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:

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

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>) = $1.68 * 10^{-8}$

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**

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	Det → the	0.1
VP → Verb NP	0.7	Verb → Cut | Ask | Find ……	0.1
VP → Verb NP PP	0.3	Prep → with | in ……	0.1
NP → NP PP	0.3	Noun → envelop | grandma | scissors | men | suits | summer | ……	0.1
NP → Det Noun	0.7		
PP → Prep Noun	1.0		

(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

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.

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

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]**

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

	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