General Instructions

This exam consists of two parts:

- **Part A** consists of math questions, multiple-choice questions, and other questions requiring *simple* answers. You can obtain a total of **24 points** in this part.
- **Part B** consists of questions requiring longer and *well-formulated* free-text answers. You can obtain an additional **8 points** in this part.

Note: The format and content of this exam was changed this year, and the exact points you can get for individual questions + grade thresholds may be adjusted for future course iterations.

Grade requirements

- To obtain a **passing grade** for the exam, you need **at least 18 points from Part A** (i.e. 75% of the total possible points from this part).
- To obtain a **higher grade** for the exam, you can earn additional points in both Part A and B.

The following table shows how your total points will map to your grade for the exam, depending on your course code:

	0–17	18–20	21–23	24–26	27–29	30–32
729G86	F	E	D	C	B	A
TDP030	U	3	3	4	4	5

Remember that **at least 18 points need to come from Part A**. If you do not reach this threshold in Part A, we will not grade Part B. In other words, you cannot compensate for missing the 18-point threshold in Part A by answering questions in Part B.

Instructions for mathematical expressions

When a question requires you to answer with a numerical expression, all expressions that evaluate to the correct answer will be considered equally correct. This means that **you do not need to simplify fractions** or evaluate them yourself. Please avoid using unnecessary round brackets/parentheses in your math expressions.

To write a fraction in Inspera, type a slash (/) on your keyboard. For example, you can write the fraction $\frac{1}{2}$ by typing “1/2” on your keyboard. You can also use the fraction symbol on the toolbar that appears when you press the “Σ” symbol in a math answer field.

Tips

- It’s a good idea to **go through the entire exam first** and familiarize yourself with the tasks. I (Marcel) will visit the exam halls within the first hour of the exam so that you can ask for clarification if anything is unclear.
- **Solve the easier questions first!** You can solve questions in any order, so you can skip questions that you struggle with and come back to them later.
- Remember that you only need to work on **Part B** if you’re aiming **for a higher grade**. If you only want a passing grade for the exam, you can skip Part B entirely.

Good luck!

For all answers requiring mathematical expressions, all answers that evaluate to the same value are considered equally correct. The "correct" answers shown here are therefore not exhaustive.

1.1 Precision and recall

Below is a confusion matrix from evaluating a classifier that predicts the genre of a song based on its lyrics. Rows correspond to gold-standard classes, while columns correspond to predicted classes. *(For example, the **highlighted** cell contains the number of times the classifier predicted "country" where the gold-standard class was "hip-hop".)*

	pop	rock	hip-hop	country
pop	1500	250	80	20
rock	435	1120	12	25
hip-hop	52	7	850	**4**
country	33	12	2	400

Based on the confusion matrix above, compute the following evaluation metrics!

a) precision with respect to "pop"

= [ ] ($\frac{1500}{1500+435+52+33}$, $\frac{1500}{(1500+435+52+33)}$, $\frac{1500}{2020}$)

b) recall with respect to "hip-hop"

= [ ] ($\frac{850}{52+7+850+4}$, $\frac{850}{(52+7+850+4)}$, $\frac{850}{913}$)

Maximum marks: 2

1.2 Most frequent class baseline I

Here are some counts of class labels from a text classification dataset:

	A	B	C
training data	255	275	190
test data	100	95	80

Given the counts above, what is the **accuracy of the most frequent class baseline** on the test data?

Answer: [ ] ($\frac{95}{100+95+80}$, $\frac{95}{(100+95+80)}$, $\frac{95}{95+100+80}$, $\frac{95}{275}$)

Maximum marks: 1

1.3 Most frequent class baseline II

What is the main reason for using a most frequent class baseline?

Select one alternative:

- To serve as a point of comparison for evaluating another classifier. ✔
- To balance the class distribution in the training data.
- To detect what the most frequent class in the dataset is.
- To estimate how well the classifier generalizes to new, unseen data.

Maximum marks: 1

1.4 Maximum likelihood estimation and smoothing

Consider a dataset of English literature, containing a total of **225000 tokens** with a vocabulary of **19000 unique words**. Assume that the following counts of unigrams and bigrams were extracted from this dataset:

warn	of	warn of	of warn
35	950	6	0

a) Estimate the following probabilities using maximum likelihood estimation (MLE) **without smoothing.**

$P(warn)$	$P(of\|warn)$
($\frac{35}{225000}$)	($\frac{6}{35}$)

b) Estimate the following probabilities using maximum likelihood estimation (MLE) **with add-one smoothing.**

$P(of)$	$P(warn\|of)$
($\frac{950+1}{225000+1\cdot 19000}$, $\frac{950+1}{225000+19000}$)	($\frac{0+1}{950+1\cdot 19000}$, $\frac{1}{950+19000}$)

Maximum marks: 4

1.5 Conditional probabilities in n-gram LMs

A **trigram language model** has been trained on a dataset using the special symbols BOS and EOS to mark sentence boundaries. Which of the following probabilities does the model use to compute the probability of the following sentence?

- *the dog ate my homework*

Mark all that apply!

- [ ] $P(the|dog\ ate)$
- [x] $P(dog|\text{BOS}\ the)$
- [ ] $P(\text{BOS}|the\ dog)$
- [ ] $P(dog|the)$
- [ ] $P(ate|dog)$
- [ ] $P(homework|my)$
- [x] $P(ate|the\ dog)$
- [ ] $P(\text{EOS}|my\ homework)$
- [x] $P(\text{EOS}|homework)$
- [ ] $P(homework|ate\ my)$

Note: You can get partial points for questions like these if you answer partially correct. The "minimum points" are always zero, i.e. you can never get negative points from a single question.

Maximum marks: 2

1.6 Entropy and perplexity

N-gram language models compute probability values for sequences of words. However, it is more common to use **entropy and perplexity** when evaluating them. Which of the following statements are true?

Mark all that apply!

- [ ] Perplexity is an example for an extrinsic evaluation metric.
- [ ] The perplexity will always be higher than the entropy. ✔
- [ ] Higher perplexity corresponds to lower probability. ✔
- [ ] The perplexity will always be a value between 0 and 1.

Maximum marks: 2

1.7 BIO tagging scheme

Below is a sentence with named entity spans ***highlighted*** and annotated with their type. For example, the span “Italian” is annotated as a named entity of type “NAT” (nationality).

The ***Italian*** *poet was called* ***Pietro Metastasio***
NAT **PER**

Convert the named entity annotations shown above to **BIO notation** by choosing the correct tag for each token!

token	tag
The	Select alternative (B-PER, I-PER, B-NAT, I-NAT, O)
Italian	Select alternative (B-PER, I-PER, B-NAT, I-NAT, O)
poet	Select alternative (B-PER, I-PER, B-NAT, I-NAT, O)
was	Select alternative (B-PER, I-PER, B-NAT, I-NAT, O)
called	Select alternative (B-PER, I-PER, B-NAT, I-NAT, O)
Pietro	Select alternative (B-PER, I-PER, B-NAT, I-NAT, O)
Metastasio	Select alternative (B-PER, I-PER, B-NAT, I-NAT, O)

Maximum marks: 1

1.8 Span-level precision/recall

Below is a sentence with named entity spans ***highlighted***, both according to a gold-standard annotation and the prediction of a named entity recognition model.

gold-standard	*The **Royal Museum of Mariemont** is a museum situated in **Mariemont**.*
predicted	*The **Royal Museum** of **Mariemont** is a museum situated in **Mariemont**.*

Based on these underlined spans, compute the following **span-level** evaluation metrics!

1. **Span-level precision:** [ ] ($\frac{1}{3}$)
2. **Span-level recall:** [ ] ($\frac{1}{2}$)

Maximum marks: 2

1.9 Distributional hypothesis

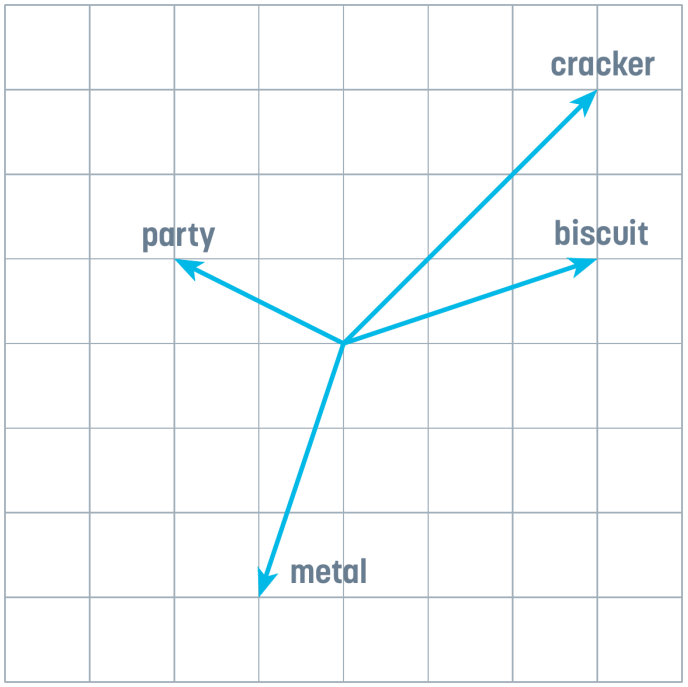
State the "distributional hypothesis".

"A word's distribution tells us something about its meaning."
"We can learn something about a word's meaning by looking at the context in which it is used."
"You shall know a word by the company it keeps."

Maximum marks: 2

NB: This kind of question will likely just give 1 point in future exams.

1.10 Cosine similarity



The illustration above shows some words along with their vectors in a two-dimensional embedding space.

Match the word pairs below to the cosine similarity of their vectors!

	–0.894	–0.316	+0.894	+1.000
cracker–biscuit	○	○	○ ✔	○
cracker–metal	○ ✔	○	○	○
metal–metal	○	○	○	○ ✔
cracker–party	○	○ ✔	○	○

Maximum marks: 2

1.11 Masked LMs

A masked language model like BERT is an example of what type of model architecture?

Select one alternative:

- encoder–decoder model
- autoregressive model
- encoder model ✔
- decoder model

Maximum marks: 1

1.12 Fertility scores

Below are sentences and their representation by some subword tokenizer, formatted as a Python list. What is the **fertility score** of the tokenizer on each sentence?

a)

sentence	Not all paleontology involves fossils
tokens	['Not', 'all', 'pal', '##eont', '##ology', 'involves', 'foss', '##ils']

- **Fertility:** [ ] ($\frac{8}{5}$, 1.6)

b)

sentence	Nothing is as cuddly as a pet porcupine
tokens	['Nothing', '_is', '_as', '_cuddly', '_as', '_a', '_pet', '_por', 'cu', 'pine']

- **Fertility:** [ ] ($\frac{10}{8}$, $\frac{5}{4}$, 1.25)

Maximum marks: 2

1.13 Prompting

A generative language model is being tested on basic world knowledge with the following prompt:

The capital of France is Paris.
The capital of Brazil is Brasilia.
The capital of Sweden is Stockholm.
The capital of Belgium is

What prompting technique is being used here?

Select one alternative:

- zero-shot prompting
- few-shot prompting ✔
- one-shot prompting
- chain-of-thought prompting

Maximum marks: 1

1.14 LLM training

Large language models are generative models that predict the next word in a sentence. In practice, they are often used as AI assistants that respond to a user's query like a chatbot. Which training step is the most important for making the LLM behave like an AI assistant?

Select one alternative:

- preference alignment
- instruction fine-tuning ✔
- zero-shot prompting
- pre-training

Maximum marks: 1

Part B Reminder

The following questions belong to **Part B** of the exam.

You only need to work on this part if you're aiming for a higher grade.

If you do not reach the passing threshold in Part A, we will not grade Part B. In other words, working on Part B does not help you pass the exam; it can only be used to obtain a higher grade.

What is expected in Part B?

The questions that follow require longer free-text answers that should be well-formulated and correctly use the relevant terminology from the course.

- Your answers should be understandable to a fellow student who took the course. That means it is not necessary to define every basic concept, just the one(s) that are asked for in the answer.
- It is not required to use up the maximum word limit for full points; a concise answer that is well-written is better than an answer that is unnecessarily long.

2.1 Static vs. contextual embeddings

What is the difference between static and contextual word embeddings? Explain in a short, *well-formulated* text of around 100–500 words.

For maximum points, your answer should:

- Explain the conceptual difference between static and contextual word embeddings.
- Name one model or algorithm that can produce each of these types of embeddings.
- Describe the functional difference (advantage/disadvantage) between static and contextual word embeddings with a relevant example.
- Use relevant terminology correctly and precisely.

Write your answer in the box below. Changes are saved automatically.

Format

We do not give example answers here. There are many ways to write a good text that addresses the points asked in the question.

Words: 0/500

Maximum marks: 4

2.2 Byte-pair encoding

What is byte-pair encoding? How does it work, and what problem(s) does it solve? Explain in a short, *well-formulated* text of around 100–500 words.

For maximum points, your answer should:

- Explain what byte-pair encoding is used for (in the context of language technology).
- Describe conceptually how byte-pair encoding works. *(“Conceptually” means that it is not necessary to write down every single step of the algorithm, as long as you can correctly describe the core idea of what the algorithm is doing.)*
- Explain why byte-pair encoding is used by discussing what problem(s) it solves.
- Use relevant terminology correctly and precisely.

Write your answer in the box below. Changes are saved automatically.

Format | B I U x_2 x^2 | I_x | | | | | | | | | Ω | | | Σ |

We do not give example answers here. There are many ways to write a good text that addresses the points asked in the question.

Words: 0/500

Maximum marks: 4