Assignment 02
N-gram Language Models

Deadline: March 20, 2022

General Rules

- It is an individual assignment, it needs to be done individually and you are not to exchange or share materials with other students.
- Use python for the solutions.
- Submit your assignment on Google Classroom as a zip file. The name of the file should be your name with your roll number(name-rollnumber.zip)
- Unless specified explicitly, do not use high-level NLP libraries, such as NLTK.
- Late submissions are not accepted.

The Corpus for this assignment should be prepared by yourself. The corpus should consist of 10 different domains and each domain should have 50 distinct files. You are supposed to implement following Python functions.

The text files are not tokenized. You need to implement a function with name *tokenize* () that takes the file path as its argument and returns the tokenized sentences.

Write a function *Ngram*() that should accept two required argument, n the order of the n-gram model & sentences and returns the n-grams .

Write a function *SentenceProb*() that should accept a sentence and returns the probability of the given sentence using Bigram model.

Write a function *SmoothSentenceProb*() that should accept a sentence and returns the probability of the given sentence using Bigram model and with Laplace smoothing.

Write a method *Perplexity*(), that calculates the perplexity score for a given sequence of sentences