Need to parse raw data

Parsing raw data...

There are non-native 1 files. Total size of samples: floor(11044/1)=11044

non-native files parsed: tofel.txt

There are native 6 files. Total size of samples: floor(11044/6)=1840

native files parsed: reddit.Australia.txt.tok.clean reddit.Canada.txt.tok.clean reddit.Ireland.txt.tok.clean reddit.NewZealand.txt.tok.clean reddit.UK.txt.tok.clean reddit.US.txt.tok.clean

all\_data size is 22084

saving to file: ../parsedData/alldata25.txt

file ../parsedData/alldata25.txt with 22084 lines was created

Finished Parsing raw data

Reading file(311 lines) ../parsedData/functionWords.txt

Reading file(22084 lines) ../parsedData/alldata25.txt

Starting feature vector classification

Reading parsed data

There are 11040 Native samples and 11044 Non-Native

Train set size 17667 - native=8860, non-native=8807

Test set size 4417 - native=2180, non-native=2237

Train data 'all feature vector' size is 311x17667

Train data size is 17667

Test data 'all feature vector' size is 311x4417

Test data size is 4417

-----------FUNC\_WORDS\_START---------------------------

Running SVM...

non-native native

precision[0.78957529 0.74371002] - tp/(tp+fp)

recall [0.73133661 0.8 ] - tp/(tp+fn)

fscore [0.75934091 0.77082873] - harmonicAvg(prec + recall)

fscore weighted =0.765010698498843 - (fscore1\*realPortion+fscore2\*realPortion)

Total accuracy =76.5225266017659%

Running Decision Tree...

non-native native

precision[0.67164179 0.65911151] - tp/(tp+fp)

recall [0.66383549 0.66697248] - tp/(tp+fn)

fscore [0.66771583 0.6630187 ] - harmonicAvg(prec + recall)

fscore weighted =0.6653975690987866 - (fscore1\*realPortion+fscore2\*realPortion)

Total accuracy =66.53837446230473%

Running NB...

non-native native

precision[0.76007162 0.75309208] - tp/(tp+fp)

recall [0.7590523 0.75412844] - tp/(tp+fn)

fscore [0.75956162 0.7536099 ] - harmonicAvg(prec + recall)

fscore weighted =0.7566241629107586 - (fscore1\*realPortion+fscore2\*realPortion)

Total accuracy =75.66221417251529%

-------------FUNC\_WORDS\_END-------------------------

Starting top x words classification

top\_words\_list size is 344

Reading parsed data

There are 11040 Native samples and 11044 Non-Native

Train set size 17667 - native=8860, non-native=8807

Test set size 4417 - native=2180, non-native=2237

Train data 'all feature vector' size is 344x17667

Train data size is 17667

Test data 'all feature vector' size is 344x4417

Test data size is 4417

------------------TOP\_WORDS\_START--------------------

Running SVM...

non-native native

precision[0.95250255 0.84871899] - tp/(tp+fp)

recall [0.83370586 0.95733945] - tp/(tp+fn)

fscore [0.88915375 0.89976288] - harmonicAvg(prec + recall)

fscore weighted =0.8943898634815786 - (fscore1\*realPortion+fscore2\*realPortion)

Total accuracy =89.47249264206475%

Running Decision Tree...

non-native native

precision[0.83793411 0.83648088] - tp/(tp+fp)

recall [0.84130532 0.83302752] - tp/(tp+fn)

fscore [0.83961633 0.83475063] - harmonicAvg(prec + recall)

fscore weighted =0.8372148753376933 - (fscore1\*realPortion+fscore2\*realPortion)

Total accuracy =83.72198324654742%

Running NB...

non-native native

precision[0.89894552 0.91078935] - tp/(tp+fp)

recall [0.91461779 0.89449541] - tp/(tp+fn)

fscore [0.90671394 0.90256885] - harmonicAvg(prec + recall)

fscore weighted =0.9046681391871505 - (fscore1\*realPortion+fscore2\*realPortion)

Total accuracy =90.46864387593389%

---------------TOP\_WORDS\_END-----------------------

duration(formatted HH:MM:SS): 00:05:28