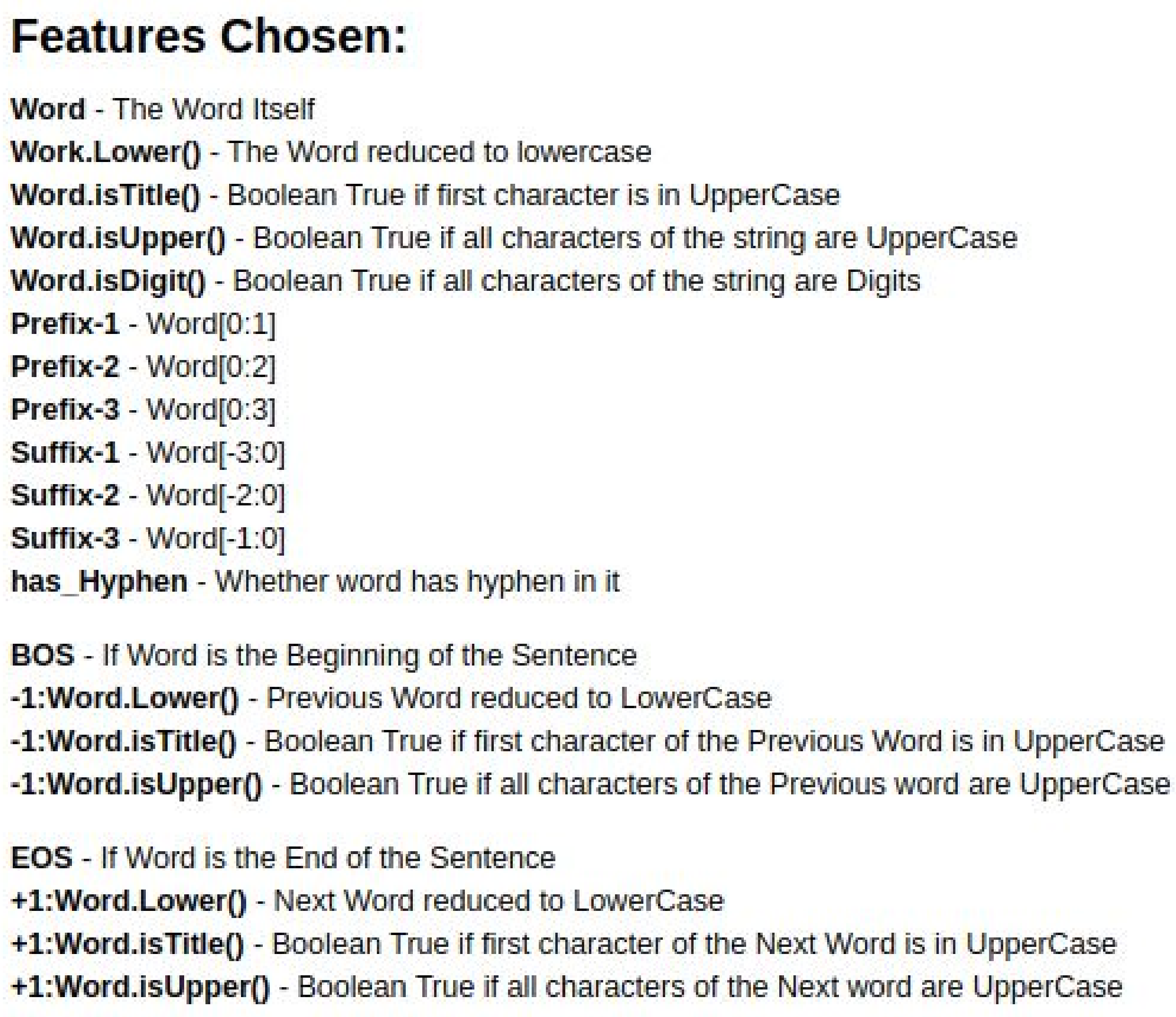
CS60075 - NATURAL LANGUAGE PROCESSING

ASSIGNMENT 2 - POS TAGGING of UNIVERSAL DEPENDENCY HINDI CORPUS

NAME: **Himanshu Mundhra**
ROLL No. : **16CS10057**



Features Chosen:

Word - The Word Itself
Work.Lower() - The Word reduced to lowercase
Word.isTitle() - Boolean True if first character is in UpperCase
Word.isUpper() - Boolean True if all characters of the string are UpperCase
Word.isDigit() - Boolean True if all characters of the string are Digits
Prefix-1 - Word[0:1]
Prefix-2 - Word[0:2]
Prefix-3 - Word[0:3]
Suffix-1 - Word[-3:0]
Suffix-2 - Word[-2:0]
Suffix-3 - Word[-1:0]
has_Hyphen - Whether word has hyphen in it

BOS - If Word is the Beginning of the Sentence
-1:Word.Lower() - Previous Word reduced to LowerCase
-1:Word.isTitle() - Boolean True if first character of the Previous Word is in UpperCase
-1:Word.isUpper() - Boolean True if all characters of the Previous word are UpperCase

EOS - If Word is the End of the Sentence
+1:Word.Lower() - Next Word reduced to LowerCase
+1:Word.isTitle() - Boolean True if first character of the Next Word is in UpperCase
+1:Word.isUpper() - Boolean True if all characters of the Next word are UpperCase

Top 10 Most Common POS Transition Features	
VERB => AUX	4.24992
PROPN => PROPN	3.52655
ADJ => NOUN	3.06800
NUM => NOUN	2.47544
DET => NOUN	2.22345
AUX => AUX	2.13554
NOUN => ADP	2.09563
PROPN => ADP	2.05048
NOUN => VERB	1.77167
VERB => SCONJ	1.69359

Top 10 Least Common POS Transition Features	
ADP => CCONJ	-1.24472
ADP => AUX	-1.24605
AUX => ADP	-1.27069
PROPN => AUX	-1.31203
DET => CCONJ	-1.34330
ADP => COMMA	-1.40946
CCONJ => AUX	-1.74072
ADJ => PRON	-1.92173
ADJ => ADP	-2.21294
DET => ADP	-2.41440

MODEL PREDICTION ON TRAINING DATA				
	Precision	**Recall**	**F1-Score**	**Support**
ADJ	1.00	1.00	1.00	570
ADP	1.00	1.00	1.00	1387
ADV	0.97	0.98	0.98	111
AUX	0.98	1.00	0.99	730
CCONJ	0.99	1.00	1.00	150
COMMA	1.00	1.00	1.00	114
DET	1.00	0.99	0.99	231
NOUN	1.00	1.00	1.00	1597
NUM	1.00	1.00	1.00	152
PART	1.00	1.00	1.00	163
PRON	1.00	1.00	1.00	431
PROPM	1.00	1.00	1.00	708
PUNCT	1.00	1.00	1.00	564
SCONJ	0.98	1.00	0.99	61
VERB	1.00	0.98	0.99	640

Accuracy Obtained on Training Data **0.996715280515044**

MODEL PREDICTION ON TESTING DATA

	Precision	Recall	F1-Score	Support
ADJ	0.69	0.74	0.71	94
ADP	0.96	0.98	0.97	309
ADV	0.71	0.48	0.57	21
AUX	0.98	0.96	0.97	139
CCONJ	1.00	1.00	1.00	25
COMMA	~	~	~	~
DET	0.82	0.89	0.85	36
NOUN	0.78	0.90	0.83	329
NUM	0.92	0.92	0.92	25
PART	0.97	1.00	0.99	33
PRON	0.92	0.83	0.87	65
PROPM	0.71	0.46	0.56	145
PUNCT	1.00	1.00	1.00	135
SCONJ	0.75	1.00	0.86	3
VERB	0.89	0.87	0.88	99

Accuracy Obtained on Testing Data **0.869684499314129**