**School of Computer Science Engineering and Technology**

Course- BTech Type- Elective

Course Code- CSET335L Course Name-Natural Language processing

Year- 2024 Semester- Even

**Lab Assignment 1. Natural Language Processing: Tokenization and Language Modeling**

This week’s lab assignment is all about implementing your first language model.

You are supposed to download a huge body of text at least (10,000 words). You can download the text from Wikipedia.

You now must create a model that computes the probabilities of the next word based on unigrams and bigrams.

Assignment: Given a starting word, you now must predict the whole sentence of length 15 using the ideas of both unigrams and bigrams.