NLP14-class document classification

Yogesh

Class	# of records (Full data)	# of Records (test –set)	Correct Prediction (test –set)	Comments/ predicted classes distribution
class A1	3958	1113	84%	
class A2	1257	0	-	Class Removed
class B	1792	412	70.6%	
class B1	1046	206	73%	
class C	554	229	3%	Class A1 :117 , class A2 :88 , B1 :10 Very little unique patterns, multi-lingual (Chinese data)
class C1	507	112	75%	
class C2	874	138	23%	Class A1: 98
class D	482	133	35%	
class D0	415	73	1.4%	Class A1 : 41 , class A2 : 19 , B : 9 Reason: multi-lingual Chinese data
class D1	399	89	64%	
class A_B	257	55	47%	
class A_C	104	24	4.1%	Very less unique patterns
class B_D	58	10	0%	Insufficient data
class misc	16	9	0%	Insufficient data

Result Statistics

- Overall Accuracy on 100% training and 100% testing:72.9%
- Training Accuracy(75% on Full data): 76.9%
- Testing Accuracy(25% on Full data):62%

Solution Details

Doc Preprocessing

- Lower _casing (words)
- custom chars pruning
- stop words pruning (nltk , tribal, custom [intent and data specific])
- word splitting “loginpassword” = “login” + “password”)
- Lemmatization of words (nltk vs spacy(gerunds))
- language id (google languid , Yandex , Microsoft language id , (python packages vs translation s/w)
- Translating (google translate :
- Pruning null docs or docs with just noise

Text

- Word sequences
- string Indexing
- Sequence matrix
- Word-sequence padding (constant sizing (i/p))
- unseen word mapping

WordEmbedding

- Glove (200D + 100D)
- Word2vec
- Fasttext (multi - lingual)
- cove
- (to try) :elmo

ANN

- Convolution Layer
- Fully connected Layer
- activation(logistic, reLU, eLu,p-reLU,Gaussian-reLU)
- Dropout
- Batch normalization
- RNN / GRU / LSTM / Bi-LSTM
- softmax - 14 class
- (to try : DocGAN , RMDL,BiDAF)
- (To try Attention , residual network ,)

Area Of Improvement

- Quality: Better labelling with clear distinction between the labels. (data dependent and labelling dependent)
- Volume: More data per class. at least, 1000 records per class would be required.
- Language translation could certainly improve accuracy