T-61.5020 Statistical Natural Language Processing
Exercises 8 — N-gram language models
Version 1.0

1. This task is recommended to do without looking ahead, so that only the information given so far affects your estimates.

   a) The task is to estimate the probability of the word following words "tuntumaan jo" (*feel already*). The possible followers are words

   – ja [*and*]
   – hyvältä [*good*]
   – kumisaapas [*rubber boot*]
   – keväältä [*(like) spring* (season)]
   – ilman [*without*]
   – päihtyneeltä [*drunk*]
   – turhalta [*vain*]
   – koirineen [*with (his) dogs*]
   – öljyiseltä [*oily*]
   – Turku [a city in Finland]

   Give a probability value for each so that they sum up to one. Compare the estimates given by yourself to ones calculated directly from a text corpus.

   b) Now you know the full beginning of the sentence, which is "Leuto sää ja soidinmenonsa aloittaneet tiaiset ovat saaneet helmikuun tuntumaan jo" (free translation: *"Mild weather and the titmice that have started their displays have made the February feel already"*). Estimate the same probabilities using this full context.

   c) What kind of knowledge would a language model need to order to match up with a human in the b) case?

   (Word used in the original sentence is found on the next page.)

2. A language model has a vocabulary of 64 000 words in base forms. We know that the word history is (1) "vuosi joka olla" (*year that be*) or (2) "tämä tehtävä vaikuttaa" (*this task appear*). Estimate the probabilities for the next word to be either "olla", "leuto", or "gorilla". Estimate both unigram and bigram probabilities using

   a) ...maximum likelihood estimates
   b) ...ML estimates with Laplace smoothing
   c) ...ML estimates with Lidstone (additive) smoothing with parameter $\lambda = 0.01$.

The training material is from Finnish version of the text in Problem 1. Preprocessed version with words converted to their base forms is available from `http://www.cis.hut.fi/Opinnot/T-61.5020/Exercises08/extra/ex8-2_data.txt`.

3. In the previous exercise we calculated separate smoothed distributions for unigrams and bigrams. However, it is more sensible to combine the estimates of n-grams of different lengths either with a back-off or an interpolated model. E.g., when we observed the probabilities of words "olla" and "leuto" in the context "vaikuttaa", the estimates were equal because there were no occurrences for either of the bigrams. Yet we know form the unigram probabilities that "olla" is much more likely to occur than "leuto", and thus we can assume that it is more likely also in an unseen context.

   Use an interpolated bigram model to calculate probabilities for the examples of the previous exercise. Smooth the bigram estimates using absolute discounting with discount parameter $D = 0.5$.

4. Calculate perplexity for the following sentence: "Kielen oppiminen on monimutkainen ja huonosti ymmärretty tapahtumaketju." Probabilities for a back-off n-gram model can be counted as follows:

$$P(w_3|w_2, w_1) = \begin{cases} T(w_1, w_2, w_3) & \text{if there exists trigram } w_1, w_2, w_3 \\ bo(w_1, w_2)P(w_3|w_2) & \text{if there exists bigram } w_1, w_2 \\ P(w_3|w_2) & \text{otherwise} \end{cases}$$

$$P(w_2|w_1) = \begin{cases} T(w_1, w_2) & \text{if there exists bigram } w_1, w_2 \\ bo(w_1)T(w_2) & \text{otherwise} \end{cases}$$

   Values for the functions T and b are given in Table 1.

n-grammi	$\log_{10}(T)$	$\log_{10}(\text{bo})$
kielen	-4.1763	-0.2917
kielen oppiminen	-2.1276	-0.0526
kielen oppiminen on	-0.4656	
oppiminen on	-0.5889	-0.001
on monimutkainen	-4.2492	-0.0697
on monimutkainen ja	-0.8876	
monimutkainen ja	-0.8660	0.0495
ja huonosti	-4.1804	-0.1415
huonosti	-4.2513	-0.1652
ymmärretty	-5.2195	-0.0870

Table 1: *Probabilities of an n-gram model trained with 30 million word corpus for the most common 64.000 words. Katz back-off with Good-Turing smoothing were used. Only the estimates relevant to the problem are shown in the table.*

The original word missing in Problem 1 was "keväältä".