European Summer School for Digital Humanities
Text Mining with Canonical Text Services
Practical Tasks 3

Alignment Tools, Fulltext Search, CTS Cloning

The following tasks can be accomplished using of of the publicly available CTS instances or the Virtual Machine demos. Mind that some of the published instances may be served using outdated software versions. If a certain feature is not working, try again using the CTS demo on the Virtual Machine.

Configuration Parameter

Learn how to use the configuration parameter by structuring a requested text passage using the generic div notation.

Delete the XML markup in a text passage that contains XML tags. If no suitable data set is available, the following request can be used
http://ctstest.informatik.uni-leipzig.de/perseus/cts/?request=GetPassage&urn=urn:cts:greekLit:tlg0007.tlg001.perseus_eng2:4.1

What kind of problem can occur, when XML tags are deleted from the text?

Combine both parameters to request the text passage without XML tags using a generic div notation.

Text Alignment

The following tasks are best done with the CTS Instance based on the Parallel Bible Corpus. Links to the data sets are available here:

http://cts.informatik.uni-leipzig.de/tools.html

Learn how you can use the Parallel Alignment Browser to align text passages across several languages.

Export the result as a CSV file. What kind of use case can you imagine for this data.

Learn how you can use the Candidate Alignment Browser to visualize the variations of text passages in different editions of one document in one language.

Create a persistent reference to your result and add it as a hyperlink to your local document.

Advanced Request Features

Request the document level CTS URNs for the *ted* CTS instance using the request GetCapabilities.

Request the same information using the suitable advanced function. If you do not see a performance difference, repeat the process using the *textgrid* CTS instance.

Learn how you can request document level meta information without having to request the text inventory.

Learn how you can generally request document structure information based on CTS URNs.

Request the citation depths and types of each text part as a compressed and uncompressed result.

Request a list of static identifiers for structural elements along with the length of their text content.

Search for every CTS URN in the *demo* CTS instance that contains the token "Tannenbaum".