ABSTRACT

The present disclosure provides a text data processing method, a text representation method, a text clustering method, and a device. According to the method, level classification is performed on text data used for training, to obtain a plurality of training sets of different data levels; and multi-stage training is performed on a pre-trained model by using a configured multi-stage training policy and the plurality of training sets of different data levels, to obtain a text representation model. Compared with performing one-time (one-stage) training on the pre-trained model based on all training data, performing multi-stage training on the pre-trained model based on the plurality of training sets of different data levels can better improve a representation capability of the pre-trained model, thereby improving a text representation capability of the trained text representation model, improving text representation quality, and further improving an effect of a downstream text processing task.

Fig. 2