Computer Engineering Department – SVNIT - Surat.
Mid Semester March - 2020
B.Tech. - IV – 8th Semester
Course: Natural Language Processing (CO426)

Date: 4th March 2020 **Time:** 16:00 hrs to 17:30 hrs **Max Marks:** 30

Q1

A **Answer the following (Any Four)** [16]

1. Define language model. Considering the closed vocabulary system, for the given text, predict the word which is expected to come after: "*to inform*" using appropriate language model.
It is to inform you, that this is a keyboard. It is to inform you, that this is a mouse. It is to inform all, that computer is machine which consist keyboard and mouse

2. Given the corpus:
   1. *Pen, pencil and eraser are stationary items.*
   2. *Pen and pencil are used for writing.*
   3. *We cannot write on mobile phone using pencil.*
   4. *Pen and pencil are used to draw sketches.*

   Name and use the appropriate frequency based method to find out whether *pencil* is closer to *mobile* or *pen*. Consider the window size as 5. Apply necessary pre-processing techniques.

3. Design a model using appropriate technique to generate the present continuous word from the given root word (Eg: *going* will be generated from word *go*). State and apply appropriate rule so that model can also handle '*n insertion*' for the words (Eg: *running* will be generated from word *run*).

4. The text documents may be related to hard or soft landscape. Based on the given set of documents, design an optimum system to identify the landscape of a given Document D.
Documents are as follows:
   D_1: *Plants, grass are beautiful* – soft landscape.
   D_2: *This drive-way contains rocks and steps* – hard landscape.
   D_3: *Trees, plants and other grass are grown in backyard* – soft landscape.
   D_4: *Steps are made up of rocks* – hard landscape.
   New Document D: *sometimes grass is found on drive-ways and in-between steps.*

5. State and explain different types of ambiguity in each level of Natural Language Processing. Also, identify the types of ambiguity in each sentence of the given text:
Nita invited Sita to her home, but she told her she had to go to work.
Nita said "Okay, but ..."

B Explain OOV rate. Find OOV rate for the following dataset given below: [2]
Training Data: *Good Morning, all we are going to plan a conference on Machine learning. All are request to kindly submit their names to class mentor for the same.*
Test Data: *Class monitor will plan the schedule for the conference.*

Q2 Answer the following (Any Three) [12]

1. Parse the sentence *"Papa ate the caviar with a spoon"* with Earley's Algorithm and also draw a parse tree based on algorithm states only.

ROOT → S

S	→ NP VP	NP	→ Papa
NP	→Det N	N	→ caviar
NP	→NP PP	N	→ spoon
VP	→VP PP	V	→ ate
VP	→ V NP	P	→ with
PP	→ P NP	Det	→ the
Det	→ a		

2. Explain the CKY Algorithm with the following example:

S → N VP
VN → N N
VP → V N
N → students | Jeff | geometry | trains
V → trains

Parse the sentence *"Jeff trains geometry students"* using the CYK algorithm.

3. Find tagging error in each of the following sentences that are tagged with the Penn Treebank tagset [Annexure-I]:
   a. I/PRP need/VBP a/DT flight/NN from/IN Atlanta/NN
   b. Does/VBZ this/DT flight/NN serve/VB dinner/NNS
   c. I/PRP have/VB a/DT friend/NN living/VBG in/IN Denver/NNP
   d. Can/VBP you/PRP list/VB the/DT nonstop/JJ afternoon/NN flights/NNS

4. Give an equation for finding the most probable sequence of part of speech (POS) tags that could be utilised by a stochastic POS tagger. You should assume a bigram model.

****