CS542 (Fall 2023) Written Assignment 2
Sequence Labeling

Due October 6, 2023

1 Hidden Markov Models

(You may find the discussion in Chapter A of the Jurafsky and Martin book helpful.)

You are given the following short sentences, tagged with parts of speech:

```
Alice/NN admired/VB Dorothy/NN
Dorothy/NN admired/VB every/DT dwarf/NN
Dorothy/NN cheered/VB
every/DT dwarf/NN cheered/VB
```

1. Train a hidden Markov model on the above data. Specifically, compute the initial probability distribution π:

y_1	NN	VB	DT
$P(y_1)$	•	•	•

The transition matrix $\mathbf{A}$:

$P(y_i \mid y_{i-1})$		y_i		
		NN	VB	DT
y_{i-1}	NN	•	•	•
	VB	•	•	•
	DT	•	•	•

And the emission matrix **B**:

$P(x_i \mid y_i)$		y_i		
		NN	VB	DT
x_i	Alice	•	•	•
	admired	•	•	•
	Dorothy	•	•	•
	every	•	•	•
	dwarf	•	•	•
	cheered	•	•	•
	<UNK>	•	•	•

Note that you should account for the unknown word <UNK>, but you don't need to account for the start symbol <S> or the stop symbol </S>. There are ways to train the probabilities of <UNK> from the training set, but for this assignment, you can simply let count(<UNK>, y) = 1 for all tags y (before smoothing). You should use add-1 smoothing on all three tables.

2. Use the forward algorithm to compute the probability of the following sentence:

   Alice cheered

   As part of your answer, you should fill in the forward trellis below:

	Alice	cheered
NN	•	•
VB	•	•
DT	•	•

3. Use the Viterbi algorithm to compute the best tag sequence for the following sentence:

   Goldilocks cheered

   As part of your answer, you should fill in the Viterbi trellis below. You should also keep track of backpointers, either using arrows or in a separate table.

	Goldilocks	cheered
NN	•	•
VB	•	•
DT	•	•

Submission Instructions

Please submit your solutions (in PDF format) to the submission box on Canvas.