Fr. C. Rodrigues Institute of Technology, Vashi
Department of Computer Engineering

Amroz Siddiqui

Course Code: CSDL7013 | Academic Year: Second Half - 2022
Course Name: Natural Language Processing Lab | Branch/Semester: Computer Engineering / VII

Name: _____

Roll No.: _____

Lab.: 04

1. **Title:** n-gram language model
2. **Objective/Aim:** To generate n-gram language model
3. **Tools/Techniques/Technology Used:** python
4. **Due Date:** Friday August 19, 2022
5. **Lab Instructors:**
   - Amroz Siddiqui
   - Ms. Padmashree

NOTE:

- There are three text files given.
- **survey.txt** to be used for exercise 1.
- The other two files, **madteaparty.txt** and **hitchhikersguidetogalaxy.txt** are for exercise 2.
- **Solve the following exercises:**
  1. *[02 Marks]* Read the file **survey.txt** given, generate the trigram model. Consider each sentence separately.
  2. *[08 Marks]* Read the given text files, remove all punctuation symbols, change all words to small case, generate trigram model for each one of them.

Checklist for exercises on **n-gram Language Model**

The exercises have been designed with specific learning objectives. The following self-assessment list should give you a fair idea of the extent to which you have learnt the expected material. Please respond to the statements after doing the exercises with ratings between 1 and 10. (1 being the lowest)

1. I know how to build find frequency of sentences in python: _____
2. I know how to find probability of words _____
3. I know how to compute conditional probability of words: _____
4. I know how to build the n-gram language model in python: _____