1. *Text Representation.* What is the document-term matrix?
   - ☐ Each term of a language is getting documented by specially trained linguists
   - ☐ The latent representation of attention heads in Transformer blocks
   - ☐ Feature space spawned by terms over documents represented as a matrix
   - ☐ Adjacency matrix of the bipartite graph of documents and the terms contained in each document

2. *Assumptions.* Which are reasonable approaches to capture information in text?
   - ☐ Words are influenced by their surrounding words in context
   - ☐ The sentiment influences the word order in conjunctions
   - ☐ Average sentence length depends on the author's specific writing style
   - ☐ Lexems depends on the anaphoric resolution

3. *Learning.* Which of these statements is true?
   - ☐ Deep learning is always preferred over rule-based approaches
   - ☐ Rule-based approaches require large amounts of labelled training data
   - ☐ The industry often prefers rule-based approaches over machine learning
   - ☐ Deep learning is considered the state of the art for many NLP tasks

4. *Named Entity Recognition.* There are multiple ways on how to encode the tokens used to train sequence classification systems, like named entity recognition systems.

   (a) Why is a simple binary scheme like EO (entity token, outside token) not a good idea?

   (b) What could be the reason that a BERT-based sequence classifier fail to learn [O B I O $\cdots$] and outputs [O I I O $\cdots$] instead?

5. *Word Embeddings.* Traditional word embedding techniques like word2vec have been popular in past years, but today are less used and contextual embedding techniques are preferred.

   (a) What are the advantages of contextual word embeddings, like BERT, over classical word embeddings?

   (b) Are there cases, where a traditional word embedding method is preferred? If yes, please provide an example.

6. *Deep Learning.* Many attention-based approaches combine as inputs a token embedding and a position embedding.
   (a) Why is there a position embedding, what is its purpose?
   (b) Are there tasks, where the position embedding is not useful? If yes, please provide an example.

7. *Word Senses.* You are asked to develop a method for disambiguation of named entities, e.g., names of celebrities, based on Tweets. For example this tweet refers to Adam Scott, the golfer and not Adam Scott, the actor.

   AmerExperience.com/Golf @ShopAmerGolf May 23

   FREE GOLF MAGAZINE Golf: Will Zalatoris, Adam Scott, Keegan Bradley among those now exempt into 2022 U.S. Open — Flipboard <url> Read for free #golfnews #golflessons #golftraining

   (a) Are there any Twitter-specific features, which can be used for this task?

   (b) Would there be other data sources that can be useful for the task?

   (c) Briefly describe how you would solve the task.

8. *Plagiarism.* Consider the university asks you to develop a system to test thesis (e.g., Bachelor and Master) for cases of plagiarism. Your system should for each thesis check, if there are plagiarised passages and mark the beginning and the end of a suspicious passage, which is then checked by human experts.

   (a) Which data sources would you consider?

   (b) What features would you use? (short list with explanation)

   (c) What method would you choose?

   (d) How well do you expect your method to work? What are the bottlenecks?

9. *Causality.* You are asked to build a system to extract causal statements from text of a manufacturing company with a lot of textual document, including technical reports. For example the sentence
"`Mechanical stress is one of the main causes of yield loss`"
should be automatically annotated as
"{`Mechanical stress`}$_{Cause}$ `is one of the main` {`causes`}$_{Cue}$ `of` {`yield loss`}$_{Effect}$".

   (a) What type of approach do you choose?

   (b) What properties do you expect for your approach? E.g., better recall/precision, better performance on longer sentence?

10. *Evaluation.* You developed a method for style transfer for German text. Given a sentence written by an arbitrary writing style, you methods outputs the "same" text as it were written by a famous author (e.g., Thomas Mann, Wolf Haas). Now you are requested to assess how well your system is working.

    (a) What evaluation methodology do you follow?

    (b) What evaluation measures do you use?

    (c) Are there known limitations in your evaluation methodology or evaluation measures?