01 **Text classification** **(3 points)**

A Naive Bayes text classifier has to decide whether the document 'Stockholm Oslo' is news about Sweden (class `S`) or news about Denmark (klass `D`).

a) List the probabilities that the classifier uses to make this decision.

b) Estimate the relevant probabilities from the following document collection using Maximum Likelihood estimation. Answer with fractions.

	document	class
1	Stockholm Oslo	`S`
2	Köpenhamn Stockholm	`D`
3	Stockholm Köpenhamn	`S`
4	Köpenhamn Oslo	`D`

c) Based on the estimated probabilities, which class does the classifier predict? Explain. Show that you have understood the Naive Bayes classification rule.

Sample answers:

a) $P(\texttt{S})$, $P(\texttt{D})$, $P(\text{Stockholm} \mid \texttt{S})$, $P(\text{Stockholm} \mid \texttt{D})$, $P(\text{Oslo} \mid \texttt{S})$, $P(\text{Oslo} \mid \texttt{D})$

b) Estimated probabilities:

$$P(\texttt{S}) = 2/4 \qquad P(\text{Stockholm} \mid \texttt{S}) = 2/4 \qquad P(\text{Oslo} \mid \texttt{S}) = 1/4$$
$$P(\texttt{D}) = 2/4 \qquad P(\text{Stockholm} \mid \texttt{D}) = 1/4 \qquad P(\text{Oslo} \mid \texttt{D}) = 1/4$$

c) The system first computes class-specific scores:

$$\begin{aligned}\text{score}(\texttt{S}) &= P(\texttt{S}) \cdot P(\text{Stockholm} \mid \texttt{S}) \cdot P(\text{Oslo} \mid \texttt{S}) \\ &= \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{1}{4} = \frac{4}{64}\end{aligned}$$

$$\begin{aligned}\text{score}(\texttt{D}) &= P(\texttt{D}) \cdot P(\text{Stockholm} \mid \texttt{D}) \cdot P(\text{Oslo} \mid \texttt{D}) \\ &= \frac{2}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{2}{64}\end{aligned}$$

The system then predicts the class with the highest score, here: `S`.

02 **Language modelling** **(3 points)**

For the 520 million word Corpus of Contemporary American English, we have the following counts of unigrams and bigrams: *your*, 883,614; *rights*, 80,891; *doorposts*, 21; *your rights*, 378; *your doorposts*, 0.

a) Estimate the probabilities $P(your)$ and $P(rights \mid your)$ using Maximum Likelihood estimation. Answer with fractions.

b) Estimate the bigram probability $P(doorposts \mid your)$ using Maximum Likelihood estimation and add-k smoothing with $k = 0.01$. Assume that the vocabulary consists of 1,254,193 unique words. Answer with a fraction.

c) What is entropy, and how can we use this measure to evaluate the quality of a language model? Give an informal explanation.

Sample answers:

a) Maximum Likelihood Estimation:

$$P(your) = \frac{883{,}614}{520{,}000{,}000} \qquad P(rights \mid your) = \frac{378}{883{,}614}$$

b) Estimation with add-k smoothing, $k = 0.01$:

$$P(doorposts \mid your) = \frac{0 + 0.01}{883{,}614 + 0.01 \cdot 1{,}254{,}193}$$

c) Entropy is a measure for how 'surprised' a language model is, on average per word, when it is presented with a text. Formally, entropy is the negative log probability of the text divided by the number of words in the text. A good language model has low entropy; a bad language model has high entropy.

03 **Part-of-Speech Tagging** **(3 points)**

The evaluation of a part-of-speech tagger produced the following confusion matrix. The marked cell gives the number of times the system tagged a word as a verb (`VB`) whereas the gold standard specified it as a noun (`NN`).

	NN	JJ	VB
NN	58	6	1
JJ	5	11	2
VB	0	7	43

a) Set up a fraction for the tagger's accuracy.

b) Set up fractions for the tagger's recall on verbs and its precision on nouns.

c) Write down another confusion matrix where accuracy is the same as in the matrix above, but where the tagger's precision on adjectives is 100%.

Sample answers:

a) $(58 + 11 + 43)/(58 + 6 + 1 + 5 + 11 + 2 + 7 + 43) = 112/133$

b) $43/(0 + 7 + 43) = 43/50$, $58/(58 + 5 + 0) = 58/63$

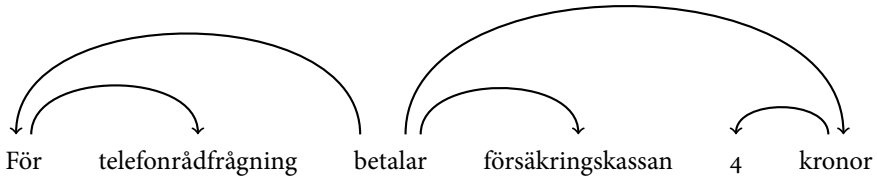
c) Example:

	NN	JJ	VB
NN	58	0	7
JJ	5	11	2
VB	7	0	43

Explanation: The tagger's correctness is the number of instances on the diagonal (112) divided by the total number of instances in the matrix (133). To maintain accuracy, we can thus take the given matrix and move the 13 misclassified instances of `JJ` to some other column, but not to the diagonal.

04 **Syntactic analysis** **(3 points)**

A transition-based dependency parser analyses the sentence *För telefonrådfrågning betalar försäkringskassan 4 kronor*. Here is the gold-standard tree for this sentence:



a) Give the initial configuration for the sentence. To represent the partial dependency tree, list the arcs contained in it.

b) Suppose now that the parser takes the transitions `SH`, `SH`, `RA`. State the new configuration. Represent the partial dependency tree as in the previous item.

c) State a complete sequence of transitions that takes the parser all the way from the initial configuration to a terminal configuration, and that recreates all arcs of the gold-standard tree.

Sample answers:

a) Initial configuration:

stack: [] **buffer:** [För, telefonrådfrågning, betalar, försäkringskassan, 4, kronor]

Partial dependency tree: no arcs

b) Configuration after `SH`, `SH`, `RA`:

stack: [För] **buffer:** [betalar, försäkringskassan, 4, kronor]

Partial dependency tree: För → telefonrådfrågning

c) `SH SH RA SH LA SH RA SH SH LA RA`

05 **Semantic analysis** **(3 points)**

Consider the following document collection:

(1) automobile wheel motor vehicle transport passenger

(2) car form transport wheel capacity carry five passenger

(3) transport London game spectator advise avoid use car

(4) London soccer tournament begin goal match

(5) Giggs score goal football tournament Wembley London

(6) Bellamy passenger football match play part goal

a) Complete the following co-occurrence matrix. Each cell is supposed to contain the number of documents in which the target word (row) co-occurs with a specific context word (column).

	context 1		context 2	
	passenger	transport	goal	match
automobile	☐	☐	☐	☐
car	☐	☐	☐	☐
soccer	☐	☐	☐	☐
football	☐	☐	☐	☐

b) Draw the target words as vectors in a two-dimensional coordinate system where the x-axis corresponds to the total number of occurrences in context 1 (*passenger*, *transport*) and the y-axis corresponds to the total number of occurrences in context 2 (*goal*, *match*).

c) How can we use these vector representations to measure the semantic similarity between the target words? What results would that give for this concrete example? Answer with a short text.

Sample answers:

a) Co-occurrence matrix:

	context 1		context 2	
	passenger	transport	goal	match
automobile	1	1	0	0
car	1	2	0	0
soccer	0	0	1	1
football	1	0	2	1

b) Vectors: *automobile* $(2, 0)$, *car* $(3, 0)$, *soccer* $(0, 2)$, *football* $(1, 3)$.

c) The semantic similarity between target words can be measured as the straight-line or cosine distance between their vector representations. With this method we would get two clusters, each of which containing words that are more similar to one another than to the words in the other cluster: *automobile*, *car*; *soccer*, *football*.