Assignment 1- Exploring N-gram Models

In this assignment, you will be exploring N-gram models. You are given a corpus comprising of text from Harry Potter books. You are required to do the following:

1. Clean the data, this step can be done as per your choice. For example, you can remove capitalizations, remove certain tokens or punctuations as per your requirement.
2. Build N-gram models for n=1, 2, … m, choose m suitably, whatever is appropriate according to your analysis. Show one sentence for each case.
3. Experiment with various smoothening methods (Add-One, Good-Turing, Kneser-Ney, Backoff, Interpolation) and report your results.
4. Calculate perplexity for each case, report any trends or observations.

You need to implement N-gram models, smoothening and perplexity functions from scratch, no libraries are allowed for these, libraries can be used for data cleaning. You need to report the best model by changing n values and smoothening.

Dataset: Please download the text from the Link

Submission: You need to submit a report containing all your experiment details and results along with code in a zip file. Name the zip file by your entry number (e.g. 2019EEXXXXX.zip). Submission is to be made via Google Classroom, Link: Classroom. Deadline 27/01/23 11:59pm.