Arif Ali
LING 572
Oct. 10th 2016
a1

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	0.7023411371237458
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 0.6903973509933775 stopping the learning at iteration 6. It performs slightly worse than naïve bayes, which could indicate perceptron is over fitting using the the dev dataset.

Question 8:

Part A:

'ARA',	'DEU',	'FRA',	'HIN',	'ITA',	'JPN',	'KOR',	'SPA',	'TEL',	'TUR',	'ZHO'
41	0	2	3	1	1	0	4	3	4	1
1	30	4	3	0	1	0	0	0	0	2
2	1	36	2	3	1	0	2	1	1	2
2	0	1	18	0	0	0	0	6	2	1
2	2	2	0	43	1	0	3	0	0	1
2	0	2	1	0	47	6	0	1	2	1
0	1	1	0	0	13	36	0	0	1	9
3	1	2	2	6	3	1	34	1	5	3
1	0	0	14	0	1	0	1	47	0	0
2	3	0	2	1	2	6	0	1	36	2
2	0	1	1	0	6	4	2	0	0	49

Part B: please see (format is language, max, min)
ARA
[('ALOT OF', 27), ('EVERY THING', 18), ('AND A', 18), ('MANY REASON', 16), ('REASON .', 16), (', AND', 16), ('TO HELP', 14), ('ANY THING', 14), ('IN ADDITION', 14), ('OF THAT', 14)]

[('. IF', -18), ('. BUT', -17), ('PEOPLE DO', -15), ('THAT PEOPLE', -14), ('HA BEEN', -14), ('TO ENJOY', -13), ('OF THEM', -12), ('ABLE TO', -12), ('ARE THE', -12), ('THE STATEMENT', -12)]
DEU
[(', THAT', 23), ('. BUT', 18), ('AND THEREFORE', 15), ('I WOULD', 14), ('THE STATEMENT', 14), (', BECAUSE', 14), ('MONEY .', 14), ('ONE HAND', 13), ('. FURTHERMORE', 13), ('HA TO', 13)]
[(', AND', -19), ('THAT IS', -14), ('WE CAN', -14), ('TIME TO', -13), ('WHEN I', -12), ('IF WE', -12), ('THEM TO', -12), ('TO MAKE', -12), (', I', -12), ('***bias_term***', -11)]

FRA
[('. INDEED', 23), ('IS A', 19), ('TO CONCLUDE', 18), ('INDEED ,', 18), ('IN FACT', 18), ('EVEN IF', 18), ('TO TAKE', 17), ('NOWADAYS ,', 16), ('EXPERIENCE .', 15), ('MONEY TO', 14)]
[('THE IDEA', -20), ('THE PEOPLE', -17), ('IN MY', -16), ('AGREE THAT', -14), ('WHICH MAKE', -13), ('WHEN I', -13), ('THERE ARE', -13), ('SHOULD BE', -12), ('TO GET', -12), ('NOT ONLY', -12)]

HIN
[('IN TODAY', 15), ('. BUT', 15), ('SAY THAT', 14), ('WANT TO', 14), ('RISK AND', 14), ('OLD AGE', 14), ('A WELL', 14), ('OF LIFE', 14), ('OF THE', 14), ('NUMBER OF', 14)]
[('. FINALLY', -15), ('BASED ON', -14), ('OUR LIFE', -14), ('HARD TO', -13), ('. IF', -12), ('BECAUSE THE', -12), ('PEOPLE WILL', -12), (', AND', -11), ('AND THE', -11), (', EVEN', -11)]

ITA
[('I THINK', 24), ('THINK THAT', 21), ('POSSIBILITY TO', 17), ('TO IMPROVE', 17), (', IN', 16), ('PEOPLE THAT', 16), ('THAT IN', 16), ('A SPECIFIC', 15), ('IN ITALY', 15), ('THE PAST', 15)]
[('. BECAUSE', -20), ('. BUT', -18), ('. THERE', -15), ('OVER THE', -13), ('BECAUSE OF', -13), ('. YOU', -12), ('SOME PEOPLE', -12), ('WHICH IS', -12), ('GOING TO', -12), ('ONE HAND', -12)]

JPN
[('IN JAPAN', 33), ('JAPAN ,', 20), (', AND', 19), ('. IF', 18), ('. THEREFORE', 17), ('IF PEOPLE', 16), ('I DISAGREE', 16), ('. FROM', 16), ('OPINION THAT', 15), ('JAPAN .', 15)]
[('ALL THE', -17), ('LIFE ,', -15), ('OF TIME', -14), ('IN A', -13), ('THEM TO', -13), ('FOR A', -13), (', THAT', -13), ('TO GIVE', -13), ('ARE A', -13), ('ADVERTISEMENT ,', -13)]

KOR
[('IN KOREA', 30), ('KOREA ,', 26), ('THESE DAY', 20), ('. HOWEVER', 20), ('EVEN THOUGH', 17), ('HOWEVER ,', 17), ('. ALSO', 17), ('SCHOOL STUDENT', 16), (', MANY', 16), ('SUCH A', 15)]
[('FOR ME', -17), ('OF THE', -15), ('THINK THAT', -15), ('OF VIEW', -15), ('IMPORTANT FOR', -14), (', TO', -14), ('THEY MAY', -14), ('THEY WILL', -14), (', BECAUSE', -13), ('TRYING NEW', -13)]

SPA
[(', IS', 22), ('THAT ARE', 18), ('OTHER HAND', 18), ('A BETTER', 17), ('PEOPLE IS', 17), (', ETC', 16), ('GOING TO', 15), ('IN THEIR', 15), (', AND', 15), ('IDEA THAT', 14)]
[('FROM THE', -17), ('PEOPLE .', -15), ('ACCORDING TO', -15), ('. IT', -15), ('AND SO', -15), ('I WANT', -14), ('BETTER THAN', -14), ('ON .', -13), ('TODAY ,', -13), ('WHICH ARE', -13)]

TEL
[('I STRONGLY', 20), ('THE ABOVE', 17), ('THE STATEMENT', 16), ('. FINALLY', 16), ('WHEN COMPARED', 15), ('MAY BE', 15), ('ALL THE', 15), ('IN THE', 14), ('EVERY ONE', 13), ('. EVERY', 13)]
[('HOWEVER ,', -18), (', YOU', -16), ('. HOWEVER', -15), ('I THINK', -15), ('DO NOT', -14), ('A WELL', -12), ('YOU WILL', -12), ('AND A', -12), ('IT CAN', -12), (', AND', -11)]

TUR
[('. BECAUSE', 25), ('CAN NOT', 18), ('THE IDEA', 17), ('MAKE U', 16), ('MUCH MORE', 16), ('START TO', 16), ('IN TURKEY', 16), ('OF THIS', 15), ('THIS WAY', 15), ('AS A', 14)]
[(', AND', -26), (', THE', -19), ('TO KNOW', -18), ('ENJOY THEIR', -16), ('LEARN FACT', -15), ('A LOT', -15), ('AGREE WITH', -15), (', BUT', -14), ('A GOOD', -14), ('IN FACT', -13)]

ZHO

[('ENJOY THE', 16), ('. TAKE', 16), ('OPINION ,', 16), ("WO N'T", 15), ('TIME ON', 14), (', THE', 14), ('DIFFERENT PEOPLE', 13), ('BUT NOT', 13), ('PEOPLE MAY', 13), ('TO TRY', 13)]
[('EVEN IF', -19), ('***bias_term***', -16), ('THAT ARE', -16), ('RISK AND', -16), ('TRYING TO', -15), ('SITUATION .', -14), ('HAVE TO', -14), ('AND THAT', -14), ('ABLE TO', -14), ('THE TIME', -14)]

Part C:

Language	Precision	Recall	F1	Weight of bias
ARA	. 626865671642	.7	. 661417322835	11
DEU	. 723404255319	. 829268292683	. 772727272727	-11
FRA	. 706896551724	. 803921568627	. 752293577982	-5
HIN	. 548387096774	. 566666666667	. 55737704918	8
ITA	. 878048780488	. 666666666667	. 757894736842	4
JPN	. 666666666667	. 677419354839	. 672	-1
KOR	. 672131147541	. 672131147541	. 672131147541	0
SPA	. 729166666667	. 573770491803	. 642201834862	-5
TEL	. 787878787879	. 8125	.8	12
TUR	. 636363636364	. 763636363636	. 694214876033	3
ZHO	.803571428571	. 692307692308	. 743801652893	-16

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. The Japanese bi-grams are interesting, because “In Japan” is one of the top tokens and the same idea happens in the Korean bi-grams. DEU (German I think) also have the bias term as one of its least common tokens. This is interested because unlike ZHO, it had the smallest prior.

The Bias features do not seem correlated with the prior probabilities from the Naïve Bayes model. The one possible exception is the relationship between Japanese and Korean. In Naive Bayes, their Prior Probabilities are identical and in the perceptron model, the two have close to the same precision, recall, and F1 along with near identical bias weight features