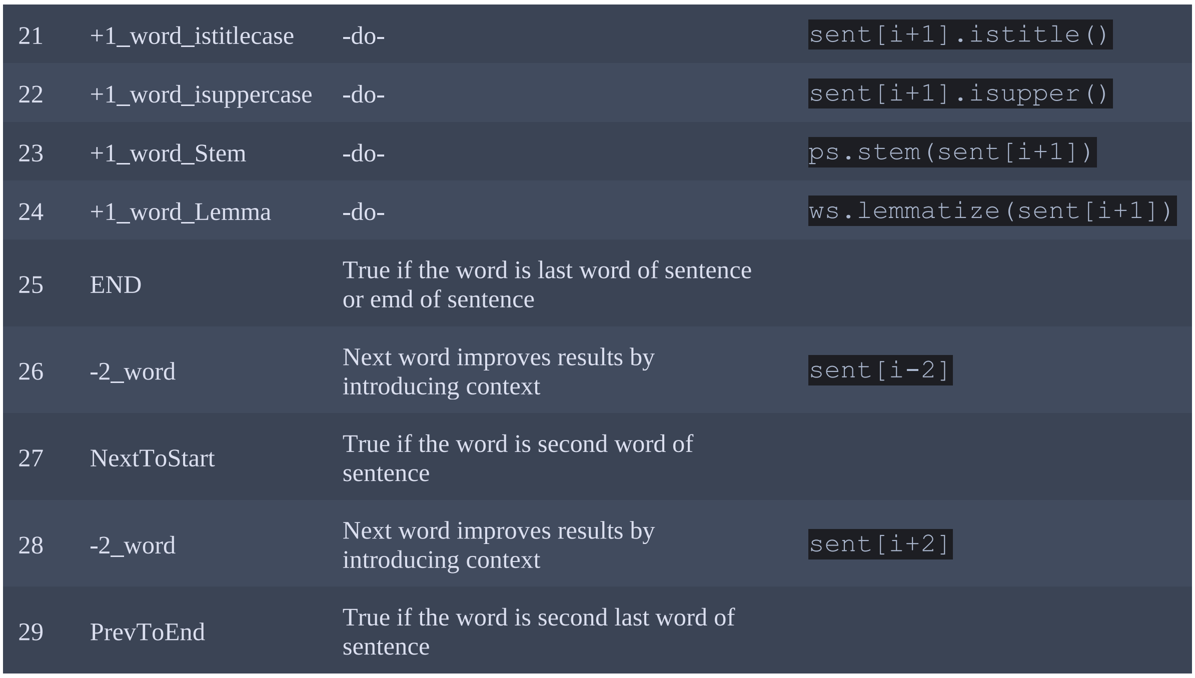
REPORT - ASSIGNMENT 4

- Feature Description :

S. No.	Feature	Convenience	Implementation
2	LowerCase	Reduces the ambiguity of case in words	`word.lower()`
3	IsUpperCase	Check if the word is all upper case, can check for emphasis in words	`word.isupper()`
4	IsTittleCase	Check if the word is all upper case, can check for emphasis in words	`word.istitle()`
5	IsDigit	Filtering digits adds in details to words and reduces chances of certain POS numerics can't be assigned	`word.isdigit()`
6	Suffix[-3:]	Extracting last 3 letters as suffix as most hindi suffixes involve 2 or 3 letters	`word[-3:]`
7	Suffix[-2:]	-do-	`word[-2:]`
8	Prefix[3:]	Extracting first 3 letters as prefix as most hindi prefixes involve 2 or 3 letters	`word[:3]`
9	Prefix[2:]	-do-	`word[:2]`
10	Stem	Extracting Stem from the word helps in removing common prefixes and suffixes to get root form of words reducing possible vocabulary	`ps.stem(word)`
11	Lemma	Extracting Lemma from the word helps in reducing possible vocabulary size and reduces ambiguity among words with same base meaning	`ws.lemmatize(word)`
12	-1_word	Previous word improves results by introducing context	`sent[i-1]`
13	-1_word_Lowercase	-do-	`sent[i-1].lower()`
14	-1_word_istitlecase	-do-	`sent[i-1].istitle()`
15	-1_word_isuppercase	-do-	`sent[i-1].isupper()`
16	-1_word_Stem	-do-	`ps.stem(sent[i-1])`
17	-1_word_Lemma	-do-	`ws.lemmatize(sent[i-1])`
18	START	True if the word is first word of sentence or begining of sentence	
19	+1_word	Next word improves results by introducing context	`sent[i+1]`
20	+1_word_Lowercase	-do-	`sent[i+1].lower()`



21	+1_word_istitlecase	-do-	`sent[i+1].istitle()`
22	+1_word_isuppercase	-do-	`sent[i+1].isupper()`
23	+1_word_Stem	-do-	`ps.stem(sent[i+1])`
24	+1_word_Lemma	-do-	`ws.lemmatize(sent[i+1])`
25	END	True if the word is last word of sentence or emd of sentence	
26	-2_word	Next word improves results by introducing context	`sent[i-2]`
27	NextToStart	True if the word is second word of sentence	
28	-2_word	Next word improves results by introducing context	`sent[i+2]`
29	PrevToEnd	True if the word is second last word of sentence	

- Hyperparameters tuned (if any) : we included 2 attributes to sklearn.crfsuite model i.e. c1 and c2 to the parameter space. We set cv = 3 i.e. we operate 3-fold cross-validation and perform a randomized search in the parameter space for fitting the model to get better f1 scores.

#		Train	Test
1.	10 most common transition features	4.627470 NUM IsDigit 3.543670 ADJ Suffix[-3:]:iwa 3.257322 VERB Suffix[-2:]:ne 3.056400 NOUN Suffix[-2:]:oM 2.796039 PRON Prefix[3:]:apa 2.752691 PRON Prefix[3:]:Apa 2.730343 PRON Prefix[2:]:Ap 2.597030 PRON Prefix[2:]:is 2.513943 PRON Prefix[2:]:ap 2.497892 NOUN bias	
2.	10 least common transition features	-1.126591 ADP -2_word:nihArane -1.276119 NOUN Prefix[2:]:ra -1.283902 X bias -1.287043 CCONJ +1:word_isuppercase -1.321439 AUX -2_word:xeKane -1.332726 NOUN Suffix[-3:]:Ina -1.346224 PROPN IsTittleCase -1.449749 AUX -2_word:jagaha -1.455890 PROPN Suffix[-2:]:oM -1.555633 NOUN IsDigit	

3.	Precision (per tag)	X 1.000 PART 1.000 CCONJ 1.000 SCONJ 1.000 ADJ 1.000 ADP 1.000 ADV 1.000 VERB 1.000 DET 1.000 COMMA 1.000 NOUN 1.000 PRON 1.000 PROPN 1.000 NUM 1.000 PUNCT 1.000 AUX 0.999 macro avg 1.000 weighted avg 1.000	X 0.000 PART 1.000 CCONJ 1.000 SCONJ 0.750 ADJ 0.658 ADP 0.964 ADV 0.643 VERB 0.904 DET 0.800 COMMA 0.000 NOUN 0.812 PRON 0.800 PROPN 0.648 NUM 0.958 PUNCT 1.000 AUX 0.949 macro avg 0.742 weighted avg 0.859
4.	Recall (per tag)	X 1.000 PART 1.000 CCONJ 1.000 SCONJ 1.000 ADJ 1.000 ADP 1.000 ADV 1.000 VERB 0.998 DET 1.000 COMMA 1.000 NOUN 1.000 PRON 1.000 PROPN 1.000 NUM 1.000 PUNCT 1.000 AUX 1.000 macro avg 1.000 weighted avg 1.000	X 0.000 PART 0.939 CCONJ 1.000 SCONJ 1.000 ADJ 0.777 ADP 0.970 ADV 0.429 VERB 0.859 DET 0.889 COMMA 0.000 NOUN 0.883 PRON 0.862 PROPN 0.562 NUM 0.920 PUNCT 0.828 AUX 0.949 micro avg 0.859 macro avg 0.742 weighted avg 0.859
5.	F-Score (per tag)	X 1.000 PART 1.000 CCONJ 1.000 SCONJ 1.000 ADJ 1.000 ADP 1.000 ADV 1.000 VERB 0.999 DET 1.000 COMMA 1.000 NOUN 1.000 PRON 1.000	X 0.000 PART 0.969 CCONJ 1.000 SCONJ 0.857 ADJ 0.712 ADP 0.967 ADV 0.514 VERB 0.881 DET 0.842 COMMA 0.000 NOUN 0.846 PRON 0.830

		PROPN 1.000 NUM 1.000 PUNCT 1.000 AUX 0.999 macro avg 1.000 weighted avg 1.000	PROPN 0.602 NUM 0.939 PUNCT 0.906 AUX 0.949 micro avg 0.859 macro avg 0.738 weighted avg 0.858
6.	Overall Accuracy	0.9998683864174783	0.8587257617728532