The Journal of Open Source Software

pyuca: a Python implementation of the Unicode Collation Algorithm

J. K. Tauber[1]

1 jktauber.com

DOI: 10.21105/joss.00021

Software

- Review
- Repository
- Archive

Summary

Collation, the sorting of strings, is an important part of computational work in corpus linguistics and digital humanities. Lexicographical sorting, however, is rarely appropriate for languages other than English. The Unicode Consortium has developed the Unicode Collation Algorithm (The Unicode Consortium (2015)) to solve this problem.

pyuca is a Python implementation of the Unicode Collation Algorithm suitable for researchers doing text processing in Python. It passes 100% of the UCA conformance tests for Unicode 5.2.0 (Python 2.7) and 6.3.0 (Python 3.3+) with a variable-weighting setting of Non-ignorable.

pyuca includes the Default Unicode Collation Element Table (DUCET) which provides a default collation suitable for many of the world's scripts.

References

The Unicode Consortium. 2015. "The Unicode Consortium. Unicode Collation Algorithm (Unicode Technical Standard #10)." http://unicode.org/reports/tr10/.