', 1)]	
Bias Weight	3	

Observations:

- The True Negative rate is much higher than any other count. This makes sense, since most samples were correctly classified, and for every class but one, a correct classification increases the true negative by 1.
- Bigrams had much more predictive power than trigrams
- I expected the bias weights to be proportional to the number of documents with a given label, but this was not the case.
- Some highly weighted features are not at all intuitive. For example, the unigram ':' (colon) was the highest weighted feature for Italian. Others were far more intuitive, such as 'in japan' being highly weighted for Japanese.
- Some languages were far easier to classify than others. Going by F1 score, the easiest language to classify was Japanese, while the hardest was Hindi.