Natural Language Processing
Lab 2

January 22, 2024

This lab sheet is to practice the concepts taught this week far with a focus on n-gram language models.

1. Write out the equation for trigram probability estimation (modifying Eq. 3.11 from SLP Chapter 3). Now write out all the non-zero trigram probabilities for the I am Sam corpus in Chapter 3 on page 4.

2. Write a program to compute unsmoothed n-grams. Use the Dr. Seuss corpus to test this.

3. Run your n-gram program on two different small corpora of your choice (you might use email text or newsgroups). Now compare the statistics of the two corpora. What are the differences in the most common unigrams between the two? How about interesting differences in bigrams?

4. Write a definition of perplexity in language modeling.

5. Add an option to your program to compute the perplexity of a test set.

6. You are given a training set of 100 numbers that consists of 91 zeros and 1 each of the other digits 1-9. Now we see the following test set: 0 0 0 0 0 3 0 0 0 0. What is the unigram perplexity?