NATIONAL INSTITUTE OF TECHNOLOGY ROURKELA

Mid Semester Examination, Spring 2022

Subject: Natural Language Processing **Subject Code: CS6314** **Full Marks: 30**

Answer all questions. Mere answer without justification will not fetch any mark.

1. Consider the following corpus of words with their frequency: 3 marks
low:5, lower:2, newest:6, widest:3
Find the vocabulary after 3 iterations of BPE (Byte Pair Encoding).

2. Consider two documents having the following content inside it: 3 marks
Doc A: the man went out for a morning walk.
Doc B: the children sat near the pond.
Find TF-IDF value for all words in document Doc A and Doc B. (Take base as 10 while calculating log values.)

3. Consider the ANN given in Figure 1 with weights and biases. 4 marks
Initial input i1=0.05 and i2=0.10 was given. After completion of a single forward pass the output of o1=0.751 and o2=0.773. But the desired output of o1 is 0.01 and o2 is 0.99. Activation function used is Sigmoid function $\sigma(x) = (1/((1 + e^{(} - x))))$. For calculating error, Squared Error function is used. Calculate the new weights of w5, w6, w1 and w2 after one back propagation. (use learning rate=0.5).

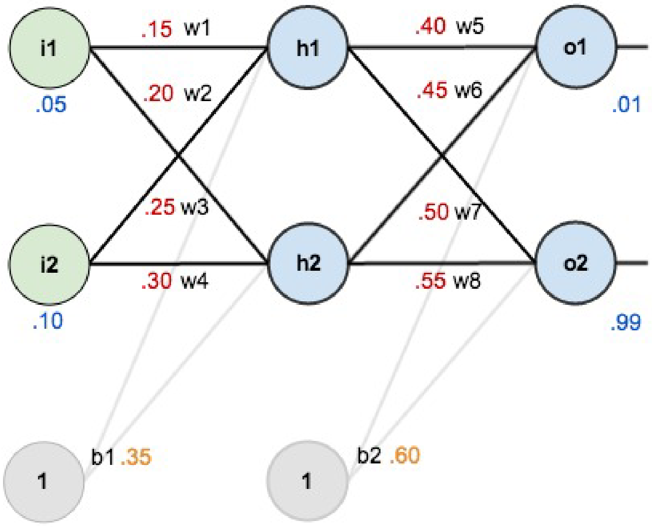

Figure 1: ANN of Question 3

4. Consider the corpus as given below. 3 marks

   Hey, I just met you, And this is crazy,
   But here’s my number, So call me, maybe?
   It’s hard to look right, At you ,
   But here’s my number, So call me, maybe?

   Using a bigram model and Laplace smoothing (add one smoothing), calculate P(my number call me).

5. Calculate minimum edit distance between the strings PLASMA” and ”ALTRUISM” using edit distance table. Show the backtrace. For insertion and deletion, use a penalty of 1 and for substitution use a penalty of 2. 3 marks

6. 1. Consider the corpus. 4 marks
I am going. We are leaving.

   Vocabulary=[I, am, going, We, are, leaving].
We are given the weights and contexts as follows. We are implementing word2vec with skipgram and negative sampling. Consider window size +/-1 and 1 negative sample for each context calculation. The

initial weights are as given, each being a 7 dimensional vector.

$$\begin{aligned} W_I &= [1111110] & C_I &= [1111101] \\ W_{am} &= [1111011] & C_{am} &= [1011111] \\ W_{going} &= [1101111] & C_{going} &= [0111111] \\ W_{We} &= [1111011] & C_{We} &= [1111101] \\ W_{are} &= [1011111] & C_{are} &= [1111011] \\ W_{leaving} &= [1111101] & C_{leaving} &= [1110111] \end{aligned}$$

With respect to the word I, assume ‘am’ is the positive word, ‘We’ is the negative word.

Calculate the Loss with respect to word I, and show one step updation of W_I , C_{am} , C_{We}. Assume that we use sigmoid function to model the probabilities as discussed in class.

7. Draw a single encoder of an Transformer as given in Vaswani. et. al. 5 marks
and explain each of the modules in the architecture. You need to specify what is the input to each module and what is the output of each module.

8. Write a function in C to implement an DFA(Deterministic Finite Automaton) 5 marks
recognizing the language 0^*1^*(any string starting with arbitrary number of zeros followed by arbitrary number of ones). The function should take a string as input, simulate the DFA on the string and print whether the string belongs to the language or not. Input always consists of 0’s and 1’s.

– – ★ – –