The exam text is in English. You can answer the questions in English or Dutch. ***Please be concise!***

The exam consists of 11 questions (4 pages). Mark each answer with the question number. The maximum number of points is indicated per question. The total maximum number of points is 100.

You have 3 hours, unless you have an extension clause, which gives you an additional 30 minutes.

Electronic devices (including calculators) are not allowed and must be stowed.

Good luck!

1. Tasks (7)

a. (4 points) What is the difference between multi-class and multi-label text classification?
b. (3 points) Why is named entity recognition typically modelled as a sequence labelling task?

2. Challenges of text data (10)

a. (3 points) What is a long tail distribution?
b. (3 points) What is sparseness in the context of the vector space model?
c. (4 points) Why does the bag-of-words model lead to a sparse feature space?

3. Text pre-processing (12)

a. (4 points) Does lemmatization increase or decrease the dimensionality of feature space for text classification? Explain why.
b. (3 points) Is stemming a language-dependent task? Motivate your answer.
c. (5 points) Compute the Levenshtein distance between ‘alone’ and ‘galore’. Show your computation. (Showing the full matrix is allowed but not required; the table with edits suffices.)

4. Data collection and annotation (10)

Consider this agreement table for a sentiment analysis annotation task:

Agreement table		Annotator 2		
		Positive	Negative	Neutral
Annotator 1	Positive	25	10	5
	Negative	0	25	15
	Neutral	5	5	10

a. (7 points) Compute Cohen's Kappa for this agreement table. Show your computation. (You can keep the last fraction of your computation as it is, without estimating the decimal numbers.)

b. (3 points) What is the meaning of Kappa = 0?

5. Text classification (14)

Consider this toy training set for a text classification task with Naïve Bayes:

Doc id	Content	Class
1	make our garden grow	relevant
2	we make the best of it	not relevant
3	together we can grow	not relevant
4	we make the best plans	not relevant

a. (3 points) What is the prior probability of the 'relevant' class?

b. (3 points) What is the vocabulary size of the training set? Assume that we do not remove stop words.

c. (4 points) Estimate *P('make', not relevant)* using the maximum likelihood estimate on the train set.

d. (4 points) Why is add-one smoothing needed when we estimate the probability of an unseen document? Provide an example test document given the toy training set for which add-one smoothing is needed.

6. Neural NLP and transfer learning (7)

You are given the task to develop a named entity recognition model for biomedical Chinese patent texts with limited labeled data (2500 items).

a. (4 points) How can transfer learning alleviate the limited data problem?

b. (3 points) What resource(s) would be needed for transfer learning for this task?

7. Evaluation (9)

Consider the following output table of an automatic classifier for 10 documents:

doc id	class assigned by classifier	ground truth class
1	A	C
2	A	A
3	B	B
4	B	A
5	C	C
6	A	A
7	D	A
8	A	D
9	B	B
10	C	A

a. (6 points) Compute (please show the fractions):
   i) the recall for the A class
   ii) the precision for the A class

b. (3 points) Which of the three labels I, O, B would you disregard in the evaluation of sequence labelling methods, and why?

8. Information Extraction (10)

Consider this text fragment:

> *The Good Place is an American fantasy comedy television series created by Michael Schur. The series premiered on September 19, 2016, on NBC. The series focuses on Eleanor Shellstrop (Kristen Bell), who wakes up in the afterlife and is introduced by Michael (Ted Danson) to "the Good Place", a highly selective Heaven-like utopia he designed, as a reward for her righteous life.*

Suppose that we are extracting entities and relations for expanding a knowledge base on television and film. We distinguish two entity types that refer to persons: ACTOR and ROLE.

a. (3 points) Which word in this text suffers from type ambiguity?
b. (3 points) Given that you have an incomplete knowledge base with actors and their roles, which learning paradigm can you use for relation extraction? (name the learning paradigm)
c. (4 points) Explain how this would work.

9. Automatic summarization (7)

In the paper by Kryściński et al (2019), "Neural text summarization: A critical evaluation" the authors analyze the ROUGE scores for a number of summarization methods. They report the following:

> *We notice that the ROUGE-1 scores vary considerably less than ROUGE-4 scores. This suggests that the models share a large part of the vocabulary on the token level, but differ on how they organize the tokens into longer phrases.*

a. (3 points) Give a written definition of ROUGE-4 (equation optional).

b. (4 points) What do the authors mean when they state that "the models share a large part of the vocabulary on the token level, but differ on how they organize the tokens into longer phrases"?

10. Authorship attribution (7)

a. (3 points) What type of words are the most effective for authorship attribution: function words or content words? Explain your answer.

b. (4 points) Complete the explanation of the unmasking method by writing down the words for the blanks (1), (2), (3) in the text below:

The so-called unmasking method builds a _______ (1) to distinguish an unknown text from the set of known documents (all by a single author). Then, it removes a predefined amount of the most important ________ (2) and iterates this procedure. If the drop in classification accuracy is _________ (3), then the unknown document was written by the examined author.

11. Domain-specific search (7)

a. (3 points) What is a controlled vocabulary?

b. (4 points) How can a controlled vocabulary help to bridge the semantic gap between CVs and job ads in a search system for recruiters? Give an example.