SimBaTex: Similarity-based Text Exploration

Daniel Witschard, Ilir Jusufi and Andreas Kerren
Linnaeus University and Linköping University, Sweden

Natural language processing in combination with visualization can provide efficient ways to discover latent patterns of similarity which can be useful for exploring large sets of text documents. In this poster, we present an overview of the interface of a work-in-progress visual analytics application, called SimBaTex, which is based on embedding technology, dynamic specification of similarity criteria, and a novel approach for similarity-based clustering. The goal of SimBaTex is to provide search-and-explore functionality to enable the user to identify items of interest in a large set of text documents by interactive assessment of both high-level similarity patterns and pairwise similarity of chosen texts.

REVEAL HIGH LEVEL SIMILARITY PATTERNS
Enable, disable or invert the specific similarity criteria to cluster each document together with the document that it is most similar to (if any) given the current settings. Assess the homogeneity of the created clusters through heatmap displays.

ASSESS DETAILED PAIRWISE SIMILARITY
Select a specific document for target-to-all comparison to reveal the pairwise similarity with the Top 10 best matches, both in compiled form and as highlighted text for close reading.