



GutenTag

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

Introduction

Practical Access: A web interface makes the tool accessible to users without a programming background.

Two-Way Avenue: It aims to create a collaborative feedback loop between Computational Linguists and Digital Humanists.

Bridging a Gap: GutenTag is a tool designed to give literary scholars direct access to advanced Natural Language Processing (NLP) techniques for analyzing the massive Project Gutenberg corpus.

01 Background

Key Concepts and Relevant Past Studies

Key Concept: Cultural Barriers

Computational Linguists develop sophisticated techniques, but these are often unknown or unavailable to the humanities community.

Digital Humanists are often interested in computational analysis but may lack the technical expertise to use state-of-the-art NLP tools.

The result is a disconnect where humanists use simple off-the-shelf tools, and linguists miss out on the challenging, interesting problems posed by literary texts.

This concept is drawn from the authors' own prior work: Hammond et al. (2013), *"A tale of two cultures: Bringing literary analysis and computational linguistics together."*

Key Concept: Literature-Specific NLP

Most NLP tools are trained on "standard" texts like newswire or web content, which have very different characteristics from literary works.

Literary texts contain unique elements like dialogue, meter, rhyme, narrative structure, and figurative language that standard tools often handle poorly.


GutenTag is explicitly designed not just to apply off-the-shelf NLP, but to become a repository for new, literature-specific modules that can address the unique challenges of the domain.

Key Concept: Surface-Level Analysis

Existing digital humanities tools like Voyant are great for surface-level analysis like word frequency counts and Key Word In Context (KWIC) displays.

GutenTag moves computational literary analysis from "counting words" to "interpreting meaning" at scale.

GutenTag enables a deeper, semantic and structural analysis by tagging complex phenomena like: quoted speech attributed to specific characters, identifying narrative elements like scenes, locations, and plot points, and stylistic features like metaphor and allusion.

The background is a solid light purple color. It features several abstract elements: in the top left, there are white and yellow circuit-like lines and shapes; in the bottom left, there are white and light blue geometric shapes resembling stacked cubes or crystals; on the right side, there are thin, winding lines in dark blue, light blue, and yellow, some ending in small circles or dots; and a grid of small, light blue dots is visible in the upper right and lower left areas.

02 Research Questions

GUTENTAG

This paper is a tool description for the GutenTag software.

It doesn't pose formal research questions because of this, but its high-level goal is:

"to create an on-going two-way flow of resources between these groups, allowing computational linguists to identify pressing problems in the large-scale analysis of literary texts, and to give digital humanists access to a wider variety of NLP tools."

What does that mean?

How can we build a user-friendly tool that lets literature experts use powerful computer analysis methods?

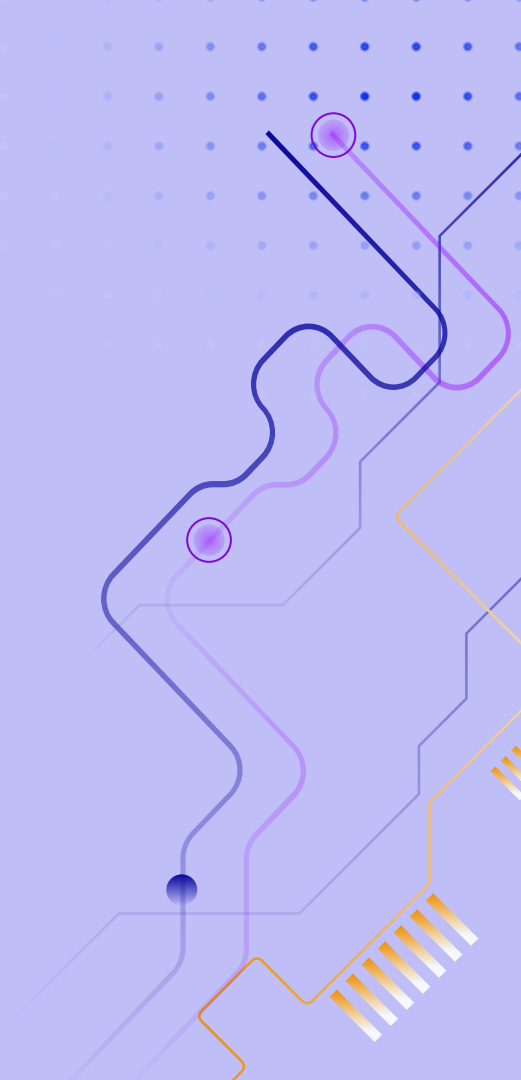
How can this tool help computer scientists and literature scholars learn from each other?

Can we automatically clean up, categorize, and tag the messy Project Gutenberg texts in a way that is useful for literary analysis?



03

Methodology





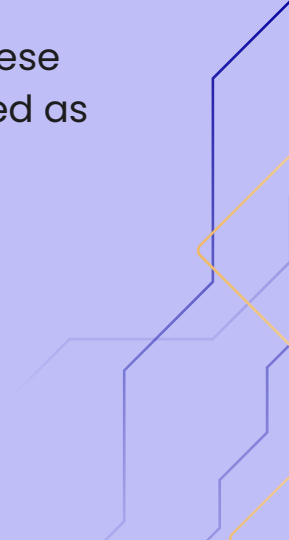
Cleaning the Text & Identifying Structures

Project Gutenberg texts have inconsistent headers/footers about copyright infringement.

They used complex heuristics and regular expressions to identify and remove this non-textual content, as well as illustrator notes and transcriber's notes.

The structure of books, plays, and poems is marked inconsistently.

GutenTag identifies structural elements before tokenization. These elements can be removed or used as tags in the final output.






Subcorpus Filter and Tagging

The full Project Gutenberg corpus is too large and diverse for most research questions.

GutenTag gives researchers the ability to create very specific filters. It uses Project Gutenberg metadata and automatically infers genre using textual clues.

A "tag" is a label for a span of text (e.g., a word, a quote, a chapter). GutenTag uses a pipeline where different "tagger" modules add layers of analysis.



04

Tools



Technologies Used

2010 PG DVD

29,557 documents, ~1.7 billion English tokens

Chosen because it is pre-proofread and/or hand-typed, avoiding OCR errors

NLP Library

Natural Language Toolkit (NLTK)

NLTK Regex Tokenizer with custom tweaks for handling abbreviations, hyphens, quotes

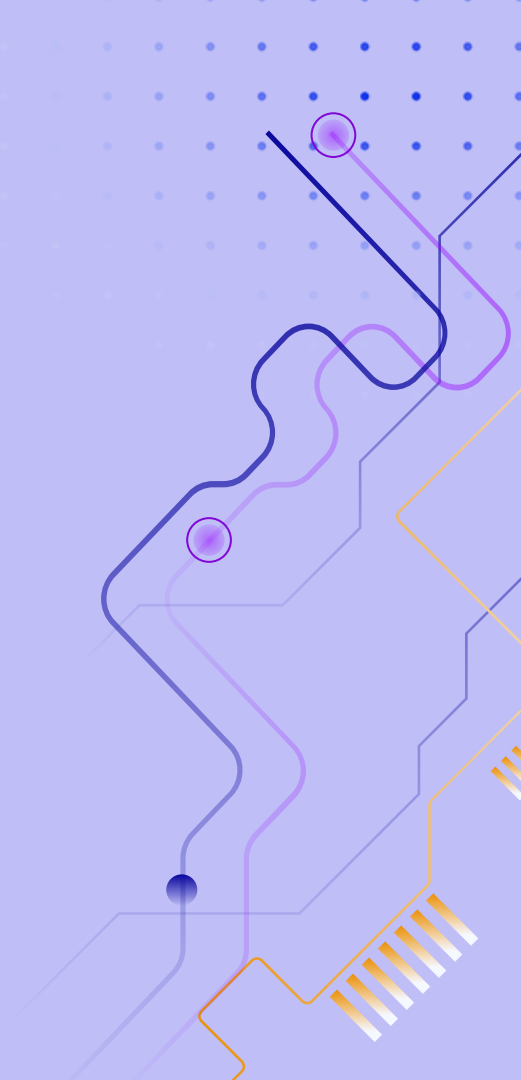
Output Format

Text Encoding Initiative (TEI) XML standard



05

Commentary



Commentary

GutenTag's most compelling feature is not any single technical module, but its foundational role as a bridging framework.

The tool successfully operationalizes a vast, challenging, and inherently "messy" domain and provides a clean, modular pipeline for deploying and testing new NLP solutions.

It directly tackles the problem of data scarcity for literature-specific tasks by providing a pre-processed, structured corpus, and its explicit design for community expansion invites collaboration.

GutenTag is a catalyst; it identifies a suite of high-impact, unsolved problems and provides the infrastructure to tackle them.

06

Quiz



Question 1

What is the primary goal of the GutenTag tool?

Answers:

- A.) To digitize new books for Project Gutenberg.
- B.) To give digital humanists access to NLP techniques for literary analysis.
- C.) To create a new social network for academics.

Question 2

What was a major challenge in preparing the Project Gutenberg texts for analysis?

Answers:

- A.) The texts were written in too many different languages.
- B.) The files were too large to download.
- C.) The formatting of headers, footers, and structural elements was highly inconsistent.

Question 3

How does GutenTag help a researcher who only wants to study 18th-century poetry?

Answers:

- A.) It manually selects the best poems.
- B.) Its Subcorpus Filter allows the creation of a custom collection based on metadata and genre.
- C.) It translates prose into poetry.

Question 4

Which technology was not a core part of GutenTag's methodology?

Answers:

- A.) Optical Character Recognition (OCR) for all texts.
- B.) The Python programming language and NLTK.
- C.) Heuristics and regular expressions for text cleaning.



Fin.

Abby Clark