BFSU\_Chinese\_MATTR is a user-friendly python script that helps to automatically work out the moving-average type–token ratio (MATTR) of texts with Chinese simplified.

Copyright © 2022-2023, LIU Dingjia, BFSU Corpus Group.

Environment Setup

Java is needed.

Python 3.7.x or later is available.

Prior to MATTR calculation, Chinese word segmentation is needed, which is based on HanLP package.

pip

pip install pyhanlp

download pyhanlp package in the Python Terminal

Run BFSU\_Chinese\_MATTR

Place your txt at the same directory of “BFSU\_Chinese\_MATTR.py”.

Rename the “test.txt” at the end of the script and set your desired window size (e.g., 500, 1000, 2000, etc).

Run the script.

A txt named “MATTR+filename” will be generated at the aforementioned directory, where the result of MATTR is included.

If you use BFSU\_Chinese\_MATTR in your work, please cite the following items:

Miaoru LIN. BFSU\_Chinese\_MATTR. BFSU Corpus Research Group. 2023.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94-100.

Levow G. The third international Chinese language processing bakeoff: word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 108–117. Sydney, Australia, July 2006. Association for Computational Linguistics.