Name: _____ Reg #: _____ Section: _____

National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Natural Language Processing**	**Course Code:**	**CS 535**
Program:	**MS(Computer Science)**	**Semester:**	**Spring 2021**
Duration:	**3 hour**	**Total Marks:**	**70**
Paper Date:	**12-July-21**	**Weight**	**70%**
Section:	**CS**	**Page(s):**	**6**
Exam:	**Final Exam**		

Instruction/Notes: Attempt the examination on the question paper and write concise answers. You can use extra sheet for rough work. Do not attach extra sheets used for rough with the question paper. Don't fill the table titled Questions/Marks.

Questions	1	2-3	4-5	6-7	8-10	11	12/13	Total
Marks	/8	/8	/15	/16	/6	/9	/8	

Q1) Answer the following multiple choice questions. Suppose you have the following training data for Näıve Bayes. Encircle correct option. [8 Marks]

I liked the movie [LABEL=+]
I hated the movie because it was an action movie [LABEL=-]
Really cool movie [LABEL=+]

A) What is the unsmoothed maximum likelihood estimate of P(+)for this data?
i. 1/3 ii. 1/2 iii. 2/3 iv. 1

(B) What is the unsmoothed maximum likelihood estimate of P(movie|+)for this data?
i. 2/17 ii. 1/5 iii. 2/7 iv. 1/2

(C) Suppose we are given an unseen input sentence "the movie". What is the joint probability P(−,the movie)?
i. 2/300 ii. 4/98 iii. 1/12 iv. 1/3

(D) What prediction will the model make on sentence "the movie"?
i. Positive ii. Negative

Q2) Calculate the TFIDF for the terms listed below for documents 1 to 3. There are 1000 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 for Doc 3 in the table below. [5 Marks]

Number of Documents Containing Terms:
_ Exam: 30
_ Fruit: 10
_ Apple: 80

	Raw Term Counts		
	Doc 1	Doc 2	Doc 3
Exam	4	54	3
Fruit	7	5	30
apple	25	34	9

Fill in the table below and show all working.

	Tf.IDF for terms in Doc 3
exam	
Fruit	
apple	

Q3) You are an English teacher and you ask your class to write a play in the style of Shakespeare. You want to score their plays using a trigram language model you computed from a corpus of all Shakespeare plays but you find that the data is too sparse and most of your students' sentences receive a score of zero. How would you use a back-off model to alleviate this problem? [3 Marks]

Q4) Following table gives co-occurrence counts based on syntactic dependencies of words. Write down context vector of the word "duty" using PPMI (Positive Pointwise Mutual Information) of words. (You can assume following table contains all words that can appear as object of a given a word. E.g. total count of words that appear as object of "assert" is 10. Sum of row counts represent total count of the word in collection. E.g. duty appears 22 times in collection. Total words in collection = N = 100) **[5 Marks]**

$$\text{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

	Object of assert	Object of assign	Object of avoid	Object of become	Modified by collective	Modified by assumed
duty	3	4	5	3	5	2
responsibility	2	2	7	4	2	7
taxes	0	0	3	0	0	1
danger	0	0	6	0	1	0
control	5	0	0	1	0	0

Q5) You are given the following training corpus: [1 + 1 + 2 + 3 + 3 = 10 Marks]

<s> I like oranges </s>
<s> oranges like I </s>
<s> We like cherries </s>
<s> I do not like cherries and oranges </s>

a) Calculate the probability of following test sentence. Include </s> in your counts just like any other token. λ_1 = trigram weight, λ_2 = bigram weight, λ_3 = unigram weight, $\lambda_1 = 0.4$, $\lambda_2 = 0.3$, $\lambda_3 = 0.3$

<s> I like bikes </s>

i. Unigram Model

ii. Bigram Model

iii. Trigram Model

iv. Trigram language model with linear interpolation.

Q6) Suppose we are training a LSTM language model for the sentence ”computers are able to see, hear, and learn”
One hot encoded vector of words is given as follows: [5 Marks]
computers = x_1: [1 0 0 0 0 0 0 0]
are = x_2: [0 1 0 0 0 0 0 0]
able = x_3: [0 0 1 0 0 0 0 0]
to = x_4 : [0 0 0 1 0 0 0 0]
see = x_5 : [0 0 0 0 1 0 0 0]
hear = x_6 : [0 0 0 0 0 1 0 0]
and = x_7: [0 0 0 0 0 0 1 0]
learn = x_8: [0 0 0 0 0 0 0 1]

Suppose the input at 7 different time stamps is as follws:
x_1 = computers, x_2 = are, x_3 = able, x_4 = to, x_5 = see, x_6 = hear, x_7 = and,

The predicted output distribution of words at different time stamps is as follows:
y_1 = [0 0.2 0.1 0.1 0.4 0.2 0 0]
y_2 = [0.1 0.2 0.3 0.3 0 0 0.1 0]
y_3 = [0 0.1 0 0.3 0.4 0.2 0 0]
y_4 = [0 0.1 0.1 0 0.6 0.2 0]
y_5 = [0 0 0.1 0 0 0.4 0.3 0.2]
y_6 = [0 0 0.1 0 0 0 0.4 0.5]
y_7 = [0 0 0.1 0 0.1 0 0.5 0.3]

Compute the cross entropy loss for this sentence.

Name: _____ Reg #: _____ Section: _____

Q7) (a) Suppose we have following language model: [4+4 = 8 Marks]

• input sequence of length 5 (lets say 5 words).
• Hidden layer units are 4.
Embedding vector size = 6
• V = vocabulary = 8

i. Draw RNN architecture diagram with dimensions of all layers and weight matrices

ii. Give the update equations for a simple RNN unit in terms of x, y, and h (input x, output y, and recurrent state h). Assume it uses tanh non-linearity.

Q7) (b) What is the role of gates in LSTM? How are gates implemented? [3 Marks]

Q8) Which of the following statements is INCORRECT? [2 Marks]

A. Recurrent neural networks can handle a sequence of arbitrary length,while feedforward neural networks can not.
B. Training recurrent neural networks is hard because of vanishing and ex-ploding gradient problems.
C. Gradient clipping is an effective way of solving vanishing gradient prob-lem.
D. Gated recurrent units (GRUs) have fewer parameters than LSTMs.

Q9) What is the probable approach when dealing with "Exploding Gradient" problem in RNNs? [2 Marks]
A) Use modified architectures like LSTM and GRUs
B) Gradient clipping
C) Dropout
D) None of these

Q10) If calculation of reset gate in GRU unit is close to 0, which of the following would occur? [2 Marks]
A) Previous hidden state would be ignored
B) Previous hidden state would be not be ignored

Q11) (a) What are problems of greedy decoding and how beam search resolves these problems? Describe in context of neural machine translation. [3 Marks]

Q11) (b) What are some advantages of neural machine translation as compared to statistical machine translation. [3 Marks]

Q11) (c) What is effect of changing beam size k on neural text generation? [3 Marks]

Q12 is only for MS students

Q12) (a) Describe some smoothing techniques used in neural language modeling? [4 Marks]

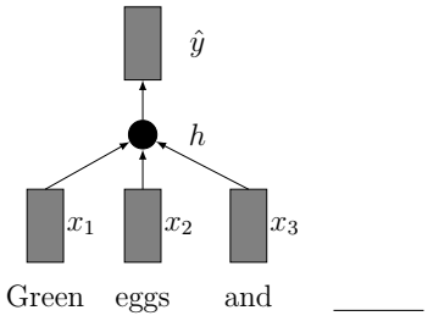
Q12) (b) What are advantages of using dense word vectors like word2vec as compared to sparse word vectors? [4 Marks]

Name: _____ Reg #: _____ Section: _____

Q13 is only for PhD students

Q13) (a) If we chose to update our word vectors when training the LSTM model on sentiment classification data, how would these word vectors differ from ones not updated during training? Explain with an example. Assume that the word vectors of the LSTM model were initialized using word2vec. [4 Marks]

Q13) (b) A feedforward neural network language model (NNLM) can be used as another architecture for training word vectors. This model tries to predict a word given the N words that precede it. To do so, we concatenate the word vectors of N previous words and send them through a single hidden layer of size H with a tanh nonlinearity and use a softmax layer to make a prediction of the current word. The size of the vocabulary is V. The model is trained using a cross entropy loss for the current word. Let the word vectors of the N previous words be x1,x2,...,xN, each a column vector of dimension D, and let y be the one-hot vector for the current word. [4 Marks]



$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}$$

$$\mathbf{h} = \tanh(W\mathbf{x} + \mathbf{b})$$

$$\hat{\mathbf{y}} = \text{softmax}(U\mathbf{h} + \mathbf{d})$$

$$J = \text{CE}(\mathbf{y}, \hat{\mathbf{y}})$$

$$CE = -\sum_i y_i \log(\hat{y}_i).$$

State two important differences between NNLM the WordToVec language model we learned in class. Explain how each might affect the word vectors learned.

Name: _____ Reg #: _____ Section: _____

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.

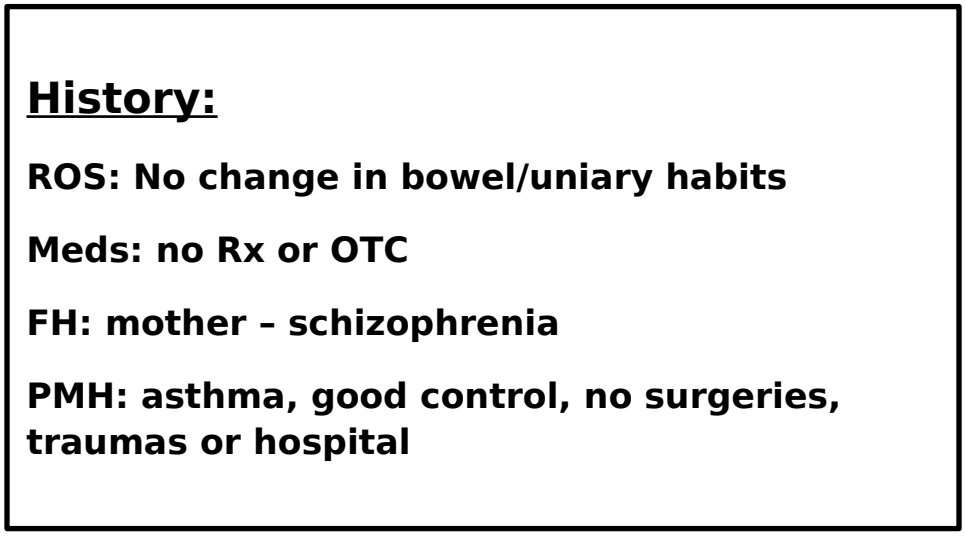


Figure 1: Sample medical note

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

Name: _____ Reg #: _____ Section: _____

National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Natural Language Processing**	**Course Code:**	**CS 535**
Program:	**BS(Computer Science)**	**Semester:**	**Spring 2018**
Duration:	**180 Minutes**	**Total Marks:**	**41**
Paper Date:	**23-May-18**	**Weight**	**50%**
Section:	**ALL**	**Page(s):**	**8**
Exam:	**Final**		

Instruction/Notes: Attempt the examination on the question paper and write concise answers. You can use extra sheet for rough work. Do not attach extra sheets used for rough with the question paper. Don't fill the table titled Questions/Marks.

Question	1-4	5-7	8-10	11-14	Total
Marks	/ 9	/ 10	/ 12	/10	/ 41

41

Q1) You are given the following corpus: **[2 + 2 = 4 Marks]**

<s> She likes green apples </s>
<s> Ali likes green apples </s>
<s> green apples are good for health </s>
<s> I like red apples </s>

a) Calculate the probability of following test sentence using trigram language model with linear interpolation. Include <s> and </s> in your counts just like any other token. λ_1 = trigram weight, λ_2 = bigram weight, λ_3 = unigram weight, $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$

<s> He likes green apples </s>

b) Calculate the probability of P(green | likes) using Kneser Ney smoothing from the corpus given above. d = discounting factor = 0.5

Name: _____ Reg #: _____ Section: _____

Q2) Suppose a language model assigns the following conditional n-gram probabilities to a 3-word test set: 1/8, 1/2, 1/6. What is the perplexity? **[2 Marks]**

Q3) $P_{continuation}(w)$ for a word is defined as follows: **[2 Marks]**

$$P_{CONTINUATION}(w) = \frac{\left|\{w_{i-1} : c(w_{i-1}, w) > 0\}\right|}{\sum_{w'} \left|\{w'_{i-1} : c(w'_{i-1}, w') > 0\}\right|}$$

a) Consider the following incomplete sentence:

"How much wood would a woodchuck chuck would if woodchuck could would chuck"
What is $|\{w_{i-1} : C(w_{i-1}\ w_i)>0\}|$ for w_i="woodchuck"?

i. 0 ii. 1 iii. 2 iv. 3

b) Which word is more likely to complete the sentence (follow the last "chuck") based on $P_{continuation}$?

i. How ii. wood iii. would iv. chuck

Q4) Which of the following word pairs, A/B, has A as a hypernym of B? [1 **Mark]**

i. Washington/The United States
ii. vehicle/car
iii. Java/programming language
iv. wheel/car
v. None of the above

Q5) Consider a trigram HMM tagger with: **[3 Marks]**

_ The set K of possible tags equal to {D, N, V}
_ The set V of possible words equal to {the, dog, barks}
_ The following parameters:

q(D|*, *) = 1	q(N|D, V) = 0.3	e(dog|N) = 0.4
q(N|*, D) = 0.5	q(STOP|N, V) = 0.6	e(barks|N) = 0.6
q(V|*, D) = 0.5	q(STOP|V, N) = 0.4	e(dog|V) = 0.1
q(V|D, N) = 0.7	e(the|D) = 1	e(barks|V) = 0.9

with all other parameter values equal to 0. Write down the set of all pairs of sequences $x_1 \ldots x_{n+1}$, $y_1 \ldots y_{n+1}$ such that the following properties hold:
_ $p(x_1 \ldots x_{n+1}, y_1 \ldots y_{n+1}) > 0$

_ $x_i \in V$ for all $i \in 1 \ldots n$

_ $y_i \in K$ for all $i \in 1 \ldots n$, and y_{n+1} = STOP

Q6) Show how following lexicalized grammar rule parameter is decomposed into 2 parameters for learning probabilities from training data. Also show how to use smoothed estimation for the decomposed parameters. . **[3 Marks]**

$q(S(saw) \rightarrow_2 NP(man)\ VP(saw))$

Q7) Write down at least two different parse trees (with different probabilities) for following sentence and PCFG. **[4 Marks]**

"The boy saw the dog in the park with the telescope"

S → NP VP 0.8
S → NP VP PP 0.2
NP → DET N 0.5
NP → NP PP 0.5
VP → V NP 1.0

PP → P NP 1.0
N → dog 0.25
N → boy 0.25
N → park 0.25
N → telescope 0.25

V → saw 1.0
P → with 0.5
P → in 0.5
DET → the 1.0

Q8) In the following gloss of different word seneses of the wrods ”bank” and ”coast” are given. Compute similarity between the words ”bank” and ”coast” uing Lesk algorithm. **[4 Marks]**

Bank$_1$: sloping land (especially the slope beside a body of water)

Bank$_2$: a financial institution that accepts deposits and channels the money into lending activities

Bank$_3$: a long ridge or pile

Bank$_4$: an arrangement of similar objects in a row or in tiers

Bank$_5$: a supply or stock held in reserve for future use (especially in emergencies)

Coast$_1$: the shore of a sea or ocean

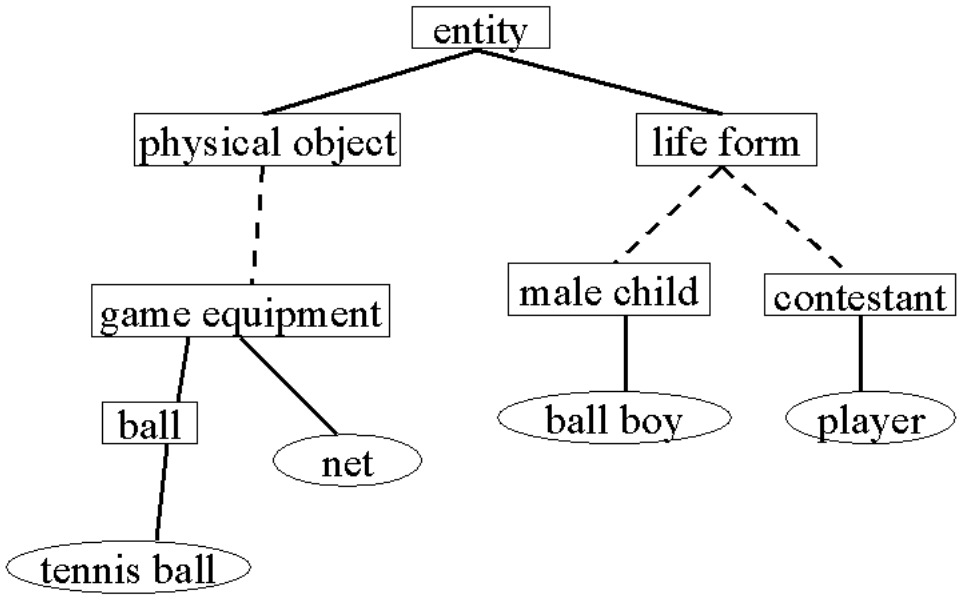
Coast$_2$: a slope down which sleds may coast

Coast$_3$: the area within view

Coast$_4$: the act of moving smoothly along a surface while remaining in contact with it

Name: _____ Reg #: _____ Section: _____

Q9) Following is a WordNet hierarchy. The probabilities of words are given in table below: **[4 Marks]**



Word	Probability
entity	0.395
Physical object	0.167
Life form	0.0231
Game equipment	0.00453
Male child	0.00153
contestant	0.00743
Ball	0.000343
Net	0.00054
Ball boy	0.000113
Player	0.000445
Tennise ball	0.000189

a) Compute path based similarity between “tennis ball” and “net”

b) Compute information content based similarity proposed by Lin (Lin Similarity function) between "ball" and "player"

Q10) a) Write down context vectors of words mango and apple using PPMI (Positive Pointwise Mutual Information) of words. **[2 Marks]**

Counts(w, context)				
	information	**data**	**sweet**	**Fitness**
Banana	**0**	**0**	**5**	**3**
Apple	**3**	**2**	**4**	**6**
Mechanical	**5**	**4**	**1**	**2**
computer	**7**	**6**	**0**	**1**

b) Following table gives co-occurrence counts based on syntactic dependencies of words. Write down context vectors of words duty and responsibility using PPMI (Positive Pointwise Mutual Information) of words. (You can assume following table contains all words that can appear as object of a given a word. E.g. total count of words that appear as object of "assert" is 10. Sum of row counts represent total count of the word in collection. E.g. duty appears 22 times in collection. Total words in collection = N = 100) **[2 Marks]**

	Object of assert	**Object of assign**	**Object of avoid**	**Object of become**	**Modified by collective**	**Modified by assumed**
duty	3	4	5	3	5	2
responsibility	2	2	7	4	2	7
taxes	0	0	3	0	0	1
danger	0	0	6	0	1	0
control	5	0	0	1	0	0

Q11) Compute value of ROUGE-2 score for following summary. **[2 Marks]**

System Generated Summary: The quake had a preliminary magnitude of 6.9. in an area so isolated there are no roads connecting it to the outside world.

Reference Summary (Human Generated Summary): The quake had a preliminary magnitude of 6.9. An earthquake in the same region in February killed 2300 people and left thousands homeless.

Q12) The first step in query focused multi document summarization is to simplify the sentences. Simplify following sentences using simple rules discussed in class. **[4 Marks]**

a) Genette's bedroom desk, the biggest disaster area in the house, is a collection of overdue library books, dirty plates, computer components, old mail, cat hair, and empty potato chip bags.
b) Robbie, a hot-tempered tennis player, charged the umpire and tried to crack the poor man's skull with a racket.
c) The car began sliding sideways, and then it hit the tree," she said
d) He died in France, as a matter of fact, and wated to be buried there.

Q13) Word occurrence in sentiment analysis matters more than word frequency. Briefly describe difference between multinomial Naïve Bayes and Boolean Multinomial Naïve Bayes for sentiment analysis. **[2 Marks]**

Q14) Give at least 5 features that can be used to resolve ambiguity in name entity recognition. **[2 Marks]**

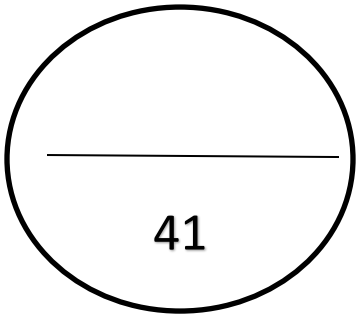
Name: _____ Reg #: _____ Section: _____

National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Natural Language Processing**	**Course Code:**	**CS 535**
Program:	**BS(Computer Science)**	**Semester:**	**Spring 2018**
Duration:	**180 Minutes**	**Total Marks:**	**41**
Paper Date:	**23-May-18**	**Weight**	**50%**
Section:	**ALL**	**Page(s):**	**8**
Exam:	**Final Solution**		

Instruction/Notes: Attempt the examination on the question paper and write concise answers. You can use extra sheet for rough work. Do not attach extra sheets used for rough with the question paper. Don't fill the table titled Questions/Marks.

Question	1-4	5-7	8-10	11-14	Total
Marks	/ 9	/ 10	/ 12	/10	/ 41



Q1) You are given the following corpus: **[2 + 2 = 4 Marks]**

<s> She likes green apples </s>
<s> Ali likes green apples </s>
<s> green apples are good for health </s>
<s> I like red apples </s>

a) Calculate the probability of following test sentence using trigram language model with linear interpolation. Include <s> and </s> in your counts just like any other token. λ_1 = trigram weight, λ_2 = bigram weight, λ_3 = unigram weight, $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$

<s> He likes green apples </s>

Solution:

P (<s> He likes green apples </s>) = P1 * P2 * P3 * P4

$= (7.6 * 10^{-3}) * (1.53*10^{-3})*(2.15*10^{-3}) *(2.15* 10^{-3})$

$= 5.37 * 10^{-11}$

P1 = $\lambda_1 *$ Count (<s> He likes) / Count(<s> He) + $\lambda_2 *$ Count (He likes) / Count(He) + $\lambda_3 *$ Count (likes) / N

b) Calculate the probability of P(green | likes) using Kneser Ney smoothing from the corpus given above. d = discounting factor = 0.5

Solution:

$$P_{KN}(w_i \mid w_{i-1}) = \frac{\max(c(w_{i-1}, w_i) - d, 0)}{c(w_{i-1})} + \lambda(w_{i-1}) P_{CONTINUATION}(w_i)$$

= ((2 – 0.5) /2) + (0.25) (0.17)

Q2) Suppose a language model assigns the following conditional n-gram probabilities to a 3-word test set: 1/8, 1/2, 1/6. What is the perplexity? **[2 Marks]**

Solution:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1 \ldots w_{i-1})}}$$

Q3) $P_{continuation}(w)$ for a word is defined as follows: **[2 Marks]**

$$P_{CONTINUATION}(w) = \frac{\left|\{w_{i-1} : c(w_{i-1}, w) > 0\}\right|}{\sum_{w'} \left|\{w'_{i-1} : c(w'_{i-1}, w') > 0\}\right|}$$

a) Consider the following incomplete sentence:

"How much wood would a woodchuck chuck would if woodchuck could would chuck"
What is $|\{w_{i-1} : C(w_{i-1}\ w_i) > 0\}|$ for w_i="woodchuck"?

i. 0 ii. 1 iii. 2 iv. 3

b) Which word is more likely to complete the sentence (follow the last "chuck") based on $P_{continuation}$?

i. How ii. wood iii. would iv. chuck

Q4) Which of the following word pairs, A/B, has A as a hypernym of B? [1 **Mark]**

i. Washington/The United States
ii. vehicle/car
iii. Java/programming language
iv. wheel/car
v. None of the above

Name: _____ Reg #: _____ Section: _____

Q5) Consider a trigram HMM tagger with: **[3 Marks]**

_ The set K of possible tags equal to {D, N, V}

_ The set V of possible words equal to {the, dog, barks}

_ The following parameters:

q(D|*, *) = 1	q(N|D, V) = 0.3	e(dog|N) = 0.4
q(N|*, D) = 0.5	q(STOP|N, V) = 0.6	e(barks|N) = 0.6
q(V|*, D) = 0.5	q(STOP|V, N) = 0.4	e(dog|V) = 0.1
q(V|D, N) = 0.7	e(the|D) = 1	e(barks|V) = 0.9

with all other parameter values equal to 0. Write down the set of all pairs of sequences $x_1 \ldots x_{n+1}$, $y_1 \ldots y_{n+1}$ such that the following properties hold:

_ $p(x_1 \ldots x_{n+1}, y_1 \ldots y_{n+1}) > 0$

_ $x_i \in V$ for all $i \in 1 \ldots n$

_ $y_i \in K$ for all $i \in 1 \ldots n$, and $y_{n+1} = STOP$

Solution:

1. * * The dog barks STOP (D N V)
2. * * The dog barks STOP (D V N)
3. * * The barks dog STOP (D N V)
4. * * The barks dog STOP (D V N)
5. * * The dog dog STOP (D N V)
6. * * The dog dog STOP (D V N)
7. * * The barks barks STOP (D N V)
8. * * The barks barks STOP (D V N)

Q6) Show how following lexicalized grammar rule parameter is decomposed into 2 parameters for learning probabilities from training data. Also show how to use smoothed estimation for the decomposed parameters. **. [3 Marks]**

$q(S(saw) \rightarrow_2 NP(man)\ VP(saw))$

Solution:

$$
\begin{aligned}
& q(\mathsf{S} \rightarrow_2 \mathsf{NP\ VP} | \mathsf{S}, \mathsf{saw}) \\
= \ & \lambda_1 \times q_{ML}(\mathsf{S} \rightarrow_2 \mathsf{NP\ VP} | \mathsf{S}, \mathsf{saw}) + \lambda_2 \times q_{ML}(\mathsf{S} \rightarrow_2 \mathsf{NP\ VP} | \mathsf{S}) \\
& q(\mathsf{man} | \mathsf{S} \rightarrow_2 \mathsf{NP\ VP}, \mathsf{saw}) \\
= \ & \lambda_3 \times q_{ML}(\mathsf{man} | \mathsf{S} \rightarrow_2 \mathsf{NP\ VP}, \mathsf{saw}) + \lambda_4 \times q_{ML}(\mathsf{man} | \mathsf{S} \rightarrow_2 \mathsf{NP\ VP}) \\
& + \lambda_5 \times q_{ML}(\mathsf{man} | \mathsf{NP})
\end{aligned}
$$

Name: _____ Reg #: _____ Section: _____

Q7) Write down at least two different parse trees (with different probabilities) for following sentence and PCFG. **[4 Marks]**

"The boy saw the dog in the park with the telescope"

S → NP VP 0.8
S → NP VP PP 0.2
NP → DET N 0.5
NP → NP PP 0.5
VP → V NP 1.0
PP → P NP 1.0
N → dog 0.25
N → boy 0.25
N → park 0.25
N → telescope 0.25
V → saw 1.0
P → with 0.5
P → in 0.5
DET → the 1.0

Q8) In the following gloss of different word seneses of the wrods "bank" and "coast" are given. Compute similarity between the words "bank" and "coast" uing Lesk algorithm. **[4 Marks]**

Bank$_1$: sloping land (especially the slope beside a body of water)

Bank$_2$: a financial institution that accepts deposits and channels the money into lending activities

Bank$_3$: a long ridge or pile

Bank$_4$: an arrangement of similar objects in a row or in tiers

Bank$_5$: a supply or stock held in reserve for future use (especially in emergencies)

Coast$_1$: the shore of a sea or ocean

Coast$_2$: a slope down which sleds may coast

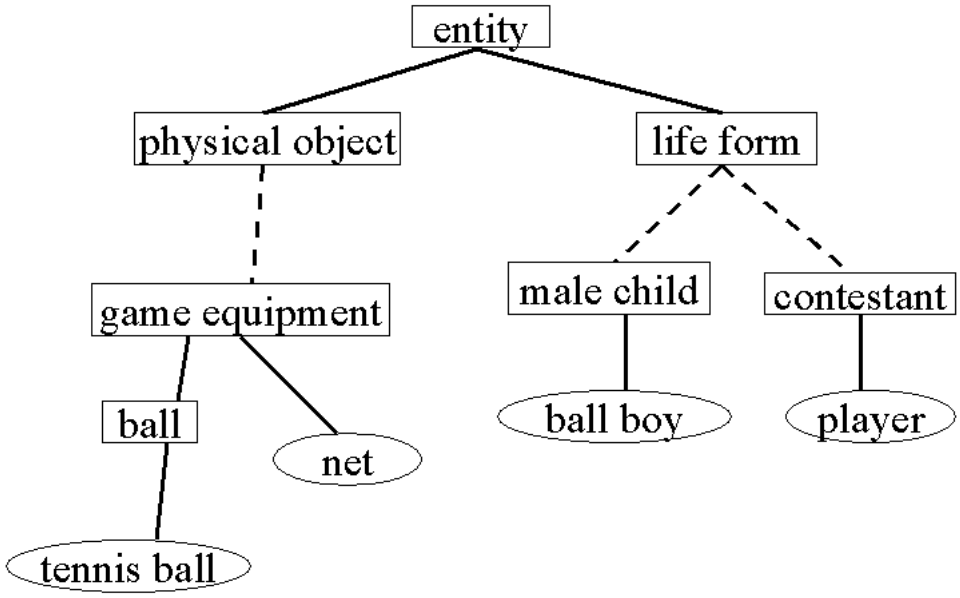
Coast$_3$: the area within view

Coast$_4$: the act of moving smoothly along a surface while remaining in contact with it

Solution:

Bank$_1$ and Coast$_2$ **= 1**

Q9) Following is a WordNet hierarchy. The probabilities of words are given in table below: **[4 Marks]**



Word	Probability
entity	0.395
Physical object	0.167
Life form	0.0231
Game equipment	0.00453
Male child	0.00153
contestant	0.00743
Ball	0.000343
Net	0.00054
Ball boy	0.000113
Player	0.000445
Tennise ball	0.000189

a) Compute path based similarity between "tennis ball" and "net"

Solution:

1/4

b) Compute information content based similarity proposed by Lin (Lin Similarity function) between "ball" and "player"

Solution:

log (0.395) / (log (0.0003)* log(0.0004))

Name: _____ Reg #: _____ Section: _____

Q10) a) Write down context vectors of words mango and apple using PPMI (Positive Pointwise Mutual Information) of words. **[2 Marks]**

Counts(w, context)				
	information	**data**	**sweet**	**Fitness**
Banana	**0**	**0**	**5**	**3**
Apple	**3**	**2**	**4**	**6**
Mechanical	**5**	**4**	**1**	**2**
computer	**7**	**6**	**0**	**1**

Solution:

Probabilities

Apple : (3/49)=0.06 0.04 0.08 0.12

PPMI : 0 0 0.42 0.73

b) Following table gives co-occurrence counts based on syntactic dependencies of words. Write down context vectors of words duty and responsibility using PPMI (Positive Pointwise Mutual Information) of words. (You can assume following table contains all words that can appear as object of a given a word. E.g. total count of words that appear as object of "assert" is 10. Sum of row counts represent total count of the word in collection. E.g. duty appears 22 times in collection. Total words in collection = N = 100) **[2 Marks]**

	Object of assert	**Object of assign**	**Object of avoid**	**Object of become**	**Modified by collective**	**Modified by assumed**
duty	3	4	5	3	5	2
responsibility	2	2	7	4	2	7
taxes	0	0	3	0	0	1
danger	0	0	6	0	1	0
control	5	0	0	1	0	0

Name: _____ Reg #: _____ Section: _____

Q11) Compute value of ROUGE-2 score for following summary. **[2 Marks]**

System Generated Summary: The quake had a preliminary magnitude of 6.9. in an area so isolated there are no roads connecting it to the outside world.

Reference Summary (Human Generated Summary): The quake had a preliminary magnitude of 6.9. An earthquake in the same region in February killed 2300 people and left thousands homeless.

Solution:

7/21

Q12) The first step in query focused multi document summarization is to simplify the sentences. Simplify following sentences using simple rules discussed in class. **[4 Marks]**

a) Genette's bedroom desk, the biggest disaster area in the house, is a collection of overdue library books, dirty plates, computer components, old mail, cat hair, and empty potato chip bags.
b) Robbie, a hot-tempered tennis player, charged the umpire and tried to crack the poor man's skull with a racket.
c) The car began sliding sideways, and then it hit the tree," she said
d) He died in France, as a matter of fact, and wated to be buried there.

Solution:

e) Genette's bedroom desk, is a collection of overdue library books, dirty plates, computer components, old mail, cat hair, and empty potato chip bags.
f) Robbie, charged the umpire and tried to crack the poor man's skull with a racket.
g) The car began sliding sideways, and then it hit the tree,"
h) He died in France, and wated to be buried there.

Q13) Word occurrence in sentiment analysis matters more than word frequency. Briefly describe difference between multinomial Naïve Bayes and Boolean Multinomial Naïve Bayes for sentiment analysis. **[2 Marks]**

Solution:

Boolean Multinomial Naïve Bayes clips word counts of all words in all documents at 1.

Q14) Give at least 5 features that can be used to resolve ambiguity in name entity recognition. **[2 Marks]**

Solution:

Identity of word

Neighboring words

Part of speech of word

Part of speech of neighboring words

Uppercase

Shape of word

Presence of hyphen

Name: _____ Reg #: _____ Section: _____

✓National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Natural Language Processing**	**Course Code:**	**CS 535**
Program:	**MS(Computer Science)**	**Semester:**	**Spring 2019**
Duration:	**180 Minutes**	**Total Marks:**	**48**
Paper Date:	**22-May-19**	**Weight**	**45%**
Section:	**CS**	**Page(s):**	**8**
Exam:	**Final**		

Instruction/Notes: Attempt the examination on the question paper and write concise answers. You can use extra sheet for rough work. Do not attach extra sheets used for rough with the question paper. Don't fill the table titled Questions/Marks.

Question	1-5	6-10	11-13	Total
Marks	/ 16	/ 20	/12	/ 48

Q1) a) Which of the following matches regexp /a(ab)*a/ **[1 Mark]**

1) abababa
✓2) aaba
3) aabbaa
4) aba
✓5) aabababa

b) Which of the following matches regexp /ab+c?/ **[1 Mark]**

✓1) abc
2) ac
3) abbb
4) bbc

c) Which of the following word pairs, A/B, has A as a hypernym of B? **[1 Mark]**

i. Washington/The United States
ii. ✓vehicle/car
iii. Java/programming language
iv. wheel/car
v. None of the above

Q2) Suppose a language model assigns the following conditional n-gram probabilities to a 3-word test set: 1/8, 1/2, 1/6. What is the perplexity? **[3 Marks]**

Solution:

$$((1/8)* (1/2)*(1/6))^{-1/3} = 4.58$$

Name: _____ Reg #: _____ Section: _____

Q3) You are given the following corpus: **[4 Marks]**

<s> She likes green apples </s>
<s> Ali likes green apples </s>
<s> green apples are good for health </s>
<s> I like red apples </s>

Calculate the probability of following test sentence using **bigram language model with Laplace** smoothing.

<s> He likes green apples for good health </s>

Solution:

P(He | <s>) = (0 + 1) / (4 + 12) = 0.0625
P(likes | He) = 0.0833
P(green | likes) = 0.214
P(apples | green) = 0.266
P(for | apples) = 0.062
P(good | for) = 0.076
P(health | good) = 0.076
P(</s> | health) = 0.15

P (<s> He likes green apples for good health </s>) = $1.68 * 10^{-8}$

Name: _____ Reg #: _____ Section: _____

Q4) $P_{continuation}(w)$ in Kneser Ney smoothing for a word is defined as follows: **[4 Marks]**

$$P_{CONTINUATION}(w) = \frac{\left|\{w_{i-1} : c(w_{i-1}, w) > 0\}\right|}{\sum_{w'} \left|\{w'_{i-1} : c(w'_{i-1}, w') > 0\}\right|}$$

a) Consider the following incomplete sentence:

"How much wood would a woodchuck chuck would if woodchuck could would chuck"

What is $|\{w_{i-1} : C(w_{i-1}\ w_i) > 0\}|$ for w_i="woodchuck"?

i. 0 ii. 1 iii. ✓2 iv. 3

b) Which word is more likely to complete the sentence (follow the last "chuck") based on $P_{continuation}$?

i. How ii. wood iii. would iv. chuck

Q5) Assume the following WordNet senses with their definitions **[2 Marks]**

cat[1]: any of several large cats typically able to roar and living in the wild

cat[2]: feline mammal usually having thick soft fur and being unable to roar

cat[3]: an informal term for a youth or man

paw: a clawed foot of an animal, especially a quadruped

mammal: any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive and nourished with milk

tiger: large feline of forests in most of Asia having a tawny coat with black stripes

man: an adult male person (as opposed to a woman)

carnivore: terrestrial or aquatic flesh-eating mammal

How is **cat[1]** related to each of the other senses – is it a homonym, a synonym, an antonym, a hyponym, a hypernym, or none of them? Note that there can be more than one relation that match.

cat[1] is a …homonym…….. of **cat[3]**

cat[1] is a …hyponym…….. of **mammal**

cat[1] is a ……hypernym…….. of **tiger**

cat[1] is a ……none…….. of **man**

cat[1] is a ……hyponym…….. of **carnivore**

Name: _____ Reg #: _____ Section: _____

Q6) Show how following lexicalized grammar rule parameter is decomposed into 2 parameters for learning probabilities from training data. Also show how to use smoothed estimation for the decomposed parameters. . **[4 Marks]**

$q(S(read) \rightarrow_2 NP(boy)\ VP(read))$

Solution:

$q(S \rightarrow NP\ VP \mid S, read) * q(boy \mid S(read) \rightarrow NP\ VP(read))$

$q(S \rightarrow NP\ VP \mid S, read) = \lambda_1 * q(S \rightarrow NP\ VP \mid S, read) + \lambda_2 * q(S \rightarrow NP\ VP)$

$q(boy \mid S(read) \rightarrow NP\ VP(read)) = \lambda_3 * q(boy \mid S(read) \rightarrow NP\ VP(read)) + \lambda_4 * q(boy \mid S \rightarrow NP\ VP) + \lambda_5 * q(boy \mid NP)$

Q7) a) Draw all possible parse tree for the sentence "Ask the grandma with scissors" by applying given PCFG. [2 Marks]

S → VP	1.0
VP → Verb NP	0.7
VP → Verb NP PP	0.3
NP → NP PP	0.3
NP → Det Noun	0.7
PP → Prep Noun	1.0

Det → the 0.1
Verb → Cut | Ask | Find …… 0.1
Prep → with | in …… 0.1
Noun → envelop | grandma | scissors | men | suits | summer | …… 0.1

(b) The rules shown above make up and example of a probabilistic grammar. What advantage such grammars have over conventional phrase structure grammars? [1 Mark]

Solution:

Ambiguity is resolved by selecting the most probable parse tree

c) Calculate probability of each parse tree. [1 Mark]

Q8) (a) Describe why production rule with zero probability are problemetic. [1 Mark]

Solution:
Such rule will make probability of entire parse tree zero.

(b) Describe one mehthod to avoid zero probabilities for lexicalized PCFGs [1 Mark]

Solution:

Smoothing

(c) 4-grams are better than trigrams for part-of-speech tagging. True or False. Justify your answer. [2 Marks]

Solution:

4-gram model will result in more zero proability issues and computational complexity will be higher. On the other hand the results will be more accurate using 4 gram model.

Q9) Suppose a corpus contains 400,000 word-tokens, and 80,000 of these are tagged as N (commn noun). The word-form cook occurs 1,000 times in the corpus, tagged either as N or V. Analysis shows that cook accounts for 0.4% of all common noun tokens in the corpus. Use Bayes forumla to calculate the probability that a given occurence of cook is tagged as N. Show your working. [2 Marks]

Solution:

P (N | cook) = P (cook | N) * P (N) / P(cook)

= (320 /80,000 * 80,000/400,000) / 1000/400,000

Q10) Given following PCFG, dry run CYK algorithm on string ”b a b ”. Show all workings. [6 Marks]

S → AB 0.3
S → BC 0.7
A → BA 0.4
A → a 0.6
B → CC 0.4
B → b 0.6
C → AB 0.5
C → a 0.5

Name: _____ Reg #: _____ Section: _____

Q11) Assume the following sentence L, in which the word **line** is in focus:

L = About three years ago, he nearly gave up because he had nothing to sell;
now his shelves are full, and towels and clothes hang from a **line** overhead.

Name: _____ Reg #: _____ Section: _____

a) Give a collocational feature vector for the word line in L, given a window size of 3 words to the left and 3 words to the right. [2 Marks]

b) Give a bag-of-words feature vector for the word line in L, given the following word feature list: [written, school, speech, row, major, hang, sell, nothing, rope, words]. [2 Marks]

Solution:
[0, 0 ,0 , 0, 0, 1, 0 , 0 , 0, 0]

Q 12) 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. [4 Marks]

Number of Documents Containing Terms:
_ reverse: 3
_ shower: 50
_ multiplex: 3

	Term Frequencies		
	Doc 1	Doc 2	Doc 3
reverse	8	10	0
shower	3	1	2
multiplex	0	8	7

Fill in the table below

	Tf.IDF for terms in documents		
	Doc 1	Doc 2	Doc 3
reverse	6.68	7	0
shower	3.4	2.3	2.9
multiplex	0	6.6	6.4

Name: _____ Reg #: _____ Section: _____

Q13) Following table gives co-occurrence counts based on syntactic dependencies of words. Write down context vector of the word duty using PPMI (Positive Pointwise Mutual Information) of words. (You can assume following table contains all words that can appear as object of a given a word. E.g. total count of words that appear as object of "assert" is 10. Sum of row counts represent total count of the word in collection. E.g. duty appears 22 times in collection. Total words in collection = N = 100) **[4 Marks]**

$$\text{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

	Object of assert	**Object of assign**	**Object of avoid**	**Object of become**	**Modified by collective**	**Modified by assumed**
duty	3	4	5	3	5	2
responsibility	2	2	7	4	2	7
taxes	0	0	3	0	0	1
danger	0	0	6	0	1	0
control	5	0	0	1	0	0

Solution:

PMI (duty | assert) = lg ((3/100) / (0.22*0.1)) = 0.447

Vector of Duty = 0.447, 1.6, 0.114, 0.77, 1.51, 0

Name: _____ Reg #: _____ Section: _____