HW 2 Answers

1. Selected threshold for the unknown words replacement is 1. The unknown words are the words whose occurrence at the document level is less than the threshold. This means that the training set has very few occurrences of these words to generate a tag of the word.
2. Total size of the vocabulary = 23183. The vocab.txt contains tab separated ‘word’ , ‘index’ , ‘count’. The < UNK > is placed first then the rest of the words are placed in descending order of the counts.
3. Total number of < unk > tokens = 20011. There are 20011 tokens/ words whose occurrence is less than threshold i.e 1.
4. Transition Parameters = 1416
5. Emission Parameters = 1043235
6. Greedy accuracy on dev data: 93.50297492562686. The greedy pos tagging is simply the greedy algorithm i.e for every word in the sentence, we consider the previous tag to be the one that has attained the maximum probability for the previous word. The greedy algorithm gives a local optimum
7. Viterbi Decoding accuracy on dev data: 94.76883613623946. The viterbi decoding is a dynamic programming technique for pos tagging. The algorithm considers all the possibilities of the previous tags to calculate the probability of a tag for the word. It then backtracks to select the sequence with the maximum probability. I.e It gives a global optimum.