The Journal of Open Source Software

Mordecai: Full Text Geoparsing and Event Geocoding

Andrew Halterman[1]

1 MIT

DOI: 10.21105/joss.00091

Software

- Review
- Repository
- Archive

Summary

Mordecai is a new full-text geoparsing system that extracts place names from text, resolves them to their correct entries in a gazetteer, and returns structured geographic information for the resolved place name. Geoparsing can be used in a number of tasks, including media monitoring, improved information extraction, document annotation for search, and geolocating text-derived event data, which is the task for which is was built. Mordecai was created to provide provide several features missing in existing geoparsers, including better handling of non-US place names, easy and portable setup and use though a Docker REST architecture, and easy customization with Python and swappable named entity recognition systems. Mordecai's key technical innovations are in a language-agnostic architecture that uses word2vec (Mikolov et al. 2013) for inferring the correct country for a set of locations in a piece of text and easily changed named entity recognition models. As a gazetteer, it uses Geonames (Geonames 2016) in a custom-build Elasticsearch database.

References

Geonames. 2016. "Geonames." http://geonames.org.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv Preprint arXiv:1301.3781*.