Corpus-DB: a Scriptable Textual Corpus Database for Cultural Analytics

Corpus-DB is a database and query framework which solves the problems of text retrieval, text cleaning, corpus compilation, and metadata aggregation that often form the first step for researchers in computational text analysis. Traditionally, scholars interested in studying a collection of texts, such as: novels set in London, Bildungsromane, sestinas, or poems written in 1889, have had to manually assemble their corpora, which can be a prohibitively laborious process. Corpus-DB gathers full texts and metadata from Project Gutenberg, the British Library, and other sources; cleans the texts; adds metadata found via Wikidata, Goodreads, and Wikipedia, and elsewhere; and provides this as a free, open, and easily scriptable API. This enables rapid prototyping of text analysis projects, as well as advanced queries of these corpora, providing easy answers to questions such as:

- What is the average Goodreads star rating of novels set in London?
- What is the median publication date for detective novels?
- What is the library of congress subject heading most predictive of number of downloads on Project Gutenberg?

The API provides a semantic query format. To retrieve all metadata about Charles Dickens texts, simply retrieve the URL:

http://corpus-db.org/api/author/Dickens, Charles

To get the full text of all Jane Austen novels:

http://corpus-db.org/api/author/Austen, Jane/fulltext

And to get metadata for all works labeled as detective and mystery stories:

http://corpus-db.org/api/subject/Detective and mystery stories

This lightning talk will briefly demonstrate the use of this tool, and suggest avenues for its further development.