Modeling of Historical Concepts of Privacy: The Challenges of Creating Datasets for Weakly Supervised Discovery with Temporally-Aware Topic Models

Natália Da Silva Perez*[1]

[1]Centre for Privacy Studies, University of Copenhagen – Denmark

Abstract

This poster will present challenges encountered in turning unstructured texts from early modern corpora into datasets for Natural Language Processing. Our project seeks to study historical conceptualizations of privacy by employing topic modeling techniques. We work on developing Natural Language Processing (NLP) tools compatible with early modern texts (c. 1500 to 1800), enabling us to: 1) discover historical terms related to privacy in different languages; 2) investigate changes in meaning over time; 3) examine their dissemination over time to diverse political contexts within Europe. In early modern textual corpora, lack of linguistic standardization, noisy texts resulting from scanned material and weird fonts, temporal and regional linguistic shifts, all require specialized NLP tools. We are working on a systematic and data-driven approach to the study of historical concepts of privacy within a broad geographical and chronological framework. This poster presents our roadblocks.

Keywords: early modern privacy, unstructured text, datasets, OCR, NLP, challenges

*Speaker