Report for Assignment 2

I have used only the tags(36) mentioned in the Penn tree bank (http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html). So I ignored all the punctuation tags and the corresponding words(punctuations). But I did use the feature 'ispunctuation' for the words that has special characters in it (like Oct.)

Baseline Model:

What is the accuracy of this baseline on all words?

91.5%

What is its accuracy on words that do not occur in the training data?

41%

Naive Bayes:

For the Naïve Bayes I added 4 punctuation tags ('' , "," ,-LRB-, -COLON). If I add those tags there is a slight increase in the accuracy (from 88.6 to 89.5).

Accuracy on the test set : 89.5 %

Accuracy on the unknown words : 32.4%

Part of the confusion matrix:

Actual	Calc. Tag	Counts
CC	CC	2220
	DT	2
	NNP	1
	POS	2
	RB	6
	FW	1
	IN	1
	POS	1
CD	CD	2884
	DT	1

	JJ	1
	NNP	23
	NNPS	7
	NNS	15
	null	131
	POS	240
DT	DT	7864
	FW	3
	IN	76
	NN	10
	NNP	103
IN	DT	1
	IN	9560
	JJ	12
	NNP	4
	NNPS	6
	RB	70
	RP	67
	CC	2
	DT	1
	FW	1
	POS	1
	TO	2
	VBG	1
	VBN	1
	VBP	1
	WDT	1
NN	CC	4
	CD	9
	JJ	268
	MD	3
	NN	11636
	NNP	166
	NNPS	36
	NNS	62
	null	127
	POS	55
	VB	172
	VBG	107
	VBN	3
	VBP	5
	IN	5
	RB	4

	VBD	4
	VBN	2
NNP	DT	6
	JJ	3
	NN	4
	NNP	7940
	NNPS	104
	POS	1
	IN	2
	PRP	2

Maximum entropy:

Analysis:

For the Single feature (ie. Here the feature is the identity of the word itself) the error rate for the Maximum entropy is 0.0847517. ie. The accuracy is 91.53.

I have tried various combinations of the existing features, for all the feature that I have used for Naïve Bayes, I got very lower accuracy which is 78%.

And for 2 of the features (word, punctuation), the accuracy is 92 %. (Error rate – 0.0823358)
For the 3 features (combination of word, punctuation and uppercase) the error rate is only slightly low - 0.0821051.

The highest accuracy that I obtained from the combination of the above features is 92 %.

Confusion matrix:

Part of the confusion matrix:

Actual	Calc. tag	Count
IN	CC	2
	DT	2
	IN	9717
	JJ	19
	NN	13
	RB	98
	TO	2
	VB	1
	VBG	1
	VBP	1

	WDT	1
JJ	CC	3
	CD	60
	DT	12
	IN	10
	JJ	4534
	JJR	4
	NN	590
	NNP	7
	NNS	2
	RB	17
	TO	1
	VB	4
	VBD	3
	VBG	5
	VBN	11
NN	CC	2
	CD	44
	IN	7
	JJ	312
	MD	3
	NN	12735
	NNS	9
	RB	3
	VB	40
	VBD	1
	VBG	6
	VBP	1
NNP	CD	734
	DT	8
	IN	86
	JJ	6741
	NN	372
	NNP	1149
	NNPS	30
	NNS	189
	PRP	2
	RB	2
	VB	95
JJ	CC	3
	CD	61
	DT	7
	IN	10

	JJ	4534
	JJR	1
	NN	1148
	NNP	7
	NNS	2
	RB	33
	TO	1
	VB	4
	VBD	3
	VBG	5
	VBN	11

The java libraries used

1. Commons-io -→for Fileutils class to read the file
2. Commons-lang3 -→ for StringUtils class to check whether the word is uppercase