GUTENTAG
An NLP-driven Tool for Digital Humanities Research in the Project Gutenberg Corpus

Definition and motivation
GutenTag is a tool for medium- and large-scale analysis of texts in the Project Gutenberg corpus. The high-level goal of the project is to create an ongoing two-way flow of resources between computational linguists and digital humanists, allowing computational linguists to identify pressing problems in the large-scale analysis of literary texts, while giving digital humanists access to a wider variety of NLP tools for exploring literary phenomena. GutenTag is intended to be a standalone software tool for non-programmers, but the source code is also available and we welcome others in the computational linguistics community to contribute to its development or adapt it as needed.

One long-term goal of the Project Gutenberg corpus is to be able to access the full functionality of GutenTag via the web. Given our diverse user base, we may need to move towards an HTML form as the user interface (GUI for GutenTag; see detail above), creating a configuration file which can be saved, loaded, and modified in a text editor. Users define the particular sub-corpus of Project Gutenberg they wish to investigate. At a lower level, users can define sets of attributes which are inserted within the text boundaries.

Tools similar to GutenTag include software for automatic analysis of texts for literary purposes such as Voyant, a literary corpus tool like PhiloLogic, general purpose NLP tool packages such as NLTK (which GutenTag is built on), and a (very) simple existing Project Gutenberg reader, Gutenberg. The overlap between these other tools and GutenTag is, however, fairly small: no existing tool offers sophisticated language analysis with literature-specific tagging appropriate for large-scale analysis. Our intent is that GutenTag will become a growing repository for NLP solutions to tasks relevant to literary analysis, and it is this wide-ranging, inherently cross-disciplinary focus that is the clearest difference between GutenTag and other tools.

References
1. http://www.gutenberg.org
3. http://www.philo.uni-jena.de/voyant/

Acknowledgments
This work was supported by the Natural Sciences and Engineering Research Council of Canada, and the MITACS Elevate program, and the University of Guelph.

© 2015 Julian Brooke, Adam Hammond, Graeme Hirst

Julian Brooke
Adam Hammond
Graeme Hirst

Other tools
Some structural tagging would likely benefit from statistical methods. Some structural tagging would likely benefit from statistical methods. Some structural tagging would likely benefit from statistical methods.

We could subdivide our four main genres into any number of sub-genres, though it might be difficult to do this without integrating content features (which might invalidate some uses of the tag). A more sophisticated classifier might be preferred, and we should integrate more features.

We built our structural tagging module by focusing on the structure of 50 texts from diverse genres (20 fiction, 10 nonfiction, 10 drama 10 poetry); this is an insufficient sample. Some structural tagging would likely benefit from statistical machine learning approaches. Other kinds of structure that would require sophisticated NLP tools include those reflecting time, location, viewpoint, topic, and narrative structure.

We use fairly sophisticated heuristics to remove this information, including certain kinds of metadata about the copyright in the United States. The work here is based on the most recently released English literature published before 1923 are included in the collection. The English portion of the corpus consists of approximately 1.7 billion tokens. Though we have tried to stay as close as possible to the TEI standard, we have omitted certain tags because we felt that they were too detailed or too challenging to deal with automatically. We would be interested in hearing feedback on other tags we should include, and on existing tags that we are handling poorly.

We could subdivide our four main genres into any number of sub-genres, though it might be difficult to do this without integrating content features (which might invalidate some uses of the tag). A more sophisticated classifier might be preferred, and we should integrate more features.

We built our structural tagging module by focusing on the structure of 50 texts from diverse genres (20 fiction, 10 nonfiction, 10 drama 10 poetry); this is an insufficient sample. Some structural tagging would likely benefit from statistical machine learning approaches. Other kinds of structure that would require sophisticated NLP tools include those reflecting time, location, viewpoint, topic, and narrative structure.

We intend to add multi-word lexical tags and to upgrade the name tagger to distinguish various types (eg. characters vs. locations). Future modules would include tagging of elements such as meter, anaphora, alliteration, onomatopoeia, foreign language, allusions, similes, and metaphor.

GutenTag uses NLTK tokenization, lemmatization, and POS tagging. Other lexical tags available are manually-built lexicon (MRC psycholinguistic database) and the General Inquirer Dictionary (GID) and a lexicicon of style built from the Project Gutenberg corpus. Users can define their own lexicons. GutenTag includes a simple name tagger and connects names and likely spans of dialogue.

GutenTag uses the popular XPLR-based Text Encoding Initiative (TEI) format as the default output format when structure (rather than just tokens) is requested, which makes it compatible with other work in the Digital Humanities.

One long-term goal of the project is to be able to access the full functionality of GutenTag via the web. Given our diverse user base, we may need to upgrade the interface to improve usability.

GutenTag is a tool for medium- and large-scale analysis of texts in the Project Gutenberg corpus. The high-level goal of the project is to create an ongoing two-way flow of resources between computational linguists and digital humanists, allowing computational linguists to identify pressing problems in the large-scale analysis of literary texts, while giving digital humanists access to a wider variety of NLP tools for exploring literary phenomena. GutenTag is intended to be a standalone software tool for non-programmers, but the source code is also available and we welcome others in the computational linguistics community to contribute to its development or adapt it as needed.