Visualizing Texts with Voyant Tools

By Subhanya Sivajothy

Image: Mhsheikholeslami (CC 4.0 BY-SA)

Session Recording and Privacy

This session is being recorded with the intention of being shared publicly via the web for future audiences.

In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.

Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.

Code of Conduct

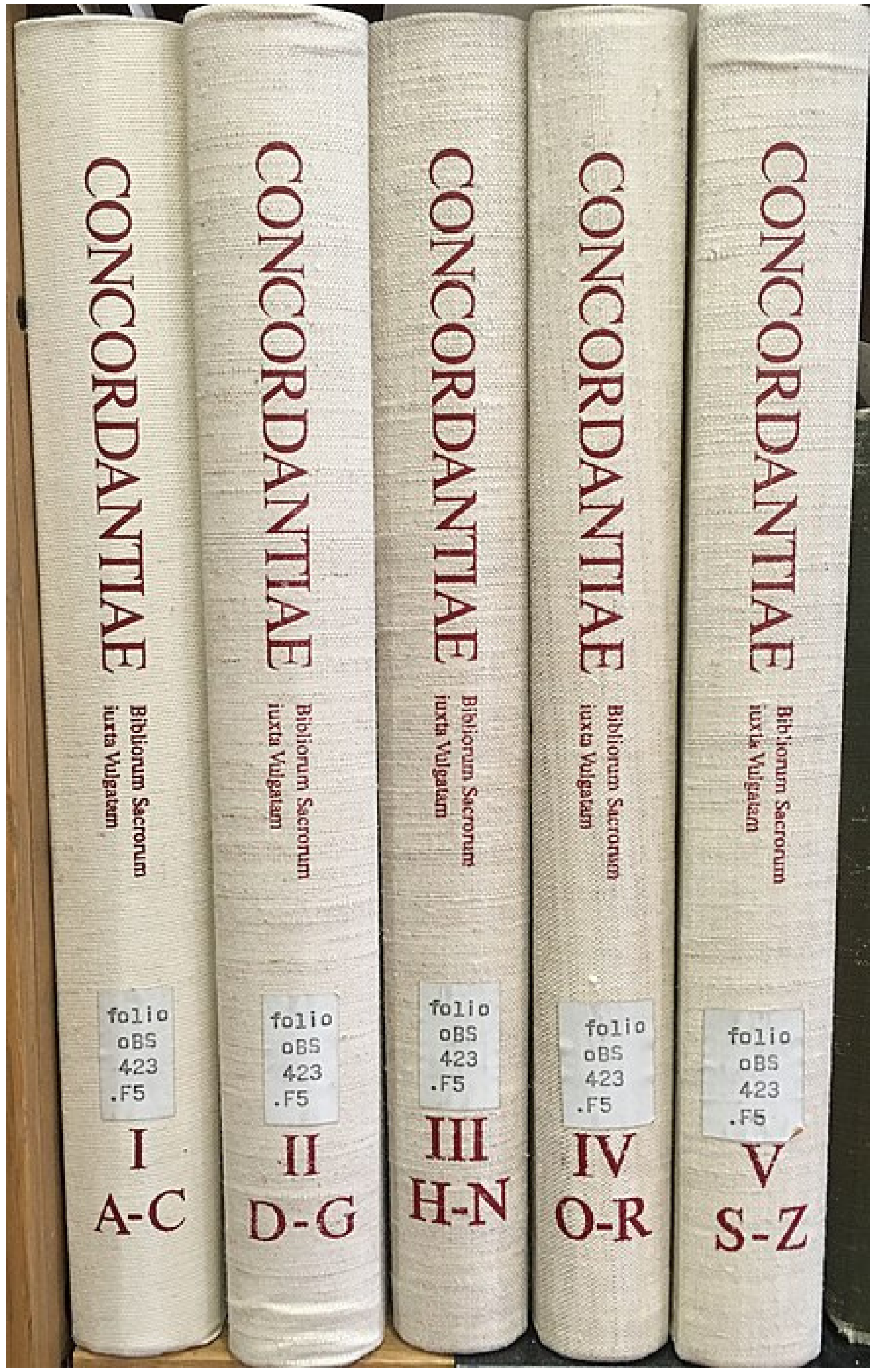
The Sherman Centre and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information:

scds.ca/events/code-of-conduct/

- Vulgate Bible Concordance (13th Century)
- Index Thomisticus (1940s)



Text Analysis

API

A specification that allows software applications to communicate with one another. An API allows client programs to access facilities within an application.

Corpus

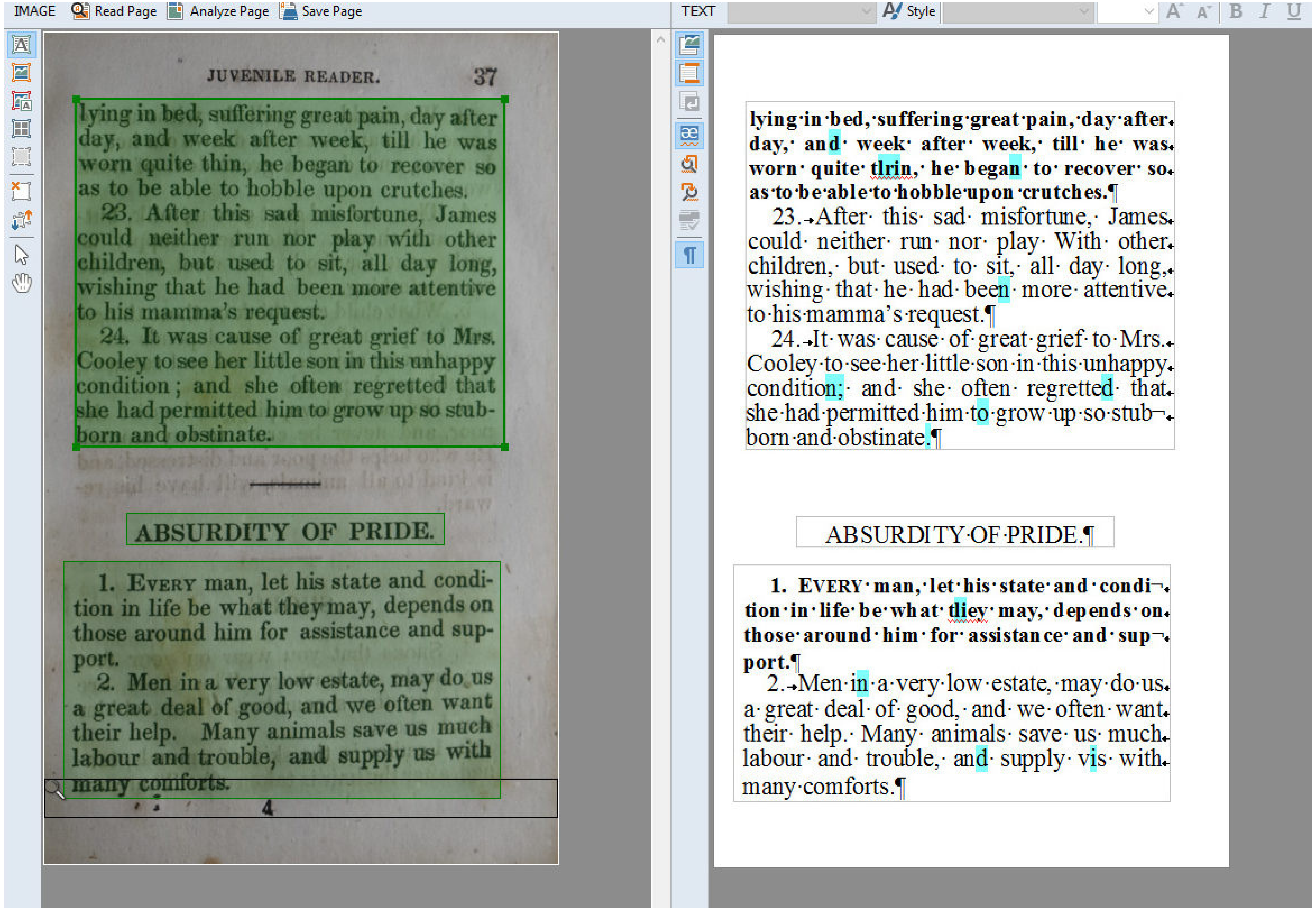
A collection of written texts, particularly the entire body of work on a subject or by a specific creator.

OCR

The use of computer technologies to convert scanned images of typewritten, printed, or handwritten text into machine-readable text.

Text Mining

The process of automatically deriving previously unknown information from written texts using computational techniques.



lying in bed, suffering great pain, day after day, and week after week, till he was worn quite thin, he began to recover so as to be able to hobble upon crutches.

23. After this sad misfortune, James could neither run nor play with other children, but used to sit, all day long, wishing that he had been more attentive to his mamma's request.

24. It was cause of great grief to Mrs. Cooley to see her little son in this unhappy condition; and she often regretted that she had permitted him to grow up so stubborn and obstinate.

ABSURDITY OF PRIDE.

1. Every man, let his state and condition in life be what they may, depends on those around him for assistance and support.

2. Men in a very low estate, may do us a great deal of good, and we often want their help. Many animals save us much labour and trouble, and supply us with many comforts.

lying in bed, suffering great pain, day after day, and week after week, till he was worn quite tlrin, he began to recover so as to be able to hobble upon crutches.

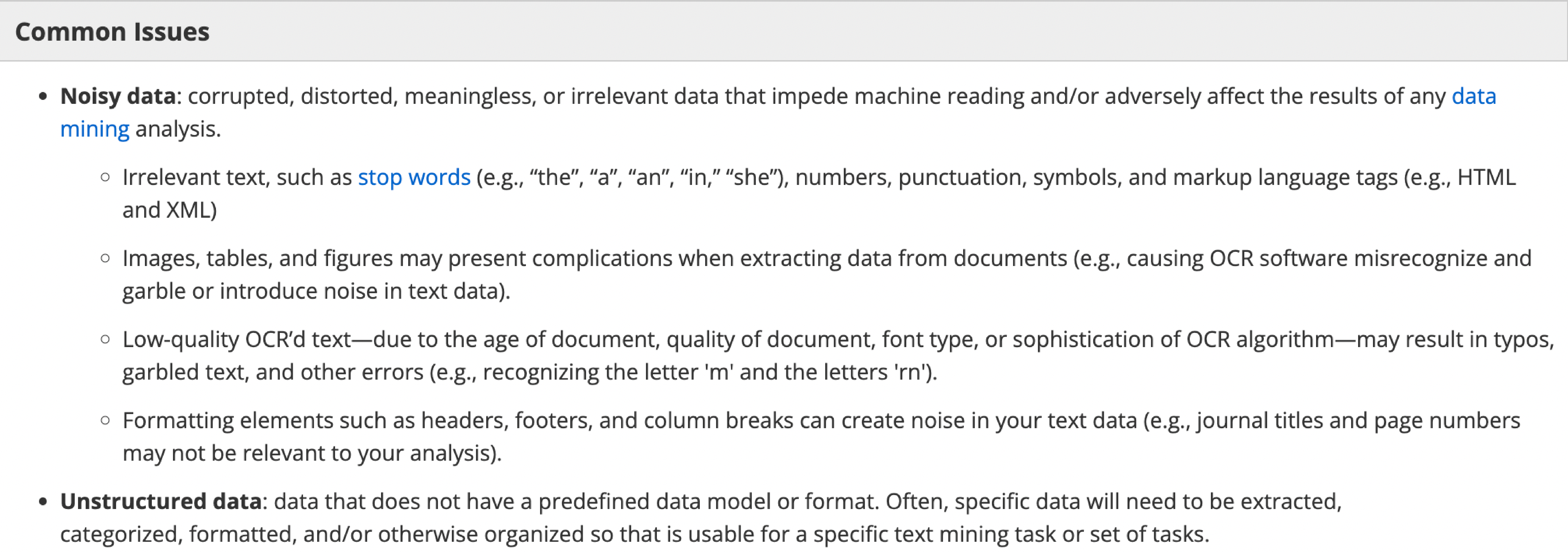
23. After this sad misfortune, James could neither run nor play With other children, but used to sit, all day long, wishing that he had been more attentive to his mamma's request.

24. It was cause of great grief to Mrs. Cooley to see her little son in this unhappy condition; and she often regretted that she had permitted him to grow up so stubborn and obstinate.

ABSURDITY OF PRIDE.

1. Every man, let his state and condition in life be what tliey may, depends on those around him for assistance and support.

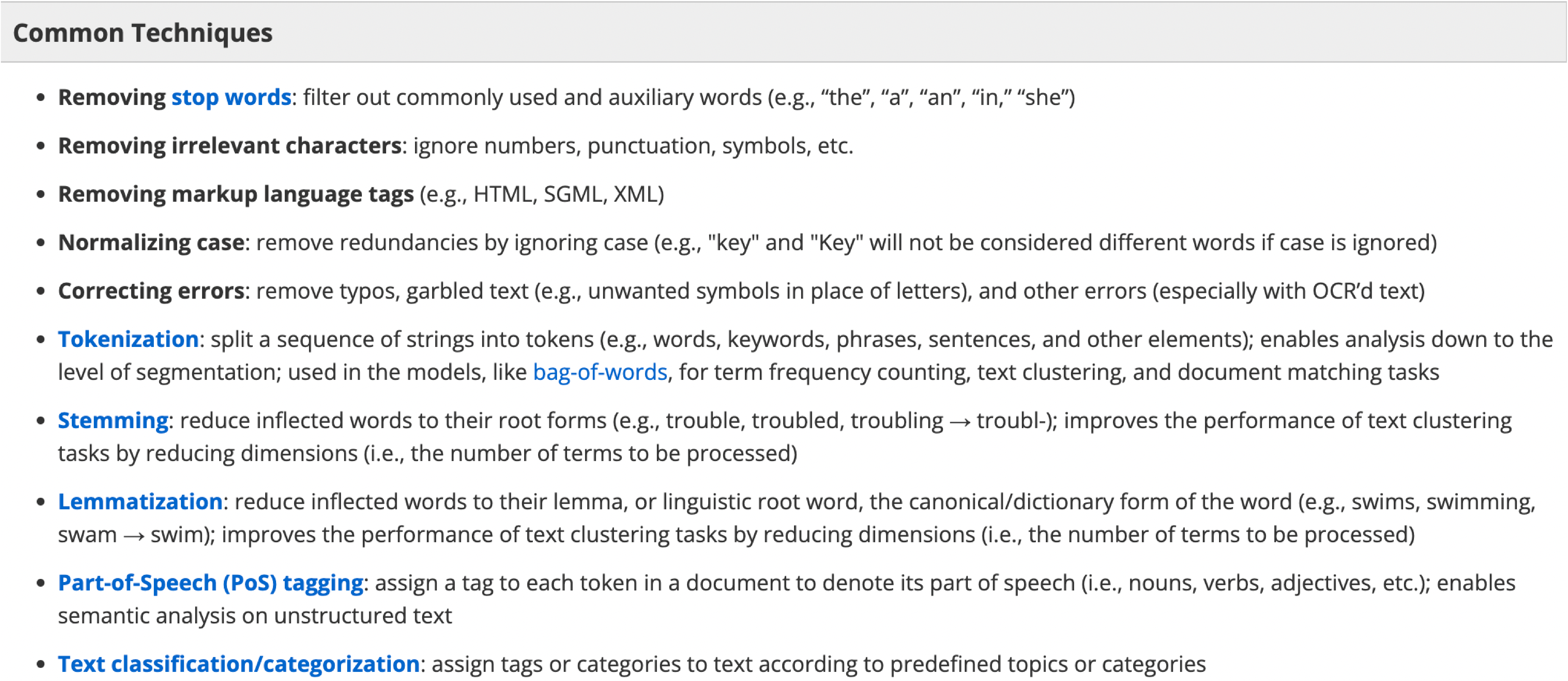
2. Men in a very low estate, may do us a great deal of good, and we often want their help. Many animals save us much labour and trouble, and supply vis with many comforts.



Common Issues

- **Noisy data**: corrupted, distorted, meaningless, or irrelevant data that impede machine reading and/or adversely affect the results of any data mining analysis.
    - Irrelevant text, such as stop words (e.g., “the”, “a”, “an”, “in,” “she”), numbers, punctuation, symbols, and markup language tags (e.g., HTML and XML)
    - Images, tables, and figures may present complications when extracting data from documents (e.g., causing OCR software misrecognize and garble or introduce noise in text data).
    - Low-quality OCR’d text—due to the age of document, quality of document, font type, or sophistication of OCR algorithm—may result in typos, garbled text, and other errors (e.g., recognizing the letter 'm' and the letters 'rn').
    - Formatting elements such as headers, footers, and column breaks can create noise in your text data (e.g., journal titles and page numbers may not be relevant to your analysis).
- **Unstructured data**: data that does not have a predefined data model or format. Often, specific data will need to be extracted, categorized, formatted, and/or otherwise organized so that is usable for a specific text mining task or set of tasks.

source: https://pitt.libguides.com/textmining/preprocessing/



Common Techniques

- **Removing stop words**: filter out commonly used and auxiliary words (e.g., “the”, “a”, “an”, “in,” “she”)
- **Removing irrelevant characters**: ignore numbers, punctuation, symbols, etc.
- **Removing markup language tags** (e.g., HTML, SGML, XML)
- **Normalizing case**: remove redundancies by ignoring case (e.g., "key" and "Key" will not be considered different words if case is ignored)
- **Correcting errors**: remove typos, garbled text (e.g., unwanted symbols in place of letters), and other errors (especially with OCR’d text)
- **Tokenization**: split a sequence of strings into tokens (e.g., words, keywords, phrases, sentences, and other elements); enables analysis down to the level of segmentation; used in the models, like bag-of-words, for term frequency counting, text clustering, and document matching tasks
- **Stemming**: reduce inflected words to their root forms (e.g., trouble, troubled, troubling → troubl-); improves the performance of text clustering tasks by reducing dimensions (i.e., the number of terms to be processed)
- **Lemmatization**: reduce inflected words to their lemma, or linguistic root word, the canonical/dictionary form of the word (e.g., swims, swimming, swam → swim); improves the performance of text clustering tasks by reducing dimensions (i.e., the number of terms to be processed)
- **Part-of-Speech (PoS) tagging**: assign a tag to each token in a document to denote its part of speech (i.e., nouns, verbs, adjectives, etc.); enables semantic analysis on unstructured text
- **Text classification/categorization**: assign tags or categories to text according to predefined topics or categories

source: https://pitt.libguides.com/textmining/preprocessing/

OpenRefine

Stopwords

words filtered out before or after processing of natural language data (text), usually words with little meaning such as “and,” “the,” “a,” “an”

Vocabulary Density

a measurement of vocabulary usage in comparison to the length of a document. Think of how many words will be read on average before a new word is encountered. (For Moby Dick, a new word appears every 12 words!)

A lower vocabulary density indicates complex text with lots of unique words, and a higher ratio indicates simpler text with words reused

Search

- `love`: match **exact term** love
- `love*`: match terms that start with the **prefix** love and then a **wildcard** as **one term**
- `^love*`: match terms that start with love as **separate terms** (love, lovely, etc.)
- `*ove`: match terms that end with the **suffix** *ove* as **one term**
- `^*ove`: match terms that end with **suffix** *ove* as **separate terms** (love, above, etc.)
- `love,hate`: match each term **separated by commas** as **separate terms**
- `love\|hate`: match terms **separated by pipes** as a **single term**
- `"love him"`: *love him* as an exact **phrase** (word order matters)
- `"love him"~0`: *love him* or *him love* **phrase** (word order doesn't matter but 0 words in between)
- `"love her"~5`: match *love* **near** *her* (within 5 words)
- `^love*,love\|hate,"love her"~5`: **combine** syntaxes