01 Text classification (3 points)

a) The evaluation of a text classifier produced the following confusion matrix. The marked cell gives the number of times the system classified a document as class `C` whereas the gold-standard class for the document was `A`.

	A	B	C
A	58	6	1
B	5	11	2
C	0	7	43

Based on this confusion matrix, compute the following values. Answer with fractions; you do not have to simplify them.

i. recall with respect to class `B`

ii. precision with respect to class `C`

b) A Naive Bayes text classifier has to decide whether the two-word document 'Stockholm Oslo' is news about Sweden (class `S`) or news about Denmark (class `D`). Estimate the relevant probabilities from the following document collection using Maximum Likelihood estimation (without smoothing). Answer with fractions.

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

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

Sample answers:

a) precision = columns, recall = rows

i. $\frac{11}{5+11+2}$

ii. $\frac{43}{1+2+43}$

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:

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

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

02 Language modelling **(3 points)**

The Corpus of Contemporary American English (COCA) is the largest freely-available corpus of English, containing approximately 560 million tokens. In this corpus we have the following counts of unigrams and bigrams:

snow	*white*	*white snow*	*purple*	*purple snow*
38,186	256,091	122	11,218	0

a) Estimate the following probabilities using maximum likelihood estimation without smoothing. Answer with fractions containing concrete numbers. You do not have to simplify the fractions.

   i. $P(\textit{purple})$ ii. $P(\textit{snow} \mid \textit{white})$

b) Estimate the following probabilities using maximum likelihood estimation with additive smoothing, $k = 0.1$. Assume that the vocabulary consists of 1,254,193 unique words. Answer with fractions containing concrete numbers. You do not have to simplify the fractions.

   i. $P(\textit{snow})$ ii. $P(\textit{snow} \mid \textit{purple})$

c) We use maximum likelihood estimation with add-k smoothing to train two trigram models on the COCA corpus: model A with $k = 0.01$, model B with $k = 0.1$. We compute the entropy of both models on the same data that we used for training. Which model has the higher entropy, and why? Answer with a short text.

Sample answers:

a) Maximum likelihood estimation without smoothing:

$$P(purple) = \frac{11218}{560 \cdot 10^6} \qquad\qquad P(snow \mid white) = \frac{122}{256{,}091}$$

b) Maximum likelihood estimation with add-k smoothing, $k = 0.01$:

$$P(snow) = \frac{38{,}186 + 0.01}{560 \cdot 10^6 + 0.01 \cdot 1{,}254{,}193}$$

$$P(snow \mid purple) = \frac{0 + 0.01}{11{,}218 + 0.01 \cdot 1{,}254{,}193}$$

c) The k values for the three rows are $k = 0.1$, $k = 0$, and $k = 1$. On the one hand, higher values of n give lower entropy. On the other hand, add-k smoothing *increases* the entropy of the training data (decreases the total probability of n-grams that occur in the data), and the more so the higher the k and the n.

03 Part-of-speech tagging **(3 points)**

a) Give at least four examples of part-of-speech categories in English. In each case, state both the name of the category and a word that falls into that category.

b) The following matrices specify (parts of) a hidden Markov model. The marked cell specifies the probability for the transition from `BOS` to `AB`.

	AB	PN	PP	VB	EOS
BOS	1/11	1/10	1/12	1/11	1/25
AB	1/11	1/11	1/11	1/10	1/14
PN	1/11	1/12	1/12	1/10	1/16
PP	1/13	1/11	1/12	1/14	1/18
VB	1/11	1/10	1/10	1/13	1/15

	she	got	up
AB	1/25	1/25	1/14
PN	1/13	1/25	1/25
PP	1/25	1/25	1/13
VB	1/25	1/14	1/19

We use the model to tag the sentence 'she got up'. Compute the probabilities for the following possible tag sequences. Answer with fractions. Which sequence gets the higher probability?

i. `PN VB AB` ii. `PN VB PP`

c) State at least three substantial differences between tagging with the hidden Markov model (based on the Viterbi algorithm) and tagging with the multi-class perceptron (based on the greedy left-to-right algorithm).

Sample answers:

a) adjective (*big*, *old*); noun (*girl*, *cat*, *tree*); verb (*run*, *eat*); pronoun (*you*, *herself*)

b) Sequence i. gets the higher probability:

$$\text{PN VB AB} = \frac{1}{10 \cdot 13 \cdot 10 \cdot 14 \cdot 11 \cdot 14 \cdot 14} = \frac{1}{10 \cdot 13 \cdot 10 \cdot 14} \cdot \frac{1}{2156}$$

$$\text{PN VB PP} = \frac{1}{10 \cdot 13 \cdot 10 \cdot 14 \cdot 10 \cdot 13 \cdot 18} = \frac{1}{10 \cdot 13 \cdot 10 \cdot 14} \cdot \frac{1}{2340}$$

c) see the comparison on the lecture slides (page 46)

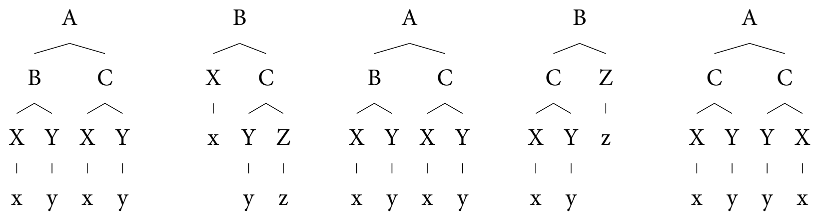
04 Syntactic analysis **(3 points)**

a) Here are all NP-rules and all VP-rules together with their probabilities from a certain probabilistic context-free grammar. State the missing numbers a, b, c.

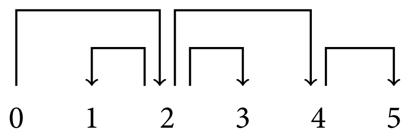
$$\text{NP} \to \text{PRP}\ \tfrac{2}{7} \quad \text{NP} \to \text{NP PP}\ a \quad \text{NP} \to \text{DT NN}\ \tfrac{2}{7} \quad \text{NP} \to \text{NN}\ \tfrac{2}{7}$$

$$\text{VP} \to \text{VB NP}\ \tfrac{2}{b} \quad \text{VP} \to \text{VB NP PP}\ \tfrac{2}{c}$$

b) Below is a small phrase structure treebank. Read off all rules whose left-hand sides are either A or B and estimate their rule probabilities using maximum likelihood estimation (no smoothing).



c) State a sequence of transitions that make the transition-based dependency parser produce the following dependency tree:



Sample answers:

a) $a = \frac{1}{7}, b = c = 4$

b) All rules whose left-hand sides are either A or B:

$$\text{A} \to \text{B C}\ \tfrac{2}{3} \quad \text{A} \to \text{C C}\ \tfrac{1}{3}$$

$$\text{B} \to \text{X Y}\ \tfrac{2}{4} \quad \text{B} \to \text{X C}\ \tfrac{1}{4} \quad \text{B} \to \text{C Z}\ \tfrac{1}{4}$$

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

05 Semantic analysis (3 points)

a) Provide an example word pair for each of the following semantic relations:

i. synonym

ii. antonym

iii. hyponym

iv. hypernym

b) Here are three signatures (glosses and examples) from Wiktionary for different senses of the word *colour*:

1 The spectral composition of visible light. *Humans and birds can perceive colour.* **2** A particular set of visible spectral compositions, perceived or named as a class. *Most languages have names for the colours black, white, red, and green.* **3** Hue as opposed to achromatic colours (black, white, and grays). *He referred to the white flag as one 'drained of all colour'.*

Based on these signatures, which of the three senses of the word *colour* does the Lesk algorithm predict in the following sentence? Ignore the word *colour*, stop words, and punctuation.

As the large flag of blue colour was raised in a highly visible spot at the top of the mountain, a light rain began to fall.

c) We read off word vectors from the following co-occurrence matrix (target words correspond to rows, context words correspond to columns):

	HuSHa'	*Ha'DIbaH*
qa'vIn	5	1
qurgh	5	5
jonta'	1	0
Dargh	1	4

Sort the four words in decreasing degree of semantic similarity (most similar to least similar) to the word *jonta'*, assuming that semantic similarity is measured as the angle between word vectors.

Sample answers:

a) Semantic relations:

   i. *purchase* is synonym to *buy*
   ii. *hot* is antonym to *cold*
   iii. *chair* is hyponym to *furniture*
   iv. *bird* is hypernym to *eagle*

b) sense 1 (match with *visible* and *light*)

c) *jonta'*, *qa'vIn*, *qurgh*, *Dargh*