Sravya Kurra
SXK180049

Q1.1)
Universal dependency is a smaller tag set compared to Penn Treebank tag set
Advantages:
Smaller tag sets give better accuracy than the larger one as we would have T^n possible sequence for larger ones
Disadvantages:
Universal dependency tag set, a smaller tag set, would result in loss of fine distinctions in parts of speech .i.e., categories and sub categories are missed, which can be seen in larger tag sets like Penn Treebank tag set

Q2.1) We need skip bigram features to increase the size of training data. The advantage of skip bigram over trigram is that we can reduce the effect noise like misspellings and effect of sparse data by skipping words.

Q2.2) We use features like capital, hyphen, prefix of word and suffix of word to predict the tag of unknown words that we can possibly encounter while testing our model

Q3.1)
Reasons to remove rare words:

1) Because they occur rare, it means that their association with the words surrounding them could be affected by noise
2) Rare words don’t play much significance in classification of word(tags). So, removing these can lead us closer to better accuracy

Q3.2)

A word has few features compared to overall features we have in feature dictionary. So, when a row is considered, we will have more 0’s than 1’s. So we use sparse matrix for X_train.
Advantages:
Using sparse matrix would save the memory and speed up the data processing.

Q4.2)
```
[['NOUN', 'NOUN', 'VERB', 'DET', 'ADJ', 'ADJ', 'NOUN', 'NOUN', 'VERB', 'ADP', 'NOUN', '.', 'NOUN', '.'], ['DET', 'X', 'X', 'X', 'X', 'X', 'NOUN', 'ADP', 'ADJ', 'NOUN', '.'], ['NOUN', 'VERB', 'ADP', 'DET', 'NOUN', 'ADP', 'DET', 'NOUN', 'NOUN', 'ADP', 'DET', 'NOUN', 'PRT', 'VERB', 'DET', 'NOUN', 'NOUN', 'ADP', 'NOUN', '.'], ['PRON', 'VERB', 'PRT', 'VERB', 'ADP', 'DET', 'NOUN', '.'], ['NOUN', 'VERB', 'VERB', 'ADP', 'NUM', '.']]
```

Q5)
5.1) 3 complete days because it’s hard to check if our output is correct

5.2) Aishwarya

Extra Credit:

Sravya Kurra
SXK180049

Q6.1)

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

Better. Tags are rightly predicted comparatively