Assignment 1(a)

POS Tagging using HMM

Group ID: 85

1. **Priyansh Singh <22b1856>**
2. **Arnav Agarwal<22b3917>**
3. **Hardik Jangir<22b3901>**
4. **Kanishk Garg<210050080>**

Problem statement:

- **Objective**: Given a sequence of words, produce the POS tag sequence using HMM-Viterbi.
- **EXAMPLE**:

  INPUT: The quick brown fox jumps over the lazy dog

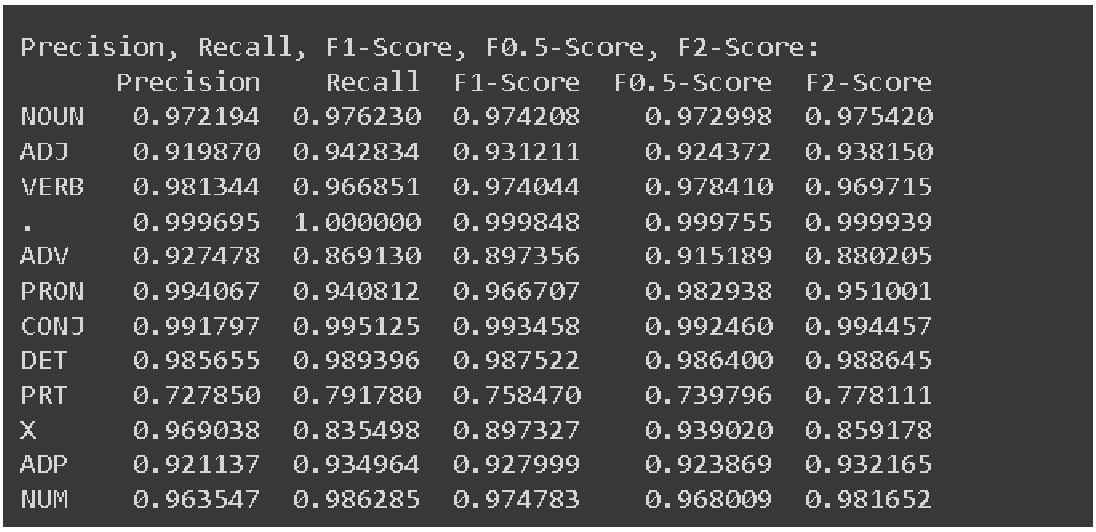
  OUTPUT: The(DET) quick(ADJ) brown(ADJ) fox(NN) jumps(VERB) over(ADP) the(DET) lazy(ADJ) dog(NN).
- **DATASET**: Brown Corpus
- Universal tag set [NUM, PRT, CONJ, DET, ADP, VERB, ADV, PRON, ADJ, X, NOUN, .]
- K fold cross validation (k=5)

DATA PROCESSING INFO (Pre-Processing):

- loaded the dataset from the NLTK **Brown Corpus**
- converted the data from an iterator to a list.

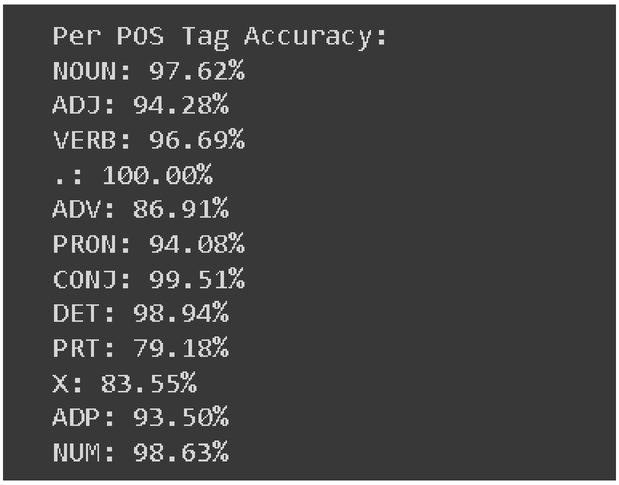
OVERALL PERFORMANCE

- Precision
- Recall:
- F-score(3 values)
  -F1-score
  -F0.5-score
  -F2-score



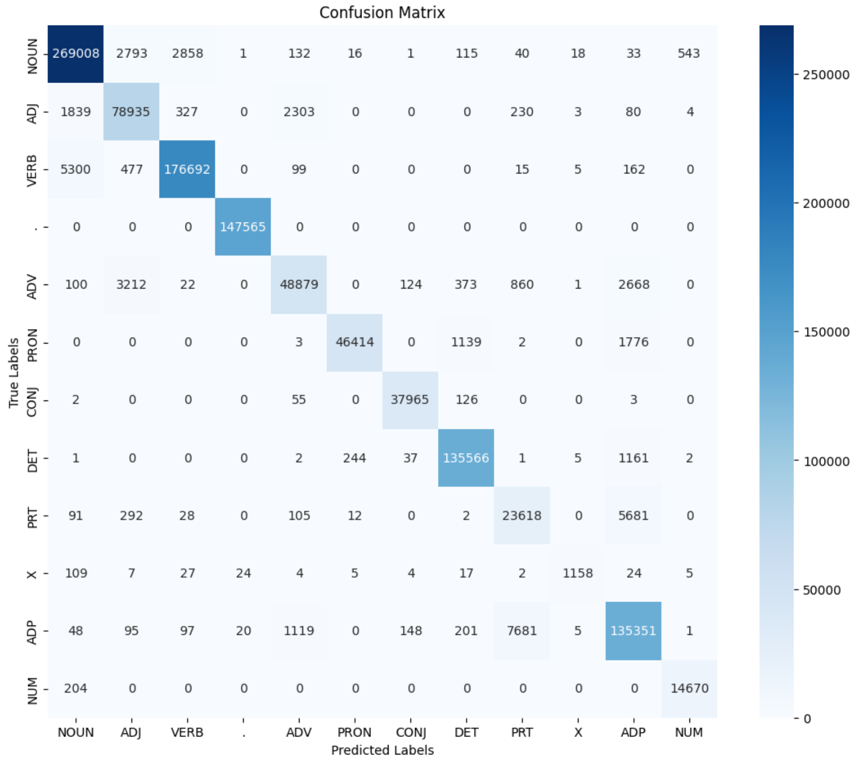
```
Precision, Recall, F1-Score, F0.5-Score, F2-Score:
      Precision    Recall  F1-Score  F0.5-Score  F2-Score
NOUN   0.972194  0.976230  0.974208    0.972998  0.975420
ADJ    0.919870  0.942834  0.931211    0.924372  0.938150
VERB   0.981344  0.966851  0.974044    0.978410  0.969715
.      0.999695  1.000000  0.999848    0.999755  0.999939
ADV    0.927478  0.869130  0.897356    0.915189  0.880205
PRON   0.994067  0.940812  0.966707    0.982938  0.951001
CONJ   0.991797  0.995125  0.993458    0.992460  0.994457
DET    0.985655  0.989396  0.987522    0.986400  0.988645
PRT    0.727850  0.791780  0.758470    0.739796  0.778111
X      0.969038  0.835498  0.897327    0.939020  0.859178
ADP    0.921137  0.934964  0.927999    0.923869  0.932165
NUM    0.963547  0.986285  0.974783    0.968009  0.981652
```

PER POS performance:

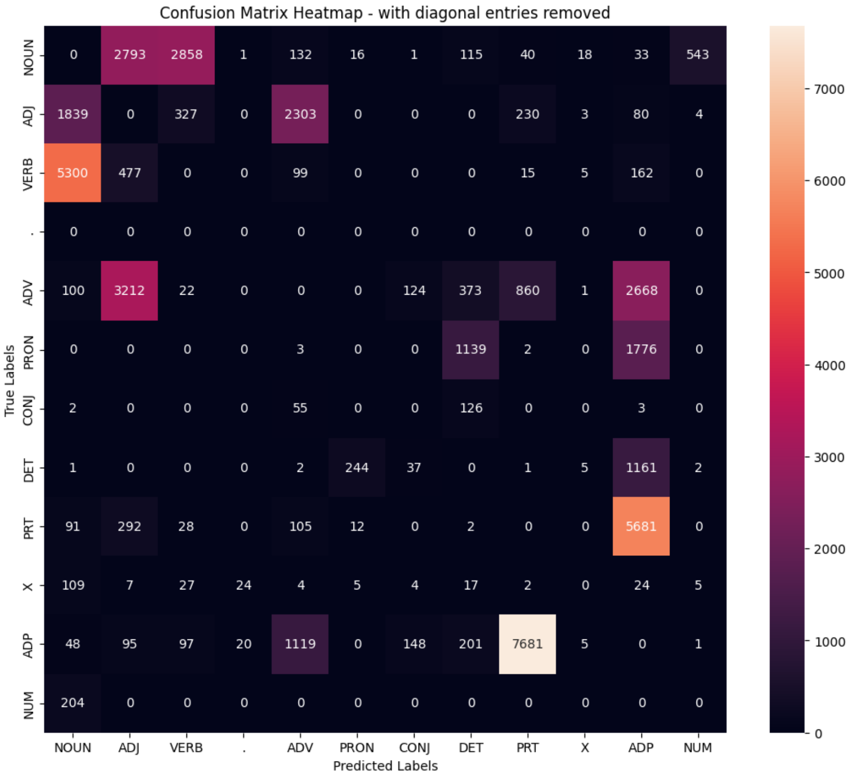


```
Per POS Tag Accuracy:
NOUN: 97.62%
ADJ: 94.28%
VERB: 96.69%
.: 100.00%
ADV: 86.91%
PRON: 94.08%
CONJ: 99.51%
DET: 98.94%
PRT: 79.18%
X: 83.55%
ADP: 93.50%
NUM: 98.63%
```

CONFUSION MATRIX



	NOUN	ADJ	VERB	.	ADV	PRON	CONJ	DET	PRT	X	ADP	NUM
NOUN	269008	2793	2858	1	132	16	1	115	40	18	33	543
ADJ	1839	78935	327	0	2303	0	0	0	230	3	80	4
VERB	5300	477	176692	0	99	0	0	0	15	5	162	0
.	0	0	0	147565	0	0	0	0	0	0	0	0
ADV	100	3212	22	0	48879	0	124	373	860	1	2668	0
PRON	0	0	0	0	3	46414	0	1139	2	0	1776	0
CONJ	2	0	0	0	55	0	37965	126	0	0	3	0
DET	1	0	0	0	2	244	37	135566	1	5	1161	2
PRT	91	292	28	0	105	12	0	2	23618	0	5681	0
X	109	7	27	24	4	5	4	17	2	1158	24	5
ADP	48	95	97	20	1119	0	148	201	7681	5	135351	1
NUM	204	0	0	0	0	0	0	0	0	0	0	14670



	NOUN	ADJ	VERB	.	ADV	PRON	CONJ	DET	PRT	X	ADP	NUM
NOUN	0	2793	2858	1	132	16	1	115	40	18	33	543
ADJ	1839	0	327	0	2303	0	0	0	230	3	80	4
VERB	5300	477	0	0	99	0	0	0	15	5	162	0
.	0	0	0	0	0	0	0	0	0	0	0	0
ADV	100	3212	22	0	0	0	124	373	860	1	2668	0
PRON	0	0	0	0	3	0	0	1139	2	0	1776	0
CONJ	2	0	0	0	55	0	0	126	0	0	3	0
DET	1	0	0	0	2	244	37	0	1	5	1161	2
PRT	91	292	28	0	105	12	0	2	0	0	5681	0
X	109	7	27	24	4	5	4	17	2	0	24	5
ADP	48	95	97	20	1119	0	148	201	7681	5	0	1
NUM	204	0	0	0	0	0	0	0	0	0	0	0

ERROR ANALYSIS:

- The values along the diagonal are significantly higher than off-diagonal values, indicating that the model performs well in tagging words correctly.
- There are significant misclassifications between ADP and other categories such as NOUN (33 times) and ADJ (80 times). This may be due to prepositions like "after" and "before," which can function as both prepositions and conjunctions or adverbs in different contexts.
- 3,212 adjectives were wrongly tagged as adverbs, while 2,303 adverbs were misclassified as adjectives. The overlap between adjectives and adverbs is expected since they are both modifiers and share similar forms
- There are 5,300 instances where a VERB was misclassified as a NOUN, and 3,858 NOUNs were tagged as VERBs.

CHALLENGES FACED:

- Dealing with the words which are not present in the corpus.
- Speeding up the Viterbi process, as it tends to be very slow for a direct approach and needs optimization

LEARNINGS:

- Learned about how applying the direct approach for calculating the probabilities can be a slow process
- Optimization/speed-up of the Viterbi algorithm by storing the word-tag count and tag count/frequency in a dictionary
- Speed up of emission probability calculation using only the tags that are possible for a word in the entire dataset
- Assigned a small value (10^-6) to any count if it is zero and is in the denominator in some expression

BENCHMARKING AGAINST CHAT GPT:

1. When does ChatGPT perform poor:
   - Cases of ambiguity for words with multiple meanings in ambiguous sentences

2. When does HMM perform better than GPT:
   - Case of structured and predictable text sequences
   - For the case of limited vocabulary, since GPT is based on a larger understanding of a language

Let's see the response to the input sentence for both by GPT and HMM:

ChatGPT RESPONSE:

POS Tags:
CHATGPT
1. The/DT old/JJ man/VBP the/DT boats/NNS
(Note: "man" here is a verb, not a noun, which can confuse the tagger.)

HMM:
[('The', 'DET'), ('old', 'ADJ'), ('man', 'NOUN'), ('the', 'DET'), ('boats', 'NOUN')]
Ambiguity: In this sentence, "man" is a verb (meaning "operate"), and "the old" is a noun phrase (meaning "old people"). This can easily be confused as a typical subject-verb-object sentence.

2. "She decided to paint the fence green."

POS Tags:
She/PRP decided/VBD to/TO paint/VB the/DT fence/NN green/JJ

HMM:
[('She', 'PRON'), ('decided', 'VERB'), ('to', 'PRT'), ('paint', 'VERB'), ('the', 'DET'), ('fence', 'NOUN'), ('green', 'ADJ')]

Ambiguity: "Green" could be tagged as an adjective (color) or mistakenly as a noun (if the tagger expects it to refer to something concrete).

3. "The bank can refuse loans in the coming months."

POS Tags:

The/DT bank/NN can/MD refuse/VB loans/NNS in/IN the/DT
coming/VBG months/NNS

HMM:
[('The', 'DET'), ('bank', 'NOUN'), ('can', 'VERB'), ('refuse', 'VERB'), ('loans', 'NOUN'), ('in', 'ADP'), ('the', 'DET'), ('coming', 'VERB'), ('months', 'NOUN')]

Ambiguity: "Bank" could refer to a financial institution (noun) or a riverbank (noun). Without full context, the word may confuse a tagger.

4. "Time flies like an arrow."

POS Tags:
Time/NN flies/VBZ like/IN an/DT arrow/NN
(Note: "flies" is a verb here, but the word "flies" can also be a noun.)

HMM:
[('Time', 'NOUN'), ('flies', 'NOUN'), ('like', 'ADP'), ('an', 'DET'), ('arrow', 'NOUN')]

Ambiguity: "Flies" could be a verb or noun, and "like" could be a preposition or verb. The sentence could be interpreted as describing how time moves, or it could mean something completely different.

5. "He turned around the corner."

POS Tags:
He/PRP turned/VBD around/RB the/DT corner/NN
"I'll meet you by the park."

HMM:
[('He', 'PRON'), ('turned', 'VERB'), ('around', 'ADV'), ('the', 'DET'), ('corner', 'NOUN')]

Ambiguity: "Turned" is a verb, but "around" can be either an adverb (as in this case) or a preposition.

6. “I'll meet you by the park.”
GPT
I/PRP 'll/MD meet/VB you/PRP by/IN the/DT park/NN

HMM
[("I'll", 'PRT'), ('meet', 'VERB'), ('you', 'PRON'), ('by', 'ADP'), ('the', 'DET'), ('park', 'NOUN')]

Ambiguity: "By" can be a preposition or an adverb, depending on context. A POS tagger might confuse it with a locative phrase
or a time phrase.

7. "Can you book a table for tonight"
GPT
Can/MD you/PRP book/VB a/DT table/NN for/IN tonight/NN

HMM
[('Can', 'VERB'), ('you', 'PRON'), ('book', 'NOUN'), ('a', 'DET'), ('table', 'NOUN'), ('for', 'ADP'), ('tonight', 'NOUN')]
Ambiguity: "Book" can be a verb (reserve a table) or a noun (the physical object).

8. After running the program, the results seemed wrong
GPT
After/IN running/VBG the/DT program/NN ,/, the/DT results/NNS seemed/VBD wrong/JJ

HMM
[('After', 'ADP'), ('running', 'VERB'), ('the', 'DET'), ('program,', 'DET'), ('the', 'DET'), ('results', 'NOUN'), ('seemed', 'VERB'), ('wrong', 'ADJ')]

Ambiguity: "Running" can be interpreted as a noun (gerund form) or as part of a verb phrase. A POS tagger might confuse these roles in complex sentence structures.

9. "The company bought a new plant last week."

GPT

The/DT company/NN bought/VBD a/DT new/JJ
plant/NN last/JJ week/NN

HMM

[('The', 'DET'), ('company', 'NOUN'), ('bought', 'VERB'), ('a', 'DET'), ('new', 'ADJ'), ('plant', 'NOUN'), ('last', 'ADJ'), ('week', 'NOUN')]

Ambiguity: "Plant" could be tagged as a noun referring to a factory or a living organism. Both are valid, but the correct tag depends on context.

VITERBI EXPLAINED:

- **Initialize State**: It maintains a list state to store the most likely tag for each word. The list T contains all possible tags
- Determine Possible Tags: If the word exists in the training data (word_tag_freq), use its possible tags; otherwise, consider all tags.
- **Calculate State Probability**: Multiply the transition and emission probabilities to get the state probability for each possible tag.
- **Choose Most Likely Tag**: Select the tag with the highest probability and append it to state.
- **Return Result**: The function returns the words paired with their most likely tags.