NATURAL LANGUAGE PROCESSING

I5GIC

Mini Project 1 – Text Generation

Select a corpus of your choice (for example, articles from Wikipedia). Separate the text corpus into 3 subsets: training (70%), validation (10%) and testing (20%). Then tokenize the corpus using split in python or the tokenizer from nltk. Limit the vocabulary size and replace the rest of the tokens as <UNK>.

Build two 4-gram language models using the training set:

a) LM1: backoff method (the computation is unsmoothed).
b) LM2: interpolation method. The computation of each n-gram term is also smoothed using add-k smoothing technique. You can conduct your own experiments to find the best values for the hyper-parameters λ's and k (use the validation set for the experiments).

I. Evaluate the two models on the test set using perplexity evaluation metric.

II. Create a text generator using each of the two models.

Write a one to two-page report the experiments and results.