Chinese Spelling Check Shared Task

In this assignment you will be required to automatically diagnose typing errors in Traditional Chinese sentences written by native Hong Kong primary students. The shared task was proposed in a Workshop of IJCNLP 2017. Details are available in this link: https://www.labviso.com/ijcnlp-nlptea-shared-task-2017/

The shared task consists of both error detection and correction. Since the language is Chinese (specifically Cantonese), this shared task may look a bit uncomfortable at first sight. Also, this task was meant to be solved with ML applications, which is why training and testing data were provided. However, you can use simple edit distance to tackle this problem as follows:

1. Build a vocabulary using a Chinese corpus. Some free Chinese corpora are available at https://www.corpus4u.org/threads/233/
2. Compute the edit distance from the erroneous word to all words in the vocabulary and select the word with the minimum edit distance as the correction. Of course, you might think of optimizing this search using efficient data structures such as tries.

You will likely face word segmentation issues. For this you can use any standard Chinese word segmenter e.g. the Stanford Word Segmenter for Chinese, available at https://nlp.stanford.edu/software/segmenter.shtml