Text Analytics Seminar Series

New Tools for Corpus Linguistics

Professor Monika Bednarek,
The University of Sydney

9th October 2023
Online (Microsoft Teams)
9.00-10.00 (UK), 19.00-20.00 (Australia)

Sign up at https://forms.office.com/e/326ZsRdpu7.
If you have any questions, please contact Emma Putland at e.putland@lancaster.ac.uk.

This talk will introduce selected new tools for Corpus Linguistics recently developed by the Australian Text Analytics Platform, focussing on analysis of quotation and identification of (near-) duplicates. Both tools are available in the form of Jupyter notebooks. After a brief general introduction to Jupyter notebooks, the talk will introduce the two new tools: The Quotation Tool finds and extracts speakers and their quotes (direct and indirect speech/thought) from news articles. The Document Similarity tool identifies similar texts in a corpus, allowing users to review them and to exclude any (near-)duplicates from the corpus. Both tools are freely-available via a clean user-interface with no/minimal coding required from users and offer interactive means to display and analyse text collections/corpora. Results can be saved and downloaded for additional qualitative analysis. The tools were developed in collaboration between the Sydney Corpus Lab and the Sydney Informatics Hub as part of our involvement in two Australian Research Data Commons research projects – the Language Data Commons of Australia HASS RDC and the Australian Text Analytics Platform.

Acknowledgments

The Language Data Commons of Australia HASS RDC and the Australian Text Analytics Platform (ATAP) projects received investment (https://doi.org/10.47486/PL074; https://doi.org/10.47486/HIR001) from the Australian Research Data Commons (ARDC). The ARDC is funded by the National Collaborative Research Infrastructure Strategy (NCRIS). The Quotation Tool notebook has been adapted from the GenderGapTracker and modified to run on a Jupyter Notebook. The Document Similarity tool uses MinHash to estimate the Jaccard similarity between sets of documents. Relevant references and links will be provided in the talk.

Jointly hosted by the ESRC Centre for Corpus Approaches to Social Science at Lancaster University (UK) and the Sydney Corpus Lab at The University of Sydney (Australia).

CASS
Corpus Approaches to Social Science

Sydney Corpus Lab