National University of Computer and Emerging Sciences, Lahore Campus

Course:	Natural Language Processing	Course Code:	CS 535
Program:	MS(Computer Science)	Semester:	Spring 2020
Duration:	3 hour 30 minutes + 30 minutes for uploading exam	Total Marks:	100
Paper Date:	6-June-20	Weight	50%
Section:	CS	Page(s):	6
Exam:	Online Final Exam		

Instructions: Handwritten solution in clear and eligible writing should be submitted.
The image of answers should be rotated at right angle and should be readable
Submit images as one combined PDF file. Name of PDF file should be your roll number and name. Write your roll number on each answer sheet.
Show complete working of each question.
Exam should be completed in 3 hour 30 minutes and you can take 30 more minutes for uploading exam. It should be submitted no later than 1:00 pm.

Questions	1	2	3-4	5	6	7	8-9	Total
Total Marks	21	6	18	6	14	15	20	100

Q1) (a) Suppose we are training a RNN language model for the sentence ”we are trying to make thinking machines”
One hot encoded vector of words is given as follows: [6 Marks]
we = x_1: [1 0 0 0 0 0 0]
are = x_2: [0 1 0 0 0 0 0]
trying = x_3: [0 0 1 0 0 0 0]
to = x_4 : [0 0 0 1 0 0 0]
make = x_5 : [0 0 0 0 1 0 0]
thinking = x_6 : [0 0 0 0 0 1 0]
machines = $x_{7:}$ [0 0 0 0 0 0 1]

Suppose the input at 6 different time stamps is as follws:
x_1 = we, x_2 = are, x_3 = trying, x_4 = to, x_5 = make, x_6 = thinking

The predicted output distribution of words at different time stamps is as follows:
y_1 = [0 0.3 0.1 0 0.4 0.2 0]
y_2 = [0.1 0.2 0.4 0.2 0 0 0.1]
y_3 = [0 0.1 0.1 0.2 0.4 0.2 0]
y_4 = [0 0.3 0.1 0 0.4 0.2 0]
y_5 = [0 0 0.1 0 0.4 0.3 0.2]
y_6 = [0 0 0.1 0 0 0.4 0.5]

Compute the cross entropy loss for this sentence.

Solution:

1/6 [(-log (0.3)) + (-log (0.4)) + (-log (0.2)) + (-log (0.4)) + (-log (0.3)) + (-log (0.5))]
=0.47

Q1) (b) Suppose we have following language model: [4+2 = 6 Marks]

• input sequence of length 7 (lets say 7 words).
• Hidden layer units are 4.
Embedding vector size = 5
• V = vocabulary = 10

i. Draw RNN architecture diagram with dimensions of all layers and weight matrices
ii. Write equations along with dimensions of all layers and weight matrices

Q1) (c) What is advantage of uing RNN for language modeling as compared to n gram based neural language model?

Give some example English sentence to motivate the advantage of RNN. The sentence should not be from lecture slides or text book and it should not match sentence of any other student in class (think abut the sentence yourself, do not google). [3 Marks]

Q1) (d) What is vanishing gradient problem in RNN?

Give some example English sentence to show the problem of vanishing gradient. The sentence should not be from lecture slides or text book and it should not match sentence of any other student in class (think abut the sentence yourself, do not google). [3 Marks]

Q1) (e) What is advantage of bi directional RNN over simple RNN. Motivate with some example of English sentence. [3 Marks]

Q2) Suppose you have made a simple spell checker based on dictionary words of English language (if a word is not present in dictionary then it is a spelling mistake). In following sentence the word ”there” is a spellig mistake. [2 + 4 = 6 Marks]

”They were playing foorball so there clothes are dirty”

Your program will not identify this spelling mistake as this word is present in dictionary. You have recently take the course of NLP.

a) Name some NLP technique that can be impleneted in your program so that this seplling mistake can be idetified and also corrected.
b) Briefly describe the technique and how it will identify the mistake.

Q3) (a) What are problems of greedy decoding and how beam search resolves these problems? Describe in context of neural machine translation. [4 Marks]

Q3) (b) What are some advantages of neural machine translation as compared to statistical machine translation. [3 Marks]

Q3) (c) What is effect of changing beam size k on neural text generation? [3 Marks]

Q4) (a) What is relation between word embeddings and neural language modeling? [4 Marks]

Q4) (b) Describe some smoothing techniques used in neural language modeling? [4 Marks]

Q5) Calculate the TFIDF for the terms listed below for documents 1 to 3. There are 10,000 documents in a collection. The number of times each of these terms occur in documents 1 to 3 as well as the number of documents in the collections are listed below. Use this information to fill in the TFIDF scores in the table below. [6 Marks]

Number of Documents Containing Terms:

_ Exam: 30
_ Fruit: 1000
_ Apple: 500

	Raw Term Counts		
	Doc 1	Doc 2	Doc 3
Exam	4	54	1
Fruit	6	5	40

apple	23	34	5

Fill in the table below and show all working.

	Tf.IDF for terms in Doc 3
exam	
Fruit	
apple	

Q6) (a) What are advantages of using dense word vectors like word2vec as compared to sparse word vectors? [4 Marks]

Q6) (b) Suppose we have multiple meanings of the word ”apple” in a corpus. At some places it is used as a fruit and in other palces it is used for company name. If we train wordToVec model on this ocrpus, will the different occurences of apple for different meanings will have different representations?

State YES / NO. Also give reason. [5 Marks]

Q6) (c) Word2Vec represents a family of embedding algorithms that are commonly used in a variety of contexts. Suppose in a recommender system for online shopping, we have information about co-purchase records for items $x_1, x_1, \ldots, x_n$ (for example, item x_i is commonly bought together with item x_j). Explain how you would use ideas similar to Word2Vec to recommend similar items to users who have shown interest in any one of the items. [5 Marks]

Q7) You are consulting for a healthcare company. They provide you with medical notes of the first encounter that each patient had with their doctor regarding a particular medical episode. There are a total of 10 million patients and cmedical notes. Figure 1 shows a sample medical note. At the time that each medical note was written, the underlying illnesses associated with the medical episode were unknown to the doctor. The company provides you with the true set of illnesses associated with each medical episode and asks you to build a model that can infer these underlying illnesses using only the current medical note and all previous medical notes belonging to the patient. The set of notes provided to you span 10 years; each patient therefore can have multiple notes (medical episodes) in that period. Each note can contain any number of tokens (see Figure 1). Some tokens (e.g. ”Meds”) occur more frequently than others in the collection of notes provided to you. You call your former teacher for advice. He tells you to first create a distributed representation of each patient note by combining the distributed representations of the words contained in the note.

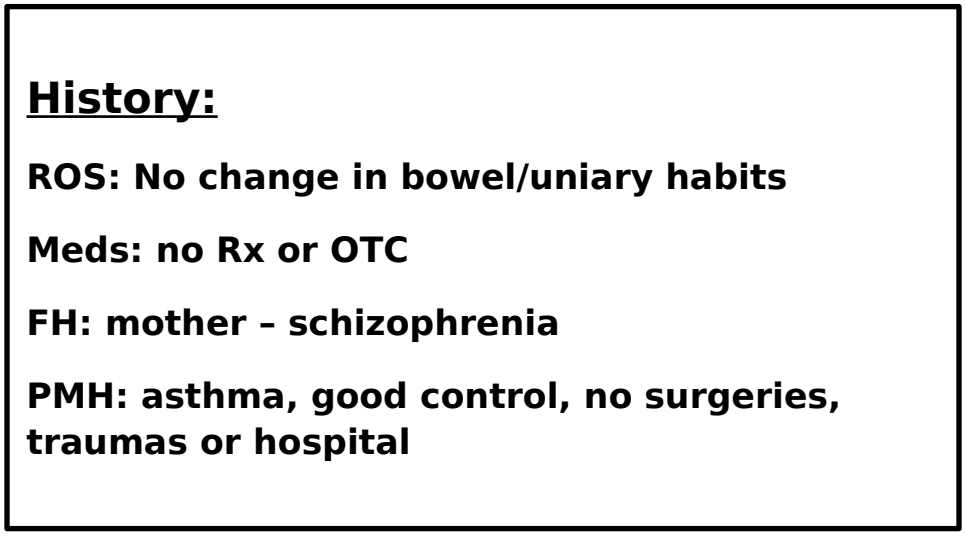

Figure 1: Sample medical note

Q 7 (a) Given the sample note provided in Figure 1, how would you map the various tokens into a distributed vector representation? [3 Marks]

Q 7 (b) How will you combine vector represntation of all words in a note for input to a neural network? [4 Marks]

Q 7 (c) You now have a distributed representation of each patient note (note-vector). You assume that a patient's past medical history is informative of their current illness. As such, you apply a recurrent neural network to predict the current illness based on the patient's current and previous note-vectors. Explain why a recurrent neural network would yield better results than a feed-forward network. In feed-forward network your input is the summation (or average) of past and current note-vectors? [4 Marks]

Q 7 (d) Your model achieves a precision score of 72% on positive cases (true positives) and a precision score of 68% on negative cases (true negatives). Confident with your initial results, you decide to make a more complex model. You implement a bidirectional deep recurrent neural network over the chronologically ordered patient note-vectors. Your new results are stellar. Your positive precision is 95% and your negative precision is 92%. You boast to your teacher that you have built an AI doctor. You coin the name Dr. AI for your model. Unfortunately, your teacher tells you that you have made a mistake. What is the mistake? [4 Marks]

Presision on positive cases = true positive / (true positive + false positive)
Presision on negative cases = true negative / (true negative + false negative)

Q8) Given the training data below, execute the following 2 steps:
Training Data:

- cat/NNS flying/VBG is/VBZ adventurous/JJ
- flying/JJ planes/NNS are/VBZ abundant/JJ
- I/PRP saw/VBZ Mary/NNP flying/VBG planes/NNS
- She/PRP planes/VBZ shelves/NNS

(a) Calculate the likelihood probabilities for each word given each POS [3 Marks]

(b) Calculate the most probable POS tage sequence for the string "flying planes". (Use bigram model for transition probabilities) [7 Marks] Show all calculations.

Q9) Given following PCFG, dry run CYK algorithm on string "x y x z ". Show all workings. [10 Marks]

S → XYZ 0.3
S → YZ 0.7
X → YX 0.4
X → x 0.6

Y → y 0.4
Y → z 0.6
Z → XY 0.5
Z → z 0.5