String similarity

The fill-in-the-blank questions are important from didactic point of view. But they can be hardly implemented, since typing errors and synonyms have to be taken into account when evaluating the student answers. To achieve this we used one of the string metrics for measuring the distance between two strings: the Damerau-Levenshtein distance, which plays an important role in natural language processing. Informally, this distance is the minimum number of single-character edits (insertion, deletion, substitution, transition) required to change one string sequence into the other. Note that this distance is a metric in mathematical sense, especially it satisfies the triangle inequality. This enables a suitable string evaluation.

To increase the quality of the assessment, we extended the function by the components: whitelist and blacklist. To have a relative measure of the difference between two strings, we convert the distance to similarity. Applying the similarity on whitelist and blacklist we define an acceptance domain for the student answers. Here we need an empirical determined threshold parameter.

Note that the presented method is - strictly speaking - not only based on the strings, but also on semantics, because the used blacklist and whitelist represent a simple semantic relation between the entries of the list.

Experiments

We aim to query for a suitable solution method by fill-in-the-blank question when given a differential equation, see subtask a). This task was used in the winter semester 2021/22 as part of a mini-test for the lecture Mathematics 2. It was completed by 53 students and all student answers were scored error-free.

We implemented a string metric in computer algebra system MAXIMA and placed the corresponding function in the Question variables field of the STACK question concerned. The corresponding XML file can be downloaded from this link.

In the bottom figure we see 18 different student answers (in German) which are positioned in a coordinate system according to both similarities and are classified without errors. In total, this task was processed 263 times. The acceptance domain for correct answers is white-marked.

Note that for this STACK question and the given whitelist resp. blacklist, only the consideration of the whitelist similarity would be sufficient for the evaluation. However, there are situations where the blacklist is necessary.

The authors would like to thank Stiftung Innovation in der Hochschullehre for supporting the project „Digitalisierung Didaktisch Denken“.

We present a method to evaluate fill-in-the-blank student answers in STACK using a string metric, which is not possible in the current version of STACK. To increase the quality of the evaluation, we use whitelist and blacklist instead of a single teacher answer. The performance of a STACK question equipped with a string metric is quantitatively demonstrated by evaluating its use in mathematics courses.