CS626 – Speech, NLP and Web

Assignment No: 2

Part-of-Speech Tagging

Prathamesh Karandikar (22M1155) **Raj Gothi (22M2160)** **Manas Jhalani (22M0806)**

Problem statement:

To implement a Part-of-Speech tagger in Python using the Hidden Markov Model

Dataset:

Brown Corpus (tagset = "universal")

Results:

1. Accuracy (5-fold cross-validation):

```
Fold 1:
Accuracy: 0.9561155018604992
Confusion Matrix:
Per POS Accuracy:
.: 0.9991192411924119
ADJ: 0.912771285475793
ADP: 0.9667686318131257
ADV: 0.8945967527282406
CONJ: 0.9913667153672466
DET: 0.9872970808764651
NOUN: 0.9469191200612893
NUM: 0.9048436963242872
PRON: 0.9841956915643245
PRT: 0.9044934221766615
VERB: 0.9449433921184411
X: 0.3768996960486322
```

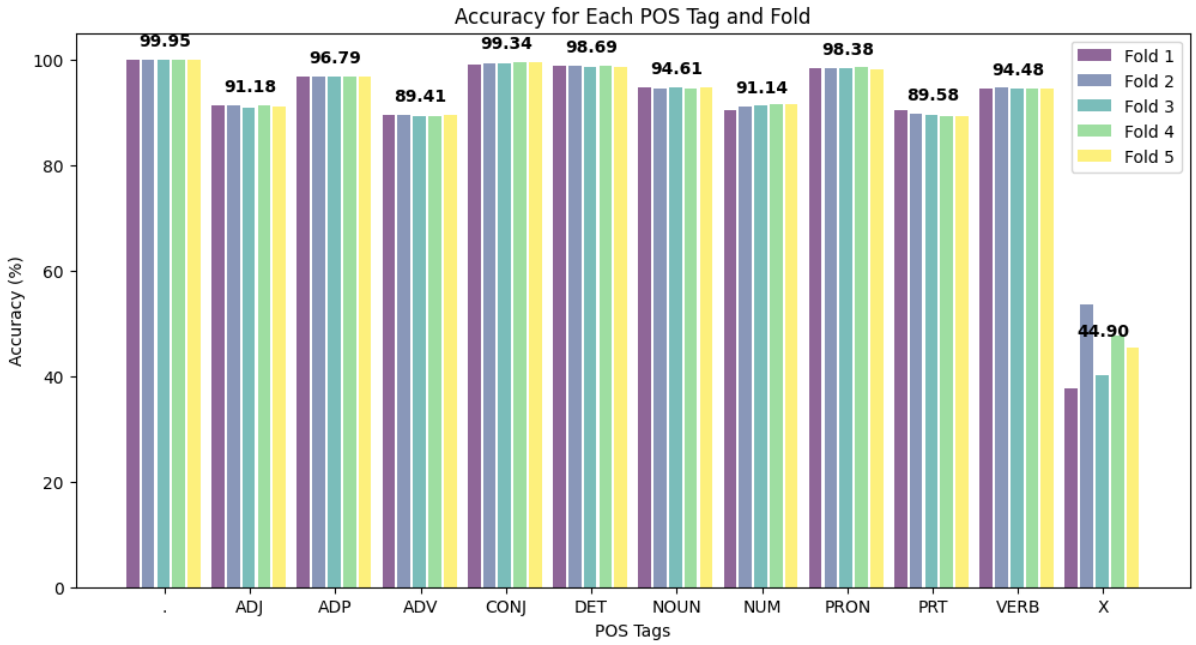
```
Fold 2:
Accuracy: 0.9561871771340925
Confusion Matrix:
Per POS Accuracy:
.: 0.9994888919176775
ADJ: 0.9121710921542713
ADP: 0.9685166706809345
ADV: 0.8955848317607263
CONJ: 0.9923861143373338
DET: 0.9869892551714056
NOUN: 0.9448284597568037
NUM: 0.9097569097569097
PRON: 0.9840242669362993
PRT: 0.8975535168195719
VERB: 0.946011392091577
X: 0.5365079365079365
```

```
Fold 3:
Accuracy: 0.9559839594472606
Confusion Matrix:
Per POS Accuracy:
.: 0.9994205862304022
ADJ: 0.9095913360856818
ADP: 0.9682727843246636
ADV: 0.8929376747542167
CONJ: 0.9927478902953587
DET: 0.9855698258045468
NOUN: 0.9470945675508496
NUM: 0.9120079391333112
PRON: 0.9836574265618535
PRT: 0.8938385389998293
VERB: 0.945383147589945
X: 0.4008438818565401
```

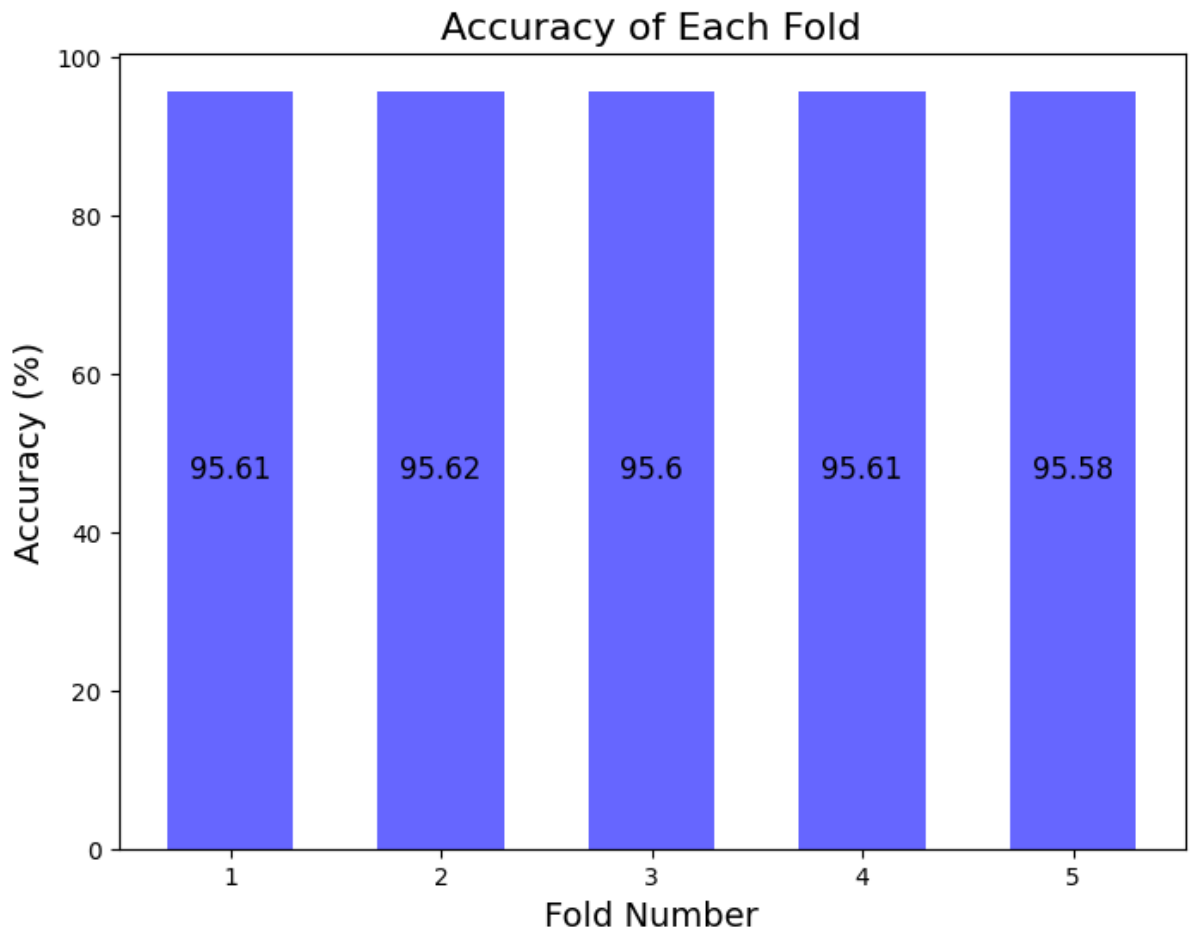
```
Fold 4:
Accuracy: 0.9560777414396014
Confusion Matrix:
Per POS Accuracy:
.: 1.0
ADJ: 0.9130487362438021
ADP: 0.9683096122835309
ADV: 0.8926396848137536
CONJ: 0.9958814932908198
DET: 0.9879923150816523
NOUN: 0.9453060475360191
NUM: 0.9141483516483516
PRON: 0.9853169268948403
PRT: 0.8915462547749543
VERB: 0.9437349364213092
X: 0.4763779527559055
```

```
Fold 5:
Accuracy: 0.9558357758893364
Confusion Matrix:
Per POS Accuracy:
.: 0.9995998532795358
ADJ: 0.9116232699976542
ADP: 0.9674706495778249
ADV: 0.8947643522954961
CONJ: 0.9945890234475651
DET: 0.9864387423717926
NOUN: 0.9463415504899948
NUM: 0.9163636363636364
PRON: 0.9819093501130666
PRT: 0.8914653784219002
VERB: 0.943807308792472
X: 0.4541832669322709
```

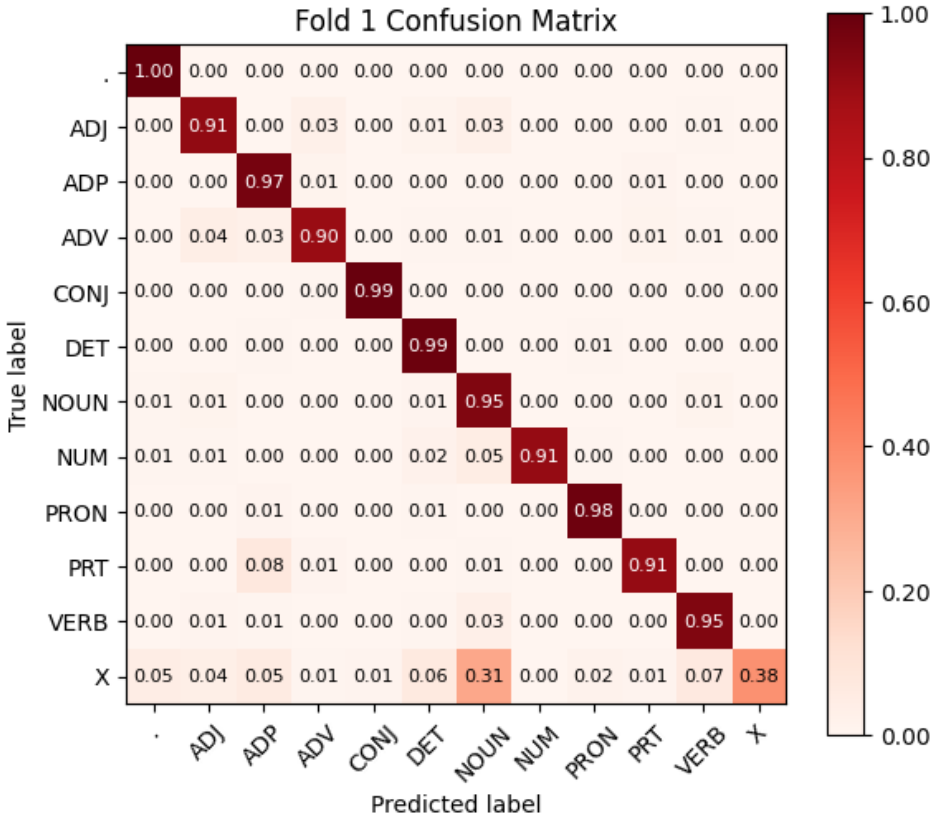
```
Mean Per POS Accuracy Across Folds:
.: 0.9995257145240055
ADJ: 0.9118411439914406
ADP: 0.9678676697360158
ADV: 0.8941046592704867
CONJ: 0.9933942473476648
DET: 0.9868574438611724
NOUN: 0.9460979490789914
NUM: 0.9114241066452993
PRON: 0.9838207324140769
PRT: 0.8957794222385835
VERB: 0.9447760354027489
X: 0.4489625468202570
```

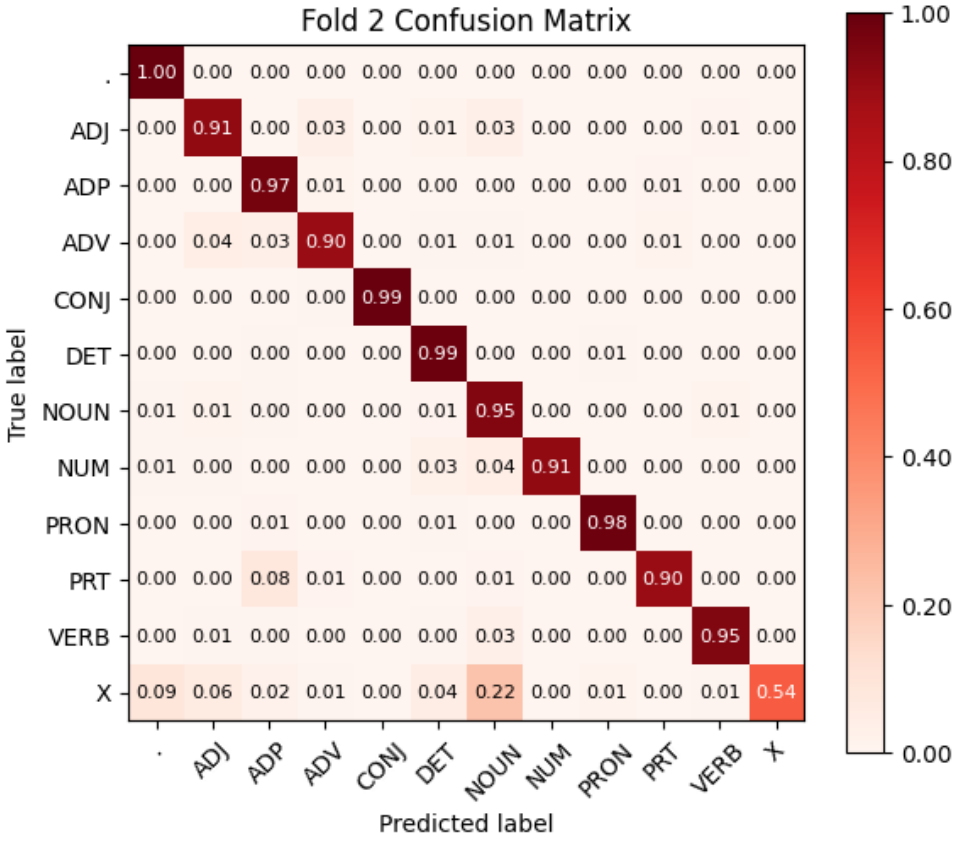


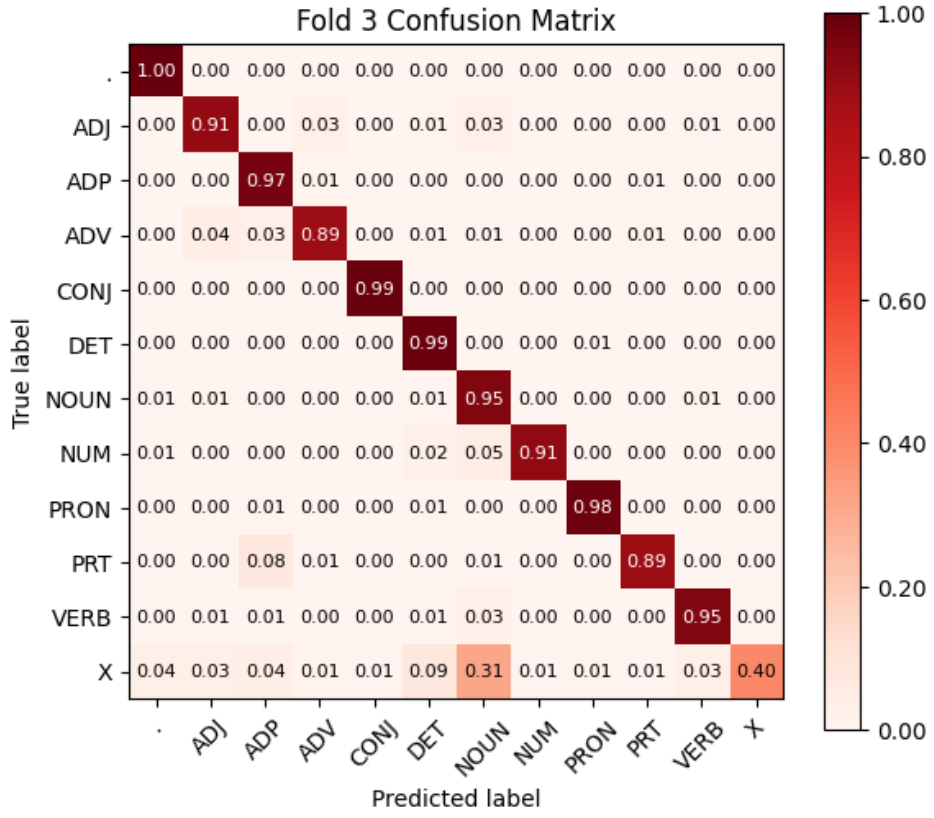
Mean Accuracy Across Folds: 0.9560400311541579

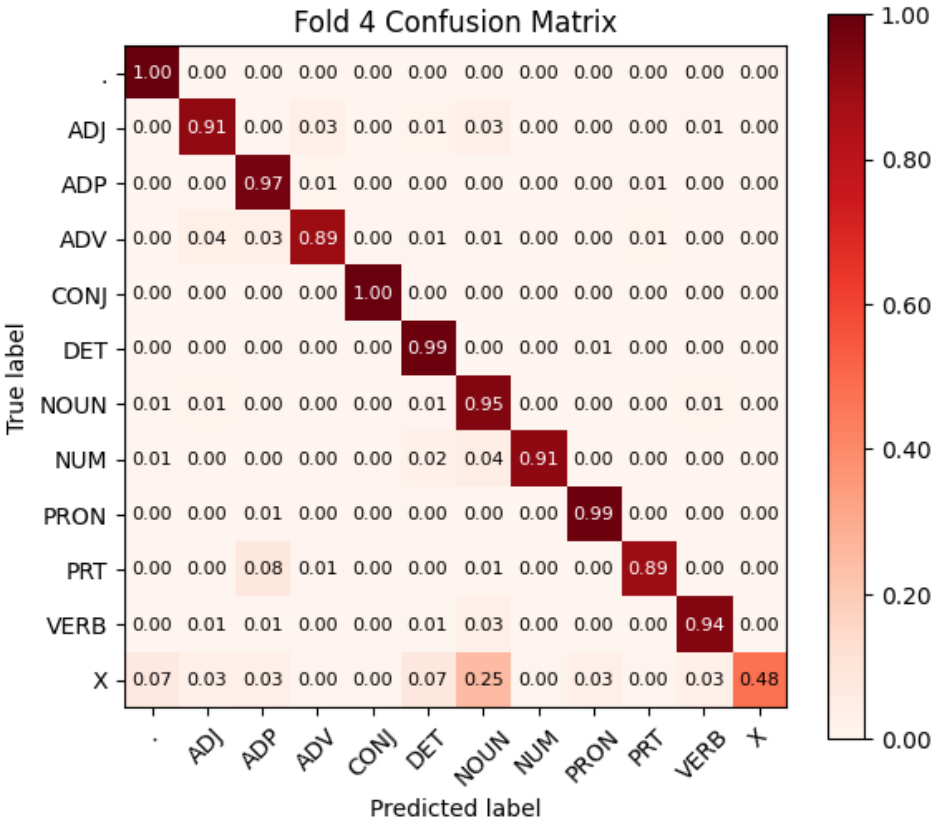


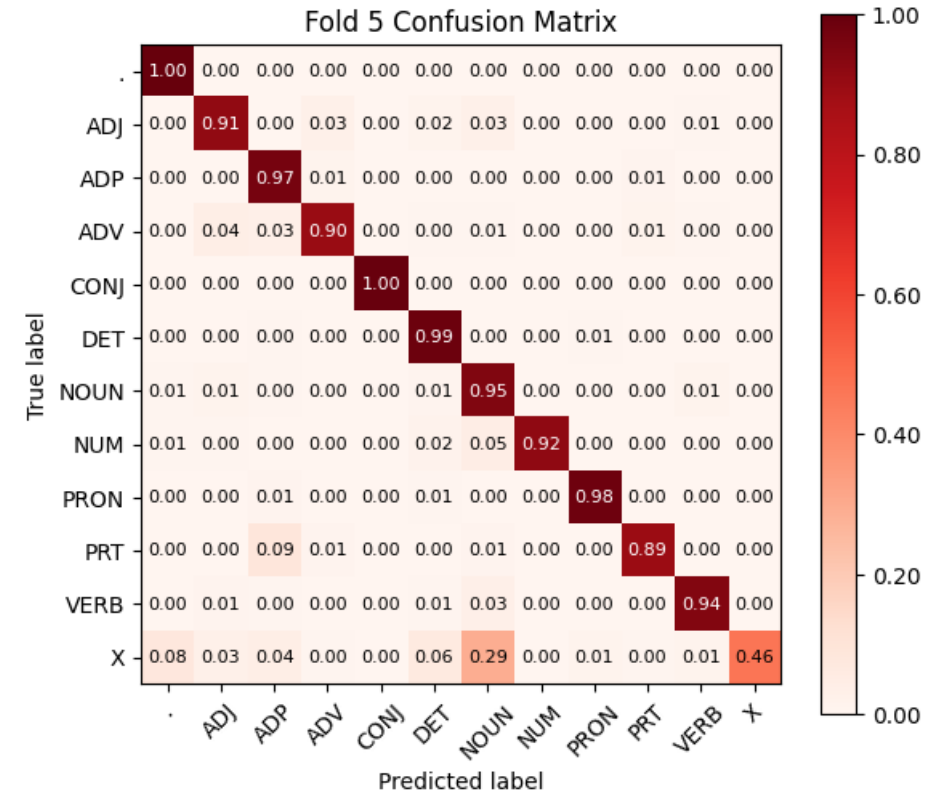
2. Confusion Matrix:

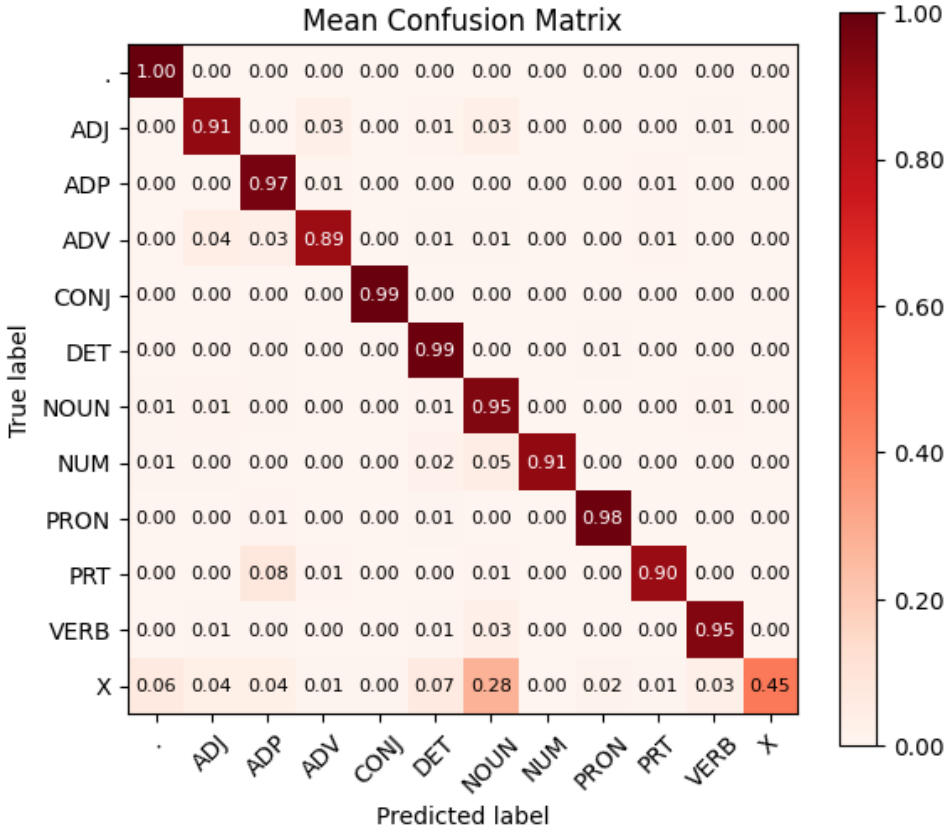





Fold 3 Confusion Matrix
True label
Predicted label
. ADJ ADP ADV CONJ DET NOUN NUM PRON PRT VERB X
1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.91 0.00 0.03 0.00 0.01 0.03 0.00 0.00 0.00 0.01 0.00
0.00 0.00 0.97 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00
0.00 0.04 0.03 0.89 0.00 0.01 0.01 0.00 0.00 0.01 0.00 0.00
0.00 0.00 0.00 0.00 0.99 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.99 0.00 0.00 0.01 0.00 0.00 0.00
0.01 0.01 0.00 0.00 0.00 0.01 0.95 0.00 0.00 0.00 0.01 0.00
0.01 0.00 0.00 0.00 0.00 0.02 0.05 0.91 0.00 0.00 0.00 0.00
0.00 0.00 0.01 0.00 0.00 0.01 0.00 0.00 0.98 0.00 0.00 0.00
0.00 0.00 0.08 0.01 0.00 0.00 0.01 0.00 0.00 0.89 0.00 0.00
0.00 0.01 0.01 0.00 0.00 0.01 0.03 0.00 0.00 0.00 0.95 0.00
0.04 0.03 0.04 0.01 0.01 0.09 0.31 0.01 0.01 0.01 0.03 0.40
1.00
0.80
0.60
0.40
0.20
0.00


Fold 4 Confusion Matrix
True label
Predicted label
. ADJ ADP ADV CONJ DET NOUN NUM PRON PRT VERB X
1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.91 0.00 0.03 0.00 0.01 0.03 0.00 0.00 0.00 0.01 0.00
0.00 0.00 0.97 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00
0.00 0.04 0.03 0.89 0.00 0.01 0.01 0.00 0.00 0.01 0.00 0.00
0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.99 0.00 0.00 0.01 0.00 0.00 0.00
0.01 0.01 0.00 0.00 0.00 0.01 0.95 0.00 0.00 0.00 0.01 0.00
0.01 0.00 0.00 0.00 0.00 0.02 0.04 0.91 0.00 0.00 0.00 0.00
0.00 0.00 0.01 0.00 0.00 0.00 0.00 0.00 0.99 0.00 0.00 0.00
0.00 0.00 0.08 0.01 0.00 0.00 0.01 0.00 0.00 0.89 0.00 0.00
0.00 0.01 0.01 0.00 0.00 0.01 0.03 0.00 0.00 0.00 0.94 0.00
0.07 0.03 0.03 0.00 0.00 0.07 0.25 0.00 0.03 0.00 0.03 0.48
1.00
0.80
0.60
0.40
0.20
0.00


Fold 5 Confusion Matrix
True label
Predicted label
. ADJ ADP ADV CONJ DET NOUN NUM PRON PRT VERB X
1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.91 0.00 0.03 0.00 0.02 0.03 0.00 0.00 0.00 0.01 0.00
0.00 0.00 0.97 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00
0.00 0.04 0.03 0.90 0.00 0.00 0.01 0.00 0.00 0.01 0.00 0.00
0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.99 0.00 0.00 0.01 0.00 0.00 0.00
0.01 0.01 0.00 0.00 0.00 0.01 0.95 0.00 0.00 0.00 0.01 0.00
0.01 0.00 0.00 0.00 0.00 0.02 0.05 0.92 0.00 0.00 0.00 0.00
0.00 0.00 0.01 0.00 0.00 0.01 0.00 0.00 0.98 0.00 0.00 0.00
0.00 0.00 0.09 0.01 0.00 0.00 0.01 0.00 0.00 0.89 0.00 0.00
0.00 0.01 0.00 0.00 0.00 0.01 0.03 0.00 0.00 0.00 0.94 0.00
0.08 0.03 0.04 0.00 0.00 0.06 0.29 0.00 0.01 0.00 0.01 0.46
1.00
0.80
0.60
0.40
0.20
0.00


Mean Confusion Matrix
True label
Predicted label
. ADJ ADP ADV CONJ DET NOUN NUM PRON PRT VERB X
1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.91 0.00 0.03 0.00 0.01 0.03 0.00 0.00 0.00 0.01 0.00
0.00 0.00 0.97 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00
0.00 0.04 0.03 0.89 0.00 0.01 0.01 0.00 0.00 0.01 0.00 0.00
0.00 0.00 0.00 0.00 0.99 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.99 0.00 0.00 0.01 0.00 0.00 0.00
0.01 0.01 0.00 0.00 0.00 0.01 0.95 0.00 0.00 0.00 0.01 0.00
0.01 0.00 0.00 0.00 0.00 0.02 0.05 0.91 0.00 0.00 0.00 0.00
0.00 0.00 0.01 0.00 0.00 0.01 0.00 0.00 0.98 0.00 0.00 0.00
0.00 0.00 0.08 0.01 0.00 0.00 0.01 0.00 0.00 0.90 0.00 0.00
0.00 0.01 0.00 0.00 0.00 0.01 0.03 0.00 0.00 0.00 0.95 0.00
0.06 0.04 0.04 0.01 0.00 0.07 0.28 0.00 0.02 0.01 0.03 0.45
1.00
0.80
0.60
0.40
0.20
0.00

3. Per POS accuracy:


```
{'.': 0.9995998532795358,
 'ADJ': 0.9116232699976542,
 'ADP': 0.9674706495778249,
 'ADV': 0.8947643522954961,
 'CONJ': 0.9945890234475651,
 'DET': 0.9864387423717926,
 'NOUN': 0.9463415504899948,
 'NUM': 0.9163636363636364,
 'PRON': 0.9819093501130666,
 'PRT': 0.8914653784219002,
 'VERB': 0.943807308792472,
 'X': 0.4541832669322709}
```


Error Analysis:

1. **'.'**: The overall accuracy for this tag is 100%.

2. **'ADJ':** This tag refers to 'adjectives' in the sentences. The overall accuracy is approximately 94%. The 6 % inaccuracies are due to words that belong to some other tags such as 'NOUN', 'ADV', 'VERB', etc. E.g., The word 'more' is present in the training corpus with tags 'ADV' and ADJ'.

Sentence: "He will be succeeded by Rob Ledford of Gainesville, who has been an assistant more than three years."

Expected Output:

```
['PRON', 'VERB', 'VERB', 'VERB', 'ADP', 'NOUN', 'NOUN', 'ADP', 'NOUN', '.',
'PRON', 'VERB', 'VERB', 'DET', 'NOUN', 'ADJ', 'ADP', 'NUM', 'NOUN', '.']
```

Output:

```
['PRON', 'VERB', 'VERB', 'VERB', 'ADP', 'NOUN', 'NOUN', 'ADP', 'ADJ', '.',
'PRON', 'VERB', 'VERB', 'DET', 'NOUN', 'ADV', 'ADP', 'NUM', 'NOUN', '.']
```

3. **'ADP':** The net accuracy for the 'adposition' tag ('ADP') is 98%. The inaccuracy of 2% is due to the words that also occur with tags 'PRT', 'ADV', etc apart from 'ADP' in the training corpus. E.g., the word 'out' has occurred with tags 'ADP', 'ADV' and 'PRT'.

Sentence: "There is a way out of this."

Expected Output:

```
['PRT', 'VERB', 'DET', 'NOUN', 'ADP', 'ADP', 'DET', '.']
```

Output:

```
['PRT', 'VERB', 'DET', 'NOUN', 'PRT', 'ADP', 'DET', '.']
```

4. **‘ADV’:** The net accuracy for the ‘adverb’ tag (‘ADV’) is about 87%. The inaccuracy of 13% is due to the words that also occur with tags ‘ADJ’ apart from ‘ADP’ in the training corpus. E.g., the word ‘most’ has occurred with tags ‘ADV’ and ‘ADJ’ with probabilities of 0.36 and 0.63 respectively.

Sentence: “The most recent film catalogue,”

Expected Output:

```
['DET', 'ADV', 'ADJ', 'NOUN', 'NOUN', '.']
```

Output:

```
['DET', 'ADJ', 'ADJ', 'NOUN', 'NOUN', '.']
```

5. **‘CONJ’:** This tag stands for the ‘conjunctions’ such as ‘and’, ‘but’, etc. The net accuracy is almost 100%. The small extent of inaccuracy is due to error in tagging the conjunctions like ‘yet’, which are available in training corpus with tag ‘ADV’, too.

Sentence: “Yet in the contemporary context this is precisely what one must not do.”

Expected Output:

```
['CONJ', 'ADP', 'DET', 'ADJ', 'NOUN', 'DET', 'VERB', 'ADV', 'DET', 'NOUN',
'VERB', 'ADV', 'VERB', '.']
```

Output:

```
['ADV', 'ADP', 'DET', 'ADJ', 'NOUN', 'DET', 'VERB', 'ADV', 'DET', 'NOUN',
'VERB', 'ADV', 'VERB', '.']
```

6. **‘DET’:** This tag stands for the ‘determiners’ such as ‘the’, ‘a’, ‘some’, etc. The net accuracy is about 98%. The 2% inaccuracy is due to errors in tagging the words like ‘that’, which are available in training corpus with tag ‘ADV’, too.

Sentence: “Yet in the contemporary context this is precisely what one must not do.”

Expected Output:

```
['CONJ', 'ADP', 'DET', 'ADJ', 'NOUN', 'DET', 'VERB', 'ADV', 'DET', 'NOUN',
'VERB', 'ADV', 'VERB', '.']
```

Output:

```
['ADV', 'ADP', 'DET', 'ADJ', 'NOUN', 'DET', 'VERB', 'ADV', 'DET', 'NOUN',
'VERB', 'ADV', 'VERB', '.']
```

7. **‘NOUN’:** The net accuracy for the ‘noun’ tag (‘NOUN’) is about 93%. The inaccuracy of 7% is due to the words which are available with tags like ‘VERB’, ‘ADJ’, etc apart from ‘NOUN’. E.g., the word ‘guides’ has also occurred in the corpus with tag ‘VERB’

Sentence: “Teaching guides are included with each record.”

Expected Output:

```
['VERB', 'NOUN', 'VERB', 'VERB', 'ADP', 'DET', 'NOUN', '.']
```

Output:

```
['VERB', 'VERB', 'VERB', 'VERB', 'ADP', 'DET', 'NOUN', '.']
```

8. **‘NUM’:** This tag is for the words indicating ‘numbers’. The net accuracy is approximately 82%. The inaccuracy is due to the unavailability of numerical symbols in the

corpus. The errors are also caused due to words like ‘one’ which are available in corpus with tags like ‘NOUN’ and ‘DET’, apart from ‘NUM’.

Sentence: “As wars go, Laos is an extremely little one.”

Expected Output:

```
['ADP', 'NOUN', 'VERB', '.', 'NOUN', 'VERB', 'DET', 'ADV', 'ADJ', 'NUM', '.']
```

Output:

```
['ADP', 'NOUN', 'VERB', '.', 'NOUN', 'VERB', 'DET', 'ADV', 'ADJ', 'NOUN', '.']
```

9. **‘PRON’:** The net accuracy for the ‘pronoun’ tag (‘PRON’) is about 95%. The inaccuracy of 5% is due to the words such as ‘that’ which are available with other tags like apart from ‘PRON’. The word ‘that’ has occurred in the corpus with tags ‘DET’, ‘ADP’ and ‘ADV’, too.

Sentence: “This is a problem that goes considerably beyond questions of salary and tenure.”

Expected Output:

```
['DET', 'VERB', 'DET', 'NOUN', 'PRON', 'VERB', 'ADV', 'ADP', 'NOUN', 'ADP',
'NOUN', 'CONJ', 'NOUN', '.']
```

Output:

```
['DET', 'VERB', 'DET', 'NOUN', 'ADP', 'VERB', 'ADV', 'ADP', 'NOUN', 'ADP',
'NOUN', 'CONJ', 'NOUN', '.']
```

10. **‘PRT’:** The net accuracy for the ‘particle’ (PRT) tag is approximately 78%. The inaccuracy is due to the errors caused due to words like ‘up’ which are available in the training corpus with tags like ‘ADP’ and ‘PRT’.

Sentence: “I have a hunch Marv Breeding might move up a notch.”

Expected Output:

```
['PRON', 'VERB', 'DET', 'NOUN', 'NOUN', 'NOUN', 'VERB', 'VERB', 'PRT', 'DET',
'NOUN', '.']
```

Output:

```
['PRON', 'VERB', 'DET', 'NOUN', 'NOUN', 'NOUN', 'VERB', 'VERB', 'ADP', 'DET',
'NOUN', '.']
```

11. **‘VERB’:** The net accuracy for the ‘verb’ (VERB) tag is approximately 94%. The inaccuracy is due to the errors caused due to verbs that also come with other tags such as ‘NOUN’, ‘ADJ’, etc. E.g., the word ‘answer’ occurs in the corpus both as noun (NOUN) and verb (VERB).

Sentence: “He selects queries or general interest to answer.”

Expected Output:

```
['PRON', 'VERB', 'NOUN', 'CONJ', 'ADJ', 'NOUN', 'PRT', 'VERB', '.']
```

Output:

```
['PRON', 'VERB', 'ADJ', 'CONJ', 'ADJ', 'NOUN', 'ADP', 'NOUN', '.']
```

12. **‘X’:** The tag X is used for words that for some reason cannot be assigned a real part-of-speech category. These words can be used as one of the above tags depending on their

meanings and the structure of the sentence. Typically, foreign words come in this category. Hence, most of the words are not present in the training corpus. This causes comparably lower accuracy (approx. 36%)

Sentence: "It would seem to represent esprit de corps run riot"

Expected Output:
`['PRON', 'VERB', 'VERB', 'PRT', 'VERB', 'X', 'X', 'X', 'VERB', 'NOUN', '.']`
Output:
`['PRON', 'VERB', 'VERB', 'PRT', 'VERB', 'DET', 'NOUN', 'NOUN', 'VERB', 'NOUN', '.']`

Strength and Weakness:

- **Strengths:**

1. The implemented POS tagger finds the part-of-speech tags of input sentence with net accuracy of about 95%
2. The tagging of unseen words in the sentence does not affect the POS tagging of the other words in the sentence.

- **Weaknesses:**

1. The implemented POS tagger could lead to wrong prediction of POS tag for unseen words, i.e. the words in given sentence that are not available in the training corpus.

Example: a.

Sentence: "My name is Raj." ('Raj' is an unseen word.)

Output: `['DET', 'NOUN', 'VERB', 'ADJ', '.']`
Expected Result: `['DET', 'NOUN', 'VERB', 'NOUN', '.']`

Example: b.

Sentence: "The September-October term jury had been charged." ('September-October' is an unseen word.)

Output: `['DET', 'ADJ', 'NOUN', 'NOUN', 'VERB', 'VERB', 'VERB', '.']`
Expected Result: `['DET', 'NOUN', 'NOUN', 'NOUN', 'VERB', 'VERB', 'VERB', '.']`

2. It leads to wrong prediction of POS tag for the words from training corpus that belong to two or more than two tags with comparable probabilities.

Example: a. The word 'to' has occurred in the training set with tags 'PRT' and 'ADP' with likelihood of almost 40% and 60%, respectively. Hence, POS tagger makes error in tagging for the word 'to'.

Sentence: "It urged that the city "take steps to remedy " this problem ."

Output:

```
['PRON', 'VERB', 'ADP', 'DET', 'NOUN', '.', 'VERB', 'NOUN', 'PRT', 'VERB',
'.', 'DET', 'NOUN', '.']
```

Expected Result:

```
['PRON', 'VERB', 'ADP', 'DET', 'NOUN', '.', 'VERB', 'NOUN', 'ADP', 'NOUN',
'.', 'DET', 'NOUN', '.']
```

Example: b. Similar to 'to', The word 'warning' has occurred in the training set with tags 'NOUN' and 'VERB' with probabilities 0.395 and 0.604, respectively.

Sentence: "Despite the warning, there was a unanimous vote to enter a candidate, according to Republicans who attended."

Output:

```
['ADP', 'DET', 'VERB', '.', 'PRT', 'VERB', 'DET', 'ADJ', 'NOUN', 'PRT', 'VERB',
'DET', 'NOUN', '.', 'ADP', 'ADP', 'NOUN', 'PRON', 'VERB', '.']
```

Expected Result:

```
['ADP', 'DET', 'NOUN', '.', 'PRT', 'VERB', 'DET', 'ADJ', 'NOUN', 'PRT', 'VERB',
'DET', 'NOUN', '.', 'ADP', 'ADP', 'NOUN', 'PRON', 'VERB', '.']
```

Learning:

1) We have gained an understanding of how to implement a Generative Hidden Markov Model (HMM) for Part-of-Speech (POS) tagging and how to train the HMM based model.
2) We now comprehend the concept of Cross-validation in Machine Learning.
3) We have learned how to calculate precision, recall, F1-score (including variations like F0.5 and F2 scores), and how to create a confusion matrix using the scikit-learn (sklearn) library.
4) We have acquired the knowledge of implementing smoothing methods to handle unseen or rare cases in statistical models.
5) We have delved into error analysis, which involves applying linguistic and language knowledge to analyze and understand the errors made by machine learning models