Kyle Tunis
Algorithms for Natural Language Processing
Fall 2017
Assignment 4

Note: I discovered a very small error in my code that was causing the weights to not be properly subtracted from after an incorrect classification. My submission contains the correct code, but the output is from the erroneous code because I didn't have time before the deadline to run the code the 11 times required to fill out all output tables correctly. I'll submit a corrected version of the writeup complete with the correct output, but I won't be able to get that in before the deadline. To be clear, the code included in my submission will not change after I submit the assignment, only the updated outputs.

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	35	25	0.6738410596026491
Full	14	14	0.7268211920529801
Without bigrams	12	12	0.7251655629139073
Without lowercasing	15	13	0.7119205298013245
Without trigrams	25	5	0.6804635761589404

4.

Language	KOR	
Pred\Gold	KOR	Not KOR
KOR	40	11
Not KOR	21	399
Precision	0.7843137254901961	
Recall	0.6557377049180327	
F1	0.7142857142857142	
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	1	

Language	JPN

Pred\Gold	JPN	Not JPN
JPN	51	25
11	388	388
Precision	0.6710526315789473	
Recall	0.8225806451612904	
F1	0.7391304347826086	
Top 10 Features	[('I', 67), ('in japan', 60), ('japan', 47), ('i think', 42), ('. if', 41), ('japan ,', 40), (', and', 38), ('there are', 37), (', but', 35), (', we', 33)]	
Bottom 10 Features	[('my sister', -2), ('study as', -2), ('specific field', -2), ('field .', -2), ('special talent', -2), ('talent and', -2), ('and habbit', -2), ('to let', -2), ('our society', -2), ('and broad', -3)]	
Bias Weight	5	

Language	SPA	
Pred\Gold	SPA	Not SPA
SPA	38	27
Not SPA	23	401
Precision	0.5846153846153846	
Recall	0.6229508196721312	
F1	0.6031746031746033	
Top 10 Features	[(', is', 40), (', and', 38), ('think that', 37), ('going to', 37), ('have to', 35), (', but', 35), ('person', 33), ('in a', 33), ('all the', 32), ('of the', 31)]	
Bottom 10 Features	[('not give', 1), ('faith', 1), ('from their', 1), ('at', 1), ('because of', 1), ('because of all', 1), ('all these', 1), ('reasons', 1), ('these reasons', 1), ('all these reasons', 1)]	
Bias Weight	4	

Language	ZHO	
Pred\Gold	ZHO	Not ZHO
ZHO	50	17
Not ZHO	15	389
Precision	0.746268656716418	
Recall	0.7692307692307693	
F1	0.7575757575757576	
Top 10 Features	[('still', 44), (', the', 43), ('just', 43), ('time on', 37), ('is a', 33), ('more and', 31), ('if we', 31), ('out', 30), ('. take', 30), ('group led', 30)]	
Bottom 10 Features	[('and concepts ant', -1), ('ant learning', -1), ('concepts ant learning', -1), ('ant learning facts', -1), ('are critical', -1), ('facts are critical', -1), ('critical in', -1), ('are critical in', -1), ('critical in learning', -1), ('in learning .', -1)]	
Bias Weight	0	

Language	TEL	
Pred\Gold	TEL	Not TEL
TEL	51	9
Not TEL	13	388
Precision	0.85	
Recall	0.796875	
F1	0.8225806451612903	
Top 10 Features	[('to the', 44), ('we can', 43), ('by', 37), ('the subject', 37), ('of the', 36), ('people enjoy', 35), ('in the', 35), ('the statement', 34), ('statement', 33), ('the concept', 33)]	
Bottom 10 Features	[('purpose of making', -1), ('making advertisements', -1), ('of making advertisements', -1), ('making advertisements .', -1), ('their products', -2), ('popular in', -2), ('tv ,', -2), ('toys are', -2), ('the tv', -2), ('; the', -2)]	

Bias Weight	10

Language	ITA	
Pred\Gold	ITA	Not ITA
ITA	41	10
Not ITA	13	398
Precision	.803921568627451	
Recall	0.7592592592592593	
F1	0.780952380952381	
Top 10 Features	[(', but', 44), (':', 38), ('i think', 38), ('I', 37), ('think', 35), ('think that', 33), (', in', 32), ('understand', 31), ('at the', 30), ('probably', 30)]	
Bottom 10 Features	[('education system will', -1), ('system will be', -1), ('be better in', -1), ('better in this', -1), ('this way .', -1), ('the students', -2), ('students ,', -2), ('public .', -2), ('to learn', -3), ('the public', -3)]	
Bias Weight	9	

Language	ARA	
Pred\Gold	ARA	Not ARA
ARA	41	12
Not ARA	19	398
Precision	0.7735849056603774	
Recall	0.6833333333333333	
F1	0.7256637168141593	
Top 10 Features	[(', and', 52), ('any', 48), ('alot', 46), ('alot of', 44), ('from', 37), ('self', 31), ('to be', 30), ('will', 29), ('his', 29), ('statment', 29)]	
Bottom 10 Features	[('of things they', -1), ('things they have', -1), ('have to be', -1), ('be done .', -1), ('done . for', -1), ('these	

	reasons', -1), ('reasons i disagree', -1), ('disagree this', -1), ('i disagree this', -1), ('disagree this statement', -1)]
Bias Weight	10

Language	FRA	
Pred\Gold	FRA	Not FRA
FRA	40	13
Not FRA	11	399
Precision	0.7547169811320755	
Recall	0.7843137254901961	
F1	0.7692307692307692	
Top 10 Features	[('think that', 44), ('is a', 42), ('...', 40), ('indeed', 39), ('. indeed', 39), ('of the', 38), (', the', 37), ('the same', 36), ('. but', 36), ('when you', 35)]	
Bottom 10 Features	[('we do have', -1), ('have developed', -1), ('do have developed', -1), ('developed the', -1), ('have developed the', -1), ('the technology', -1), ('developed the technology', -1), (', perhaps', -2), ('having the', -2), ('first world', -3)]	
Bias Weight	0	

Language	HIN	
Pred\Gold	HIN	Not HIN
HIN	14	14
Not HIN	16	425
Precision	0.5	
Recall	0.4666666666666667	
F1	0.4827586206896552	
Top 10 Features	[('. so', 46), ('which', 46), ('then', 37), ('can be', 37), ('has', 36), ('. they', 36), ('of life', 35), ('old age', 33), ('various',	

	32), ('also', 32)]
Bottom 10 Features	[('in one specific', 1), ('specific subject', 1), ('one specific subject', 1), ('specific subject is', 1), ('is better than', 1), ('better than having', 1), ('than having broad', 1), ('academic subjects', 1), ('many academic subjects', 1), ('academic subjects .', 1)]
Bias Weight	5

Language	DEU	
Pred\Gold	DEU	Not DEU
DEU	29	5
Not DEU	12	410
Precision	0.8529411764705882	
Recall	0.7073170731707317	
F1	0.7733333333333334	
Top 10 Features	[(', that', 50), ('able to', 39), ('important to', 38), ('statement', 35), ('you', 32), (', because', 32), ('often', 31), ('get', 31), ('an', 31), ('it is', 31)]	
Bottom 10 Features	[('be useful', -2), ('have different', -2), ('the school', -2), ('have broad', -2), ('at many', -2), ('many situations', -2), ('the examination', -2), ('the interview', -2), ('it would', -3), ('example ,', -3)]	
Bias Weight	3	

Language	TUR	
Pred\Gold	TUR	Not TUR
TUR	44	22
Not TUR	11	395
Precision	0.6666666666666666	
Recall	0.8	

F1	0.7272727272727272
Top 10 Features	[('. because', 56), ('can not', 44), ('. if', 41), ('lots of', 40), ('being', 35), ('the life', 34), ('about', 33), ('turkey', 33), ('can be', 32), ('this', 31)]
Bottom 10 Features	[('reasons , i', -1), ('affirm', -1), ('can affirm', -1), ('i can affirm', -1), ('affirm i', -1), ('can affirm i', -1), ('affirm i completely', -1), ('i completely agree', -1), ('agree with this', -1), ('with this statement', -1)]
Bias Weight	4

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 a very highly 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 Telugu, while the hardest was Hindi.