

METHODS AND DEVICES FOR QUANTIFYING TEXT SIMILARITY

ABSTRACT

The present disclosure provides methods and devices for quantifying text similarity. In an embodiment, there is provided a device for quantifying text similarity that comprises: a processor; and a memory including computer program code. The memory and the computer program code configured to, with the processor, cause the device to: obtain a plurality of shortest operation paths for correcting an optical correction recognition (OCR) text string with an edit text string, wherein each of the plurality of shortest operation paths includes one or more edit pairs, each of the one or more edit pairs denoting an operation performable to a character of the OCR text string during correction by the edit text string; determine a plurality of similarity scores, each of the plurality of similarity scores corresponding to one of the plurality of shortest operation paths, wherein each of the plurality of similarity scores is determined by summing historical similarity scores of the one or more edit pairs of each of the plurality of shortest operation paths; and select a minimum one of the plurality of similarity scores to quantify text similarity between the OCR text string and the edit text string.

(Figure 2)