NLP Assignment - HW3 - DIVYA PORWAL - DXP170020

Writeup Question 1.1:
The advantage and disadvantage of using smaller tagset are as follows :
Advantage:
A lot of tags would mean a lot of parameters to learn which can get very complex.
Having a small tagset helps reduce complexity by learning less parameters.
More tags allow more ambiguity, hence having less tags helps in unambiguousness.

Disadvantage :
We might not have a better model to account for the various tags during testing.

Writeup Question 2.1:
We need to add skip bigrams because trigrams alone will be insufficient to properly identify and tag every word in every sentence. Skip bigrams allow the ability to tag words by allowing to solve the problem with less information as compared to trigrams. This can be used in a similar way as HMM Smoothing

Writeup Question 2.2:
We study the suffixes because one of the best ways to tell the part of speech of an unknown English word is its suffix and it is an easy feature to generate!
We can learn a separate suffix model $p(t|ln-j+1 . . . ln)$ that predicts a tag based on a word suffix of length j. Also, we can use it to study the emission probabilities. The suffix model gives us an estimate for $p(t|w) \approx p(t|ln-j+1 . . . ln)$

Writeup Question 3.1:
We remove rare features because we cannot learn good weights for the rare words. Hence we need frequent words for good weights.
We remove rare words because the association between them and other words is dominated by noise and hence they do not serve the purpose of part of speech tagging.
They also lead to additional computations and increased complexity.

Writeup Question 3.2:
Sparse data structures like Compressed Sparse Row matrix allow us to store only non zero values assuming the rest of them are zeroes. Hence this saves a lot of memory and computing time.Dense matrices store every entry in the matrix. Sparse matrices only store the nonzero entries. Sparse matrices don’t have a lot of extra features, but they are useful as they can be compressed easily since most of the entries are 0.

Writeup Question 4.1:

```
['Apple', 'Inc.', 'is', 'an', 'American', 'multinational', 'technology', 'company', 'headquartered', 'in', 'Cupertino', ',', 'California', '.']
['NOUN', 'NOUN', 'VERB', 'DET', 'ADJ', 'ADJ', 'NOUN', 'NOUN', 'VERB', 'ADP', 'NOUN', '.', 'NOUN', '.']

['The', 'grand', 'jury', 'commented', 'on', 'a', 'number', 'of', 'other', 'topics', '.']
'DET', 'X', 'X', 'VERB', 'ADP', 'DET', 'NOUN', 'ADP', 'ADJ', 'NOUN', '.']

['Congress', 'is', 'in', 'a', 'standoff', 'with', 'the', 'Trump', 'administration', 'over', 'its', 'refusal', 'to', 'share', 'a', 'whistle-blower', 'complaint', 'with', 'lawmakers', '.']
['NOUN', 'VERB', 'ADP', 'DET', 'NOUN', 'ADP', 'DET', 'NOUN', 'NOUN', 'ADP', 'DET', 'ADJ', 'PRT', 'VERB', 'DET', 'NOUN', 'NOUN', 'ADP', 'NOUN', '.']

['I', 'want', 'to', 'go', 'to', 'a', 'restaurant', '.']
['PRON', 'VERB', 'PRT', 'VERB', 'ADP', 'DET', 'NOUN', '.']

['IBM', 'was', 'founded', 'in', '1924', '.']
['NOUN', 'VERB', 'VERB', 'ADP', 'NUM', '.']
```

Writeup Question 5.1:
5 days

Writeup Question 5.2:
NO

Writeup Question 6.1:
The predictions are far better than the previous predictions here as listed follows :

```
['Apple', 'Inc.', 'is', 'an', 'American', 'multinational', 'technology', 'company', 'headquartered', 'in', 'Cupertino', ',', 'California', '.']
['NOUN', 'NOUN', 'VERB', 'DET', 'ADJ', 'ADJ', 'NOUN', 'VERB', 'VERB', 'ADP', 'NOUN', '.', 'NOUN', '.']

['The', 'grand', 'jury', 'commented', 'on', 'a', 'number', 'of', 'other', 'topics', '.']
['DET', 'ADJ', 'NOUN', 'VERB', 'ADP', 'DET', 'NOUN', 'ADP', 'ADJ', 'NOUN', '.']

['Congress', 'is', 'in', 'a', 'standoff', 'with', 'the', 'Trump', 'administration', 'over', 'its', 'refusal', 'to', 'share', 'a', 'whistle-blower', 'complaint', 'with', 'lawmakers', '.']
['NOUN', 'VERB', 'ADP', 'DET', 'NOUN', 'ADP', 'DET', 'NOUN', 'NOUN', 'ADP', 'DET', 'ADJ', 'PRT', 'VERB', 'DET', 'ADJ', 'NOUN', 'ADP', 'NOUN', '.']
```

```
['I', 'want', 'to', 'go', 'to', 'a', 'restaurant', '.']
['PRON', 'VERB', 'PRT', 'VERB', 'ADP', 'DET', 'NOUN', '.']

['IBM', 'was', 'founded', 'in', '1924', '.']
['NOUN', 'VERB', 'VERB', 'ADP', 'NUM', '.']
```