Language modelling: Exam practice

> **Note**
>
> This document gives *examples* for tasks similar to those that will appear on the digital written exam. The solutions are provided at the end of this document, in case you first want to try solving the tasks yourself. This is not meant as an exhaustive list; the exam may also feature other types of tasks/questions than the ones shown here.

Task 2.1

Consider a dataset of children's stories, containing a total of 3850000 tokens with a vocabulary of 52800 unique words. The following counts of unigrams and bigrams were extracted from this dataset:

she	*thought*	*she thought*	*thought she*
27751	3633	309	145

Estimate the following probabilities using maximum likelihood estimation without smoothing.

P(*she*)	P(*she* | *thought*)

Task 2.2

Using the same unigram and bigram counts as in Task 2.1, estimate the following probabilities using maximum likelihood estimation with add-k smoothing, using $k = \frac{1}{5}$.

P(*she*)	P(*thought* | *she*)

Task 2.3

A bigram language model has been trained on a dataset using the special symbols BOS and EOS to mark sentence boundaries. How does the model compute the probability of the following sentence? Answer with a formula.

my cat sleeps

Task 2.4

You train two language models on the same training set, one without smoothing, and one with add-k smoothing. Afterwards, you compute the perplexity of the two language models on the training set. Which of the following statements are true? Select all that apply!

- ☐ This perplexity value is a good indicator of the model's ability to generalize to new, unseen texts.
- ☐ The model without smoothing will have higher perplexity than the model with add-k smoothing.
- ☐ The model without smoothing will have lower perplexity than the model with add-k smoothing.
- ☐ The perplexity will always be a value ≥ 1.

Task 2.5

A unigram language model has been trained with maximum likelihood estimation, resulting in the following probabilities:

P(*languages*)	**P(*likes*)**	**P(*Lisa*)**	**P(*loves*)**
$\frac{2}{500}$	$\frac{10}{500}$	$\frac{1}{500}$	$\frac{8}{500}$

What is the entropy of this unigram language model on the following sentence?

Lisa loves languages

(You can ignore sentence boundary markers for the purpose of this task.)

Solutions begin on the next page!

Solutions

Task 2.1

P(*she*)	P(*she* | *thought*)
$\frac{27751}{3850000}$	$\frac{145}{3633}$

Task 2.2

P(*she*)	P(*thought* | *she*)
$\frac{27751+\frac{1}{5}}{3850000+\frac{1}{5}\times 52800}$	$\frac{309+\frac{1}{5}}{27751+\frac{1}{5}\times 52800}$

Task 2.3

$$P(\text{my} \mid \text{BOS}) \times P(\text{cat} \mid \text{my}) \times P(\text{sleeps} \mid \text{cat}) \times P(\text{EOS} \mid \text{sleeps})$$

Task 2.4

- [ ] This perplexity value is a good indicator of the model's ability to generalize to new, unseen texts.
- [ ] The model without smoothing will have higher perplexity than the model with add-k smoothing.
- [x] The model without smoothing will have lower perplexity than the model with add-k smoothing.
- [x] The perplexity will always be a value ≥ 1.

Note

The first item could be true if the perplexity value was computed on a separate *test* set; the question explicitly stated that we compute the perplexity on the *training* set, which gives us no information about the model's ability to generalize to new, unseen texts.

For perplexity vs. smoothing, see the lecture slides.

Task 2.5

$$-\frac{1}{3}\log\left(\frac{1}{500} \times \frac{8}{500} \times \frac{2}{500}\right)$$