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:	**60 Minutes**	**Total Marks:**	**20**
Paper Date:	**26-Feb-18**	**Weight**	**13%**
Section:	**ALL**	**Page(s):**	**4**
Exam:	**Sessional 1 Solution**		

Q1) You are given the following corpus: [5 + 5 = 10 Marks]

<s> I am Sam </s>
<s> Sam I am </s>
<s> I am Sam </s>
<s> I do not like green eggs and Sam </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,
λ_1 = 0.5, λ_2 = 0.3, λ_3 = 0.2

<s> I like green eggs </s>

P (like | <s> I) = (0.5) (0) + 0.3(0) + 0.2 (1/25) = 1/125
P(green | I like) = 77/250
P(eggs | like green) = 101 / 125
P(</s> | green eggs) = 4/ 125

P (<s> I like green eggs </s>) = (1/125) (77/250) (101 / 125) (4 / 125) = $6.37 * 10^{-7}$

Name: _____ Reg #: _____ Section: _____

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

P(Sam | am) = (2 – 0.5) / 3 + 1/3(2/14) = 0.55

Q3) a) Let S = { a , b, c } the sample space. Suppose you are given following bigram probabilities P(b | a) = 0.125, P(a | c) = 0.25, P(c | c) = 0.25, P(a | a) = 0.25, P(c | b) = 0.125, P(a | b) = 0.25, P(<s> | c) = 0.1, P(<\s> | c) = 0.1, P(b | <s>) = 0.1, P(<\s> | a) = 0.1

Can you compute P(b | c) from the information given. If yes what is P(b | c) ? [2 + 3 = 5 Marks]

Yes, P (b | c) = 1 – (0.25 + 0.25 + 0.1 + 0.1) = 1 – 0.7 = 0.3

b) What is perplexity of bigram distribution from part (a) if computed against following data

<s> b c a <\s>

= ¼ (log (0.1) + log (0.125) + log (0.25) + log (0.1)) = -2.91
Perplexity = $2^{-l} = 2^{2.91} = 7.52$

Name: _____ Reg #: _____ Section: _____

Q4) Detect most probable real word spelling error from following sentence. Use Noisy Channel Model with bigram probabilities as given in Table 1. Probabilities of 1 character spelling mistakes are also given in Table 2. Assume all words with 1 edit distance from the words in test sentence are given in Table 1. α = Probability of word being correct = 0.95. **Show all calculations** [5 Marks]

<s> Three off the </s>

	Three	**off**	**the**	**there**	**tree**	**of**	**then**	**</s>**	**<s>**
Three	0.0001	0.0002	0.003	0.003	0.003	0.21	0.003	0.003	0.003
off	0.003	0.003	0.0003	0.003	0.003	0.003	0.003	0.003	0.003
the	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.01	0.01
there	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
tree	0.003	0.003	0.003	0.003	0.004	0.003	0.003	0.001	0.003
of	0.003	0.003	0.22	0.003	0.003	0.003	0.003	0.003	0.003
then	0.003	0.003	0.003	0.003	0.075	0.003	0.003	0.003	0.003
<s>	0.004	0.003	0.003	0.01	0.003	0.003	0.003	0.003	0.003
</s>	0.003	0.003	0.003	0.003	0.04	0.003	0.003	0.003	0.003

Table 1: Bigram Probabilities. P(of | Three) = 0.21

x | w	**P (x | w)**
re | er	0.03
Th | T	0.02
ff | f	0.05
e | ew	0.05

Table 2: x is spell error and w is correct

Solution:

P (<s> Tree off the </s>) = $5.4 * 10^{-13}$

P (<s> Three off the </s>) = $2.7 * 10^{-2}$
P (<s> Three of the </s>) = $9.24 * 10^{-8}$

The probability of “Three of the” is highest so spelling error is “off” instead of “of”

Name: _____ Reg #: _____ Section: _____