Lab Class NLP:I

By May 12, 2025, solutions for the following exercises have to be submitted (Discord or email): 1, 2, 3, 4.

Exercise 1 : Datasplits

When training and evaluating a ranking model, the dataset is usually separated into three "splits", **train-**, **test-**, and **validation split**.

(a) What is each of these splits used for?

(b) Why is the data split?

(c) Why do we need to separate evaluation splits? That is, why do we need separate test- and validation splits?

Exercise 2 : Spam Detection

You developed a spam detector that classifies whether an email is "spam" (positive) or "ham" (negative). To test your spam detector, you evaluate it on the last 100 emails you got. Out of these, 10 were spam. Your detector classifies only 7 emails correctly as spam but manages to classify 92 emails correctly overall.

(a) Draw the confusion matrix.

(b) Compute the Accuracy, Precision, Recall, and F1 scores.

(c) Give one example of an NLP task that requires high precision and one that requires high recall.

Exercise 3 : Evaluation

You conducted an annotation campaign on humor. Each annotator was given 6 texts and rated each on a 4-point scale from 1 (not funny) to 4 (hilarious). Since you have a fine grasp of the concept yourself, you also annotated each text as the "truth" label. The results of your campaign are as follows:

Text	**Annotator**						**Inference**		
	A	B	C	D	E	Truth	Majority	Mean	Median
1	1	2	1	1	1	1			
2	2	2	4	2	3	2			
3	2	3	1	2	2	2			
4	4	3	4	3	3	4			
5	1	1	1	1	2	1			
6	3	2	4	2	2	2			
Accuracy						–			
Micro-Precision						–			
Macro-Precision						–			
Micro-Recall						–			
Macro-Recall						–			

(a) Assess the performance of each annotator by calculating the accuracy, and the precision and recall (micro- and macro- averaged) of the annotations.

(b) Assume the annotator would assign labels randomly, what accuracy would you expect in such case?

(c) Infer the final annotation for each text from the votes of the five annotators (A-E) by calculating:

  (c1) the majority vote

  (c2) the mean of votes

  (c3) the median of votes

(d) Assess the performance of the different aggregation strategies (majority vote, mean, and median) by calculating the effectiveness metrics of the inferred annotations. Which one would you choose and why?

(e) Draw the precision-recall curve for the micro-averaged measures.

Exercise 4 : Regression vs. Classification

What is the difference between regression and classification tasks? Come up with 2 examples for each type of task and describe how they can be modelled as a regression/classification task.