Project Title: Creating a Historical Sentiment Lexicon Using Johnson's Dictionary and HPC

Lead supervisor: Dr Emily Middleton, School of English, University of Leeds (E.J.L.Middleton@leeds.ac.uk)

Executive Summary: This project will build a historical sentiment lexicon using the data from *Johnson's Dictionary Online* < https://johnsonsdictionaryonline.com> with Google's BERT model: the dictionary's 50,000 headwords, definitions and examples will be used to calculate a sentiment score for each word, which can then be applied to historical texts. The aim is to try out different ways of developing scores and test the usefulness on articles from thousands of historical newspapers using HPC. While there have been some small-scale efforts to build lexica specific to datasets, this project represents a major step forward in making sentiment analysis historically rigorous.

Project Description: Sentiment analysis has its roots in marketing and business use, but is increasingly of interest to Digital Humanities research. Simply put, sentiment analysis is applied to text to produce visualisations of sentiment or emotion at different levels of scale (within a text, within thousands of texts, on a timeline), and different levels of granularity (positive/negative, or anger/fear/joy/disgust): while this has obvious uses in marketing (does someone like this product?) it can be used, for example, to analyse whether reviews of books that are now popular were as complimentary at the time they were published (like *Dracula*, which for many years was presumed to have been negatively received, but has more recently been shown to have been popular even initially), pull apart speeches by writers like Charles Dickens and consider the nuance of rhetorical strategies to raise money or gain sympathy for a political cause, or visualise the emotional structure of a novel like *A Tale of Two Cities*. Platforms like Gale Digital Scholar Lab include sentiment analysis tools which encourage use across their vast collections, thus suggesting the possibility of getting meaningful results at scale, and the rise of Chat-GPT means scholars without technical training can produce visualisations with carefully-worded prompts, but the majority of available sentiment vocabularies are based on social media: in one case, specifically tweets relating to a UN conference (AFINN). Currently, it is difficult to present defensible historical analysis using sentiment tools, given the mismatch between the lexica they are based on and the language of literary texts: this is true even for modern fiction. A large part of the problem is having a training set large enough to produce meaningful results, and research has argued that manual coding is superior to any currently-available lexicon. No attempt on this scale has been made before: while individuals have used a small set of novels to develop lexica, this project represents a major collaboration with *Johnson's Dictionary Online* to get access to their data to create a historically rigorous sentiment lexicon, and Gale has agreed to make it possible to import the new lexicon into their tool for immediate use, to tie in with the 250th anniversary of the Declaration of Independence, and the launch of new eighteenth-century document collections. Johnson's dictionary was the first major attempt to define meaning in the English language, and was not properly superseded until more than one hundred years later. It remains one of the most influential dictionaries in history, and HPC makes it possible to work with the 50,000 words and their definitions to create a sentiment vocabulary, as well as test its usefulness on large datasets such as Gale's nineteenth-century newspapers, containing millions of words.

Schedule:

Week 1: Familiarising with dataset, cleaning, pre-processing (e.g. identifying words to exclude).
Week 2: Investigating existing methods (e.g. Hugging Face guides).
Week 3: Trialling small-scale approach with a subset of the data to assess viability.
Week 4: Lexicon creation of full dataset using HPC.
Week 5: Testing the lexicon on historical newspaper dataset using HPC.
Week 6-7: Refining and rerunning.
Week 8: Writing up a report for showcase, outlining the data preparation, steps, and results.