Lecture "Foundations: Language Models"

Prof. Dr. D. Klakow

Due: 9.12.05

Exercise

In this exercise you should built your own small piece of language model software. You can use any programming language (e.g. perl).

1. Download one of the books from http://www.gutenberg.org/. Remove the parts that do not belong to the actual text. Take the first 90% of the data as training data and the rest for testing.

2. Create a list of all unique words in the complete corpus (training+test).

3. Count the frequency of all words in the training corpus. Words that only occur in the test but not in the training data will get frequency 0.

4. Write a small tool that reads the frequencies created in the previous step as an argument and the test corpus. It should implement absolute discounting for a unigram language model and calculate the probability for each word from the test corpus. Based on these probabilities you can calculate perplexity.

5. How does perplexity depend on the discounting parameter d. Make a graph.