European Summer School for Digital Humanities
Text Mining with Canonical Text Services
Practical Tasks 4
Canonical Text Miner

The following tasks can be done using the demo on the Virtual Machine or the publicly available demo that is available here:

http://ctstm.informatik.uni-leipzig.de:8080/ctstm/vis/

First Steps

Check whether or not the Canonical Text Miner is online. If not, you can install it using the pre configured installation configuration and run the installation task from session 7.

Analyze the different tools in the Canonical Text Miner Graphical User Interface and answer the following questions and tasks.

Request Information

Visualize the top 300 tokens in text corpus. Learn how you can request token frequencies on text corpus level and document level.

Learn how you can switch between the visualisation and the raw data output.

Visualize how token frequencies of one of the top 300 tokens changed over the course of time.

Learn how you can change the time scale. Request the information with daily, monthly and yearly granularity.

Check, how the information is formatted as raw data request.

Visualize the context of the token "education". What happens, when the context size is changed from 5 to 3.

Topic Models can be used to investigate a generic thematic structure of a text collection. Use the Topic Cloud Tool to browse this thematic structure for the given data set.

Token neighbourhood can be visualized using a graph structure. Build an interesting word graph and learn, how the recursion depth and number of outgoing edges influences this graph.

Share your graph visualization with one of the other participants by sending a bookmark link.

[Hier eingeben]

Learn how you can search CTS URNs for a given text passages on document level and on text passage level.

Investigate how the selection of different index techniques influences the result.

Create a ball of tokens that visualizes text collection based token frequencies based on size and color of the elements.

Learn how you can choose different ranges of minimum and maximum token frequencies and experiment until you find a ball that looks good for you. It is advised to not visualize token balls with more than 500 tokens because the calculation is done in your browser.

Final Task

Open your local office document or create a new one. Describe your result in a short passage and add a hyperlink reference to the token ball, for example as a footnote.

[Hier eingeben]