CS60075 - NATURAL LANGUAGE PROCESSING
ASSIGNMENT-4 POS-TAGGING
NAME : KANISHK SINGH
ROLL : 17CS30018

(1) FEATURES:

Word - The Word Itself
Word.Lower() -   The Word reduced to lowercase
Word.isTitle() -   True if first character is in UpperCase
Word.isUpper() - True if all characters are Uppercase
Word.isDigit() -   True if all characters are Digits
Suffix-3 - Word[-3:]
Prefix-3 - Word[:3]
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() - True if previous word's first character is in uppercase
-1:Word.isUpper() - 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() - True if next word's first character is in uppercase
-1:Word.isUpper() - True if all characters of the next word are uppercase

(2) MOST AND LEAST COMMON TRANSITIONS

```
print_10_most_common(x_train,y_train,"Training-set")
print_10_least_common(x_train,y_train,"Training-set")
```

```
Top 10 Most Common POS Transition Features:
-------------------- Training-set --------------------
ADJ     =>  NOUN    3.99639
PROPN   =>  PROPN   3.91982
VERB    =>  AUX     3.88266
NOUN    =>  VERB    2.71304
NOUN    =>  ADP     2.63396
DET     =>  NOUN    2.54572
NUM     =>  NOUN    2.53846
ADJ     =>  VERB    2.33121
PROPN   =>  ADP     2.28136
NOUN    =>  NOUN    2.17695
```

```
Top 10 Least Common POS Transition Features:
-------------------- Training-set --------------------
COMMA   =>  ADP      -1.34458
ADJ     =>  PRON     -1.41543
DET     =>  CCONJ    -1.47929
ADP     =>  AUX      -1.49491
ADP     =>  CCONJ    -1.62925
ADP     =>  COMMA    -1.68510
ADJ     =>  ADP      -1.80175
AUX     =>  ADP      -1.80568
CCONJ   =>  AUX      -1.92248
DET     =>  ADP      -2.49674
```

(3) MODEL PREDICTION ON TRAINING DATA

```
           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       0.99      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
      PROPN       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
          X       1.00      1.00      1.00         2

   accuracy                          1.00      7611
  macro avg       1.00      1.00      1.00      7611
weighted avg       1.00      1.00      1.00      7611

precision:  0.9968742916302722
recall:     0.9968466692944422
f1-score:   0.9968472460628419
accuracy:   0.9968466692944422
```

(4) MODEL PREDICTION ON TEST DATA

```
             MODEL PREDICTION ON TEST DATA
------------------------------------------------------
              precision    recall  f1-score   support

        ADJ       0.67      0.79      0.73        94
        ADP       0.95      0.98      0.96       309
        ADV       0.71      0.48      0.57        21
        AUX       0.94      0.95      0.95       139
      CCONJ       1.00      1.00      1.00        25
        DET       0.86      0.89      0.88        36
       NOUN       0.77      0.90      0.83       329
        NUM       1.00      0.92      0.96        25
       PART       1.00      0.97      0.98        33
       PRON       0.87      0.85      0.86        65
      PROPN       0.65      0.44      0.53       145
      PUNCT       1.00      0.84      0.92       135
      SCONJ       0.50      0.67      0.57         3
       VERB       0.86      0.82      0.84        99

   accuracy                          0.85      1458
  macro avg       0.84      0.82      0.83      1458
weighted avg       0.85      0.85      0.85      1458

precision:  0.8513723498048027
recall:     0.8511659807956105
f1-score:   0.8466764320151978
accuracy:   0.8511659807956105
```