Parsing xml trees and identifying semantic structures

Semantic trees

Souce / Tree EXPRESSIONS
Astronomia nova as a structured text composed by an author - Johannes Kepler
Intended structures
Our text corpora are structured according to the textual divisions of the author: in *astronomia nova* we have
  - parts
  - chapters
  - paragraphs
  - sentences

Implicit structures

e.g.

for sentences: spacy dependendy graphs, use only tree structure

example: for each sentence as a node in the text tree there is a link to a dependecy graph in pandas

goal: a dataframe of the entire ext + implicit structure that spacy provides

Then there are other textual structures related to those structures

  - figures
  - marginals

Semantic content tree

distringuish between propositions and others

Inferences are defined as relations between proposition

- start with the proposition: each sentence refers to one proposition
- note: there are exceptions, e.g. geometrical constructions are not true sentences but construction rules

FIRST GOAL

(1) Convert text corpus into an element tree lxml.etree (equivalent to xml tree)
First applied to Testcase

Literature

- https://lxml.de/tutorial.html
- https://docs.python.org/3/library/xml.etree.elementtree.html

Testcase: parsing xml page

(2) import html page

https://academic.oup.com/clinchem/advance-article/doi/10.1093/clinchem/hvaa029/5719336

(2a) Export page for archive including date, url

(3) parse the xml of that page into element tree

(4) Identify elements by categories like

- authors:
  - author
  - affiliation
- submission
  - submission date
  - revised date
  - accepted date
- reference list

(5) create a new semantic tree with these three categories and create nodes with pointers to those elements with attached meanings. Derive reference information like xpath.