Assignment 2

March 4, 2024

1 Language Modelling

In this question, we'll use n-gram probabilities for Language modelling. Download the pre-processed dataset from https://www.kaggle.com/datasets/moxxis/harry-potter-lstm. You can use the first 10,000 words for this question.

1. Preprocess and tokenize the dataset using NLTK. (1 mark)
2. Refer NLTK documentation and fit two bigram language models on the text: MLE and Kneser-Ney discounting. (2 marks)
3. Use the beginning words 1. "Harry Potter" and 2. "Dumbledore" to generate text using both the language models. Keep the maximum text length as 20. (1 mark)

The above language modelling approaches are greedy approaches (the predicted next word is the word with highest conditional probability). But it is possible that this greedy decoding may be sub-optimal. Hence better decoding strategies have been proposed in literature. One popular decoding strategy is beam search. Beam search is a tree-based search strategy similar to BFS. In BFS, we expand every child node, however in Beam search, we expand only top k most probable children. The generated text is the text with the highest probability.

4. To Implement Beam Search, you would have to find the top k most probable words given some context. Implement a function for this. (1 mark) Hint: You may use the lm.vocab variable to your advantage.
5. Implement Beam search. Use the MLE Language model trained previously. (3 marks)
6. Repeat part 3 using Beam Search with k=2 and depth=10. Find the 5 generated texts with highest probability for each of the 2 beginning phrases. (2 marks)

Note that you can calculate probability of generated text while doing beam search itself.