**CSCS 366 – Introduction to Natural Language Processing**

Assignment 2 Deadline: April 14th 11:59PM

Total Marks: 5 Roll No.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

You are given *“J. K. Rowling - Harry Potter 4 - The Goblet of Fire.txt”* file. Write a python program that reads the file, takes value of and creates gram model. Make sure to convert all the words to lower casing. Do not use any other preprocessing (i.e, no stemming, no lemmatization, no stopwords removal). Then do the following:

1. Create unigram language model. Then generate 100 words using seed/first word “he” as input
2. Create bigram language model. Then generate 100 words using seed/first words “He had” as input
3. Create trigram language model. Then generate 100 words using seed/first words “He had awoken”
4. Write a program for interpolation from unigram to trigram. Use weights 0.5 for trigram, 0.4 for bigram, 0.1 for unigram. Then generate 100 words using seed/first words “He had awoken”

**Submission Guidelines:**

* Submit Jupyter Notebook with all four outputs above.
* Do Not use any existing library to create language models. Write your own classes/methods/functions to achieve this.
* For tokenization, you can use space and punctuations. Do not use any existing tokenizer (such as nltk or spacy).
* Submit on Moodle.
* Submit all the files so that I can run your code and reproduce the results.