# IEMS5723 Assignment 1

## The Problem

A Language Model captures the characteristics of an article and allows us to make prediction of the next words appearing in a sequence. Estimate the probability of word(s) in case-insensitive manner with Maximum Likelihood Estimation (MLE) for (a) uni-gram and (b) bi-gram.

## The Input

The input file “input.txt” has less than 500 words. Only punctuation marks comma(,) and full-stop(.) would appear but they should be ignored.

## The Output

The output file “output.txt” should contains the top 5 highest probability (in alphabetical order in case items have the same probability) uni-grams and bi-grams with their probability (2 decimal places).

## Submission

A runnable Python program file should be submitted. NLTK is the only 3rd-party library allowed in this assignment and it is optional.

## Sample Input

This is a sample input, marking input is not going to be as-short-as this sample input.

## Sample Output

Uni-gram:

0.19: input

0.12: is

0.12: sample

0.12: this

0.06: a

Bi-gram:

1.00: a sample

1.00: as-short-as this

1.00: be as-short-as

1.00: going to

1.00: marking input